SISSA

Scuola Internazionale Superiore
di Studi Avanzati

International School for Advanced Studies

Statistical mechanics of samples, efficient representations and criticality

Ryan John Abat Cubero

A Dissertation
Presented to the Faculty of SISSA
in Candidacy for the Degree of
Doctor of Philosophy

Supervisors: Matteo Marsili and Yasser Roudi

October 2018

This page has been intentionally left blank.

Abstract

Broad distributions appear frequently in empirical data obtained from natural systems even in seemingly unrelated domains. The emergence of power law distributions is so ubiquitous that it has puzzled scientists across disciplines. To understand its origin is thus crucial to understand the mechanisms from which it transpires. In this thesis, we present an information theoretic perspective on the origin of broad distributions. Guided by the principle that learning from the data is equivalent to an optimal coding problem, we show, through various viewpoints, that broad distributions manifest when the sample is maximally informative on the underlying data generating process. Furthermore, by working under the Minimum Description Length (MDL) principle, we show that the origin of broad distributions – a signature of statistical criticality – can be understood precisely as a second order phase transition with the coding cost as the order parameter. This formulation then allows us to find the neurons in the brain that contain relevant representations during spatial navigation. Taken together, this thesis suggests that statistical criticality emerges from the efficient representation of samples of a complex system which does not rely on any specific mechanism of self-organization to a critical point.

This page has been intentionally left blank.

List of Publications

1. **RJ Cubero**, M Marsili and Y Roudi.
   Finding informative neurons in the brain using Multi-Scale Relevance
   arXiv preprint arXiv:1802.10354
   2018

2. **RJ Cubero**, J Jo, M Marsili, Y Roudi, J Song.
   Minimally sufficient representations, maximally informative samples and Zipf's law
   arXiv preprint arXiv:1808.00249
   2018

3. **RJ Cubero**, M Marsili and Y Roudi
   Minimum Description Length codes are critical
   Entropy 2018, 20(10), 755
   https://doi.org/10.3390/e20100755

The following works were done during the PhD period but are not discussed in this thesis:

1. C Torrini, **RJ Cubero**, E Dirkx, L Braga, H Ali, G Prosdocimo, MI Gutierrez, C Collesi, D Licastro, L Zentilin, M Mano, M Vendruscolo, M Marsili, A Samal, M Giacca
   Common regulatory pathways mediate activity of microRNAs inducing cardiomyocyte proliferation
   (Submitted to *Cell Reports*)

2. E Dirkx, A Raso, H el Azzouzi, **RJ Cubero**, S Olieslagers, M Huibers, R de Weger, S Siddiqi, S Moimas, C Torrini, L Zentilin, L Braga, P da Costa Martins, S Zacchigna, M Giacca, L De Windt
   A microRNA program controls the transition of cardiomyocyte hyperplasia to hypertrophy and stimulates mammalian cardiac regeneration
   (*in preparation*)

This page has been intentionally left blank.

Acknowledgment

> Luck only comes to those whose minds are ready for it...
>
> *Dr. Jinky B. Bornales*

Wrapping up this dissertation had been truly rewarding. Yet, the road toward its completion has been both frustrating and exhausting. This thesis stands as a testament of my four-year journey, not only as a budding scientist but also as an individual. Through highs and lows, I have made countless discoveries, revelations and realizations. As I culminate this chapter in my academic career, I would like to take this opportunity to thank the people from whom I have taken strength, encouragement and inspiration:

To Matteo and Yasser, for giving me the opportunity to work under their guidance and dedicated involvement without which, this thesis would have been far from being accomplished.

To Matteo, most especially, for the wisdom, for teaching me to see the world in a different light and for introducing me to the teachings of Nichiren Buddhism which have given my life a sense of purpose and direction.

To Prof. Thierry Mora and Prof. Amos Maritan, for taking their time to examine my thesis and raise thought-provoking comments.

To my friends and colleagues in the International Center for Theoretical Physics in Trieste: Erica, Antonio, Valerio, Alberto, Matteo, Asja, Clelia, Anjan, Sophie, Andrea, Claudio and Florence, for making Trieste and ICTP a place I can always call home.

To my friends and colleagues in the Kavli Institute for Systems Neuroscience in Trondheim, Norway: Claudia, Nicola, Ben, Daniel, Stojan, Piero, Mauro, Ivan, Soledad, Tuce, Bartul and Karoline, for the company which made Trondheim a warmer place that it usually is.

To Silvia and Ariel, for those times that you have helped me to understand $\hat{H}[k]$.

To Martine, Elise, Erling and Anne, for providing me not just a place to stay in Trondheim but a home.

To Olav and Martine, most especially, for the companionship, the fun conversations and the songs from the second floor.

To my parents Eden and Jun and my parents-in-law Paola and Piero, for your never-ending support and understanding.

And to Lorenzo, for keeping my esteem high when everything seems so wrong. You have become an anchor to my sanity and I will always be grateful for sharing the rest of my life with you.

- Ryan

This page has been intentionally left blank.

Contents

List of Commonly-Used Abbreviations

ADn	anterodorsal nucleus
AEP	Asymptotic Equipartition theorem
CD	Contrastive Divergence
HD	head direction
IB	Information Bottleneck
IN	informative neuron
ISI	interspike interval
LVN	locally-variating neuron
MCMC	Markov chain Monte Carlo
MDL	Minimum Description Length
mEC	medial entorhinal cortex
MSR	multiscale relevance
NML	normalized maximum likelihood
ON	overlapping neuron
PoS	Post-subiculum
RBM	restricted Boltzmann machine
RN	relevant neuron
SK	Sherrington-Kirkpatrick

This page has been intentionally left blank.

1

Introduction

Our current understanding of Nature has been borne out of our ability to collect data from observations and controlled experiments. In fact, our understanding of the physical principles that govern matter and its interactions has been deepened thanks to highly controlled experiments where Nature is constrained to a point where outcomes depend only on a few variables. This has allowed us to discover relations that can be described using elegant mathematical models. However, such level of control is absent in data coming from observations of Nature itself. From the genome of an organism, the species composition of an ecosystems to the neural activity in the brain – such observations are not an outcome of any controlled experiment. We see living systems as they naturally are – noisy and complex. Yet, these uncontrollable experiments gave rise to a striking universal phenomenon – the emergence of broad distributions.

Indeed, it is not infrequent to find empirical data which exhibits broad frequency distributions even in the most disparate domains. Broad distributions manifest in the fact that if outcomes are ranked in order of decreasing frequency of their occurrence, then the rank frequency plot spans several orders of magnitude on both axes. Fig. 1.1 reports few cases (see caption for details), but many more have been reported in the literature (see e.g. [1, 2, 3, 4, 5, 6, 7]). A straight line in the rank plot corresponds to a power law frequency distribution, where the number of outcomes that are observed k times behave as $m_k \sim k^{-\mu-1}$ (with $1/\mu$ being the slope of the rank plot). Yet, as Fig. 1.1 shows, empirical distributions are not always power laws, even though they are broad nonetheless.

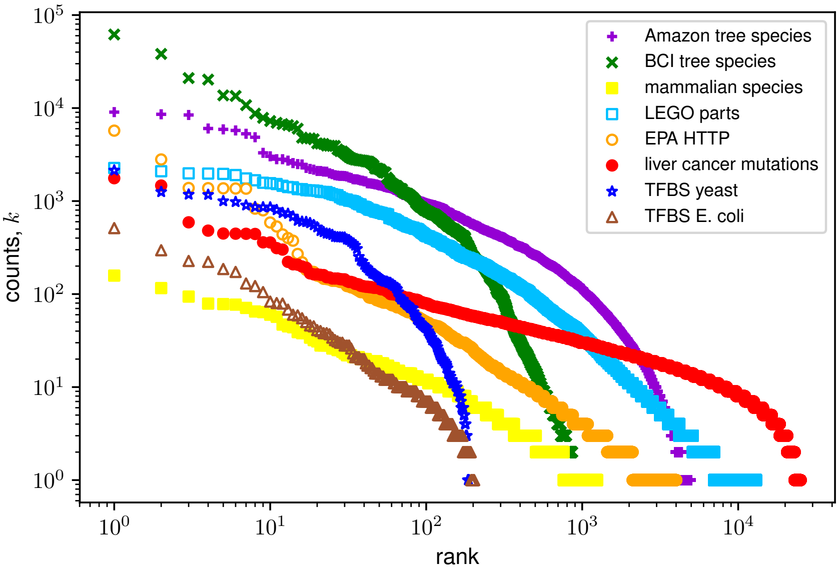

FIGURE 1.1. **Rank plot of the frequencies across a broad range of datasets.** Log-log plots of rank versus frequency from diverse datasets: survey of 4,962 species of trees sampled from the Amazonian lowlands [8], survey of 1,053 species of trees sampled across a 50 hectare plot in the Barro Colorado Island (BCI), Panama [9], the number of species across 1,247 order of mammals that have existed in the last few tens of thousands of years as compiled in Ref. [10], counts indicating the inclusion of each 13,001 LEGO parts on 2,613 distributed toy sets [11], the number of bytes of data received as a result of 4,011 web requests (HTTP) from the Environmental Protection Agency WWW server from 29-30 August 1995 [12], the number of observed mutations across 25,137 genes from patients with liver cancer taken from the Catalogue of Somatic Mutations in Cancer (COSMIC) [13] and the number of genes that are regulated by each of the 203 transcription factors (TFs) in E. coli [14] and 188 TFs in *S. cerevisiae* (yeast) [15] through binding with transcription factor binding sites (TFBS).

The emergence of power laws in empirical observations of natural systems is reminiscent of an underlying critical phenomena which is well-understood in physical systems. Perhaps, one of the great lessons from statistical physics is that the whole can be more than the sum of its parts. Indeed, no matter the underlying details of the physical system, criticality and scale invariance emerge through a collective behavior of the physical system's components which depends only on a few essential attributes such as the dimensionality and the underlying symmetries governing the system. For such physical systems, the critical point can be reached by tuning the control parameters to the right settings. Upon reaching the critical point, the physical system is found at phase transition point separating two regimes: a regime where noise dominates and another regime where order dominates.

However, in natural systems, while complex phenomena emerges from the interaction between many dynamical components (or variables), the situation can be rather different as the system is inherently not an equilibrium statistical system and the control parameters may not be well-defined from the onset. Nevertheless, an emerging hypothesis [16], inspired by statistical mechanics, has been put forward which argues that the probability of finding a natural system in some particular configuration is mathematically equivalent to that of an equilibrium statistical distribution which sits at a critical point, i.e., that the natural system is fine-tuned to operate near the critical point. Evidences supporting this hypothesis have been shown through a series of studies in many natural systems [16, 17, 18, 19, 20]. Countless mechanisms have also been advanced explaining this behavior (see e.g. Refs. [1, 2, 21, 22, 23, 24]). However, the fact that broad distributions are so widespread across disparate systems suggests that a universal mechanism with which the system is able to find itself in a critical point may not exist just as much as microscopic details giving rise to signatures of criticality in physical systems are not unique.

Furthermore, critical points in models estimated from empirical data are surprisingly "special" in the sense that they occupy a small region of the parameter space [25]. Hence, it is rather surprising to find these systems sitting in these regions. As such, counter-arguments have been put forward which does not require the fine-tuning of parameters [26, 27, 28, 29]. These hypothesis trace the origin of statistical criticality to relevant variables in the system that are not observed but are nonetheless interacting with the ones that are visible.

In this thesis, the central question that we shall address is that of the ubiquity of broad distributions found in natural systems. The main hypothesis that we will defend in what shall follow is that it is possible to trace an information theoretic origin of broad distributions that is independent of any specific mechanism. In particular, we will argue that broad distributions arise when the data is mostly informative about the underlying generative process, i.e., the data is expressed in terms of the *relevant variables*. This idea has been suggested by Marsili, Mastromatteo and Roudi (2013) [30] and was later grounded by Haimovici and Marsili (2015) [31] on the basis of Bayesian model selection. In this thesis, however, we will give an even more solid foundation to this idea in a purely information theoretic setting. This hypothesis rests on the assumption that the only information we have is the frequencies (or the fraction of times) with which different outcomes are observed in the sample. Nonetheless, if this hypothesis is correct, then we should be able to find broad distributions when the samples are efficiently represented. Furthermore, we should be able to use this principle to extract efficient representations, i.e., relevant variables, in high-dimensional data. These are the lines along which the thesis shall progress.

More precisely, we shall show in Chapter 3 that broad distribution arises from an efficient representation of the data, i.e., when the samples are maximally informative about how the data was generated. We shall see that the construction allows us to measure the information content of a sample in a setting where the data generating process is unknown and the only information we have is the empirical frequency distribution. Furthermore, this construction will allow us to characterize the samples that maximize this information content. We shall clarify this finding by relating it to other approaches such as point estimate statistics, information bottleneck method and asymptotic equipartition principle. We shall also see that the large deviation of the mostly informative samples are also mostly informative at a different resolution with which we see the data.

If this construction is correct, then we should be able to find samples as maximally informative when they are represented in an efficient manner. In Chapter 4, we consider a framework called the *minimum description length* which allows for efficient representation of samples. By studying different statistical models (Dirichlet model, independent and pairwise dependent spin models, and restricted Boltzmann machines), we shall show that the samples that achieve optimal compression have broad distributions and are mostly informative. Also, looking at large deviation properties of such efficiently represented samples will allow us to provide a clear understanding of the origin of criticality in a very precise way.

Furthermore, if our hypothesis is correct, then we should be able to use our construction as guiding principles to find neurons from different brain regions of rodents whose representations (responses with respect to some external correlate, e.g., position or head direction) are relevant to spatial navigation as done in Chapter 5. We shall see that our construction allows to identify *relevant neurons* which are only those whose responses are not only informative about the navigational correlate but allow for efficient decoding of such relevant external correlate.

While each chapter will have its own conclusion, we shall close this thesis with a synthesis of what we have shown and provide directions for future investigations in the Conclusions and present all the basic notions, pertinent derivations and research methods in the Appendix.

This page has been intentionally left blank.

2

Statistical criticality: A survey

In this chapter, we shall first introduce the general setup of the problem that we shall address in the succeeding chapters. Doing so will allow us to discuss the concept of *statistical criticality* – the emergence of power law frequency distributions in empirical data – and different hypotheses describing its origin.

2.1 Setting up the scene

Throughout this thesis, we shall consider a sample (or data) $\hat{s} = (s^{(1)}, \ldots, s^{(N)})$ of N observations where each data point $s^{(\ell)}$ are drawn from a countable and finite set χ. Here, we assume that each observation $s^{(\ell)}$ in the sample are outcomes of N independent experiments which are done in the same conditions, so the order of the observations is irrelevant[1]. Mathematically, this is equivalent to each s as being independent and identically distributed draws from an unknown distribution $p(s)$[2]. This sampling procedure then defines a *generative process*.

[1] (ℓ) indexes the observation in the sample. We shall drop this notation whenever it is not necessary.

[2] An intuitive discussion about sampling is presented in Chapters 3 and 9 of Ref. [32]. In brief, the setup is mathematically related to sampling balls labeled by S integers with replacement from an urn with a countably large number of balls. The labels s, when we have a very large sample, have a limiting distribution $p(s)$. Hence, whenever we say "drawing from a distribution", we shall have the urn analogy in mind.

In some cases, the observation $\boldsymbol{s}$ can be multivariate. In this case, each observation can be described as n-dimensional vector, i.e., $\boldsymbol{s} = (s_1, \ldots, s_n)$ where each element s_i can either be a spin variable ($s_i = \pm 1$) or a binary variable ($s_i = 0, 1$) which can interact with each other.

In any case, for a given sample, $\hat{s} = (s^{(1)}, \ldots, s^{(N)})$, we shall define k_s as the number of observations in $\hat{s}$ for which $s^{(\ell)} = s$, i.e., the *frequency* of s. The number of states s that occur k times will be called as the *degeneracy* denoted by m_k. Of course, both k_s and m_k depend on the sample $\hat{s}$ and follow the obvious relation $\sum_k k m_k = N$. Furthermore, one can obtain an empirical estimate of a function $g(s)$ of the observation from the sample. Such an empirical estimate shall be denoted by a hat, i.e., $\hat{g}(s)$. For example, the empirical estimate of the distribution is given by $\hat{p}(s) = k_s/N$ which tends to $p(s)$ when $N \to \infty$.

2.2 Statistical mechanics of samples

It has been argued (see for instance Refs. [16, 20]) that it is possible to construct a statistical mechanics approach of a single sample. The core idea is to construct a distribution $q(s)$ such that it matches the statistics of the empirical data, i.e., $q(s) = \hat{p}(s)$. This construction allows us to make a connection between samples and statistical physics models and therefore, criticality.

To make this point clearer and more pedagogical, let us consider, without loss of generality, a system comprising n spin variables where each observation[3] (or spin configuration or spin state) $\boldsymbol{s} = (s_1, \ldots, s_n)$ is drawn from a distribution $p(\boldsymbol{s})$. With this, one can choose some functions $\mathcal{T}_\mu(\boldsymbol{s})$ of the spin variables $\boldsymbol{s}$ called observables or *statistics* for which the empirical averages $\hat{\mathcal{T}}_\mu(\boldsymbol{s})$ can be measured from the sample[4].

With the assumed choice of the statistics $\{\mathcal{T}_\mu(\boldsymbol{s})\}$, one then constructs the distri-

[3] Here, we note that the observations $\boldsymbol{s} \in \chi$ and the size of the state space is $S = 2^n$.

[4] For example, one can consider the spin $\mathcal{T}_i(\boldsymbol{s}) = s_i$ (with n of such statistics) or product of two different spins $\mathcal{T}_{i,j}(\boldsymbol{s}) = s_i s_j$ (with $\frac{n(n-1)}{2}$ of such statistics) and thus, one realizes that the empirical averages, $\hat{\mathcal{T}}_i(\boldsymbol{s}) = \frac{1}{N}\sum_{\ell=1}^{N} s_i^{(\ell)}$ and $\hat{\mathcal{T}}_{i,j}(\boldsymbol{s}) = \frac{1}{N}\sum_{\ell=1}^{N} s_i^{(\ell)} s_j^{(\ell)}$, are the mean and correlation of the spin variables respectively.

bution $q(s)$ which does not incorporate any other assumption of information that we do not have. This is done through the principle of maximum entropy [33, 34]. In the maximum entropy approach, one finds the distribution $q(\boldsymbol{s})$ which maximizes the entropy

$$H[\boldsymbol{s}] = -\sum_{\boldsymbol{s}\in\chi} q(\boldsymbol{s}) \log q(\boldsymbol{s}) \tag{2.1}$$

constrained to reproduce the empirical averages of the statistics

$$\hat{\mathcal{T}}_\mu(\boldsymbol{s}) = \langle \mathcal{T}_\mu(\boldsymbol{s})\rangle_q \equiv \sum_{\boldsymbol{s}\in\chi} \mathcal{T}_\mu(\boldsymbol{s}) q(\boldsymbol{s}). \tag{2.2}$$

This amounts to setting the derivatives of the functional

$$\mathcal{F} = -\sum_{s\in\chi} q(\boldsymbol{s}) \log q(\boldsymbol{s}) + \lambda_0 \left[\sum_{\boldsymbol{s}\in\chi} q(\boldsymbol{s}) - 1\right] + \sum_\mu \lambda_\mu \left[\sum_{\boldsymbol{s}\in\chi} \mathcal{T}_\mu(\boldsymbol{s}) q(\boldsymbol{s}) - \hat{\mathcal{T}}_\mu(\boldsymbol{s})\rangle\right] \tag{2.3}$$

with respect to the distribution $q(\boldsymbol{s})$ to zero which yields the solution

$$q(\boldsymbol{s}) = \frac{\exp\left(-E_{\boldsymbol{s}}\right)}{Z} \tag{2.4}$$

where $E_{\boldsymbol{s}} = \sum_\mu \lambda_\mu \mathcal{T}_\mu(\boldsymbol{s})$ and $Z = \sum_{\boldsymbol{s}\in\chi} \exp(-E_{\boldsymbol{s}})$ is a normalization constant[5].

Now, introducing a parameter β to the distribution $q(\boldsymbol{s})$ in Eq. (2.4), i.e.,

$$q_\beta(\boldsymbol{s}) = \frac{\exp\left(-\beta E_{\boldsymbol{s}}\right)}{Z(\beta)}, \tag{2.5}$$

allows one to “tilt” the distribution of a configuration $\boldsymbol{s}$. In this form, the distribution resembles the Gibbs-Boltzmann distribution with $E_{\boldsymbol{s}}$ as the energy, $\beta = 1/T$ is the

[5] For example, most cases in literature (see Ref. [16] and the examples therein for examples) choose the observables to be the first $\mathcal{T}_i(\boldsymbol{s}) = s_i$ and second moments $\mathcal{T}_{i,j}(\boldsymbol{s}) = s_i s_j$ of the spin variables. In this case, the maximum entropy distribution can be written as

$$q(s) = \frac{\exp\left(\sum_i h_i s_i + \sum_{i,j} J_{i,j} s_i s_j\right)}{Z}$$

with h_i as a bias parameter in the spin s_i and $J_{i,j}$ as an effective interaction parameter between spins s_i and s_j.

inverse temperature and

$$Z(\beta) = \sum_{\boldsymbol{s}\in\chi} \exp\left(-\beta E_{\boldsymbol{s}}\right) \tag{2.6}$$

is the partition funtion which normalizes the distribution in Eq. (2.5). Ref. [16] took this distribution as a statistical mechanics problem[6] and compute the specific heat in the usual way.

With the formulation in Eq. (2.5), one can calculate the energy $E_{\boldsymbol{s}}$ of each configuration $\boldsymbol{s}$ and aggregate each configuration into energy levels. In particular, the partition function in Eq. (2.6) can be expressed as

$$\begin{aligned} Z(\beta) &= \sum_{\boldsymbol{s}\in\chi} \left[\int dE\delta(E - E_{\boldsymbol{s}})\right] \exp\left(-\beta E\right) &(2.7)\\ &= \int dE \left[\sum_{\boldsymbol{s}\in\chi} \delta(E - E_{\boldsymbol{s}})\right] \exp\left(-\beta E\right) &(2.8)\\ &= \int dE W(E) \exp\left(-\beta E\right) &(2.9) \end{aligned}$$

where $W(E) = \sum_{\boldsymbol{s}\in\chi} \delta(E - E_{\boldsymbol{s}})$ corresponds to the degeneracy at each energy level E. Thus, one can define a Boltzmann entropy as

$$S_B(E) = \log W(E) \tag{2.10}$$

whose average corresponds to the entropy in Eq. (2.1). Here, we shall assume that the thermodynamic limit where $n \to \infty$ exists [16, 20] where both the energy E and the entropy $S_B(E)$ scales with n[7]. In this limit, the energy levels become dense and thus, instead of calculating the degeneracy $W(E)$ in each level, one can instead consider

[6] With this construction, by varying β, one can interpolate between low and high temperature phase. At infinite temperature, $\beta = 0$, all configurations are equally probable while at zero temperature, $\beta \to \infty$, the distribution concentrates on configurations that minimizes the energy $E_{\boldsymbol{s}}$ i.e., the ground state configurations which, depending on the details of the system, can be hard to define. Notice that when $\beta = 1$, one recovers the distribution estimated from the sample.

[7] The existence of this limit is never always guaranteed. If this limit exists, this corresponds to making the size S of the state space χ very, very large.

the number of configurations $\mathscr{W}(E)$ with energy less than E, i.e.,

$$\mathscr{W}(E) = \sum_{\boldsymbol{s}\in\chi} \Theta(E - E_{\boldsymbol{s}}) \tag{2.11}$$

where $\Theta(x)$ is the Heaviside step function[8]. Since the step function is the integral of the delta function, one can integrate the partition function in Eq. (2.9) by parts which results to

$$\begin{aligned} Z(\beta) &= \beta \int dE \mathscr{W}(E) e^{-\beta E} && (2.12)\\ &= \beta \int dE e^{-\beta E + \log \mathscr{W}(E)} && (2.13)\\ &= \beta \int dE e^{-\beta E + \mathscr{S}_B(E)} && (2.14) \end{aligned}$$

where $\mathscr{S}_B(E) = \log \mathscr{W}(E)$ takes the form of an entropy[9]. Notice that the energy E and the entropy $\mathscr{S}_B(E)$ are both extensive quantities and hence, we can define the energy and entropy per spin as $\epsilon = E/n$ and $\mathscr{s}(\epsilon) = \mathscr{S}_B(E)/n$. With this, we can write the partition function as

$$Z(\beta) = n\beta \int d\epsilon e^{-n\phi(\epsilon)} \tag{2.15}$$

where

$$\phi(\epsilon) = \beta\epsilon - \mathscr{s}(\epsilon) \tag{2.16}$$

is the free energy. Notice that in the thermodynamic limit, the partition function in Eq. (2.15) will be dominated by the value ϵ^* for which $\phi(\epsilon^*)$ is a saddle point which

[8] The Heaviside step function $\Theta(x)$ is defined as

$$\Theta(x) = \begin{cases} 1 & x > 0 \\ 0 & x < 0 \end{cases}$$

[9] Note that when the size n of the system is small, $W(E)$ and $\mathscr{W}(E)$ are both badly behaving functions, i.e., $W(E)$ is singular while $\mathscr{W}(E)$ has visible steps. However, in the thermodynamic limit, both functions become smooth. In this limit, it is better to work with $\mathscr{W}(E)$ since we avoid the problem of properly binning the energy E.

is given by the condition

$$\left.\frac{\partial\phi(\epsilon)}{\partial\epsilon}\right|_{\epsilon=\epsilon^*} = 0 \Longrightarrow \frac{d\mathscr{s}(\epsilon^*)}{d\epsilon^*} = \beta \geq 0. \tag{2.17}$$

And thus, we can expand the free energy $\phi(\epsilon)$ around ϵ^* to find

$$\phi(\epsilon) = \phi(\epsilon^*) - \frac{1}{2}\frac{d^2\mathscr{s}(\epsilon)}{d\epsilon^2}(\epsilon-\epsilon^*)^2 + \cdots \tag{2.18}$$

with which the partition function can now be re-cast as

$$Z(\beta) \simeq n\beta e^{-n\phi(\epsilon^*)} \int d\epsilon \exp\left[\frac{n}{2}\frac{d^2\mathscr{s}(\epsilon)}{d\epsilon^2}(\epsilon-\epsilon^*)^2\right]. \tag{2.19}$$

With this, one can think of the energy per spin as being drawn from (approximately) a Gaussian distribution with mean ϵ^* and variance

$$\left\langle(\delta\epsilon)^2\right\rangle = \frac{1}{n}\left[-\frac{d^2\mathscr{s}(\epsilon)}{d\epsilon^2}\right]^{-1} \tag{2.20}$$

which is related to the specific heat by the fluctuation-dissipation theorem as

$$c(\beta) = \frac{1}{\beta^2}\left\langle(\delta\epsilon)^2\right\rangle. \tag{2.21}$$

In this construction, the parameter $\beta = 1$ is an extremely special point. First, $\beta = 1$ corresponds to a point where we recover the distribution $q(\boldsymbol{s})$ in Eq. (2.4), i.e., the point which matches the chosen statistics of the empirical data. At this point, Ref. [20] argues that the free energy in Eq. (2.16) becomes $\phi(\epsilon) = \epsilon - \mathscr{s}(\epsilon)$ and thus, the configurations $\boldsymbol{s}$ with energy ϵ^* that dominate the integral in Eq. (2.15), i.e. where the free energy $\phi(\epsilon)$ is at a minimum, obey the relation

$$\frac{d\mathscr{s}(\epsilon^*)}{d\epsilon^*} = 1 \tag{2.22}$$

implying that with such configurations, the energy ϵ^* and entropy $\mathscr{s}(\epsilon^*)$ have a linear relationship. The concept of configurations dominating the integral in Eq. (2.15) introduces the concept of “typicality”. By the Asymptotic Equipartition Principle (AEP), at large n, such dominating states are observed with the same log probability. However, in this case, the curvature of the free energy vanishes, i.e., $\frac{d^2 s(\epsilon^*)}{d\epsilon^{*2}} = 0$ and thus, the usual ideas of “typicality” breaks down resulting to large variations in the observed log probability which is mathematically equivalent to a diverging specific heat. Indeed, because the second derivative $\frac{d^2 s(\epsilon)}{d\epsilon^2} = 0$, then following Eq. (2.21), the specific heat $c(\beta = 1)$ diverges. These are clear signatures of a second order phase transition and $\beta = 1$ is a *critical point*[10]. Finally, $\beta = 1$ corresponds to the well-celebrated Zipf’s law [35] which can be written as[11]

$$m_k \sim k^{-2} \tag{2.23}$$

where m_k is the degeneracy of finding a state with frequency k in the sample $\hat{s}$.

As it turned out, the critical point for diverse natural systems [17, 18, 19, 16, 20] where power law frequency distributions have been observed sits at $\beta \approx 1$. This allowed Ref. [16] to suggest that such systems exhibit statistical criticality because they operate close to a critical point. A number of dynamical mechanisms have been put forward to explain the emergence of criticality through fine-tuning of parameters [21, 24, 36]. The most notable one is that in Ref. [23] wherein a collection of living systems “communicate” to each other resulting to adaptation or evolution in order to cope in a complex and changing environment from which self-tuning to criticality is attained. In particular, each individual constructs an “internal representation” of the environment with a trade-off between accuracy and flexibility so as to best respond to

[10] The construction above can be extended to non-equilibrium systems as shown in Ref. [20] where one does not have a notion of a partition function. Here, one starts by assigning an energy to every configuration $\boldsymbol{s}$ as the negative logarithm of the distribution $q(\boldsymbol{s})$. One employs similar calculations as above and eventually, one can make the correspondence with an equilibrium thermodynamic system with which the divergence of the specific heat is a consequence of the large fluctuations in the log-probability of the configurations.

[11] To see this, consider a sample $\hat{s}$ and the empirical distributions $\{\hat{p}(\boldsymbol{s})\}$. One can define the energy $E_{\boldsymbol{s}} = -\log \hat{p}(\boldsymbol{s})$ as in Ref. [16] and the entropy $S_B(E_{\boldsymbol{s}}) = \log W(E_{\boldsymbol{s}})$ as in Eq. (2.10) where $W(E_{\boldsymbol{s}})$ is the number of observations $\boldsymbol{s}^{(\ell)}$ in $\hat{s}$ having energy $E_{\boldsymbol{s}}$. Thus, the energy of the system is $E_{\boldsymbol{s}} = -\log(k/N)$ while $W(E_{\boldsymbol{s}}) = k m_k$ and consequently, $S_B(E_{\boldsymbol{s}}) = \log(k m_k)$. Thus, the linear energy-entropy relationship results to a Zipf’s law for the multiplicities as in Eq. (2.23).

the external conditions.

However, finding the system at a critical point as discussed above has been argued to be surprisingly “special”. Indeed, Mastromatteo and Marsili (2011) [25] observed that, at the critical point, inferred models concentrate on a small region of the parameter space. This entails that such critical systems have to constantly fine-tune its parameters to stay within the critical region.

Counter-arguments have since been presented which traces the origin of statistical criticality without the need for fine-tuning of parameters. In Ref. [26], they have shown that a linear energy-entropy relationship from which a power law distribution emerges can be attained when a relevant hidden variables that is interacting with the observed variables (or spins) of the system is marginalized out of the system. In particular, they have argued that if the observed system is adapted to the variations of the hidden variables, i.e., if the variations of the hidden variable is broad but not necessarily a power law, then statistical criticality emerges in the thermodynamic limit. This result has been generalized in Refs. [27, 28] which shows that a broad distribution of energies (i.e., negative logarithm of the distribution) and consequently, a large specific heat can emerge under fairly general circumstances of high-dimensional hidden variable models. This construction of statistical criticality due to hidden variables has been shown to hold in a neuron model with common input [37] as well as in several empirical data [27, 28]. Of noteworthy is Ref. [29] where they have shown that signatures of criticality can arise solely by subsampling a correlated system. While there are still no concrete mechanical mechanisms of attaining criticality with hidden variables, a number of mechanisms have been put forward that attain criticality without the need of parameter fine-tuning [38, 39].

The capacity of a system to perform complex computation has been conjectured to be enhanced when operating near a critical point separating a noise-dominated region where information transmission and storage is corrupted and an order-dominated region where information modification (i.e., adaptations) due to changing environment are hindered [40, 41]. As such, natural systems should be able to efficiently represent the external environment [23] allowing for a trade-off between accuracy – how detailed the environment should be represented – and flexibility – how complex the resulting representation is. Hence, it is only natural to look for origins of statistical criticality in such representations. In the next chapters, we shall contribute to this debate on the origin of statistical criticality by following information theoretic arguments which do not need any mechanism of parameter fine-tuning.

This page has been intentionally left blank.

3

Efficient representations exhibit statistical criticality

When the sample is generated as independent draws from a parametric distribution, $f(s|\theta)$, one can draw a sharp distinction between what is noise and what is *useful* information, i.e., that part of the data that can be used to estimate the generative model through the maximum entropy principle as discussed in Chapter 2.2. Useful information in concentrated in (minimal) sufficient statistics $\hat{\mathcal{T}}(s)$ [42], which are those variables whose empirical values suffice to fully estimate the model's parameter [33, 34]. This chapter aims at drawing the same distinction in the case where the data is discrete and the model is not known. In this case, we show that the information on the generative model is contained in the empirical distribution of frequencies, i.e. the fraction of times different outcomes are observed in the sample. Hence, of the total information contained in the sample, which is quantified in the entropy of the distribution of outcomes, the amount of useful information is quantified by the entropy of the frequency distribution. As argued in Ref. [31], while the former entropy is a measure of *resolution*[1]. – i.e. of the number of bits needed to encode a singe observation – the latter

[1] For example, stocks in the financial market can be classified by their SIC (Standard Industrial Classification) code using different number of digits, gene sequences can be defined in terms of the

quantifies *relevance* – i.e. maximal number of bits each observation carries about the generative model.

The first aim of this chapter is to present a simple derivation of this result. We look for a model-free characterization of a sample similar to that offered by sufficient statistics in the case of parametric models. We do this by searching for *minimally sufficient representations* in terms of hidden features, requiring that conditional to the latter, the sample contains no more information to estimate the generative model. This identifies the frequencies as the hidden features and the entropy of the frequency distribution – i.e. the relevance – as the measure of information content of the sample.

As shown in Refs. [30, 31], samples that maximize the relevance at a fixed resolution – that we shall call *maximally informative samples* henceforth – exhibit *statistical criticality* [16]. This implies that the number of outcomes that are observed k times in the sample behaves as

$$m_k \sim k^{-\mu-1}. \tag{3.1}$$

Here, the exponent μ encodes the trade-off between resolution and relevance: a decrease of one bit in resolution translates to an increase of μ bits in relevance. Hence, the case $\mu = 1$, which corresponds to the celebrated Zipf's law [35], encodes the *optimal trade-off*, since further decrease in resolution delivers an increase in relevance that is not compensated by an information loss.

It is interesting that several systems that are meant to encode efficient representations follow Zipf's law. This is the case, for example, for the frequency of words in language [35], the antibody binding site sequences in the immune system [43, 18] and spike patterns of population of neurons [20]. In each of this cases, the r^{th} most frequent value of the relevant variables (words, binding site sequences or spike patterns) occurs with frequency proportional to $1/r$.

The finding that maximally informative samples exhibit statistical criticality, suggests [30, 31] that broad distributions arise from an optimization principle of maximal

sequence of the bases or in terms of the sequence of amino acids they code for, organisms can be classified based on different taxonomic levels, etc relevance, i.e. that the data is expressed in terms of relevant variables. For example, the distribution of a population within a country exhibits a broad distribution when expressed in terms of city sizes [38], but a much narrower one when expressed in terms of ZIP codes. The same principle may explain why broad frequency distributions occur in other systems, such as natural images [44] or the distribution of firms by SIC codes [45]. Furthermore, this also opens the way to attempt at extracting relevant variables from high dimensional data, as e.g. in [46, 47] and in Chapter 5, or to shed light on the principles underlying deep learning [48].[2]

The second aim of this chapter is to clarify the relation of the findings above with other approaches. First, we show that maximally informative samples can be derived from the Information Bottleneck [51] like approach, when the frequency is taken as the output variable. We, then, use the thermodynamic analogy of samples as done in Chapter 2.2 and proposed in Ref. [16, 17] to show that if a typical samples is maximally informative, the large deviation in the information cost also correspond to maximally informative samples. Furthermore, Zipf's law arises as the most likely maximally informative sample, under a non-informative prior. Finally, we comment on the relation between our results and those of Refs. [26] and [28] that propose necessary conditions for the occurrence of Zipf's law and conclude with few general comments.

3.1 Parametric models: Sufficiency

Throughout this chapter, we shall be interested in addressing the question of *optimal data reduction*. However, before we delve into the case of a model-free setting, i.e., when $p(s)$ is unknown, we first consider the instance when the distribution $p(s) = f(s|\theta)$ were a known family of distribution parametrized by θ. In this context, data

[2] In Ref. [48], they have trained deep belief networks [49] (a network architechture similar in construction to restricted Boltzmann machines as in Chapter 4.2.3 but with multiple hidden layers) on the MNIST data [50]. They have found that deep learning maps the data into a hierarchy of maximally informative samples at different resolution, depending on the depth of the hidden layer. In addition, the layer whose distribution is closest to Zipf's law was found to exhibit the best generation performance, i.e., the statistics of the generated samples follow that of the training samples.

reduction amounts to asking that any estimation on the parameter θ should depend on the sample $\hat{s}$ only through some summarization of the data.

If $\hat{\mathcal{T}}(s)$ is a function (or a *statistic*) of the sample, then we can construct a Markov chain $\theta \to \hat{s} \to \hat{\mathcal{T}}(s)$ which describes a generating process. With this, we mean that we can generate the sample $\hat{s}$ from the parametric distribution $f(s|\theta)$ and then, we generate the statistic $T(\hat{s})$ from $\hat{s}$. By the data processing inequality (see Appendix A.1.4) which states that no clever manipulation of the sample $\hat{s}$ can increase the information content of $\hat{s}$, we have that

$$I(\theta, \hat{\mathcal{T}}(s)) \leq I(\theta, \hat{s}). \tag{3.2}$$

If this equality holds, then no information about the sample is lost and the statistic $\hat{\mathcal{T}}(s)$ provides a summary of the generating process, i.e., $\hat{\mathcal{T}}(s)$ can be used to estimate the parameter θ as well as generate the sample $\hat{s}$. Indeed, we say that the statistic $\hat{\mathcal{T}}(s)$ is *sufficient* for the parameter θ if it contains all the information in the sample $\hat{s}$ about the parameter θ[3]. Equivalently, this tells us that, conditional on $\hat{\mathcal{T}}(s)$, the sample $\hat{s}$ does not contain any information on the parameter θ, i.e., the conditional distribution $f(\hat{s}|\hat{\mathcal{T}}(s) = t)$ is independent of θ and we can write $f(\hat{s}|\theta) = g(\hat{\mathcal{T}}(s)|\theta)p(\hat{s})$ for some functions g and p. This means that since the conditional distribution no longer depends on the parameter θ, then the particular values of the sample $\hat{s}$ do not constrain the parameter θ in any way. This said, the sufficient statistic $\hat{\mathcal{T}}(s)$ provides a representation of θ in the sample $\hat{s}$ and optimal statistical estimations can then be designed around $\hat{\mathcal{T}}(s)$ (as is done with the maximum entropy principle in Chapter 2.2).

While the sufficient statistic $\hat{\mathcal{T}}(s)$ provides an efficient representation through a dimensional reduction, $\hat{\mathcal{T}}(s)$ may not be an *optimal* data representation. Indeed, any function of the sufficient statistic is also a sufficient statistic. Hence, if a sufficient statistic $\hat{\mathcal{T}}^*(s)$ is a function of *every* other sufficient statistic $\hat{\mathcal{T}}(s)$, then $\hat{\mathcal{T}}^*(s)$ is a *minimally sufficient statistic* relative to the parametric distribution $f(s|\theta)$. This implies

[3] Take as an example a sample $\hat{s}$ where each of the N observation $s^{(i)}$ are independently drawn from $f(s|\theta)$. Then, $\hat{\mathcal{T}}(s) = \sum_{i=1}^{N} s^{(i)}$ is a sufficient statistic for the following discrete parametric distributions: *(i)* a Bernoulli distribution, $f(s|\theta) = \theta^s(1-\theta)^{1-s}$, $s \in 0, 1$, with parameter θ and *(ii)* a Poisson distribution, $f(s|\theta) = \frac{\theta^s e^{-\theta}}{s!}$, $s \in \{0, 1, \ldots\}$, with parameter θ.

that we can construct a Markov chain

$$\theta \to \hat{\mathcal{T}}^*(s) \to \hat{\mathcal{T}}(s) \to s \tag{3.3}$$

of the data generating process. Thus, a minimal sufficient statistic provides an *optimal representation* about θ in the sample $\hat{s}$.

The minimal sufficient statistic has an intuitive explanation in terms of partitioning the observation space χ. In particular, a sufficient statistic $\hat{\mathcal{T}}(s)$ can be realized as a labeling of disjoint partitions of χ such that the observations s and s' belong to the same partition if and only if $\hat{\mathcal{T}}(s) = \hat{\mathcal{T}}(s')$. A minimal sufficient statistic is then a sufficient statistic which provides the *coarsest possible partitioning* of the observation space χ[4]. That is, given a partitioning of χ corresponding to a minimal sufficient statistic $\hat{\mathcal{T}}^*(s)$, the statistic resulting from taking the union of two distinct partitions is no longer sufficient. Such a partitioning is called the *minimal sufficient partition* [52] and is unique even if the minimal sufficient statistic is not. As such, any sufficient statistic induces a partitioning which is a refinement of the minimal sufficient partition.

3.2 Minimally sufficient representations

In this section, we ask the same question of *optimal data reduction* in a model-free setting. By model-free, we mean that we assume there is a data generating process (see Fig. 3.1) but the underlying model which generated the data is unknown. The answer to this question also allows us to quantify the amount of information that a sample $\hat{s}$ contains on the generative process. This will allow us to ask, in the next section, which are the samples that are mostly informative, i.e. that contain a maximal amount of information on the generative process.

Here, we shall be guided by the principle that learning from the sample $\hat{s}$ is equiv-

[4] This can be done by defining a relation $\sim$ which is reflexive ($s \sim s$), symmetric ($s \sim s' \Leftrightarrow s' \sim s$) and transitive (if $s \sim s'$ and $s' \sim s''$, then $s \sim s''$) such that $s \sim s'$ for $s, s' \in \chi$ if and only if $\frac{f(s|\theta)}{f(s'|\theta)} = H(s, s')$, i.e., the likelihood ratio does not depend on the parameter θ for some function H. This relation $\sim$ then induces a partition in the observation space χ. Then, any statistic $\hat{\mathcal{T}}^*$ such that $\hat{\mathcal{T}}^*(s)$ is constant for all the observations s in a partition is a minimal sufficient statistic.

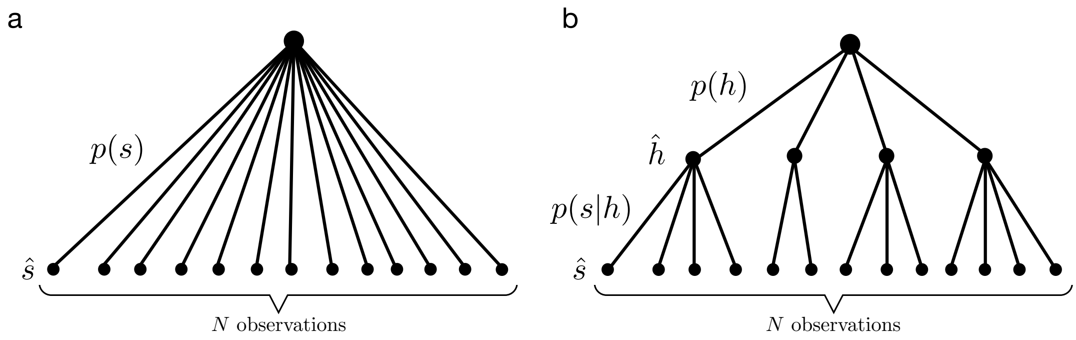

FIGURE 3.1. **An illustration of the data generating process when the model is unknown.** (**a**) $p(s)$ is the generative process. The points in the lowest layer represent the samples $\hat{s} = (s^{(1)}, \ldots, s^{(N)})$ of N observations which are independent draws from a distribution $p(s)$ with $s^{(i)} \in \chi$ and χ is a finite and countable alphabet. (**b**) $p(s|h)p(h)$ is the generative process. The sample $\hat{s}$ (represented by the points in the lowest layer) can also be obtained by first, generating a hidden feature $h^{(i)} \in \mathcal{H}$ which is drawn independently from a distribution $p(h)$ with $\mathcal{H}$ being a discrete set. Using the generated h, the value of $s^{(i)}$ can then be obtained as a draw from a distribution $p(s|h)$. Here, we have grouped together observations $s^{(i)}$ that were generated from the same hidden feature h (represented by the points in the middle layer).

alent to an optimal coding problem. This entails that encoding the sample $\hat{s}$ requires (see Appendix A.2.3)

$$\hat{H}[s] = -\sum_{s\in\chi} \hat{p}(s) \log \hat{p}(s) \tag{3.4}$$

$$= -\sum_{s\in\chi} \frac{k_s}{N} \log \frac{k_s}{N} \tag{3.5}$$

$$= -\sum_{k} \frac{k m_k}{N} \log \frac{k}{N}. \tag{3.6}$$

In order to address the first question, we search for a representation of the form

$$p(s) = \sum_{h\in\mathcal{H}} p(s|h)p(h) \tag{3.7}$$

in terms of a set of discrete features $h \in \mathcal{H}$. In this representation, $p(s|h)p(h)$ is the data generating process (see Fig. 3.1b). In other words, this amounts to thinking of the sample as being obtained by first, generating a *hidden feature* $h^{(i)}$ for each $i \in [1, N]$ independently from $p(h)$ and then, the value of $s^{(i)}$ from $p(s|h^{(i)})$. The representation should be

(i) accurate in the sense that it should accurately describe the statistics of the sam-

ple, i.e., the likelihood of the sample should be maximal. In addition to this, we require that

(i) the variable h should provide a *minimally sufficient* description of the data. This means that h should provide a concise representation of s in such a way that $p(s|h)$ should not contain any useful information on the generative process. This requirement amounts to requiring that conditional on h, $p(s|h)$ must satisfy the maximum entropy principle and should not depend on s. Such a requirement is an analog to ancillarity in statistics [42].

The question we address is, given a sample $\hat{s}$, what are the *hidden features* $\hat{h}$, i.e., what are $p(s|h)$ and $p(h)$ that satisfy these requirements?

The solution p^* to this problem can already be guessed from the outset by realizing that the likelihood is maximized when $p(s) = k_s/N$, i.e. the empirical distribution, where k_s is the number of points in the sample with $s^{(i)} = s$. This realization satisfies requirement *(i)*. Therefore, one can guess that the hidden feature is $h = k_s$. Indeed, the distribution of s conditional on $h = k_s$

$$p^*(s|k) = \frac{1}{m_k}\delta_{k,k_s} \tag{3.8}$$

does not depend on s and is a maximum entropy distribution, i.e., all observations s having the same frequency $k_s = k$ are assigned with equal probability. Here, m_k is the number of states for which $k_s = k$ and

$$p^*(k) = \frac{km_k}{N}. \tag{3.9}$$

It is instructive to derive this result through a direct calculation. In order to fulfill the requirement *(i)* above, we write

$$p(\hat{s}) = \sum_{h^{(1)},\dots,h^{(N)}} \exp\left\{\sum_{i=1}^{N} \log\left[p(s^{(i)}|h^{(i)})p(h^{(i)})\right]\right\}. \tag{3.10}$$

Notice that we can introduce a partition $\mathscr{G}$ of the integers $\{1,\ldots,N\}$ into $|\mathscr{H}|$ sets wherein $I_h = \{i : h^{(i)} = h\}$. In each partition I_h, there are $n_{s,h}$ observations for which $s^{(i)} = s$ and $h^{(i)} = h$. In doing so, we can express Eq. (3.10) as

$$p(\hat{s}) = \sum_{\mathscr{G}} \exp\left\{\sum_{s,h} n_{s,h} \log[p(s|h)p(h)]\right\} \tag{3.11}$$

$$= \sum_{\mathscr{G}} \exp\left[-N\hat{H}_{\mathscr{G}}[s,h] - ND_{KL}(\hat{p}_{\mathscr{G}}\|p)\right] \tag{3.12}$$

where the sum is taken on all possible partitions $\mathscr{G}$ of the indices $\{1,\ldots,N\}$[5]. $\hat{H}_{\mathscr{G}}[s,h]$ is the entropy of the empirical joint distribution[6]

$$\hat{p}_{\mathscr{G}}(s,h) = \frac{n_{s,h}}{N} \tag{3.13}$$

and $D_{KL}(\hat{p}_{\mathscr{G}}\|p)$ is the Kullback-Leibler distance between this empirical joint distribution and the generative process $p(s,h) = p(s|h)p(h)$.

Since $\hat{H}_{\mathscr{G}}[s,h] = \hat{H}[s] + \hat{H}_{\mathscr{G}}[h|s]$, the partitions that dominate the sum in Eq. (3.12) are those for which $h^{(i)} = h^*(s^{(i)})$ is a function of $s^{(i)}$, because then, $\hat{H}_{\mathscr{G}}[h|s] = 0$[7]. Hence, we have

$$\hat{H}_{\mathscr{G}}[s,h] = \hat{H}[s] = -\sum_s \frac{k_s}{N}\log\frac{k_s}{N} \tag{3.14}$$

and

$$\hat{p}_{\mathscr{G}}(s,h) = \frac{k_s}{N}\delta_{h,h^*(s)}. \tag{3.15}$$

Furthermore, the generative process, $p(s|h)p(h)$, that minimizes $D_{KL}(\hat{p}_{\mathscr{G}}\|p)$ is the one with $p^*(s|h)p^*(h) = \hat{p}_{\mathscr{G}}(s,h)$. Therefore, once a function $h^*(s)$ is singled out, only the term $\mathscr{G}^*$ that corresponds to this partition survives in the sum of Eq. (3.12). Since $D_{KL}(\hat{p}_{\mathscr{G}}\|p^*) = 0$, one recovers that the likelihood of the data attains indeed the maximal value $p^*(\hat{s}) = e^{-N\hat{H}[s]}$.

[5] For $N = 2$ and $h = 0, 1$, the possible partitions are $\mathscr{G} = (\{1,2\},\emptyset),\ (\{1\},\{2\}),(\{2\},\{1\})$ and $(\emptyset,\{1,2\})$.

[6] $\hat{H}_{\mathscr{G}}[s,h]$ can be regarded as the cost of encoding an extended sample $\hat{d} = (\hat{s},\hat{h})$.

[7] The fact that $\hat{H}_{\mathscr{G}}[h|s] = 0$ implies that the hidden features do not induce an additional coding cost of the sample $\hat{s}$.

The function $h^*(s)$ is determined by imposing the requirement *(ii)* that the representation should be minimally sufficient. This amounts to requiring that

$$p^*(s|h) = \frac{k_s \delta_{h,h^*(s)}}{\sum_{s'} k_{s'} \delta_{h,h^*(s')}} \tag{3.16}$$

does not contain any useful information about the generative process. For all s such that $h^*(s) = h$, the amount of information gained on s by the knowledge of $p^*(s|h)$ is given by $\log n_h - H[p^*(\cdot|h)]$, where $n_h = \sum_s n_{s,h}$ is the number of s with $h^*(s) = h$. When p^* is a distribution of maximal entropy $H[p^*(\cdot|h)] = \log n_h$ and the gain is maximal. This is equivalent to requiring that $p^*(s|h)$ does not depend on s, which implies that $k_s = k_{s'}$ whenever $h^*(s) = h^*(s')$. This includes, in particular, the trivial choice $h^*(s) = s$ or any other function for which $h^*(s) \neq h^*(s')$ for some s and s' with $k_s = k_{s'}$. Therefore,

$$H[h^*] \geq \hat{H}[k]. \tag{3.17}$$

The representation is *minimally* sufficient if this inequality is saturated[8][9], i.e. if $h^*(s) = k_s$ or a strictly monotonic function thereof, which leads to Eq. (3.8). The amount of information stored in the hidden features, is then given by

$$\hat{H}[k] = -\sum_k \frac{km_k}{N} \log \frac{km_k}{N} \tag{3.18}$$

as claimed in Refs. [31, 30]. Notice that $\hat{H}[k] = I(k,s)$ is precisely the mutual information between the sample and the *hidden features*. In addition, Eq. (3.14)

[8] When the inequality in Eq. (3.17) is saturated, one obtains the coarsest partition on the data points where we do not lose information about the generative process. Such a partition is unique. This is analogous to the Lehmann-Scheffé partitions of the minimal sufficient statistics [52, 53]. Indeed, an refinement to this coarsest partition is a sufficient representation.

[9] Furthermore, when the inequality in Eq. (3.17) is saturated, then the cost of encoding the hidden features is minimal and thus, optimal.

implies that

$$\begin{aligned} \hat{H}[s] &= -\sum_k \frac{km_k}{N} \log \frac{k}{N} \qquad (3.19) \\ &= \hat{H}[k] + \hat{H}[s|k] \qquad (3.20) \end{aligned}$$

In other words, the number of bits needed to describe a point in the sample can be divided in the number of bits $\hat{H}[k]$ needed to describe the features and the number of bits $\hat{H}[s|k] = \sum_k p^*(k) \log(m_k)$ that quantifies noise, i.e., the amount of information in the sample $\hat{s}$ that cannot be used to estimate the underlying model.

Notice that the partition that survives in the sum of Eq. (3.12) aggregates together the data points

$$I_k = \{i : \; k_{s_i} = k\} \qquad (3.21)$$

according to the frequency k with which they occur in the sample. Therefore, the *a priori* symmetry of $p(\hat{s})$ under permutations of the sample points (or over partitions) is broken *a posteriori* in the optimization procedure. Loosely speaking, $\mathcal{I}^*$ encodes the structure of the hidden features that can be learned from the sample. The only situations where this symmetry is not broken are when either $k_s = 0$ or 1 for all states, or when $k_s = N$ for $s = s^*$ and $k_s = 0$ otherwise. In both cases the $\mathcal{K}$ partition contains one set only. The non-informativeness of these samples is reflected in the fact that $\hat{H}[k] = 0$ in both cases.

Eqs. (3.8) and (3.9) represent an ideal limit. In practical implementations, $p(s|h)$ is *restricted* to a parametric form – as e.g. in restricted Boltzmann machines – that generally does not allow to reach this limit. Yet, the representation that emerges from taking such a parametric form is expected to come close to this limit. This suggest that the emergent representation should be a *sparse* one, meaning that $p(s|h)$ is a very peaked distribution. Also, regularization is expected to bring the model further away from the ideal limit, i.e. to favor less sharp (or more disordered) distributions.

In summary, as a result of the optimization on the partition and on the data generating process, we obtain that the representation that contains the maximal information on the generative process – that we call *minimally sufficient* – is the one in terms of frequencies $\mathcal{K}$ and that the hidden features are the frequencies k_s of the states themselves. Accordingly, the amount of information that is extracted from the data coincides with the entropy $\hat{H}[k]$ of the frequencies, as claimed in Refs. [31, 30]. This is evident, in information theoretic terms because in the absence of prior information to distinguish the states s, the frequency k_s with which different outcomes s occur is the only statistics where the dependence structure between different points s_i can be stored. In [31], $\hat{H}[s]$ has been termed *resolution* since it quantifies the amount of detail that the sample contains on the state s of the system under study, and $\hat{H}[k]$ has been termed *relevance* since it quantifies the amount of information the sample contains on the distribution $p(s)$[10], as also observed in [30].

3.3 Maximally informative samples

As we have seen in Fig. (figure with the Amazon data), the same sample can be represented at different levels of resolution $\hat{H}[s]$ in Eq. (3.19) and consequently, contain different amount of information on the generating process, which is quantified by $\hat{H}[k]$ in Eq. (3.18). The relation between these quantities can be visualized in a plot of $\hat{H}[k]$ as a function of $\hat{H}[s]$, as the latter varies between 0 and $\log N$. In one extreme, all the observations are different, i.e., $s^{(i)} \neq s^{(j)}$, $\forall\, i \neq j$, such that $m_{k=1} = N$ and $m_{k'} = 0$, $\forall k' > 1$. Hence, one finds $\hat{H}[s] = \hat{H}[k] = 0$. On the other extreme, all the data points are equal, i.e., $s^{(i)} = s$, $\forall i = 1, \ldots, N$ such that $m_k = 0$ for $k = 1, \ldots, N-1$ and $m_{k=N} = 1$. With this limiting case, while $H[s] = \log N$, one finds that $H[k] = 0$. In both of these extreme cases, we do not learn anything from the sample. In between, $\hat{H}[k]$ follows a bell-shaped curve as $\hat{H}[s]$ varies between 0 and $\log N$. This curve is bounded from above by the line $\hat{H}[k] = \hat{H}[s]$, that is attained when $m_k = 0$ or 1 for all values of k. This is because k_s is a function of s and the data processing inequality imposes that it cannot contain more information that s itself.

[10] In Ref. [31], it has been shown that $\hat{H}[k]$ corresponds to the number of parameters of a model $p(\hat{s}|\theta)$ that can be estimated using the sample $\hat{s}$ without overfitting.

Varying the resolution $\hat{H}[s]$ changes the distribution of the degeneracy m_k and consequently, the relevance $\hat{H}[k]$. Hence, we call $\hat{s}$ a maximally informative sample if m_k^* is such that the relevance $\hat{H}[k]$ is maximal at a given resolution $\hat{H}[s] = H_0$. This implies the maximization of the functional

$$\mathcal{F} = \hat{H}[k] + \mu(\hat{H}[s] - H_0) + \lambda \left(\sum_k k m_k - N \right) \tag{3.22}$$

over m_k, where the Lagrange multipliers μ and λ are adjusted to enforce the conditions $\hat{H}[s] = H_0$ and $\sum_k k m_k = N$, respectively. The solution to this optimization problem is rather non-trivial as m_k are positive integers. However, a suitable upper bound to the maximal $\hat{H}[k]$ can be obtained by approximating m_k as positive real numbers. As shown in [30, 31], this amounts to finding m_k^* such that $\frac{d\mathcal{F}}{dm_k}|_{m_k^*} = 0$ and one finds that the maximally informative samples have a characteristic power law frequency

$$m_k^* \approx c k^{-1-\mu} \tag{3.23}$$

where c is a normalization constant such that $\sum_k k m_k^* = N$. A lower bound to the maximal $\hat{H}[k]$ can be obtained by thinking that m_k as being drawn from a Poisson distribution with mean n_k and solve the optimization problem by finding n_k such that the functional $\mathcal{F}$ in Eq. (3.22) is maximized. It turns out (as done in Ref. [31]) that n_k will have to satisfy the equation

$$n_k \mathcal{L}'(n_k) = \lambda - (\mu + 1) \log \frac{k}{N} - \mathcal{L}(n_k) \tag{3.24}$$

which can be solved numerically as in Fig. 3.2 where

$$\mathcal{L}(n_k) = \int_0^1 dz \frac{e^{-n_k z} - 1}{\log(1 - z)}. \tag{3.25}$$

Fig. 3.2 compares the curve obtained for random samples (dashed lines) obtained as random draws of N balls in L boxes, with most informative samples (full lines), which are those that are obtained by maximizing $\hat{H}[k]$ over m_k, at fixed $\hat{H}[s]$ and sample size N.

In the rightmost part, $\hat{H}[k]$ increases as $\hat{H}[s]$ decreases with a slope $-\mu$ that is the same exponent of the frequency distribution in Eq. (3.23), and the Lagrange multiplier in the constrained optimization of $\hat{H}[k] + \mu\hat{H}[s]$. As mentioned in Ref. [48], μ quantifies the trade-off between resolution $\hat{H}[s]$ and relevance $\hat{H}[k]$ in the sense that a decrease of Δ bits in $\hat{H}[s]$ grants an increase of $\mu\Delta$ bits in $\hat{H}[k]$. Therefore, the point $\mu = 1$, that corresponds to Zipf's law, marks the limit beyond which further reduction in the resolution implies losses in accuracy. This point is the one that achieves maximal $\hat{H}[s] + \hat{H}[k]$ and the sample is optimally represented.

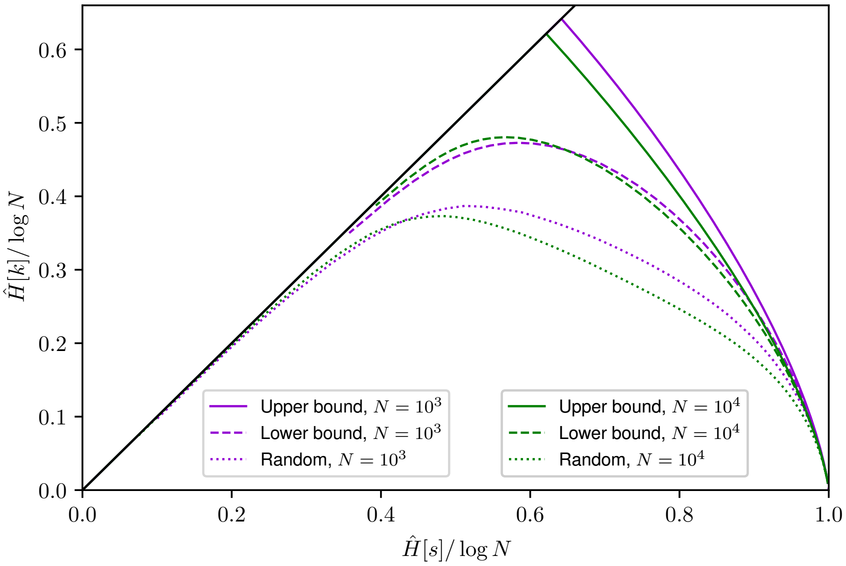

FIGURE 3.2. **Plot of $\hat{H}[k]$ versus $\hat{H}[s]$ for mostly informative samples and for typical samples, for $N = 10^3$ and 10^4.** For maximally informative samples, we report both the upper bound and lower bound. For random samples we compute the averages of $\hat{H}[s]$ and $\hat{H}[k]$ over 10^7 realizations of random distributions of N balls in L boxes, with L varying from 2 to 10^7. Here, each box corresponds to one state $s = 1, \ldots, L$ and k_s is the number of balls in box s. The full black line represents the limit $\hat{H}[k] = \hat{H}[s]$ which is obtained when $m_k \leq 1$ for all k. Points above this line are ruled out by the data processing inequality.

3.4 Criticality in point estimate statistics

As noted earlier, the occurrence of power law distributions in samples – sometimes termed *statistical criticality* – is reminiscent of critical phenomena in statistical physics. As discussed in Ref. [25], a similar relation between most informative samples and criticality also arises in the standard setting of point estimate in statistics.

Let us consider the case where the sample comes from a parametric family $p(s) = f(s|\theta)$ and let us ask what would most informative samples look like. In this setting, the amount of information learned on θ from the sample $\hat{s}$ can be quantified in the Kullback-Leibler divergence between the posterior distribution $f(\theta|\hat{s})$ and the prior $p_0(\theta)$

$$D_{KL}\left(f(\theta|\hat{s})\|p_0(\theta)\right) = \int d\theta f(\theta|\hat{s}) \log \frac{f(\theta|\hat{s})}{p_0(\theta)}. \tag{3.26}$$

When we have a reasonable amount of data, i.e., $N \gg 1$, with which we calculate the likelihood distribution, $f(\hat{s}|\theta)$, a straightforward calculation yields (see Appendix B.1)

$$D_{KL}\left(f(\theta|\hat{s})\|p_0(\theta)\right) \simeq \frac{k}{2} \log \frac{N}{2\pi e} - \log p_0(\hat{\theta}) + \frac{1}{2} \log \det \hat{L}(\hat{\theta}) + \mathcal{O}(1/N) \tag{3.27}$$

where k is the number of parameters (i.e. the dimension of θ), $\hat{\theta}$ is the likelihood estimate of θ and $\hat{L}(\theta)$ is the Hessian matrix of the likelihood at θ with the matrix elements $L_{a,b} = \langle \frac{\partial^2 \log f(s|\theta)}{\partial \theta_a \partial \theta_b} \rangle_\theta$ defined by an expectation over the parametric model $f(s|\theta)$. For exponential models of the type

$$f(s|\theta) = \frac{1}{Z} e^{\sum_\mu \theta_\mu \mathcal{T}_\mu(\boldsymbol{s})} \tag{3.28}$$

where $\mathcal{T}(\boldsymbol{s})$ are monomials of the spin variables $\boldsymbol{s}$ and $\hat{L}(\boldsymbol{\theta})$ coincides with the Fisher Information matrix which has the nature of a generalized susceptibility matrix:

$$L_{a,b} = \frac{\partial^2}{\partial \theta_a \partial \theta_b} \log Z = \frac{\partial \langle \mathcal{T}_a \rangle}{\partial \theta_b}. \tag{3.29}$$

The first term in Eq. (3.27) encodes the fact that since the error on θ typically scales as $1/\sqrt{N}$ there are $\frac{1}{2} \log N$ bits learned for each component of θ. This term is independent of the data. Disregarding the prior, the last term suggests that most informative samples are those for which the susceptibility is maximal, that typically occur at critical points if one interprets $f(s|\theta)$ as a model in statistical physics (see also Ref. [25]).

This result reveals the rich behaviors in the space of parameters that appear at the critical point. In particular, Ref. [25] found that a high susceptibility implies that there is a highly dense region in the parameter space where the parameter values cannot be distinguished on the basis of the samples. This implies that, at criticality, one finds a parametric model that is flexible enough to account for the wide variability of the data given the sample size.

3.5 Relation with the Information Bottleneck method

In this section, we shall consider a generic task in unsupervised learning. In this setting, each data point v is produced from an unknown data generating process t and we wish to extract a representation s of the data points that can shed light on t. For example, in a data clustering task, data $\hat{v}$ consists in a series of N different objects v_i. The task is that of grouping these points into classes, by attaching a label s_i to each point, that may highlight features of the generating process t, such as the similarities and differences among data points. Another example is that in deep learning where the data point v can represent patterns that a deep neural network aims at learning and s the state of one of the layers in the architecture [48]. Viewed as a Markov chain, the data generating process corresponds to

$$t \to v \to s. \tag{3.30}$$

A formal approach to the task of finding an optimal representation s^* consists in looking for the association $p(s^*|v)$ that solves the problem

$$s^* = \arg\max_{p(s|v)} \left[I(t,s) - \beta I(s,v) \right], \tag{3.31}$$

where the first term of the optimization function is the information that the representation s contains on the generative process and the second term penalizes redundant representations. Eq. (3.31) is very close to the Information Bottleneck (IB) method

[51] in spirit. The main difference is that in the IB, t and v are given, so IB deals with the *supervised* learning task of determining the representation s that encodes the relation between t and v in an optimal manner. Instead, here, t is unknown and we are interested in the *unsupervised* task of learning an optimal representation of the data. As long as t is unknown, Eq. (3.31) remains a formal restatement of the problem[11].

As suggested in Chapter 3.3, the maximally informative samples can be obtained from Eq. (3.31) if we replace t with the empirical frequency $t \simeq \hat{p}_s = k_s/N$. Therefore, the first term in Eq. (3.31) becomes

$$I(t,s) = I(k,s) = \hat{H}[k] - \hat{H}[k|s] = \hat{H}[k] \tag{3.32}$$

where the last equality derives from the fact that k_s is a function of s and consequently, $\hat{H}[k|s] =$. The second term in Eq. (3.31) is minimal when s is a function of v and $p(v|s)$ is a maximum entropy distribution, i.e. when $p(v|s) = 1/k_s$. Hence, $I(s,v) = \hat{H}[s] - \hat{H}[s|v] = \hat{H}[s]$ is the resolution (see also Ref. [55] for IB approaches where $I(s,v)$ is replaced by $\hat{H}[s]$ at the outset). Taken together, this and Eq. (3.32) turn Eq. (3.31) into the constrained maximization of $\hat{H}[k]$ subject to a constraint on $\hat{H}[s]$ as in Eq. (3.22) in Chapter 3.3.

The substitution $t \to \hat{p}$ in the Markov chain in Eq. (3.30) amounts to the statement that conditional on s, v contains no information on t. Indeed, $I(v,\hat{p}|s) = 0$ because $\hat{p}$ is a function of s. This is equivalent to reversing the Markov chain in Eq. (3.30) as $t \to s \to v$[12], so that t becomes the generative model of s. This suggests that Eq. (3.31) seeks the representations with optimal generation ability. Interestingly, Ref. [48] have suggested that the outstanding generation performance of deep learning can be explained by the observation that deep neural networks extract maximally informative representations s from data presented in the visible layer v.

[11] For application of IB to unsupervised learning problems, such as geometric data clustering, see e.g. Ref. [54].

[12] This argument parallels the definition of sufficient statistics [56]: When $t = f(\cdot|\theta)$ then the Markov chain in Eq. (3.30) reads $\theta \to v \to s$. If s is chosen as a sufficient statistics for θ then, conditional on s, the data v do not contain any information on θ. This implies that the chain can be reversed, i.e. $\theta \to s \to v$.

In summary, maximally informative samples are the solution of an optimization problem similar to IB, with the important difference that while IB is a *supervised* learning scheme, maximally informative samples are the outcome of an *unsupervised* learning task. Indeed, the IB addresses the issue of maximally compressing an input v to transmit relevant information to reconstruct a given output t [57], whereas the definition of maximally informative samples takes the frequency k of the internal representations s as output features. While relevance is defined with respect to the output in the IB, the approach discussed here quantifies relevance with respect to internal criteria. We remark, in this respect, that the first term $I(k,s) = \hat{H}[k]$ in the optimized function does not depend at all on the relation between s and v, but only on the distribution of the former. Note also that, in contrast to the rate-distortion curves typical of IB where relevance $I(t,s)$ is an increasing function of channel capacity $I(s,v)$, here the relation is not monotonic. The decreasing part of this relation, that corresponds to $\beta < 0$, is due to finite sample size and it characterizes the under sampling regime.

3.6 Thermodynamics of samples and large deviations

As discussed in Chapter 2.2, it is possible to construct a statistical mechanics approach of a sample. Recalling the arguments, the core idea is that the Gibbs-Boltzmann distribution

$$q(s) = \frac{e^{-E_s/T}}{Z} \tag{3.33}$$

of a system, where each micro-state (i.e., the observation) s has an energy E_s, coincides with the empirical distribution

$$q(s) = \hat{p}(s) = \frac{k_s}{N} \tag{3.34}$$

for $T = Z = 1$. Consequently, the energy E_s can now be defined as

$$E_s = -\log(k_s/N). \tag{3.35}$$

Each energy level $E_k = -\log \frac{k}{N}$ has degeneracy m_k, so one can define a Boltzmann entropy

$$S_B(E_k) = \log W(E_k) \tag{3.36}$$

in terms of the degeneracy $W(E_k) = m_k$ of the energy levels E_k.

For a system defined at the micro-scale by a discrete set of energy levels $E \in \mathcal{E}$, each with degeneracy $W(E)$, we can also define a description in terms of macro-states, corresponding to the distribution[13]

$$p_\beta(s) = \frac{1}{Z(\beta)} e^{-\beta E_s}, \qquad Z(\beta) = \sum_{E \in \mathcal{E}} W(E) e^{-\beta E}, \tag{3.37}$$

as in statistical mechanics. This construction corresponds to the maximum entropy distribution on micro-states s with a given value of the average energy

$$\langle E \rangle_\beta = \sum_s p_\beta(s) E_s. \tag{3.38}$$

We have seen in Chapter 2.2 that this construction allows one to define *statistical criticality* in a sample, by relating it to anomalous fluctuations at second order phase transitions in a precise way [16, 17, 18, 19, 20]. Specifically, the critical point in the (fictitious) inverse temperature parameter β can be located where the "specific heat" (i.e. the variance of the variable E_s) diverges in an infinite system, or attains its maximum in a finite system. Ref. [16] shows that Zipf's law is peculiar in that it corresponds to statistical criticality at $\beta = 1$, i.e. for the original data.

We'd like to relate these observations to resolution and relevance, as described in previous sections. First, we observe that the energy $E_s = -\log(k_s/N)$ in Eq. (3.35) has a natural interpretation as coding cost, i.e. the number of bits needed to represent

[13] The microscopic description applies, in statistical mechanics, to closed systems where the energy remains constant whereas the macro-state description applies to open systems in contact with the environment where the energy fluctuates. In this case, the micro-canonical ensemble corresponds to observations that are seen the same number of times in the sample while the macro-canonical ensemble corresponds to the sample with a fixed size N.

state s. The typical coding cost that we expect per data point is given by the resolution

$$\hat{H}[s] = -\sum_s e^{-E_s} E_s. \tag{3.39}$$

Among all distributions with a given average coding cost $\hat{H}[s] = -\sum_s p(s)E_s$, the distribution $p(s) = e^{-E_s}$ is the one of *maximal entropy*, i.e. the one that embodies no other information on s except for the coding cost E_s (See Appendix A.2.3). This is a different way of stating our requirement in Chapter 3.2 that the features $h = e^{-E}$ should contain all relevant information.

The construction discussed in Chapter 2.2 and in Refs. [16, 17] suggests an interpretation in terms of large deviations: if we generate a new sample $\hat{\tilde{s}}$ from the distribution $q(s) = k_s/N$, we typically expect that the coding cost per sample point attains a value close to the resolution $\hat{H}[s]$. In this section, we shall instead be interested in cases when the coding cost per sample point is $U = \sum_s \tilde{k}_s E_s/N \neq \hat{H}[s]$, i.e., when the coding cost is *atypical*. We shall also be interested in characterizing the properties of the *typical* samples with an *atypical* coding cost. To address this, we turn to large deviations theory [56] (see also Appendix A.3) which shows that the probability to find a sample with a coding cost per sample $U \neq \hat{H}[s]$ is given by

$$P\left\{\frac{1}{N}\sum_s \tilde{k}_s E_s = U\right\} \sim e^{-ND_{KL}(p_\beta || q))} \tag{3.40}$$

where $p_\beta(s) = \frac{1}{Z(\beta)}e^{-\beta E_s}$ is the Gibbs-Boltzmann distribution as given in Eq.(3.37) with $Z(\beta) = \sum_{E\in\mathcal{E}} W(E)e^{-\beta E}$ and $W(E)$ is the degeneracy of the energy level E. The large deviation parameter β controls the atypicality of the coding cost of the samples that are typically realized under Eq. (3.37) and is adjusted so that

$$\sum_s p_\beta(s)E_s = D_{KL}(p_\beta||q) + S_G(\beta) = U \tag{3.41}$$

where

$$S_G(\beta) = -\sum_s p_\beta(s)\log p_\beta(s) = \beta U + \log Z(\beta) \tag{3.42}$$

is the Gibbs-Shannon entropy of the macro-state. Finally, we expect that the state s occurs $\tilde{k}_s \approx Np_\beta(s)$ times in samples with an anomalous coding cost U.

Note that the Gibbs-Shannon entropy S_G in Eq. (3.42), as a function of U, is a different function from the Boltzmann entropy $S_B(E)$ in Eq. (3.36) and in Chapter 2.2: $S_G \simeq \hat{H}[\tilde{s}]$ is an estimate of the resolution of the new sample, whereas $S_B(E)$, as suggested in [48], is a measure of the noise at the energy level E, i.e. it measures the residual uncertainty about the state s once the energy level $E = E_s$ is known. Indeed, the average degeneracy measures the amount of useless information in the sample, i.e.,

$$\langle S_B \rangle_{\beta=1} = \sum_{E \in \mathcal{E}} W(E) e^{-E} \log W(E) \tag{3.43}$$

$$= \hat{H}[s|E] \tag{3.44}$$

$$= \hat{H}[s] - \hat{H}[E] \tag{3.45}$$

$$= \hat{H}[s] - \hat{H}[k] \tag{3.46}$$

where we used that $\hat{H}[E] = \hat{H}[k]$ because $E_s = -\log(k_s/N)$ is a function of k_s[14]. Therefore, samples with a minimal $\langle S_B \rangle_{\beta=1}$ for a given resolution are those with maximal relevance $\hat{H}[k]$.

While macro-states are obtained by *maximizing* $S_G = -\sum_s p(s) \log p(s)$ over $p(s)$ for a given distribution $W(E)$ of energy levels and at fixed U, the most informative samples are obtained by *minimizing* $\langle S_B \rangle_{\beta=1}$ over $W(E)$, keeping $p(s) = e^{-E_s}$, at fixed resolution

$$\langle E \rangle_{\beta=1} = \sum_{E \in \mathcal{E}} W(E) e^{-E} E = \hat{H}[s]. \tag{3.47}$$

Let us now show that this results in a linear behavior $S_B(E) = \log W(E) \simeq \mu E +$

[14] For a thermodynamic system, both S_B and $S_G = \hat{H}[s]$ are extensive, i.e. proportional to the system size V. For a system with a minimal energy level spacing ΔE, $H[E] \leq -\log \Delta E + \text{const.}$ Typically ΔE vanishes as an inverse power of the system size, which means that $\hat{H}[E] \sim \log V$ is sub-extensive. In these cases, the two descriptions (the micro-canonical and the canonical ensembles) are equivalent.

const. In order to do this, we introduce Eq. (3.47) and

$$\sum_{E \in \mathcal{E}} W(E) = \Omega \tag{3.48}$$

$$\sum_{E \in \mathcal{E}} W(E) e^{-E} = 1 \tag{3.49}$$

as constraints in the minimization of Eq. (3.43). Therefore, most informative samples are those that minimize the functional

$$\mathcal{F}[W] = \sum_{E \in \mathcal{E}} W(E) \left\{ e^{-E} \left[\log W(E) - \nu - \mu E \right] + \lambda \right\} \tag{3.50}$$

where the Lagrange multipliers mu, ν and λ need to be adjusted in order to enforce the three constraints in Eqs. (3.47), (3.48) and (3.49) respectively. By finding the degeneracy $W^*(E)$ such that $\frac{d\mathcal{F}[W]}{dW(E)}|_{W^*(E)} = 0$, one finds the solution to be

$$W^*(E) = e^{\nu + \mu E - \lambda e^E}. \tag{3.51}$$

Notice that λ introduces a cutoff in the distribution of energies, i.e. $W^*(E) \approx 0$ for $E \gg -\log \lambda$. $W^*(E)$ has its maximum at energy $E_{\max} = \log(\mu/\lambda)$. For $E \leq E_{\max}$, the entropy $S(E) = \log W^*(E) \simeq \nu + \mu E$ is well approximated by a linear behavior, as claimed above.

This behavior of $W^*(E)$ corresponds to the power law behavior Eq. (3.23) in m_k^*. In order to show this, consider the limit $N \to \infty$ where energy levels become dense. Then the number $W^*(E)dE$ of energy levels in $[E, E+dE)$ should match the number $m_k^* dk$ of states s observed with frequency in the corresponding interval for $k = Ne^{-E}$. From this,

$$m_k^* \simeq W^*(E) \left| \frac{dE}{dk} \right| \Bigg|_{E=-\log(k/N)} \tag{3.52}$$

which implies that $W^*(E) \simeq k m_k$. Therefore the relation $\log W^*(E) = \mu E + c$ corresponds to Eq. (3.23).

The average degeneracy $\langle S_B \rangle_{\beta=1}$ is a *convex* monotonic increasing function of

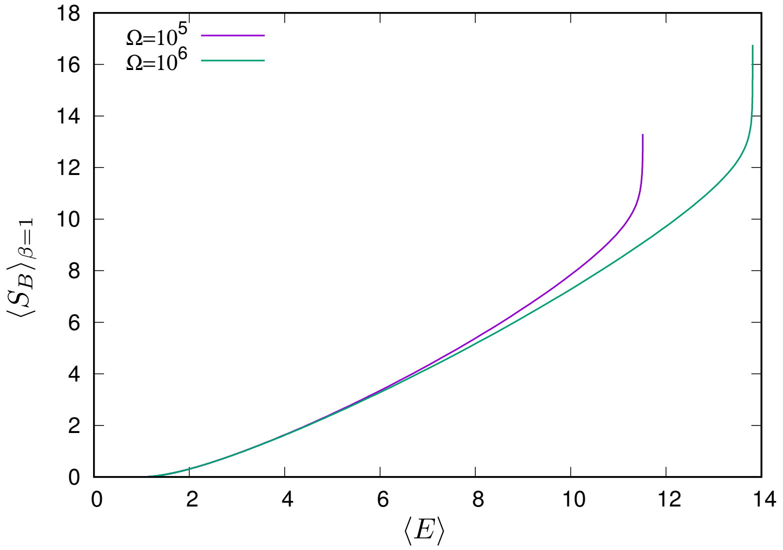

FIGURE 3.3. **Relation between entropy $\langle S_B\rangle_{\beta=1}$ and energy $\langle E\rangle$ for mostly informative samples arising from the minimization of $\langle S_B\rangle_{\beta=1}$ given the constraints in Eqs. (3.48 - 3.49), for $\Omega = 10^5$ and 10^6.**

$\hat{H}[s]$, as shown in Fig. 3.3. This contrasts with the relation between S_G and U, which is *concave*. Indeed, the Lagrange multiplier μ enforcing the constraint on energy cannot be thought of as an inverse temperatures (e.g. high (low) μ corresponds to high (low) "energy" $\hat{H}[s]$, contrary to what happens in statistical mechanics). Rather μ encodes the trade-off between resolution $\hat{H}[s]$ and relevance $\hat{H}[k]$ – or noise $\langle S_B\rangle_{\beta=1}$ – in maximally informative samples.

As discussed above, moving away from $\beta = 1$ corresponds to probing the large deviation regime of samples with an anomalously small or large coding cost. It is interesting to observe that large deviations preserve the linear behavior of $\log W(E)$ in Eq. (3.51) w.r.t. E. Indeed, this corresponds to a linear transformation of energy levels $\tilde{E}_s = -\log(\tilde{k}_s/N) \simeq \beta E_s + \text{const}$ with degeneracy $\tilde{W}(\tilde{E}) \simeq \tilde{\mu}\tilde{E} + \text{const}$ with $\tilde{\mu} = \mu/\beta$. Therefore, large deviations of maximally informative samples are typically realized as maximally informative samples at a different resolution.

We finally observe that the specific heat

$$C_V(\beta) = \frac{d^2}{d\beta^2}\log Z(\beta) = \left\langle (E - \langle E\rangle_\beta)^2 \right\rangle \tag{3.53}$$

whose divergence signals statistical criticality in Ref. [16], also coincides with the

Fisher Information of the parametric distribution p_β as in Eq. (3.29). For maximally informative samples (where the Kullback-Leibler divergence in Eq. (3.27) is at a maximum), this defines the non-informative (Jeffreys) prior $p_0(\beta) = \mathcal{N}\sqrt{C_V(\beta)}$ on β, with $\mathcal{N}$ a normalizing constant[15]. Futhermore, for maximally informative samples, C_V attains a maximum at $\beta = \mu$, so when $\mu = 1$ the maximum coincides with the point where inference is made ($\beta = 1$), as observed in Ref. [16]. The same construction can be carried out for the distribution of states s in maximally relevant samples, in terms of the parameter μ. Again, the Fisher Information coincides with the variance of energy levels, which is maximal at $\mu = 1$. This shows that the point $\mu = 1$, that corresponds to Zipf's law, is the most likely value of μ under the non-informative (Jeffreys) prior. In this sense, Zipf's law is the most likely behavior that we can expect from a maximally informative sample.

3.7 Statistical criticality and hidden variables

Schwab, Nemenman and Mehta (2014) [26] argue that Zipf's law arises from the presence of hidden variables. They considered $n \gg 1$ independent (discrete) variables $\boldsymbol{s} = (s_1, \ldots, s_n)$ drawn independently from the same probability $p(s|h)$ that depends on a variable h. Under these conditions, this sequence $\boldsymbol{s}$ satisfies the Asymptotic Equipartition Property (AEP) [56]. The AEP states that there is a typical set $\mathcal{A}_n^{(h)}$ such that (asymptotically as $n \to \infty$) *(i)* all samples in $\mathcal{A}_n^{(h)}$ have the same probability

$$-\log p(\boldsymbol{s}|h) = -\sum_{i=1}^{n} \log p(s_i|h) \simeq E(h) = -n \sum_{s} p(s|h) \log p(s|h) \tag{3.54}$$

and *(ii)* the probability to draw a sample in $\mathcal{A}_n^{(h)}$ is very close to one, and, because of this, *(iii)* the number of typical samples is equal to $|\mathcal{A}_n^{(h)}| \approx e^{E(h)}$. If one defines the entropy $S(h)$ as the logarithm of this number, then one has that $S(h) = \log |\mathcal{A}_n^{(h)}| \approx$

[15] It is called non-informative because it is invariant under reparametrization and because, for $N \to \infty$, this is the prior that maximizes the Kullback-Leibler divergence between the posterior and the prior [58], i.e. the one for which the data brings in the maximal amount of information. (See also Appendix B.1)

$E(h)$.

In the absence of a hidden variable h that may induce a variation of the energy $E(h)$, we expect that the logarithm of the frequency $-\log(k_s/N)$ will not have large variations and will concentrate on a certain value. In order to observe a broad distribution of frequencies, as in Zipf's law, a variation of $E(h)$ is needed. Hence, hidden variables are a necessary condition for the occurrence of Zipf's law. Indeed, the probability (density) to observe a value E of $-\log p(s|h)$ is

$$p(E) = \int dh p(h) \sum_s \delta(E + \log p(s|h)) p(s|h) \tag{3.55}$$

$$\simeq \int dh p(h) e^{S(h)-E(h)} \delta(E - E(h)) \tag{3.56}$$

where we used the AEP to substitute the sum over all s with the sum only on $s \in \mathcal{A}_n(h)$, which are those sequences for which[16] $-\log p(s|h) = E(h)$. Notice that $S(h) \sim E(h) \sim n$ are extensive functions, and the integral on h would be dominated by a single point h^* for n large. However $S(h) \approx E(h)$ by the AEP, which implies that the distribution $p(E)$ remains broad and it does not concentrate. This argument for sufficiency is corroborated by convincing numerical experiments for few examples of models in the exponential family. Aitchison, Corradi and Latham (2016) [28] argue that a broad, quasi uniform distribution $p(E)$ is a general sufficient condition for Zipf's law. While they convincingly show that this condition holds in several empirical data and models, this directly follows from the definition $E_s = -\log k_s$ of energy levels.

From the point of view of the present discussion, the presence of hidden variables is clearly a necessary condition for a sample to have a minimally interesting structure, not necessarily in the form of Zipf's law. Our main point is that a sufficient condition for the emergence of power law frequency distributions is that the sample should be maximally informative. These samples are those that achieve an optimal trade-off between resolution and relevance, i.e. that store maximal information on the genera-

[16] Indeed, typical sequences are exactly defined as $\mathcal{A}_n^{(h)} = \{s : E(h) \leq -\log p(s|h) < E(h) + \epsilon\}$ for any $\epsilon > 0$, and the AEP states that $1 = \sum_s p(s|h) \simeq \sum_{s\in\mathcal{A}_N^{(h)}} p(s|h) \simeq e^{-E(h)}|\mathcal{A}_n^{(h)}|$, hence $S(h) \equiv \log |\mathcal{A}_n^{(h)}| = E(h)$.

tive model, at a given resolution. In particular, Zipf's law emerges in samples where the resolution is reduced up to the point where the increase in relevance engendered by a further reduction in resolution, does not compensate the loss in resolution. For mostly informative samples where $W^*(E) \sim e^{\mu E}$, as shown in the previous section, so $p^*(E) \sim e^{(\mu-1)E}$ which implies that a flat distribution in energy is achieved exactly at $\mu = 1$. At this particular point, the energy spectrum is used as efficiently as possible (see Ref. [48]). When $\mu > 1$, high energy states overweight low energy ones, whereas when $\mu < 1$, the distribution of energy levels in the sample is skewed on low energy states.

3.8 Conclusions

The aim of this chapter is to clarify the derivation and nature of the relevance $\hat{H}[k]$, recently introduced in [31, 30], as a measure of the useful information that a sample contains on the generative model. We do this by relating our approach to the standard approach employed in parametric statistics. This allows one to characterize the properties of maximally informative samples and the trade-off they embody between resolution and relevance. This offers a different explanation of the widespread occurrence of statistical criticality [1] that is independent of any self-organization mechanisms [59, 21]. In particular, we find that Zipf's law characterizes the statistics of maximally informative samples at the optimal trade-off between resolution and relevance. We believe that this finding, besides its appeal as a simple rationale for the occurrence of Zipf's law in many domains [35, 43, 18, 20], also offers a guideline for extracting efficient representations form high dimensional data.

This page has been intentionally left blank.

4

Minimum Description Length (MDL) codes are critical

From the previous chapter, we have suggested that broad distributions, as exhibited in a vast number of real world domains (see e.g. [1, 2]), arise from efficient representations. By efficient representations, we mean that the data has sampled the relevant variables which are those that carry the maximal amount of information on the generative process [30, 31, 60]. Such Maximally Informative Samples (MIS) are those for which the entropy of the frequency with which outcomes occur – called *relevance* in [31, 60] – is maximal at a given resolution, which is measured by the number of bits needed to encode the sample. MIS exhibit power law distributions with the exponent μ governing the tradeoff between resolution and relevance [60, 48]. This argument for the emergence of broad distributions is independent of any mechanism or model. A direct way to confirm this claim is to check that samples generated from models that are known to encode efficient representations are actually maximally informative. In this line, Ref. [48] found strong evidence that MIS occur in the representations that deep learning extracts from data. This chapter explores the same issue in efficient coding as defined in Minimum Description Length [61].

Regarding empirical data as a message sent from Nature, we expect it to be expressed in an efficient manner if relevant variables are chosen. This requirement can be made quantitative and precise, in information theoretic terms, following Minimum Description Length (MDL) principle [61]. MDL seeks the optimal encoding of data generated by a parametric model with unknown parameters (see Section 4.1). MDL derives a probability distribution over samples that embodies the requirement of optimal encoding. This distribution is the Normalized Maximum Likelihood (NML).

This chapter studies the NML as a generative process of samples and studies both its typical and atypical properties. In a series of cases, we find that samples generated by NMLs are typically close to being maximally informative, in the sense of Ref. [60], and that their frequency distribution is typically broad. In addition, we find that NMLs are critical in a very precise sense, because they sit at a second order phase transition that separates typical from atypical behavior. More precisely, we find that large deviations, for which the resolution attains atypically low values, exhibit a condensation phenomenon whereby all N points in the sample coincide. This is consistent with the fact that NML correspond to efficient coding of random samples generated from a model, so that codes achieving higher compression do not exist. Large deviations enforcing higher compression force parameters to corners of the allowed space where the model becomes deterministic.

4.1 Minimum Description Length and the Normalized Maximum Likelihood

In this section, we shall introduce the ideas behind the minimum description length (MDL) principle.

4.1.1 Kolmogorov complexity: The ideal MDL

As in the previous chapter, the question we wanted to address here is still of the optimal data reduction. Given a sample, the task is to learn the regularities of this sample and use these regularities to compress it, i.e., to find a short description. Of course, the description depends on the description method we wish to adapt.

The choice of the description method as general-purpose computing language (like Fortran or C) leads to the definition of the *Kolmogorov complexity* [62]. The Kolmogorov complexity measures the length of the shortest program that reproduces the sample and stops after that. This means that a sample that is highly regular needs a shorter program to reproduce it and consequently, the sample has a lower Kolmogorov complexity.

One may argue that the resulting Kolmogorov complexity will depend on the computing language adapted. However, because one can always construct a program in one language which translates another program which uses a different language with a constant description length (a result which is called the *invariance theorem*), the Kolmogorov complexity between two different programming language differs only by a constant provided that the sample is large enough.

However, the Kolmogorov complexity cannot be calculated[1]. This problem calls for practical implementation of finding the minimum description length.

4.1.2 Normalized Maximum Likelihood: the practical MDL

The MDL principle circumvents the non-computability of the Kolmogorov complexity by constraining the set of possible model distributions (or programs) with which one can calculate the model complexity with [64]. In the MDL, one chooses to work with *prefix-free codes* – uniquely and instantaneously decodable codes that satisfy the Kraft inequality (see Appendix A.2). It turns out that, under prefix-free codes, the optimal length of codeword is equal to the negative logarithm of the probability distribution. Hence, specifying the length of the codeword allows to identify the distribution and

[1] The non-computability of the Kolmogorov complexity stems from the fact that one needs to construct a program which iterates over all possible programs that reproduces the sample and returns all the programs that halted. However, because of the halting problem [63], no algorithm is able to possibly predict whether the program will eventually finish running or not. Hence, even if we have a short program that reproduces the sample which may be a good candidate for the shortest of such a program, there may always be a number of shorter programs which can reproduce the sample and which we do not know if they will ever stop.

vice versa[2]. This is the main insight in the MDL principle. Specifically, *learning from data is equivalent to data compression* [61]. In turn, data compression is equivalent to assigning a probability distribution over the space of samples. This section provides a brief derivation of this distribution[3].

In an information theoretic perspective, one can think of the sample, $\hat{s}$, as a message generated by some source (e.g. Nature) that we wish to compress as much as possible. This entails translating $\hat{s}$ in a sequence of bits[4]. A code is a rule that achieves this for any $\hat{s} \in \chi^N$ and its efficiency depends on whether frequent patterns are assigned short codewords or not. Conversely, any code implies a distribution $P(\hat{s})$ over the space of samples and the cost of encoding the sample $\hat{s}$ under the code P is given by [56]

$$E = -\log P(\hat{s}) \tag{4.1}$$

bits (assuming logarithm base two)[5]. Optimal compression is achieved when the code P coincides with the data generating process [56].

Consider the situation where the data is generated as independent draws from a parametric model $f(s|\theta)$. If the value of θ were known, then the optimal code would be given by $P(\hat{s}) = \prod_i f(s^{(i)}|\theta) \equiv f(\hat{s}|\theta)$. MDL seeks to derive P in the case where θ is not known[6]. This applies, for example, to the situation where $\hat{s}$ is a series of experiments or observation aimed at measuring the parameters θ of a theory.

In hind sight, i.e. upon seeing the sample, the best code is $f(\hat{s}|\hat{\theta})$, where $\hat{\theta}(\hat{s})$ is the maximum likelihood estimator for θ, and it depends on the sample $\hat{s}$. Therefore, one can define the *regret* $\mathcal{R}$ as the additional encoding cost that one needs to spend to

[2] Chapter 6.5 of Ref. [56] makes an analogy between data compression and gambling.

[3] We refer the interested reader to Refs. [61, 65] for a more detailed discussion of MDL.

[4] It is natural to work in bits because, intuitively, this corresponds to the minimum number of true or false questions needed to recover the message.

[5] The correspondence between the encoding cost E and the probability distribution $P(\hat{s})$ is made clear in Appendix A.2.

[6] Indeed, MDL aims at deriving efficient coding under f irrespective of whether $f(s|\theta)$ is the "true" generative model or not. This allows one to compare different models and choose the one providing the most concise description of the data.

encode the sample, $\hat{s}$, if one uses the code $P(\hat{s})$ to compress $\hat{s}$, i.e.,

$$\mathcal{R} = -\log P(\hat{s}) - \min_{\theta} \left[-\log f(\hat{s}|\theta) \right]. \tag{4.2}$$

Notice that $\min_\theta \left[-\log f(\hat{s}|\theta) \right] = -\log f(\hat{s}|\hat{\theta}(\hat{s}))$. $\mathcal{R}$ is called regret of P relative to f for sample $\hat{s}$ because it depends both on P and on $\hat{s}$.

MDL derives the optimal code, $\bar{P}(\hat{s})$, that minimizes the regret, assuming that for any P the source produces the worst possible sample [61]. The solution [66]

$$\bar{P}(\hat{s}) = \frac{f(\hat{s}|\hat{\theta}(\hat{s}))}{\sum_{\hat{x} \in \chi^N} f(\hat{x}|\hat{\theta}(\hat{x}))}. \tag{4.3}$$

is called the Normalized Maximum Likelihood (NML)[7]. The optimal regret is given by

$$\bar{\mathcal{R}} = \log \sum_{\hat{s} \in \chi^N} f(\hat{s}|\hat{\theta}(\hat{s})) \tag{4.4}$$

which is known in MDL as the *parametric complexity*[8]. For models in the exponential family, Rissanen showed that the parametric complexity is asymptotically given by [67]

$$\bar{\mathcal{R}} \simeq \frac{k}{2} \log \frac{N}{2\pi} + \log \int \sqrt{\det L(\theta)} d\theta + \mathcal{O}(1) \tag{4.5}$$

where $L(\theta)$ is the Fisher information matrix with the matrix elements defined by an expectation $L_{ij}(\theta) = -\langle \frac{\partial^2 \log f(s|\theta)}{\partial \theta_i \partial \theta_j} \rangle_\theta$ over the parametric model $f(s|\theta)$ (see Appendix C.1 for a simple derivation). The NML code is a *universal code* because it achieves a compression per data point which is as good as the compression that would be achieved with the optimal choice of θ. This is easy to see, because the regret $\bar{\mathcal{R}}/N$ per data point vanishes in the limit $N \to \infty$, hence the NML code achieves the same compression as $f(\hat{s}|\hat{\theta})$.

[7] Note that the NML is not the only formulation of the MDL principle. One other formulation is the crude, two-part MDL formulation where the description length is such that the coding cost $-\log f(\theta|\hat{s} = -\log f(\hat{s}|\theta) - \log p_0(\theta)$ of posterior distribution is minimized. The first term is the description of the data while the second term is the description of the parameters of the model.

[8] Notice that $e^{\bar{\mathcal{R}}}$ can be seen as a partition sum. Hence, throughout the chapter, we shall refer to the parametric complexity as the *UC partition function*.

Notice also that the optimal regret, $\bar{\mathcal{R}}$, in Eq. (4.4) is independent of the sample $\hat{s}$. It indeed provides a measure of complexity of the model f that can be used in model selection schemes. Note that the optimal regret, $\bar{\mathcal{R}}$, in Eq. (4.4) is not only due to the number of parameters, k, in the model as in the first term of $\bar{\mathcal{R}}$[9] but also due to the functional form of the parametric model as encapsulated by the second term of $\bar{\mathcal{R}}$. This implies that more complex models, i.e., models with larger parametric complexity $\bar{\mathcal{R}}$ will require more encoding costs than less complex (or simple) ones. Thus, in a model selection perspective, whenever a sample $\hat{s}$ occurs with the same likelihood, $f(\hat{s}|\hat{\theta}(\hat{s}))$ under different models, one chooses a model which is the least complex – a principle known as the Occam's razor.

For exponential families, MDL procedure penalizes models with a cost which equals the one obtained in Bayesian model selection [70] under a Jeffreys prior. Indeed, considering $\bar{P}(\hat{s})$ as a generative model for samples, one can show that the induced distribution on θ is given by Jeffreys prior (see Appendix C.1).

4.2 NML codes provide efficient representations

In this section we consider $\bar{P}$ as a generative model for samples and we investigate its typical properties for some representative statistical models.

4.2.1 Dirichlet model

Let us start by considering the Dirichlet model distribution $f(s|\theta) = \theta_s$, $\forall s \in \chi$. The parameters $\theta_s \geq 0$ are constrained by the normalization condition $\sum_{s\in\chi}\theta_s = 1$. Let $S = |\chi|$ denote the cardinality of χ and define, for convenience, $\rho = N/S$ as the average number of points per state. Because each observation is mutually independent,

[9] Notice that other parametric model selection approaches only penalizes the number of parameters, k, e.g. AIC $= -2\log f(\hat{s}|\hat{\theta}(\hat{s})) + 2k$ for the Akaike information criterion [68] and BIC $= -2\log f(\hat{s}|\hat{\theta}(\hat{s})) + k\log N$ for the Bayesian information criterion [69].

the likelihood of a sample $\hat{s}$ given $\theta = (\theta_1, \ldots, \theta_S)$ can be written as

$$f(\hat{s}|\theta) = \prod_{s\in\chi} \theta_s^{k_s}, \tag{4.6}$$

where k_s is the the number of times that the state s occurs in the sample $\hat{s}$. From here, it can be seen that $\hat{\theta}_s = k_s/N$ is the maximum likelihood estimator for θ_s. Thus, the universal code for the Dirichlet model can now be constructed as

$$\bar{P}(\hat{s}) = e^{-\bar{\mathcal{R}}} \prod_{s\in\chi} \left(\frac{k_s}{N}\right)^{k_s} \tag{4.7}$$

which can be read as saying that for each s, the code needs $-k_s \log(k_s/N) + \bar{\mathcal{R}}/N$ bits. In terms of the frequencies, $\{k_1, \ldots, k_S\}$, the universal codes can be written as

$$\bar{P}(k_1, \ldots, k_S) = e^{-\bar{\mathcal{R}}} \frac{N!}{\prod_{s\in\chi} k_s!} \prod_{s\in\chi} \left(\frac{k_s}{N}\right)^{k_s} \delta\left(\sum_{s\in\chi} k_s - N\right) \tag{4.8}$$

wherein the multinomial coefficient, $\frac{N!}{\prod_{s\in\chi} k_s!}$, counts the number of samples with a given frequency profile $k_1, \ldots, k_S$. In order to compute the optimal regret $\bar{\mathcal{R}}$, we have to evaluate the partition function

$$\begin{aligned} e^{\bar{\mathcal{R}}} &= \frac{N!}{N^N e^{-N}} \int_{-\pi}^{\pi} \frac{d\mu}{2\pi} e^{i\mu N} \left[\sum_{k_1=0}^{\infty} \frac{k_1^{k_1} e^{-k_1} e^{-i\mu k_1}}{k_1!}\right] \cdots \left[\sum_{k_S=0}^{\infty} \frac{k_S^{k_S} e^{-k_S} e^{-i\mu k_S}}{k_S!}\right] &(4.9) \\ &= \frac{N!}{N^N e^{-N}} \int_{-\pi}^{\pi} \frac{d\mu}{2\pi} e^{i\mu N} \left[\mathcal{n}(i\mu)\right]^S &(4.10) \\ &\simeq \sqrt{2\pi N} \int_{-\pi}^{\pi} \frac{d\mu}{2\pi} e^{S\Phi(i\mu)} &(4.11) \end{aligned}$$

where

$$\Phi(z) = \rho z + \log \mathcal{n}(z) \tag{4.12}$$

and

$$\mathcal{n}(z) = \sum_{k=0}^{\infty} \frac{k^k e^{-(1+z)k}}{k!}. \tag{4.13}$$

The integral in Eq. (4.11) is dominated by the value where the function Φ attains its saddle point value $z^*(\rho)$, which is given by the condition

$$\frac{d\Phi}{dz} = \rho - \langle k \rangle_z = 0 \tag{4.14}$$

where the average $\langle \ldots \rangle_z$ is taken with respect to the distribution

$$q(k|z) = \frac{1}{\mathcal{N}(z)} \frac{k^k e^{-(1+z)k}}{k!}. \tag{4.15}$$

Gaussian integration around the saddle point leads then to

$$e^{\bar{\mathcal{R}}} \simeq \sqrt{\rho} \frac{e^{S\Phi(z^*(\rho))}}{\sqrt{\langle k^2 \rangle_{z^*} - \langle k \rangle_{z^*}^2}} \tag{4.16}$$

where we used the identity $\Phi''(z) = -\left[\langle k^2 \rangle_z - \langle k \rangle_z^2\right]$.

The distribution Eq. (4.8) can also be written introducing the Fourier representation of the delta function

$$\bar{P}(k_1, \ldots, k_S) = \frac{N! e^{-\bar{\mathcal{R}}}}{N^N e^{-N}} \int_{-\pi}^{\pi} \frac{d\mu}{2\pi} e^{i\mu N} \prod_{s \in \chi} \frac{k_s^{k_s} e^{-(1+i\mu)k_s}}{k_s!}. \tag{4.17}$$

For typical sequences $k_1, \ldots, k_S$, the integral is also dominated by the value $\mu = -iz^*(\rho)$ that dominates Eq. (4.11), which means that the distribution factorizes as

$$\bar{P}(k_1, \ldots, k_S) \simeq \prod_{s \in \chi} q(k_s|z^*). \tag{4.18}$$

This means that the NML is, to a good approximation, equivalent to S independent draws from the distribution $q(k|z^*)$ or, equivalently, that the distribution $q(k|z^*)$ is the one that characterizes typical samples. This is fully confirmed by Fig. 4.1A, which compares $q(k|z^*)$ with the empirical distribution of k_s drawn from $\bar{P}$. For large k, we find $q(k|z^*) \sim e^{-z^* k}/\sqrt{k}$, which shows that the distribution of frequencies is broad, with a cutoff at $1/z^*$. This underlying broad distribution is confirmed by Fig. 4.1B which shows the dependence of the degeneracy m_k with the frequency k.

In the regime where $\rho \gg 1$ and k is large, the cutoff extends to large values of k and we find $z^*(\rho) \simeq \frac{1}{2\rho}$ (see Appendix C.2.1). Also, the parametric complexity can be computed explicitly via Eq. (4.5) in this regime, with the result

$$\bar{\mathcal{R}} \simeq \frac{S}{2}(1 + \log \rho) + \frac{1}{2}\log(2\rho), \qquad \rho \gg 1. \tag{4.19}$$

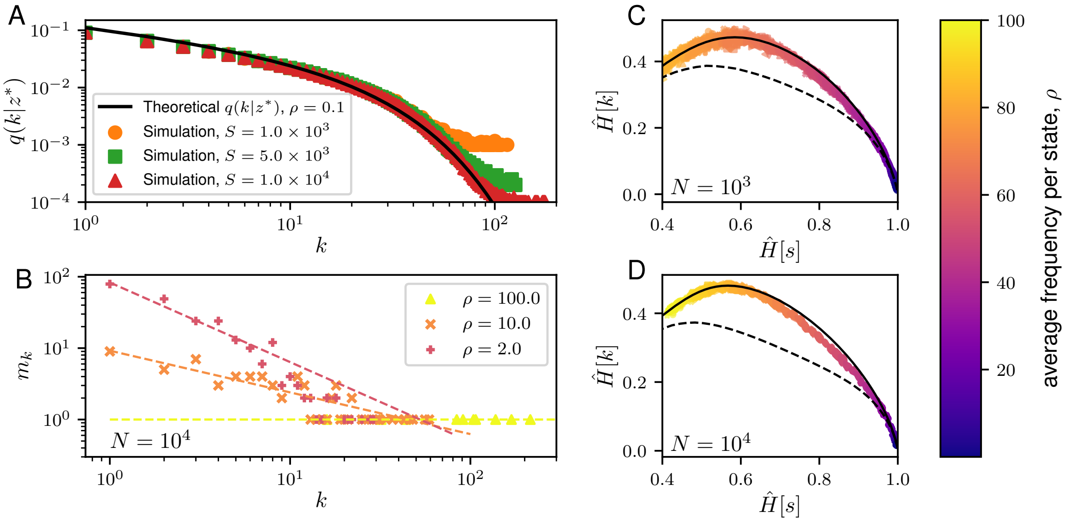

FIGURE 4.1. **Properties of the typical samples generated from the NML of the Dirichlet model.** (A) A plot showing the frequency distribution of the typical samples of the Dirichlet NML code. Given S, the cardinality of the state space, χ, with $S = 1.0 \times 10^3$ (orange dots), 5.0×10^3 (green squares), and 1.0×10^4 (red triangles), we compute the average frequency distribution across 100 generated samples from the Dirichlet NML of size $N = 10S$ such that the average frequency per state, ρ, is fixed. This is compared against the theoretical calculations (solid black line) for $q(k|z^*)$ in Eq. (4.15). (B) Plot showing the degeneracy, m_k, of the frequencies, k, in a representative typical sample of length $N = 10^3$ generated from the Dirichlet NML code with average frequencies per spike: $\rho = 100$ (yellow triangle), $\rho = 10$ (orange x-mark) and $\rho = 2$ (red cross). The corresponding dashed lines depict the best-fit line. (C-D) Plots of $\hat{H}[s]$ versus $\hat{H}[k]$ for the typical samples of the Dirichlet NML code. For a fixed size of the data, N ($N = 10^3$ in B and $N = 10^4$ in C), we have drawn 100 samples from the Dirichlet NML code varying ρ, ranging from 2 to 100. The results are compared against the $\hat{H}[k]$ and $\hat{H}[s]$ for maximally informative samples (MIS, solid black line) and random samples (dashed black lines). For the MIS, the theoretical lower bound is reported [31]. For the random samples, we compute the averages of $\hat{H}[s]$ and $\hat{H}[k]$ over 10^7 realizations of random distributions of N balls in L boxes, with L ranging from 2 to 10^7. Here, each box corresponds to one state $s = 1, \ldots, L$ and k_s is the number of balls in box s. Note that all the calculated values for $\hat{H}[k]$ and $\hat{H}[s]$ are normalized by $\log N$.

The coding cost of a typical sample is given by

$$\begin{aligned} E &= -\log \bar{P}(\hat{s}) && (4.20)\\ &= -\sum_{s\in\chi} k_s \log \frac{k_s}{N} + \bar{\mathcal{R}} && (4.21)\\ &= N\hat{H}[s] + \bar{\mathcal{R}}. && (4.22) \end{aligned}$$

The number of samples with encoding cost E can be computed in the following way. The number of samples that correspond to a given degeneracy m_k of the states that occurs $k_s = k$ times in $\hat{s}$, is given by

$$\frac{N!}{\prod_k (km_k)!}. \qquad (4.23)$$

Therefore, the number of samples with coding cost E is

$$\begin{aligned} W(E) &= \sum_{\{m_k\}\in\mathcal{M}(E)} \frac{N!}{\prod_k (km_k)!} && (4.24)\\ &= \sum_{\{m_k\}\in\mathcal{M}(E)} e^{\log N! - \sum_k \log(km_k)!} && (4.25)\\ &\sim \sum_{\{m_k\}\in\mathcal{M}(E)} e^{N\hat{H}[k]}, \qquad \rho \gg 1 && (4.26) \end{aligned}$$

where $\mathcal{M}(E)$ is the set of all sequences $\{m_k\}$ that are consistent with samples in χ^N and satisfy Eq. (4.22). The last expression assumes $\log M! \simeq M\log M - M$, which is reasonable for $M = km_k \gg 1$, i.e. when $\rho \gg 1$. In this regime we expect the sum over $\mathcal{M}(E)$ to be dominated by samples with maximal $\hat{H}[k]$. Indeed, Fig. 4.1C and D show that samples drawn from $\bar{P}$ achieve values of $\hat{H}[k]$ close to the theoretical maximum, especially in the region $\rho \gg 1$.

4.2.2 A model of independent spins

In order to corroborate our results for the Dirichlet model, we study the properties of the universal codes for a model of independent spins, i.e. a paramagnet. For a single spin, $s = \pm 1$, in a local field h, the probability distribution is given by

$$P(s|h) = \frac{e^{sh}}{2\cosh h}. \tag{4.27}$$

Thus for a sample $\hat{s}$ of size N,

$$P(\hat{s}|h) = e^{[Nmh - N\log(2\cosh h)]} \tag{4.28}$$

where $m = \frac{1}{N}\sum_{i=1}^{M} s^{(i)}$ is the local magnetization. The maximum likelihood estimate for h is $\hat{h}(m) = \tanh^{-1} m$, hence the universal code for a single spin can be written as

$$\bar{P}(\hat{s}) = e^{N[mh(m) - \log(2\cosh h(m))] - \bar{\mathcal{R}}} \tag{4.29}$$

where $\bar{\mathcal{R}} \simeq \frac{1}{2}\log\frac{\pi N}{2}$ (see Appendix C.2.2). Note that a sample with a magnetization m can be realized by considering the permutation of the up-spins ($s = 1$, where there are $\ell = \frac{N+Nm}{2}$ of such spins) and the permutation of the down-spins ($s = -1$, where there are $N - \ell$ of such spins). Consequently, the magnetization for samples drawn from $\bar{P}$ has a broad distribution given by the arcsin law (see Appendix C.2.2)

$$\begin{aligned} \bar{P}(m) &= \binom{N}{\frac{N-Nm}{2}} e^{N\left[m\tanh^{-1} m - \log\left(2\cosh\left(\tanh^{-1} m\right)\right)\right] - \bar{\mathcal{R}}} && (4.30) \\ &\simeq \frac{1}{\pi\sqrt{1-m^2}}, \qquad m \in [-1, 1]. && (4.31) \end{aligned}$$

It is straightforward to see that the model of a single spin is equivalent to a Dirichlet model with two states $\chi = \{-1, +1\}$. In terms of the number ℓ of up-spins, using $m = \frac{2\ell - N}{N}$, the NML for a single spin can be written as

$$\bar{P}(\ell) = e^{-\bar{\mathcal{R}}}\binom{N}{\ell}\left(\frac{\ell}{N}\right)^{\ell}\left(1 - \frac{\ell}{N}\right)^{N-\ell}. \tag{4.32}$$

The NML for a paramagnet with n independent spins reads as

$$\bar{P}(\ell_1, \ldots, \ell_n) = e^{-n\bar{\mathcal{R}}}\prod_{i=1}^{n}\binom{N}{\ell_i}\left(\frac{\ell_i}{N}\right)^{\ell_i}\left(1 - \frac{\ell_i}{N}\right)^{N-\ell_i}. \tag{4.33}$$

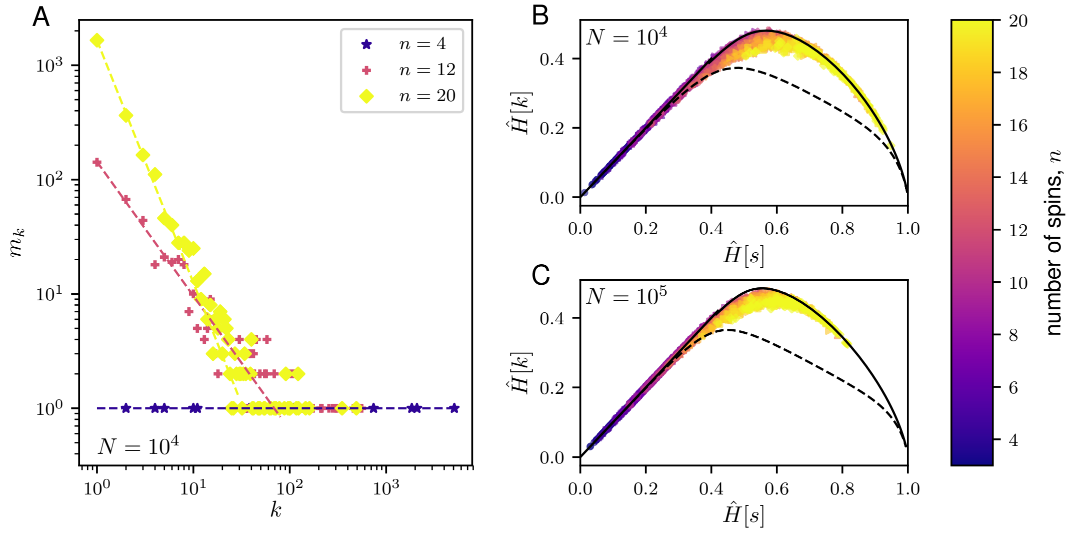

FIGURE 4.2. **Properties of typical samples for the NML codes of the paramagnet**. (A) Plots showing the degeneracy, m_k, of the frequencies, k, in a representative typical sample of length $N = 10^4$ generated from the NML of a paramagnet with different number of independent spins: $n = 4$ (blue star), $n = 12$ (red cross) and $n = 20$ (yellow diamond). The corresponding dashed lines depict the best-fit line. (B-C) Plots of the $\hat{H}[k]$ versus $\hat{H}[s]$ of the typical samples generated from the paramagnet NML code for varying sizes of the data, $N = 10^4$ (B) and $N = 10^5$ (C), and for varying number of spins, n, ranging from 3 to 20. Given N and n, we compute the $\hat{H}[k]$ and $\hat{H}[s]$ over 100 realizations of the NML code of a paramagnet. The results are compared against the $\hat{H}[k]$ and $\hat{H}[s]$ for maximally informative samples (solid black line) and random samples (dashed black line) as described in Fig. 4.1. Note that all the calculated $H[k]$ and $H[s]$ are normalized by $\log N$.

Fig. 4.2 reports the properties of the typical samples of the NML of a paramagnet. We observed that the frequency distribution of typical samples is broad (Fig. 4.2A) and that typical samples attain values of $H[k]$ very close to the maximum for a given value of $\hat{H}[s]$ (Fig. 4.2B,C). As the size N of data increases, the NML enters the well-sampled regime where $\hat{H}[k] \simeq \hat{H}[s]$, indicating that the data processing inequality [56] is saturated. In this regime, typical samples are those which maximize the entropy $\hat{H}[s]$.

4.2.3 Graphical models

This section extends our findings to systems of interacting variables and discuss the properties of typical samples drawn form the corresponding NML distribution. We shall first consider models in which the observed variables are interacting either directly (Sherrington-Kirkpatrick model) and then restricted Boltzmann machines, where the variables interact indirectly through hidden variables.

Sherrington-Kirkpatrick model

In this section, $s = (s_1, \ldots, s_n)$ is a configuration of n spins $s_i \in \{\pm 1\}$. In the Sherrington-Kirkpatrick (SK) model, the distribution of s, considers all interactions up to two-body

$$P(s|\boldsymbol{J}, \boldsymbol{h}) = \frac{1}{Z(\boldsymbol{J}, \boldsymbol{h})} \exp\left[\sum_i h_i s_i + \sum_{i<j} J_{ij} s_i s_j\right], \qquad s = (s_1, \ldots, s_n) \quad (4.34)$$

where the partition function

$$Z(\boldsymbol{J}, \boldsymbol{h}) = \sum_{s_1 = \pm 1} \cdots \sum_{s_n = \pm 1} \exp\left[\sum_i h_i s_i + \sum_{i<j} J_{ij} s_i s_j\right] \quad (4.35)$$

is a normalization constant which depends on the pairwise couplings, $\boldsymbol{J}$ with $J_{ij} = J_{ji}$ being the coupling strength between s_i and s_j, and external local fields, $\boldsymbol{h}$. Thus, given a sample, $\hat{s} = (s^{(1)}, \ldots, s^{(N)})$ of N observations, the likelihood reads as

$$P(\hat{s}|\boldsymbol{J}, \boldsymbol{h}) = \exp\left[N \sum_i h_i m_i + N \sum_{i<j} J_{ij} c_{ij} - N \log Z(\boldsymbol{J}, \boldsymbol{h})\right]. \quad (4.36)$$

where $m_i = \frac{1}{N} \sum_{l=1}^{N} s_i^{(l)}$ and $c_{ij} = \frac{1}{N} \sum_{l=1}^{N} s_i^{(l)} s_j^{(l)}$ are the magnetization and pairwise correlation respectively. Note that all the needed information about the SK model is encapsulated in the free energy, $\phi(\boldsymbol{J}, \boldsymbol{h}) = \log Z(\boldsymbol{J}, \boldsymbol{h})$. Indeed, the maximum likelihood estimators for the couplings, $\hat{\boldsymbol{J}}$, and local fields, $\hat{\boldsymbol{h}}$, are the solutions of the self-consistency equations

$$\frac{\partial \phi(\boldsymbol{J}, \boldsymbol{h})}{\partial h_i} = m_i, \quad \frac{\partial \phi(\boldsymbol{J}, \boldsymbol{h})}{\partial J_{ij}} = c_{ij}, \qquad i, j = 1, \ldots, n. \quad (4.37)$$

The universal codes for the SK model then reads as

$$\bar{P}(\hat{s}) = \exp\left[N\left(\sum_i \hat{h}_i m_i + \sum_{i<j} \hat{J}_{ij} c_{ij} - \phi(\boldsymbol{J}, \boldsymbol{h})\right) - \bar{\mathscr{R}}\right]. \quad (4.38)$$

However, unlike for the Dirichlet model and the paramagnet model, the UC partition function, $e^{\bar{\mathcal{R}}}$, for the SK model is analytically intractable[10].To this, we resort to a Markov chain Monte Carlo (MCMC) approach to sample the universal codes (See Appendix C.3.1). Figs. 4.3A and C shows the properties of the typical samples drawn from the universal codes of the SK model in Eq. (4.38).

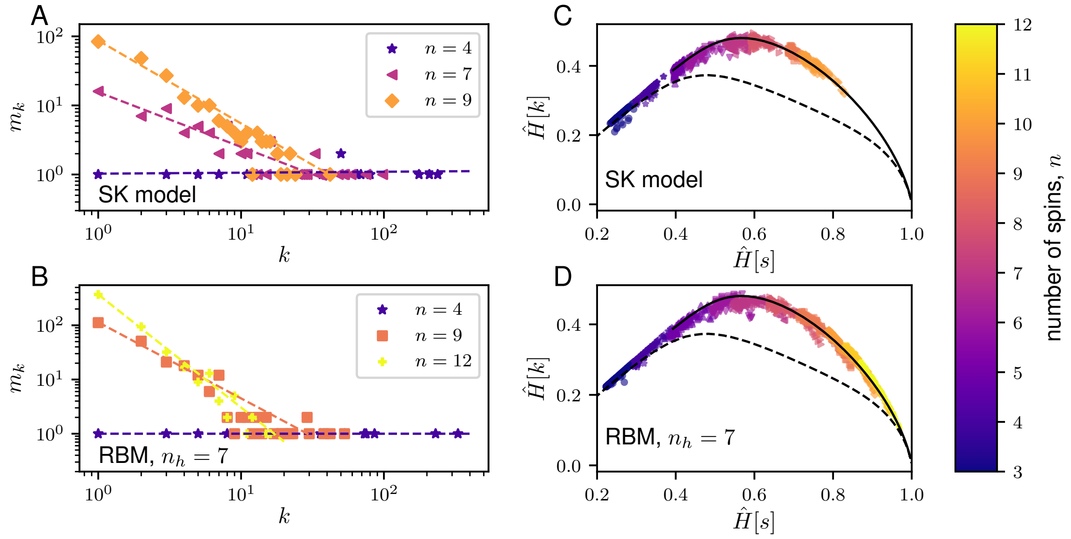

FIGURE 4.3. **Properties of typical samples for the NML codes of two graphical models: the Sherrington-Kirkpatrick (SK) model and the restricted Boltzmann machine (RBM)**. Left panels (A,C) show plots of the degeneracy, m_k, of the frequency, k, for representative typical samples generated from the NML codes for the SK model (A) and the RBM given a number of hidden variables, $n_h = 7$ (B) for different number of (visible) spins, n. The corresponding dashed lines show the best-fit lines. On the other hand, right panels (B,D) show plots of the $\hat{H}[k]$ versus $\hat{H}[s]$ of the typical samples drawn from the NML codes for the SK model (B) and the RBM with $n_h = 7$ (D) for $N = 10^3$ and for varying number of spins, n ranging from 3 to 12. Given N and n of a graphical model, we compute the $\hat{H}[k]$ and $\hat{H}[s]$ for 100 samples drawn from the respective NML codes through an Markov chain Monte Carlo (MCMC) approach (see Appendix C.3.1). Note that for the RBM, varying n_h do not qualitatively affect the observations made in the main text. As before, the $\hat{H}[k]$ and $\hat{H}[s]$ are normalized by $\log N$ and the typical NML samples are compared against maximally informative samples (solid black line) and random samples (dashed black line) as described in Fig. 4.1.

Restricted Boltzmann machines

We consider a restricted Boltzmann machine (RBM) wherein one has a layer composed of n_v independent visible boolean units, $\boldsymbol{v} = (v_1, \ldots, v_{n_v})$, which are interacting with n_h independent hidden boolean units, $\boldsymbol{h} = (h_1, \ldots, h_{n_h})$, in another layer

[10] For SK models which possess some particular structures, a calculation of the UC partition function has been done in Ref. [71].

where $v_i, h_i = 0, 1$. The probability distribution can be written down as

$$P(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{w})) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_{i=1}^{n_v} a_i v_i + \sum_{j=1}^{n_h} b_j h_j + \sum_{i=1}^{n_v}\sum_{j=1}^{n_h} v_i w_{ij} h_j\right) \tag{4.39}$$

where the partition function

$$Z(\boldsymbol{\theta}) = \sum_{v_1=0,1} \cdots \sum_{v_{n_v}=0,1} \sum_{h_1=0,1} \cdots \sum_{h_{n_h}=0,1} \exp\left(\sum_{i=1}^{n_v} a_i v_i + \sum_{j=1}^{n_h}\left(b_j + \sum_{i=1}^{n_v} v_i w_{ij}\right) h_j\right) \tag{4.40}$$

is a function of the parameters, $\boldsymbol{\theta}$, with w_{ij} is the interaction strength between v_i and h_j, $\boldsymbol{a}$ and $\boldsymbol{b}$ are the local fields acting on the visible $\boldsymbol{v}$ and hidden $\boldsymbol{h}$ units respectively. Because the hidden units, $\boldsymbol{h}$, are mutually independent, we can factorize and then marginalize the sum over the hidden variables, $\boldsymbol{h}$, to obtain the distribution of a single observation, $\boldsymbol{v}$, as

$$P(\boldsymbol{v}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left[\sum_{i=1}^{n_v} a_i v_i + \sum_{j=1}^{n_h} \log 2\cosh\left(\sum_{i=1}^{n_v} v_i w_{ij} + b_j\right)\right]. \tag{4.41}$$

Then, the probability distribution for a sample, $\hat{\boldsymbol{v}} = (\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(N)})$, of N observations is simply

$$P(\hat{\boldsymbol{v}}|\boldsymbol{\theta}) = \prod_{k=1}^{N} p(\boldsymbol{v}^{(k)}|\boldsymbol{\theta}). \tag{4.42}$$

The parameters, $\hat{\boldsymbol{\theta}}$, can be estimated by maximizing the likelihood using the Contrastive Divergence (CD) algorithm [72, 73] (see Appendix C.3.2). Once the maximum likelihood parameters, $\hat{\boldsymbol{\theta}}$, have been inferred, then the universal codes for the RBM can be built as

$$\bar{P}(\hat{\boldsymbol{v}}) = e^{-\bar{\mathcal{R}}} \prod_{k=1}^{N} p(\boldsymbol{v}^{(k)}|\hat{\boldsymbol{\theta}}). \tag{4.43}$$

Also, like in the SK model, the UC partition function, $e^{\bar{\mathcal{R}}}$, for the RBM cannot be solved analytically. To this, we also resort to a MCMC approach to sample the universal codes (See Appendix C.3.1). Figs. 4.3B and D shows the properties of the typical samples drawn from the universal codes of the RBM in Eq. (4.43). Taken together, we see that even for models that incorporate interactions, the typical samples of the NML *i)* have broad frequency distributions and *ii)* they achieve values of $\hat{H}[k]$ close to the maximum, given $\hat{H}[s]$. Due to computational constraints, we only present the results for $N = 10^3$ however, we expect that increasing N will only shift the NML towards the well-sampled regime.

4.3 Large deviations of the universal codes exhibit phase transitions

In this section, we focus on the distribution of the resolution $\hat{H}[s]$ for samples $\hat{s}$ drawn from $\bar{P}$. We note that

$$\hat{H}[s] = \frac{1}{N}\sum_{i=1}^{N} \log \frac{k_{s^{(i)}}}{N}$$

has the form of an empirical average, hence we expect it to attain a given value for typical samples drawn from $\bar{P}$. This also suggests that the probability to draw samples with resolution $\hat{H}[s] = E$ different from the typical value has the large deviation form $P\{\hat{H}[s] = E\} \sim e^{-NI(E)}$, to leading order for $N \gg 1$. In order to establish this result and to compute the function $I(E)$, as in Refs. [74, 75], we observe that

$$P\{\hat{H}[s] = E\} = \sum_{\hat{s}} \bar{P}(\hat{s})\delta\left(\hat{H}[s] - E\right) \tag{4.44}$$

$$= \int_{-\infty}^{\infty} \frac{N dq}{2\pi} \sum_{\hat{s}} \bar{P}(\hat{s}) e^{iqN(\hat{H}[s]-E)}, \tag{4.45}$$

where we used the integral representation of the δ function and $\bar{P}(\hat{s})$ is the NML distribution in Eq. (4.3).

Upon defining

$$\sum_{\hat{s}} \bar{P}(\hat{s}) e^{iqN\hat{H}[s]} = e^{N\phi(iq)}, \tag{4.46}$$

let us assume, as in the Gärtner–Ellis theorem [74], that $\phi(iq)$ is finite for $N \gg 1$ for all q in the complex plane. Then Eq. (4.45) can be evaluated by a saddle point integration

$$P\{\hat{H}[s]=E\} = \int_{-\infty}^{\infty} \frac{N d\alpha}{2\pi} e^{-N[i\alpha E-\phi(i\alpha)]} \tag{4.47}$$

$$\sim e^{-N[\beta E-\phi(\beta)]}, \tag{4.48}$$

where we account only for the leading order. β is related to the saddle point value $q^* = -i\beta$ that dominates the integral and it is given by the solution of the saddle point condition

$$E = \frac{d}{d\beta}\phi(\beta). \tag{4.49}$$

Eq. (4.48) shows that the function $I(E)$ is the Legendre transform of $\phi(\beta)$, i.e.

$$I(E) = -\beta E + \phi(\beta) \tag{4.50}$$

with $\beta(E)$ given by the condition (4.49), as in the Gärtner–Ellis theorem [74]. Further insight and a direct calculation from the definition in Eq. (4.46) reveals that Eq. (4.49) can also be written as

$$E = \sum_{\hat{s}} \bar{P}_\beta(\hat{s})\hat{H}[s] \tag{4.51}$$

which is the average of $\hat{H}[s]$ over a "tilted" probability distribution [74]

$$\bar{P}_\beta(\hat{s}) = \bar{P}(\hat{s})e^{N\left\{\beta\hat{H}[s]-\phi(\beta)\right\}}, \tag{4.52}$$

hence β arises as the Lagrange multiplier enforcing the condition $\hat{H}[s] = E$. Conversely, when $\beta(E)$ is fixed by the condition Eq. (4.51), samples drawn from $\bar{P}_\beta$ have $\hat{H}[s] \simeq E$. In other words, $\bar{P}_\beta$ describes how large deviations with $\hat{H}[s] = E$ are realised. Therefore typical samples that realize such large deviations can be obtained by sampling the distribution $\bar{P}_\beta(\hat{s})$ in Eq. (4.52).

4.3.1 Illustrative example: Dirichlet model

As an illustration, we look at the large deviation properties of samples drawn from the NML of the Dirichlet model. Without loss of generality, we shall label the states in χ as integers $\{1, \ldots, N\}$ and consider the case where the state $s = 1$ contains a fixed number, n_1, of observations in a sample $\hat{s}$ and the remaining $N - n_1$ observations are distributed across the remaining $S - 1$ states as in Fig. 4.4.

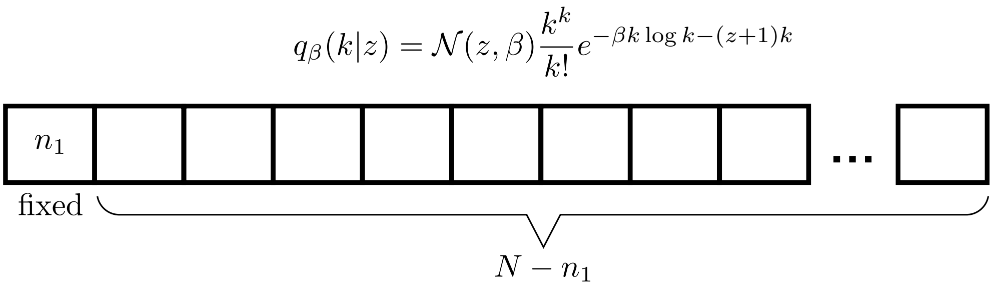
$q_\beta(k|z) = \mathcal{N}(z,\beta)\frac{k^k}{k!}e^{-\beta k \log k-(z+1)k}$
n_1
...
fixed
$N - n_1$

FIGURE 4.4. **An illustrative example of the sample properties exhibiting large deviations in the Dirichlet model.** We consider a case where the state $s = 1$ contains n_1 samples and the remaining $N - n_1$ samples are distributed in the remaining $S - 1$ states according to $q_\beta(k|z)$.

In this situation, the large deviation distribution in Eq. (4.52) can be written down as

$$\bar{P}_\beta(\hat{s}) = \frac{e^{\bar{\mathcal{R}}}}{Z(\beta)}\left(\frac{n_1}{N}\right)^{n_1}\prod_{s=2}^{S}\left(\frac{k_s}{N}\right)^{k_s} e^{\beta N \hat{H}[s]}. \tag{4.53}$$

In terms of the frequencies, $\{n_1, k_2, \ldots, k_S\}$, the large deviations to the universal codes can be expressed as

$$\begin{aligned}\bar{P}_\beta(n_1, k_2, \ldots, k_S) = \frac{e^{-\bar{\mathcal{R}}}}{Z(\beta)}\frac{N!}{n_1!\prod_{s=2}^{S}k_s!}\left(\frac{n_1}{N}\right)^{n_1}\prod_{s=2}^{S}\left(\frac{k_s}{N}\right)^{k_s} \\ \times e^{\beta N \hat{H}[s]}\delta\left(n_1 + \sum_{s=2}^{S}k_s - N\right)\end{aligned} \tag{4.54}$$

Following a similar calculation in Section 4.2.1, one finds that the partition function,

$Z(\beta)$ can be calculated as

$$\begin{aligned} Z(\beta) = {} & \frac{e^{-\bar{\mathcal{R}}} N!}{N^{N+\beta} e^{-N}} \frac{n_1^{n_1} e^{-\beta n_1 \log n_1}}{n_1!} \int_{-\pi}^{\pi} \frac{d\mu}{2\pi} e^{i\mu(N-n_1)} \left[\sum_{k_2=0}^{\infty} \frac{k_2^{k_2} e^{-\beta k_2 \log k_2 - (1+i\mu)k_2}}{k_2!} \right] \\ & \times \cdots \left[\sum_{k_S=0}^{\infty} \frac{k_S^{k_S} e^{-\beta k_S \log k_S - (1+i\mu)k_S}}{k_S!} \right] \qquad (4.55) \\ = {} & \frac{e^{-\bar{\mathcal{R}}} N!}{N^{N+\beta} e^{-N}} \frac{n_1^{n_1} e^{-\beta n_1 \log n_1}}{n_1!} \int_{-\pi}^{\pi} \frac{d\mu}{2\pi} e^{i\mu(N-n_1)} \left[\mathcal{N}_\beta(i\mu) \right]^{S-1} \qquad (4.56) \\ \simeq {} & \frac{e^{-\bar{\mathcal{R}}} \sqrt{2\pi N}}{N^{\beta}} \frac{n_1^{n_1} e^{-\beta n_1 \log n_1}}{n_1!} \int_{-\pi}^{\pi} \frac{d\mu}{2\pi} e^{(S-1)\Phi(i\mu)} \qquad (4.57) \end{aligned}$$

where

$$\Phi_\beta(z) = \rho' z + \log \mathcal{N}_\beta(z) \qquad (4.58)$$

with $\rho' = (N - n_1)/(S-1)$ and

$$\mathcal{N}_\beta(z) = \sum_{k=0}^{\infty} \frac{k^k}{k!} e^{-\beta k \log k - (1+z)k}. \qquad (4.59)$$

As in Section 4.2.1, the integral in Eq. (4.57) is dominated by the saddle point value $z^*_\beta(\rho')$ of $\Phi_\beta(z)$ given by the condition

$$\frac{d\Phi_\beta(z)}{dz} = \rho' - \langle k \rangle_{\beta,z} = 0 \qquad (4.60)$$

where the average $\langle \ldots \rangle_{\beta,z}$ is taken with respect to the large deviation distribution (see Fig. 4.4)

$$q_\beta(k|z) = \frac{1}{\mathcal{N}_\beta(z)} \frac{k^k}{k!} e^{-\beta k \log k - (1+z)k}. \qquad (4.61)$$

In the regime where $\rho' \gg 1$ and k large (similar to the calculations in Appendix C.2.1), one finds that the normalization $\mathcal{N}_\beta(z)$ can be simplified further as

$$\begin{aligned} \sum_{k=0}^{\infty} \frac{k^k}{k!} e^{-\beta k \log k - (1+z)k} & \simeq \sum_{k=0}^{\infty} \frac{1}{\sqrt{2\pi k}} e^{-\beta k \log k - (1+z)k} \qquad (4.62) \\ & = \int_0^{\infty} \frac{e^{-\beta k \log k - (1+z)k} dk}{\sqrt{2\pi k}}. \qquad (4.63) \end{aligned}$$

When $\beta = 0$, we recover the results of the typical NML codes of the Dirichlet model as in Appendix C.2.1. Notice, as well, that when $\beta > 0$, the normalization can be calculated, although numerically. Hence, we find that k_s can be considered as independent draws from the same distribution $q_\beta(k|z)$. However, when $\beta < 0$, the integrand blows up at infinity and thus, the distribution $q_\beta(k|z)$ in Eq. (4.61) is non-normalizable. Hence, in this regime, $q_\beta(k|z) = 0$ when $k > 0$. Consequently, $\langle k \rangle_{\beta,z} = 0$ in Eq. (4.60) and thus, $n_1 = N$, i.e., the observations localize on a single state, $\bar{s}$[11].

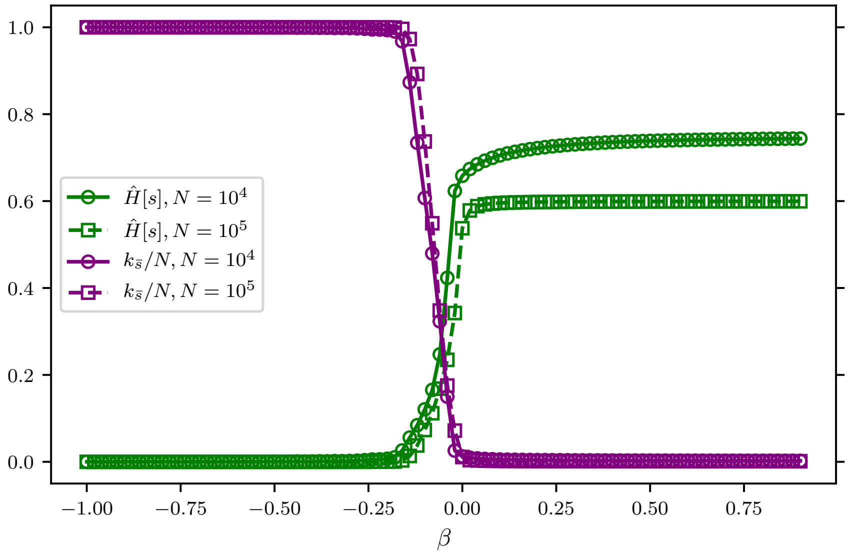

FIGURE 4.5. **Typical realizations of large deviations from the NML code of the Dirichlet model**. For a fixed parameter, β ranging from $\beta = -1$ to $\beta = 1$, samples are obtained from $\bar{P}_\beta$ in Eq. (4.52) for varying length of the dataset, N ($N = 10^4$ in solid lines with circle markers and $N = 10^5$ in dashed lines with square markers). The resolution $\hat{H}[s]$ normalized by $\log N$ (in green lines) and the maximal frequency $k_{\bar{s}}$ normalized by N (in purple lines) are calculated as an average over 100 realizations of $\bar{P}_\beta$ given β. The point $\beta = 0$ corresponds to the typical samples that are realized from the Dirichlet NML code in Eq. (4.8).

These results are corroborated by numerical simulation in Fig. 4.5 which shows that, for Dirichlet models, samples obtained from $\bar{P}_\beta$ exhibit a sharp transition at $\beta = 0$. The resolution (see green lines in Fig. 4.5) sharply vanishes for negative values of β as a consequence of the fact that the distribution *localizes* to samples where almost all outcomes coincide, i.e., $s_i = \bar{s}$. This is evidenced by the fact that the maximal

[11] Note that this condensation phenomena is not similar to the condensation occurring in many statistical mechanical systems (e.g., Bose-Einstein condensation or condensation in supersaturated solutions) where the condensation is observed in the average. The type of condensation discussed here is, instead, a condensation in the *fluctuations*, i.e., the condensation is observed as a rate event similar to that being studied in Refs. [76, 77].

frequency $k_{\bar{s}} = \max_s k_s$ approaches N very fast (see purple lines in Fig. 4.5). In other words, $\beta = 0$ marks a *localization* transition where the symmetry between the states in χ is broken, because one state $\bar{s}$ is sampled an extensive number of times $k_{\bar{s}} \propto N$.

4.3.2 Generic model

The localization behavior discussed in the previous section is generic whenever the underlying model $f(s|\theta)$ itself localizes for certain values $\bar{\theta}$ of the parameters, i.e. when $f(s|\bar{\theta}) = \delta_{s,\bar{s}}$. In order to see this, notice that, in general, we can write

$$f(\hat{s}|\hat{\theta}(\hat{s})) = \prod_s f(s|\hat{\theta}(\hat{s}))^{k_s}. \tag{4.64}$$

Thus, by inserting the identity $e^{-N\hat{H}[s]+N\hat{H}[s]}$, the NML distribution in Eq. (4.3) can be re-cast as Thus, by inserting the identity $e^{-N\hat{H}[s]+N\hat{H}[s]}$, the NML distribution in Eq. (4.3) can be re-cast as

$$\bar{P}(\hat{s}) = e^{-N\hat{H}[s]-ND_{KL}(\hat{p}||\hat{\theta})-\bar{\mathcal{R}}} \tag{4.65}$$

where $\hat{p}_s = k_s/N$ is the empirical distribution and

$$D_{KL}(\hat{p}||\hat{\theta}) = \sum_s \hat{p}_s \log \frac{\hat{p}_s}{f(s|\hat{\theta}(\hat{s}))} \tag{4.66}$$

is a Kullback-Leibler divergence.

Now, we observe that

$$e^{N\phi(\beta)} = e^{-\bar{\mathcal{R}}} \sum_{\hat{s}} e^{-(1-\beta)N\hat{H}[s]-ND_{KL}(\hat{p}||\hat{\theta})} \tag{4.67}$$

$$\geq e^{-\bar{\mathcal{R}}} \sum_{\hat{s}} e^{-(1-\beta)N\hat{H}[s]-ND_{KL}(\hat{p}||\theta_0)} \tag{4.68}$$

$$= e^{-\bar{\mathcal{R}}} \tag{4.69}$$

where the inequality in Eq. (4.68) derives from the fact that $\hat{\theta}(\hat{s})$, the maximum likelihood estimator for sample $\hat{s}$, is replaced by a generic value θ_0 and consequently,

$D_{KL}(\hat{p}||\hat{\theta}) \leq D_{KL}(\hat{p}||\theta_0)$. The equality in Eq. (4.69), instead, derives from the choice $\theta_0 = \bar{\theta}$ such that $f(s|\bar{\theta}) = \delta_{s,\bar{s}}$. Under this choice, only the term corresponding to "localized" samples where $s^{(i)} = s_0$ for all points in the sample, survive in the sum on $\hat{s}$. For such localized samples, $\hat{H}[s] = D_{KL}(\hat{p}||\theta_0) = 0$, hence Eq. (4.69) follows.

Because of the logarithmic dependence of the regret $\bar{\mathcal{R}}$ on N (see Eq. (4.5)), Eq. (4.69) implies that, for all β,

$$\phi(\beta) \geq \bar{\mathcal{R}}/N \simeq 0 \tag{4.70}$$

for $N \gg 1$. Given that $\hat{H}[s] \geq 0$ in Eq. (4.51), then $E \geq 0$ and therefore, Eq. (4.49) implies that $\phi(\beta)$ is a non-decreasing function of β. In addition, $\phi(0) = 0$ by Eq. (4.46). Taken together, these facts require that $\phi(\beta) = 0$ for all values $\beta \leq 0$. On the other hand, for $\beta > 0$, the function $\phi(\beta)$ is analytic with all finite derivatives, which corresponds to higher moments of $\hat{H}[s]$ under $\bar{P}_\beta$. Therefore, $\beta = 0$, which corresponds to the typical behavior of the NML, coincides with a second order phase transition point because the function $\phi(\beta)$ exhibits a discontinuity in the second derivative. In terms of $\bar{P}_\beta(\hat{s})$, the phase transition separates a region ($\beta \geq 0$) where all samples $\hat{s}$ have a finite probability from a region ($\beta < 0$) where only one sample, the one with $s^{(i)} = \bar{s}, \forall i$, has non-zero probability and $\hat{H}[s] = 0$.

The phase transition is a natural consequence of the fact that NML provide efficient coding of samples generated from $f(s|\theta)$. It states that codes $\bar{P}_\beta$ that achieve a compression different from the one achieved by the NML only exist for higher coding costs. Codes with lower coding cost only describe non-random samples that correspond to deterministic models $f(s|\bar{\theta}) = \delta_{s,\bar{s}}$.

4.4 Conclusion

The aim of this chapter is to elucidate the properties of efficient representations of data corresponding to universal codes that arise in MDL. Taking NML as a generative model, we find that typical samples are characterized by broad frequency distributions and that they achieve values of the relevance which are close to the maximal possible

$\hat{H}[k]$.

In addition, we find that samples generated from NML are critical in a very precise sense. If we force NML to use less bits to encode samples then the code localizes on deterministic samples. This is a consequence of the fact that if there were codes that require fewer bits, then NML would not be optimal.

This contributes to the discussion on the ubiquitous finding of *statistical criticality* [1, 16] by providing a clear understanding of its origin. It suggests that *statistical criticality* can be related to a precise second order phase transition in terms of large deviations of the coding cost. This phase transition separates random samples that span a large range of possible outcomes (the set χ in the models discussed above) from deterministic ones, where one outcome occurs most of the time. The phase transition is accompanied by a *spontaneous symmetry breaking* in the permutation between samples. The frequencies of outcomes in the symmetric phase ($\beta \geq 0$) are generated as independent draws from the same distribution, that is sharply peaked for $\beta > 0$ as can be checked in the case of the Dirichlet model. Instead, for $\beta < 0$, only one state is sampled. In the typical case, $\beta = 0$, the symmetry between outcomes is weakly broken, as there are outcomes that occur more frequently than others. At $\beta = 0$, the samples maintain the maximal discriminative power over outcomes. This type of phase transitions in large deviations is very generic, and it occurs in large deviations whenever the underlying distribution develops fat tails (see e.g. Ref. [75]).

This leads to the conjecture that broad distributions arise as a consequence of efficient coding. More precisely, broad distributions arise when the variables sampled are relevant, i.e., when they provide an optimal representation. This is precisely the point which has been made in Refs. [30, 31, 60]. The results in the present chapter add to these a new perspective whereby maximally informative samples can be seen as universal codes.

This page has been intentionally left blank.

5

Finding relevant neurons in the brain using MultiScale Relevance (MSR)

In the previous chapters, we have discussed guiding principles towards extracting relevant information from high dimensional datasets. Within this guiding principle, identifying the maximally informative samples is equivalent to looking at efficient representations which reveals the data generating process. A natural place to look for such efficient representations is in the brain. Neurons in the brain are subjected to energetic constraints [78] as they communicate with each other to understand what is happening in the external world and make behavioral decisions from the perceived stimuli. Thus, neurons must be able to extract relevant representations of the sensory input coming from downstream neurons and transmit these representations into decodable responses to upstream neurons.

Here, we explore the applicability of the theoretical framework discussed in the previous chapters by proposing a novel non-parametric, model-free method for characterizing the dynamical variability of a neural spiking across different time scales and consequently, for selecting relevant neurons – i.e. neurons whose response patterns represent information about the task or stimuli – that *does not require knowledge* *of external correlates*. This featureless selection is done by identifying neurons that have broad and non-trivial distribution of spike frequencies across a broad range of time scales. The proposed measure – called *Multiscale Relevance (MSR)* – allows an experimenter to rank the neurons according to their information content and relevance to the behavior probed in the experiment.

We illustrate the method by applying it to data on spatial navigation of freely roaming rodents in Refs. [79] and [80], that reports the neural activities of 65 neurons simultaneously recorded from the medial entorhinal cortex (mEC) and nearby regions, and 746 neurons in the anterodorsal thalamic nucleus (ADn) and post-subiculum (PoS), respectively. In all cases, we find that neurons with low MSR also coincide with those that contain no information on covariates involved in navigation, but that the opposite is not true. We find that some neurons with high MSR also contain significant relevant information for spatial navigation, some relative to position, some to head direction (HD) but often on both space and HD. These findings corroborate the recent conjecture of multiplexed coding [81] both in the mEC [82] and in the thalamus [83]. We observe that MSR correlates to different degrees with different measures that have been introduced to characterize specific neurons. More specifically, we find strong correlation between MSR and measures of sparse representations of external correlates. Furthermore, we show that the neurons in mEC with highest MSR have spike patterns that allow an upstream decoder "neuron" to discern the organism's state in the environment. Indeed, the top most relevant neurons (RNs), according to MSR, decode spatial position (or HD) just as well as the top most spatially (or HD) informative neurons (INs). In addition, we find that this decoding efficiency can not solely be due to local variations [84, 85]. Emphasizing again that the MSR does not rely on any information about space or HD and is calculated only from the timing on spikes, the correlation with spatial or HD information suggests a role for MSR as an unsupervised method for focusing on information-rich neurons without knowing *a priori* what covariate/s those neurons represent.

5.1 Information Extraction in Neuroscience: Lessons from the Medial Entorhinal Cortex

Much of the progress in understanding how the brain processes information has been made by identifying firing patterns of individual neurons that correlate significantly with the variations in the stimuli and the behaviors. These approaches have led to the discovery of V1 cells in the primary visual cortex [86], the A1 cells in the auditory cortex [87], the head direction (HD) cells in the anterodorsal nucleus (ADn) of the thalamus [88, 89], and the place cells in the hippocampus [90].

More recently, such tuning profile approaches have led to the characterization of neurons in the medial entorhinal cortex (mEC) and have revealed single cell representations that are essential for navigation. In particular, these approaches have found that the mEC and its nearby brain regions houses neurons that exhibit spatially selective firing (e.g., *grid cells* and *border cells*) which provides the brain with a locational representation of the organism and provides the hippocampus with its main cortical inputs. For example, grid cells have spatially selective firing behaviors that form a hexagonal pattern which spans the environment where the rat freely explores as in the left plots of Fig. 5.1 and in Fig. 5.2a. Apart from spatial information, grid cells can also be attuned to the HD especially in deeper layers of the mEC [91]. These cells altogether provide the organism with an internal coordinate system which it then uses for navigation. And hence, together with the information coming from speed cells [92] i.e. neurons having responses that are linearly correlated with speed, these representations allow the mEC to represent the displacement of the organism in the spatial environment.

Because the responses of these neurons in the mEC have distinct features and are preserved across different spatial contexts, subsequent studies have selected neurons based on imposed structural assumptions on the tuning profile of neurons with respect to an external correlate. In particular, different measures of information (as in Eq. (D.3)), response selectivity (as in Eq. (D.5)) and tuning curve symmetries (e.g., the grid score in Eq. (D.6) which quantifies the hexagonality of the firing responses of a neuron in the 2D spatial map and the mean vector length in Eq. (D.7) which quantifies how attuned to a single head direction the neuron's responses are) have been employed and consequently, the classification of cells are based on how statistically significant these measures are with respect to null distributions, i.e., when the neuron fires at random but still preserving the total number of observed firing events in the duration of the experiment.

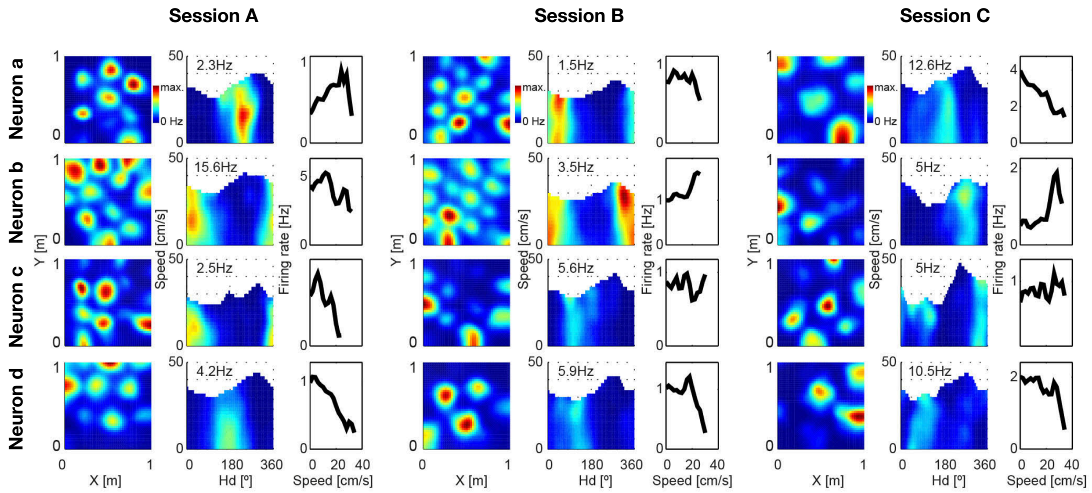

FIGURE 5.1. **Broken symmetries and conjunctive response profiles of neurons in the medial entorhinal cortex**. Response profiles of four representative mEC neurons that respond to a combination of different external correlates across three open field sessions are shown. In each session, the plots on the left show color-coded spatial firing rate maps. The scale bar is shown to the right of the first map. Notice the hexagonal patterns in the firing behavior of these neurons indicating that these are grid cells. However, notice as well that the peak firing rates do not necessarily form a regular hexagon. Furthermore, the peak firing rates are different in each grid field. The plots in the middle show the firing rate as a function of the head direction (HD, x-axis, in degrees) and running speed (y-axis, in cm/s). The maximum firing rate is indicated at the top portion for each map. The same color code were used as the spatial firing rate maps. Notice that while most of the neurons fire at a preferred HD, one neuron (Neuron c in Session C) seems to have a multimodal HD preference. The plots on the right show the firing rate as a function of running speed. Notice that some of these representative neurons have firing behaviors that are not linear with respect to the running speed (see e.g. Neuron d in Sessions A and B and Neuron a in Session C). Reprinted from Ref. [92] with permission from Springer Nature Customer Service Centre GmbH.

However, the organization of the brain and the responses of individual neuron are hardly this simple and intuitive. In particular, recent developments in understanding the spatial representation in the mEC have taught us that such approaches has its limits as illustrated in Fig. 5.1. First, neurons may break the symmetries of the tuning curves when representing navigational information through shearing [93] (i.e., the grid fields , field-to-field variability or simply by the constraints of the environment [94]. Second, the same neuron may respond to a combination of different behavioral covariates, such as position, HD and speed in spatial navigation [91, 82]. Finally, and most importantly, neurons may encode a particular behavior in ways that are unknown to the experimenter and that are not related to covariates typically used or to *a priori* features. Indeed, these limitations have resulted to a large population of neurons in the mEC that remain to be classified. With the advent of next-generation techniques that have allowed us to probe an increasing number of neurons in behaving animals, these limitations are exacerbated which calls for guiding principle towards better information extraction and neuron selection.

Under such circumstances, one can still make progress by focusing on the temporal structure of neural firing. Variations present in the spikes offer neurons with a large capacity for information transmission [95, 96, 97]. These variations, as captured by metrics describing the interspike intervals (see Appendix D.2), have been shown to be different for functionally distinct neurons in the cortex [84, 85, 98] and have been utilized to classify neurons in the subiculum [99] and in the mEC [100, 101]. However, such measures of variations are either very local or hardly take into account the temporal dependencies and time scales of natural stimuli. Hence, in the succeeding sections, we shall develop a method based on the discussion in Chapter 3 for identifying neurons that are relevant for the efficient representation of the animal's behavior by taking advantage of the variations in the dynamical responses of a neuron across a broad range of time scales.

5.2 Multiscale relevance

We consider a population composed of n neurons in an animal whose activities were simultaneously observed up to a time, t_{obs}. From hereon, we shall focus our attention to a single neuron within this population. The activity of this neuron is recorded and stamped by the spike times $\{t_1, \ldots, t_M\}$ where $t_1 < t_2 < \ldots \leq t_M \leq t_{obs}$ and M is the total number of observed spikes. By discretizing the time into T bins of duration Δt, a spike count code, $\{k_1, k_2, \ldots, k_T\}$, can be constructed where k_s denotes the number of spikes recorded from the neuron in the s^{th} time bin $B_s = [(s-1)\Delta t, s\Delta t)$ $(s = 1, 2, \ldots, T)$.

Fixing Δt allows us to probe the neural activity at a fixed time scale. Yet, rather than using Δt to measure time resolution, we adopt an information theoretic measure, given by

$$\hat{H}[s] = -\sum_{s=1}^{T} \frac{k_s}{M} \log_M \frac{k_s}{M}, \tag{5.1}$$

where $\log_M(\cdot) = \log(\cdot)/\log M$ indicates logarithm base M (in units of Mats). $\hat{H}[s]$ corresponds to the amount of information that one gains on the timing of a randomly chosen spike by knowing the index s of the bin it belongs to[1].

We argue that $\hat{H}[s]$ provides an intrinsic measure of resolution, contrary to Δt which refers to particular time scales that may vary across neurons. For example, there is a value Δt_- such that for all $\Delta t \leq \Delta t_-$, all time bins either contain a single spike or none, i.e. $k_s = 0, 1$ for all s. All these values of Δt correspond to the same value of the intrinsic resolution $\hat{H}[s] = 1$. Likewise, there may be a value Δt_+ such that for all $\Delta t \geq \Delta t_+$, all spikes of the neuron fall in the same bin. All $\Delta t \geq \Delta t_+$ then correspond to the same value $\hat{H}[s] = 0$ of the resolution, as defined here. In other words, $\hat{H}[s]$ captures the *resolution* on a scale that is fixed by the available data.

Given a resolution $\hat{H}[s]$ (corresponding to a given Δt), we can now turn to characterize the dynamic response of the neuron. The only way in which the dynamic state of the neuron in bin s can be distinguished from that in bin s' is by its activity. If the number of spikes in the two bins is the same ($k_s = k_{s'}$) there is no way to distinguish

[1] With no prior knowledge, a spike can be any of the M possible spikes, so its *a priori* uncertainty is of $\log_2 M$ bits. The information on which bin s the spike occurs, reduces the number of choices from M to k_s and the uncertainty to $\log_2 k_s$ bits. Averaging the information gain $\log M - \log k_s$ over the *a priori* distribution of spikes and dividing by $\log M$, yields Eq. (5.1). It is also worth to stress that Eq. (5.1) does not refer to the estimate of the entropy of a hypothetical underlying distribution p_s from which spikes are drawn. This would not make much sense, because it is well-known that the *naïve* estimate of the Shannon entropy in Eq. (5.1) obtained with the maximum likelihood estimator $\hat{p}_s = k_s/M$ suffers from strong biases [102, 103].

the dynamic state of the neuron in the two bins, at that resolution[2]. Therefore, one way to quantify the richness of the dynamic response of a neuron is to count the number of different dynamic states it undergoes in the course of the experiment. A proxy of this quantity is given by the variability of the spike frequency k_s, that again can be measured in terms of an entropy

$$\hat{H}[k] = -\sum_{k=1}^{\infty} \frac{km_k}{M} \log_M \frac{km_k}{M}. \tag{5.2}$$

where m_k indicates the number of time bins containing k spikes[3], so that km_k/M is the fraction of spikes that fall in bins with $k_s = k$. $\hat{H}[k]$ takes the form of an information theoretic measure of the information each spike contains on the dynamic state of the neuron at a given resolution[4]. Ref. [31] shows that $\hat{H}[k]$ measures the complexity of the variability in the sense that $\hat{H}[k]$ correlates with the number of parameters a model would require in order to describe properly the dataset, without overfitting. Hence, following Chapter 3.2, we shall call $\hat{H}[s]$ as *resolution* and $\hat{H}[k]$ as *relevance*.

In the current context, the reason for this choice can be understood as follows. In a given task or behavior, different neurons can have activities that are more or less related to the behavioral or neuronal states that are being probed in the experiment. Neurons that are *relevant* for encoding the animal's behavior or task are expected to display variation on a wide range of dynamical states, i.e. to have a large $\hat{H}[k]$. On the contrary, neurons that are not involved in the animal's behavior are expected to visit relatively fewer dynamical states, i.e. to have a lower $\hat{H}[k]$.

Notice that for very small binning times $\Delta t \leq \Delta t_-$ (when each time bins contains at most one spike, i.e. $m_{k=1} = M$ and $m_{k'} = 0,\ \forall\ k' > 1$) we find $\hat{H}[k] = 0$ (and $\hat{H}[s] = 1$). At the opposite extreme, when $\Delta t \geq \Delta t_+$ and $H[s] = 0$, we have all spikes

[2] One may argue that, if the activity in the previous bins $s - 1$ and $s' - 1$ differs considerably, then the dynamic state in bin s and s' may also be considered different. We take the view that this distinction is automatically taken into account when considering larger bins (i.e. $\Delta t \to 2\Delta t$).

[3] m_k satisfies the obvious relation $\sum_{k=0}^{\infty} km_k = \sum_{s=1}^{T} k_s = M$.

[4] Again, the knowledge of the associated dynamical state, i.e. the spike frequency k of the bin it belongs to, provides information to identify the timing of a spike by reducing the number of possible choices from M to km_k, which is the number of spikes in bins with the same dynamical state k. The information gain is given by $\hat{H}[k]$.

in the same bin, i.e. $m_k = 0$ for all $k = 1, 2, \ldots, M-1$ and $m_M = 1$. Therefore again we find $\hat{H}[k] = 0$. Hence, no information on the relevance of the neuron can be extracted at time scales smaller than Δt_- or larger than Δt_+. At intermediate scales $\Delta t \in [\Delta t_-, \Delta t_+]$, $\hat{H}[k]$ takes non-zero values[5], that we take as a measure of the relevance of the neuron for the freely-behaving animals being studied, at time scale Δt.

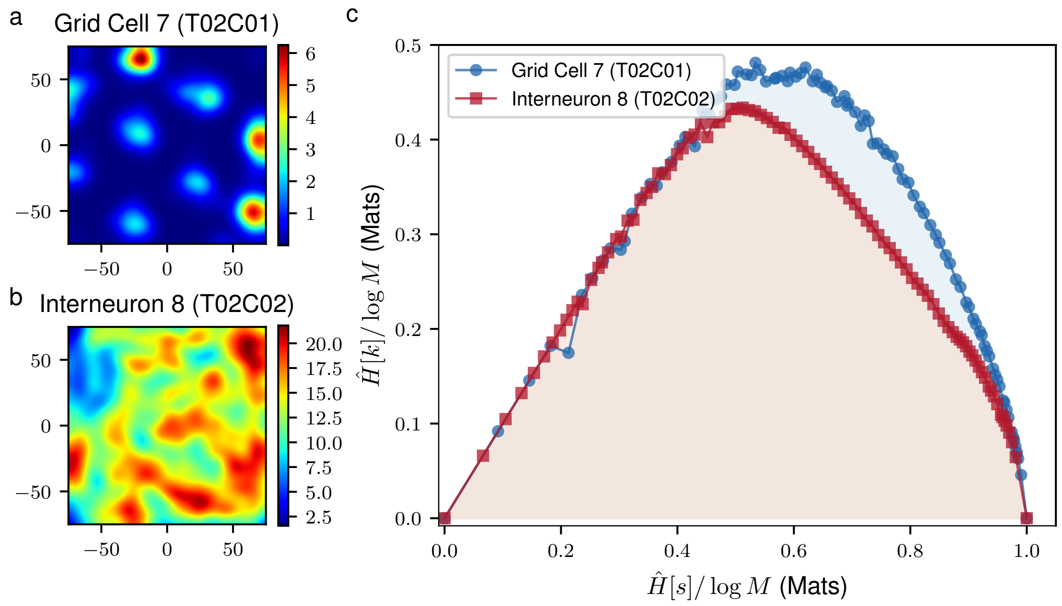

FIGURE 5.2. **Proof of concept of the MSR as a relative information content measure**. The smoothed firing rate maps of a grid cell (**a**) and an interneuron (**b**) in the mEC illustrates the spatial modulation of neural activity. Panel **c** shows the curves traced by the grid cell (green) and interneuron (red). Each point, $(\hat{H}[s], \hat{H}[k])$, in this curve corresponds to a fixed binning time, Δt, with which we see the corresponding temporal neural spike codes.

Yet, the relevant time scale Δt for a neuronal response to a stimulus may not be known *a priori* and/or the latter may evoke a dynamic response that spans multiple time scales. For this reason, we vary the binning time Δt thereby inspecting multiple time scales with which we want to see the temporal code. As we vary Δt, we can trace a curve in the $\hat{H}[s]$-$\hat{H}[k]$ space for every neuron in the sample. Neurons with broad distributions of spike frequencies across different time scales will trace higher curves in this space and in turn, will cover larger areas under this curve (see Fig. 5.2c). Henceforth, we shall call the area under this curve as the *multiscale relevance* (MSR),

[5] Chapter 3.3 shows that, for a given value of the resolution $\hat{H}[s]$, $\hat{H}[k]$ takes its maximal value for distributions $m_k \sim k^{-\mu-1}$ with a power law behavior, where μ varies between 1 and ∞ as $\hat{H}[s]$ varies.

$\mathcal{R}_t$. The *relevant neurons* (RNs), those with high values of $\mathcal{R}_t$, are expected to exhibit spiking behaviors that can be well-discriminated by upstream neural information processing units over short and long time scales and thus, are expected to be *relevant* to the encoding of higher representations.

MSR is designed to capture non-trivial structures in the spike stemming from the variations in spike rates. As such, it is expected to correlate with other measures characterizing temporal structure, such as bursty-ness and memory [104] and the coefficient of local variation [84, 85]. We have observed that, in synthetic data with given characteristics, MSR captures both the bursty-ness, memory and local variations of a time series (see Fig. D.1a,b). In addition, we find, in both synthetic and real data, a negative relation between MSR and spike frequency (i.e. M), which is partly associated with bursty-ness.

As a proof of concept of the MSR for featureless neural selection, we considered two neurons recorded simultaneously from the medial entorhinal cortex (mEC) in Ref. [79] – a grid cell (T02C01) and an interneuron (T02C02) – both of which were measured from the same tetrode and thus, are in close proximity in the brain region. Grid cells, as discussed in Chapter 5.1, have spatially selective firing behaviors that form a hexagonal pattern which spans the environment where the rat freely explores as in Fig. 5.2a. These cells altogether provide the organism with an internal map which it then uses for navigation. On the other hand, interneurons, as in Fig. 5.2b, are inhibitory neurons which are still important towards the formation of grid cell patterns [105, 106, 107] but have much less spatially specific firing patterns. Intuitively, as the mEC functions as a hub for memory and navigation, grid cells, which provide the brain with a representation of space, should be more relevant for an upstream information processing "neuron" (possibly the place cells in the hippocampus) in encoding higher representations compared to interneurons. Indeed, the grid cell traces higher curves in the $\hat{H}[s] - \hat{H}[k]$ space as in Fig. 5.2c and thus defines a larger area compared to the interneuron.

5.3 Results

Following the observations in Fig. 5.2, we sought to characterize the temporal firing behavior of the 65 neurons which were simultaneously recorded from the mEC and its nearby regions of a freely-behaving rat as it explored a square area of length 150 cm [79]. This neural ensemble, as functionally categorized in Ref. [79], consisted of 23 grid cells, 5 interneurons, 1 putative border cell and 36 unclassified neurons, some of which had highly spatially attuned firing and nearly hexagonal spatial firing patterns [79, 108, 109]. This dataset was chosen among the multiple recording sessions performed in Ref. [79] as this contained the most grid cells to be simultaneously recorded.

These results were then corroborated by characterizing the temporal firing behaviors of the 746 neurons which were recorded from multiple anterior thalamic nuclei areas, mainly the anterodorsal nucleus (ADn), and subicular areas, main the postsubiculum (PoS) of 6 different mice across 31 recording session while the mouse explored a rectangular area of dimensions 53 cm × 46 cm [80]. This data was chosen as these heterogeneous neural ensemble contained a number of *head direction (HD) cells* which are neurons that are highly attuned to HD.

Before showing the results on these data sets, we note that the the MSR is a robust measure. To establish this, we compared the MSRs computed using only the first half of the data to that computed from the second half. We obtained very similar results, confirming that the MSR is a reliable measure that can be used to score neurons (see Fig. D.2a).

5.3.1 MSR captures information on functionally relevant external correlates

As the mEC is crucial to spatial navigation, we sought to find whether the wide dynamical variations of neural firing as captured by the MSR would contribute towards a representation of the animal's spatial organization, in one way or another. Different measures relating the spatial position, $\mathbf{x}$, with neural activity had been employed in the literature to characterize spatially specific neural discharges, like the Skaggs-McNaughton spatial information, $I(s, \mathbf{x})$ defined in Eq. (D.3) and in Ref. [110], spatial sparsity measure, $sp_{\mathbf{x}}$ defined in Eq. (D.5) and in Refs. [111, 112] and grid score, g, defined in Eq. (D.6) and in Refs. [91, 109, 113, 114].

Apart from spatial location, HD also plays a crucial role in spatial navigation. The mean vector length, R (Eq. (D.7) in Section D.1.4) is commonly used as a measure of HD selectivity of the activity of neurons. However, this measure assumes that there is only one preferred HD in which a given neuron is tuned to. Hence, we calculated two measures – the HD information, $I(s, \theta)$, and HD sparsity, sp_{θ} – inspired by the spatial information and spatial sparsity to quantify the information and selectivity of neural firing to HD respectively. These measures ought to detect non-trivial and multimodal HD tuning which may also be important in representing HD in the brain [82].

Fig. 5.3 reports the spatial information (a) and the HD information (c) as a function of the MSR for each neuron in the mEC data. Figs. 5.3b and d report the spatial firing rate maps and HD tuning curves for the top five RNs (left panel) and non-RNs (right panel) by MSR score, respectively (See also Figs. D.4 and D.5). We observed that non-RNs had very non-specific spatial and HD discharges as indicated by their sparsity scores (Figs. 5.3b and d, See also Figs. D.4 and D.5) whereas RNs had a broader range of spatial and HD sparsity (Figs. 5.3b and e, See also Figs. D.4 and D.5).

While we have observed that the MSR has a negative relation with the spike frequency (i.e. M), an analysis of the residual MSR revealed that the logarithm of the spike frequency (i.e., $\log M$) could not explain all of the variations in the MSR for the neurons in the mEC. We have seen that the residual MSRs (with respect to $\log M$) appeared to be correlated with spatial and HD information (see Figs. D.2b-d).

Although local variations, as measured by L_V, could still capture spatial and HD information (see Figs. D.3a and b, respectively), we observed that the strength of correlation was stronger for MSR than for L_V. While there is a positive correlation

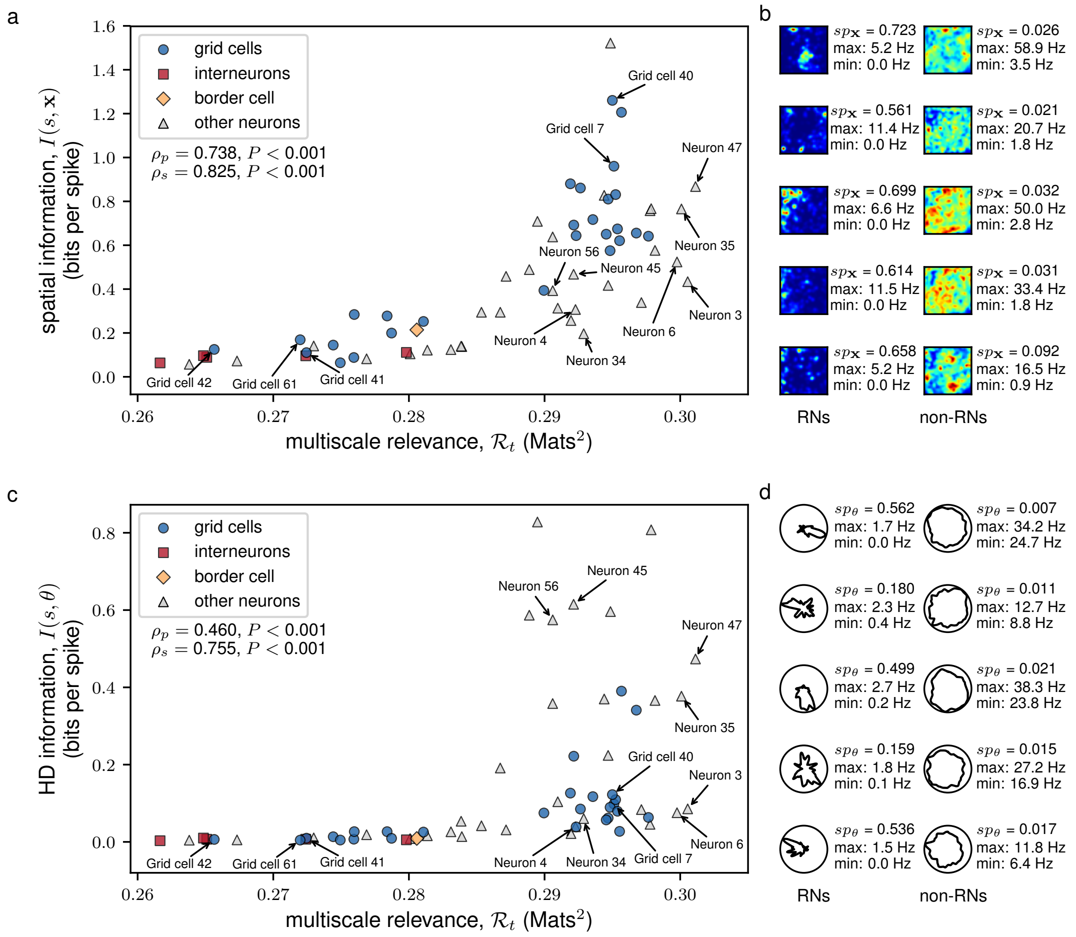

FIGURE 5.3. **The MSR identified neurons that are spatially and head directionally informative**. A scatter plot of the MSR vs. the spatial (HD) information is shown in **a** (**c**). The shapes of the scatter points indicate the identity of the neuron according to Ref. [79]. The linearity and monotonicity of the multiscale relevance and the information measures were assessed by the Pearson's correlation, ρ_p, and the Spearman's correlation, ρ_s, respectively. Information bias was measured by a bootstrapping method, i.e., calculating the average of the spatial or head directional information of 1000 randomized spike trains. The spatial firing rate maps (HD tuning curves) of the 5 most relevant neurons (RNs) and the 5 most irrelevant neurons (non-RNs) are shown together in panel **b** (**d**) together with the calculated spatial sparsity, $sp_{\mathbf{x}}$, (HD sparsity, sp_θ) and maximum and minimum firing.

between L_V and the MSR (see Fig. D.2e), we found that local variations could not explain what is captured by the MSR. In addition, the residual MSRs (with respect to L_V) were observed to still be correlated with spatial or HD information (see Figs. D.2f and g).

We found that *(i)* Neurons with high spatial information or high HD information also had high MSR, but the converse was not true. While there were high RNs that responded exquisitely to space (grid cells 7 and 40) or HD (neurons 45 and 56) alone, the majority (e.g. neurons 35 and 47) encoded significantly both spatial and HD information. Secondly, we found that *ii)* Neurons with low MSR had both low spatial and low HD information (Figs. 5.3c and f), but again, the converse was not true (e.g. neurons 4 and 34). Finally *iii)* we found that some neurons, for example, neurons 3 and 6, despite having some spatial and HD sparsity as indicated in their rate maps (Figs. 5.3b and e), had relatively low spatial and HD information but were both identified to be RNs by MSR. This high MSR suggests that perhaps these neurons responded to different correlates involved in navigation different from spatial location or HD.

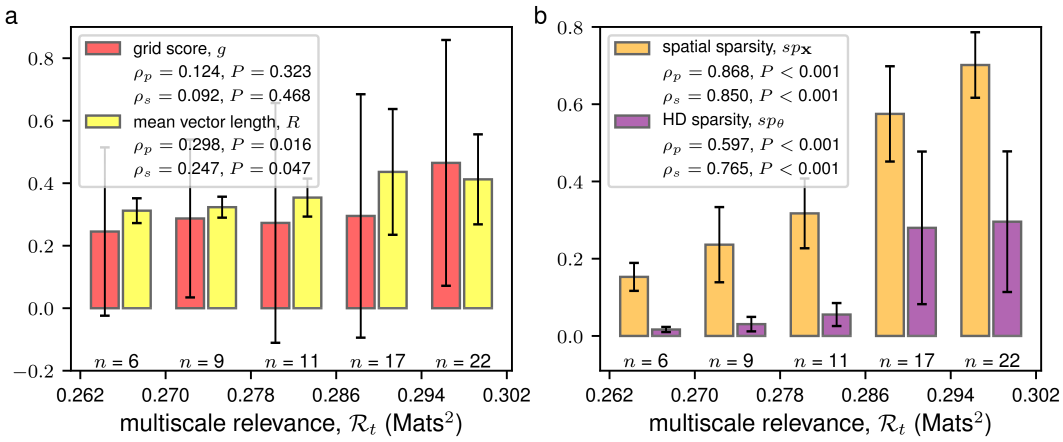

FIGURE 5.4. **The MSR identified neurons with spatially and head directionally selective discharges**. Bar plots depict the mean (height of the bar) along with the standard deviation (black error bars) of the grid score (red) and Rayleigh mean vector length (yellow) in panel **a**, and the spatial sparsity (orange) and HD sparsity (purple) in panel **b** for each neuron in the mEC within the relevance range as indicated. The relevance range was determined by equally dividing the range of the calculated MSR into 5 equal parts. The number of neurons whose MSRs fall within a relevance range is indicated below each bar. The linearity and monotonicity between the MSR and the different spatial and HD quantities were quantified using the Pearson's correlation, ρ_p, and the Spearman's correlation, ρ_s, respectively.

Many of the grid cells were spotted as RNs, but not all. For example, grid cells 41, 42 and 61, that had a significant grid score, had a low MSR and a low spatial information. This indicated that different measures correlate differently with MSR. Fig. 5.4 reports the distribution of the other four measures analyzed in this study conditional to different levels of MSR. Fig. 5.4a shows that grid score maintains a large variation across all scales of the MSR, with a moderate increase in its average. A similar behavior was observed in Fig. 5.4a for the mean vector length.

Spatial sparsity and HD sparsity, instead, showed a significant correlation with the MSR as seen in Fig. 5.4b. The observation that RNs with highly sparse firing may have either low mean vector lengths or low grid scores was an indication that a non-trivial variabilities in firing behaviors need not to obey the imposed symmetries of the tuning curves.

Following the observations in the mEC, we turned to other regions in the brain – the thalamus – and found out whether the non-trivial variability in the neural spiking captured functionally relevant external correlates. To this, we analyzed the neurons in the ADn and PoS areas of freely behaving and navigating rodents. These regions are known to contain cells that robustly fire when the animal's head is facing a specific direction [88, 89] and is believed to be crucial to the formation of grid cells in the mEC [91, 114, 115]. Thus, we sough to find whether the variability as measured by the MSR contain signals of HD tuning. We observed that, in all of the 6 mice that were analyzed, the neurons having HD specific firing, i.e., neurons having high HD sparsity and high mean vector lengths, were RNs (see Fig. D.6). Focusing on a subset of neurons of Mouse 12 (in Fig. D.6a) that were simultaneously recorded in a single session (Session 120806), we observed, as in Figs. 5.5a,b, that HD attuned neurons were RNs. However, the HD alone may not explain the structure of the spike frequencies of these neurons [116]. Hence, we also sought to find whether some of these neurons are spatially tuned. As seen in Figs. 5.5d,e, we found that some of the RNs were also modulated by the spatial location of the mouse. These results were also consistent for a subset of neurons of Mouse 28 (in Fig. D.6f) that were simultaneously recorded in a single session (Session 140313) as in Fig. 5.6.

To assess whether the variations in the spike frequencies, as characterized by the MSR, contained information about external stimuli relevant to navigation, we resampled the spike count code of the neurons in the mEC such that only spatial information, or only HD information, or both spatial and HD information were incorporated. This resampling of the neural spiking was done by generating synthetic spikes assuming a non-homogeneous Poisson spiking with rates taken from the computed spatial firing rate maps and HD tuning curves (see Section D.1.5). These assumptions were able to recover the original rate maps as seen in Figs. 5.7b and c. Here, we focused our

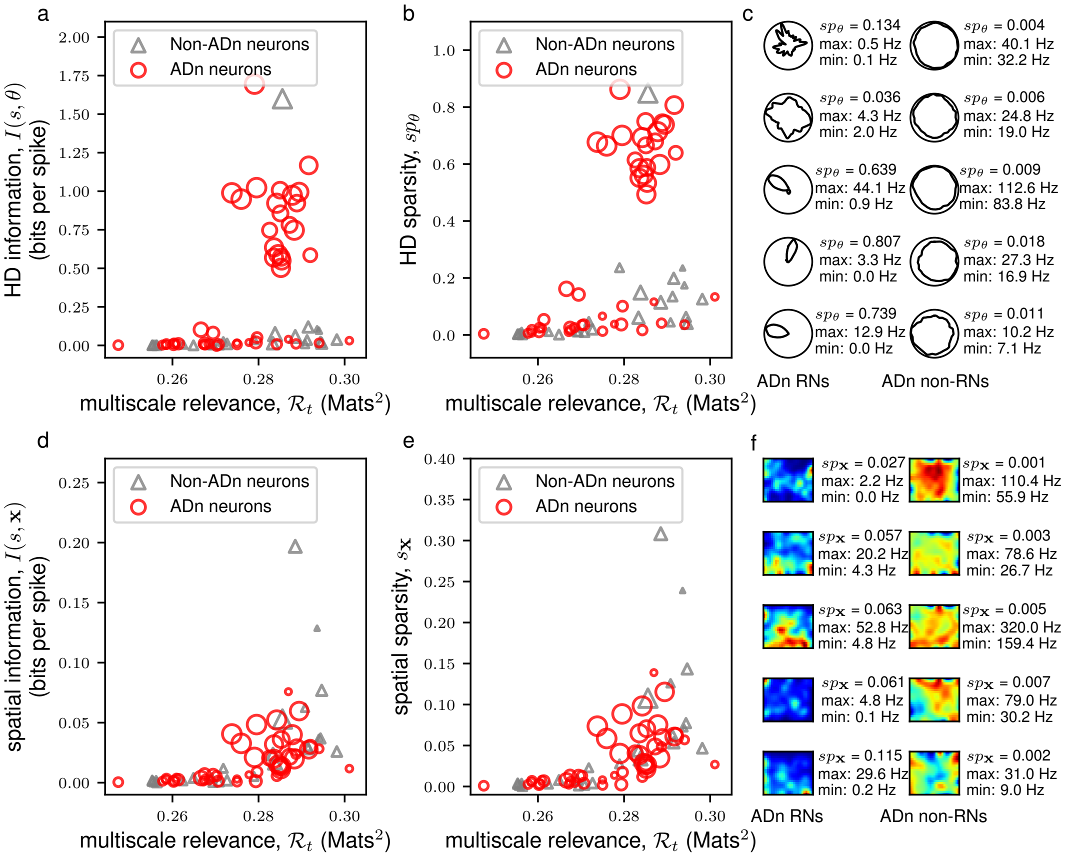

FIGURE 5.5. **MSR of neurons from the anterodorsal thalamic nucleus (ADn) of Mouse 12 from a single recording session (Session 120806)**. A scatter plot of the multiscale relevance vs. the HD (spatial) information is shown in **a** (**d**). This plot is supplemented by a scatter plot between the MSR and HD (spatial) sparsity shown in **b** (**e**). The sizes of the scatter points reflect the mean vector length of the neural activity where the larger scatter points correspond to a sharp preferential firing to a single direction. The HD tuning curves (spatial firing rate maps) of the 5 most relevant neurons (RNs) and the 5 most irrelevant neurons (non-RNs) are shown together in panel **c** (**f**) together with the calculated HD sparsity, sp_θ, (spatial sparsity, $sp_\mathbf{x}$) and maximum and minimum firing.

attention on *mEC Neuron 47* in the mEC data which had the highest MSR and also had both high spatial and high HD information.

By resampling solely the spatial firing rate map as in Fig. 5.7d, we saw a decrease in the MSR despite having as much spatial information as the original code. When HD information was incorporated into the resampled spike frequencies, assuming the factorization of the firing probabilities due to position and HD, more structure would be added onto the spiking activity of the resampled neuron. Hence, we expected to see an increase in the MSR as observed for *Neuron 47* which increased almost up to the MSR for the original code. These findings support the idea that the temporal structure of the spike counts of the neuron, as measured by the MSR, come from its tuning

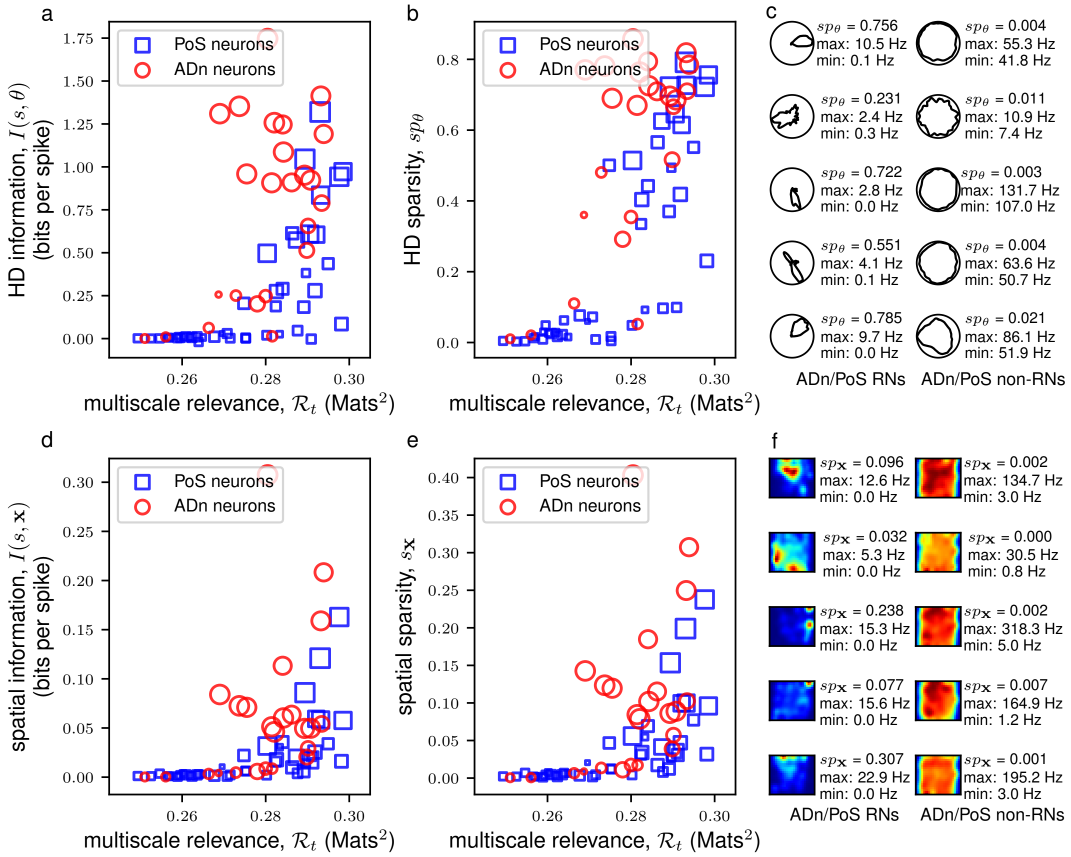

FIGURE 5.6. **MSR of neurons from the anterodorsal thalamic nucleus (ADn) and post-subiculum (PoS) of Mouse 28 from a single recording session (Session 140313)**. A scatter plot of the MSR vs. the HD (spatial) information is shown in **a** (**d**). This plot is supplemented by a scatter plot between the MSR and HD (spatial) sparsity shown in **b** (**e**). The sizes of the scatter points reflect the mean vector length of the neural activity where the larger scatter points correspond to putative head direction cells while the shapes of the scatter points indicate the region where the neuron units were recorded in Refs. [80, 117]. The HD tuning curves (spatial firing rate maps) of the 5 most relevant neurons (RNs) and the 5 most irrelevant neurons (non-RNs) are shown together in panel **c** (**f**) together with the calculated HD sparsity, sp_θ, (spatial sparsity, $sp_\mathbf{x}$) and maximum and minimum firing.

profiles for both position and HD.

We also assessed which cells among the neurons in the mEC have MSRs that could be explained well by the spatial information and thus, were highly spatially attuned. We resampled the spatial firing rate maps of each of the neurons in the mEC data (see Section D.1.5). The difference between the original and resampled MSR, $\mathscr{R}_t^{\text{original}} - \mathscr{R}_t^{\text{resampled}}$, was then computed from the resampled spikes. When the variations in the spike frequencies could be explained by the spatial firing fields, we expected this difference to be close to zero. As seen in Fig. 5.7e, we found that neurons having either high spatial information tended to have differential MSRs close to zero. We

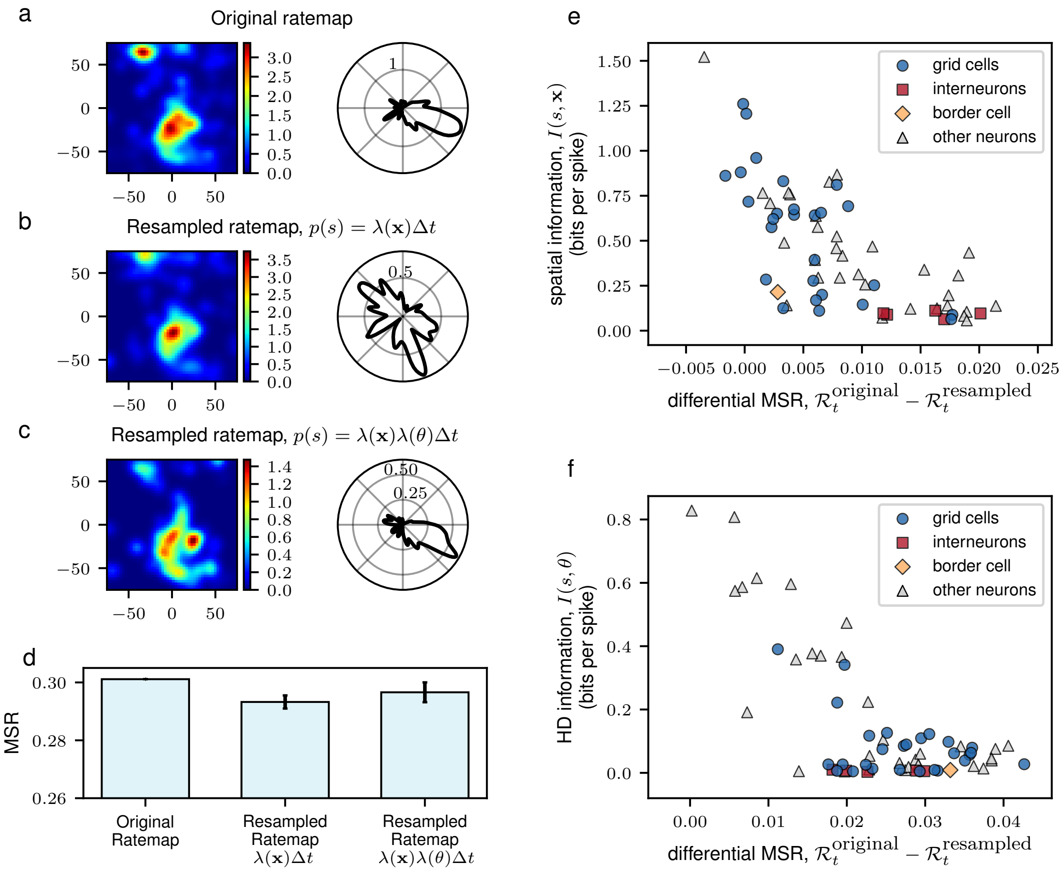

FIGURE 5.7. **The MSR is a measure of information content of the neural activity**. Resampling the firing rate map using spatial position only or in combination with HD resulted to a firing activity that closely resembled the actual firing pattern of mEC Neuron 47. Compared to the original firing rate maps in **a**, the spatial (left panels) and HD (right panels) firing rate maps were recovered by the resampling procedure in **b** and **c**. The result for a single realization of the resampling procedure is shown. (**d**) Bar plots show the MSR calculated from the original spiking activity of the neuron and the resampled rate maps. The mean and standard deviation of 100 realizations of the resampling procedure is reported. Scatter plot between the difference of the MSRs of the original spikes and of the synthetic spikes, resampled using only positional information (only HD information), for each neuron and the spatial information is shown in **e** (**f**).

also observed that most of the neurons having low differential MSRs were grid cells. The same observations could be drawn when resampling the HD tuning curves of each of the neurons in the mEC data. In particular, we also found that neurons having high HD information had differential MSRs close to zero as in Fig. 5.7f.

Taken altogether, these results suggest that the MSR can be used to identify the interesting neurons in a heterogeneous ensemble. The proposed measure is able to capture the non-trivial spike frequency distribution across multiple scales whose structure is highly influenced by external correlates that modulate the neural activity. Indeed, these analyses show that the MSR is able to capture information content of the neural spike code.

5.3.2 Relevant neurons decode external correlates as efficiently as informative neurons

We found in the previous section that neurons with low MSR had low spatial or HD information while higher MSR could indicate low or high values of spatial or HD information. In this section, we show that despite this, high MSR can still be used to select neurons that decode position or HD well. In other words, although high MSR can imply low spatial or HD information, in terms of population decoding, the highly RNs (selected based on only spike frequencies) performs equally well compared to the highly informative neurons (INs, selected using the knowledge of the external covariate).

In order to understand whether MSR could identify neurons in mEC whose firing activity allows the animal to identify its position, we compared the decoding efficiency of the 20 neurons with the highest MSR (top RNs) with that of the 20 neurons with the highest spatial information (top spatial INs) wherein the two sets overlap on 14 neurons (see Fig. D.7a).

To this end, we employed a Bayesian approach to positional decoding wherein the estimated position at the j^{th} time bin, $\hat{\mathbf{x}}_j$, is determined by the position, $\mathbf{x}_j$, which maximizes an *a posteriori* distribution, $p(\mathbf{x}_j|\mathbf{s}_j)$, conditioned on the spike pattern, $\mathbf{s}_j$, of a neural ensemble within the j^{th} time bin i.e.,

$$\hat{\mathbf{x}}_j = \arg\max_{\mathbf{x}_j} p(\mathbf{x}_j|\mathbf{s}_j) = \arg\max_{\mathbf{x}_j} p(\mathbf{s}_j|\mathbf{x}_j)p(\mathbf{x}_j) \tag{5.3}$$

where the last term is due to Bayes rule, $p(\mathbf{s}_j|\mathbf{x}_j)$ is the likelihood of a spike pattern, $\mathbf{s}_j$, given the position, $\mathbf{x}_j$, which depends on a given neuron model and $p(\mathbf{x}_j)$ is the positional occupation probability which can be estimated directly from the data. Fig. 5.8a shows that the top RNs decoded just as efficient as the top spatial INs. It can also

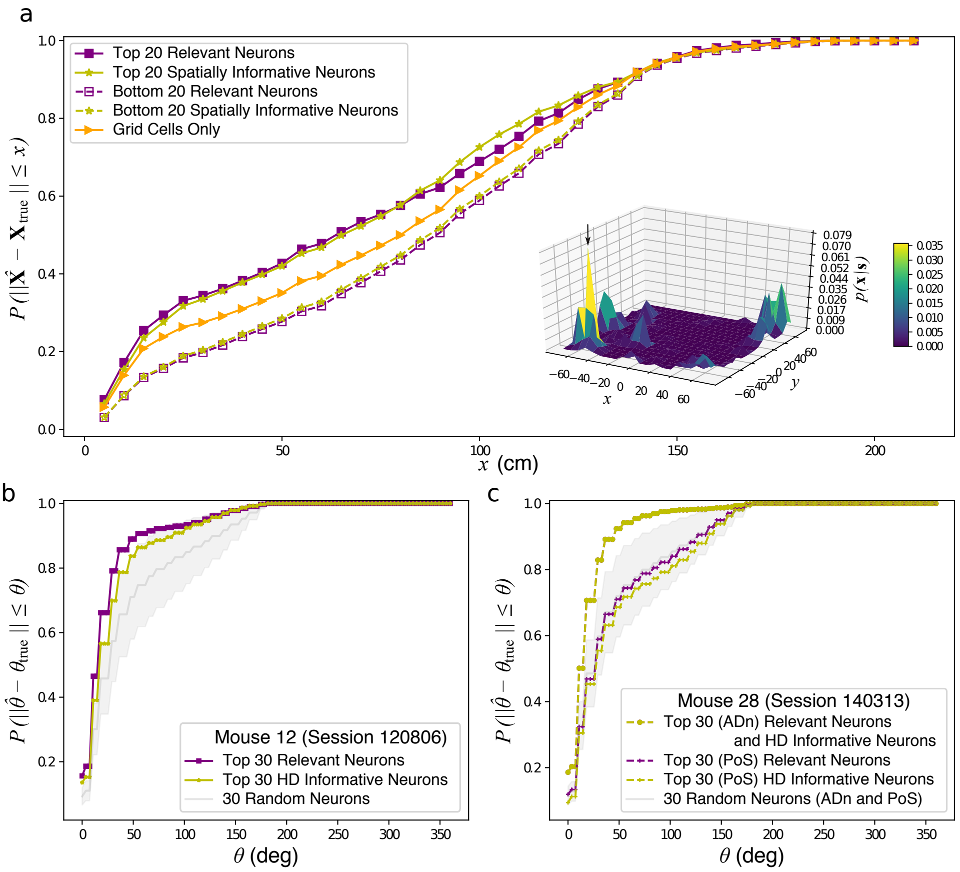

FIGURE 5.8. **Positional decoding of RNs and INs in the mEC and HD decoding of the RNs and INs in the ADn of Mouse 12 and the ADn and PoS of Mouse 28 under a single recording session**. Panel **a** shows the cumulative distribution of the decoding error, $\|\hat{\mathbf{X}} - \mathbf{X}_{true}\|$, for the RNs (solid violet squares) and spatially INs (solid yellow stars) neurons as well as for the non-RNs (dashed violet squares) and non-INs (dashed yellow stars). Spatial decoding was also performed for the 27 grid cells in the mEC data (solid orange triangles). The low positional decoding efficiency at some time points can be traced to the posterior distribution, $p(\mathbf{x}|\mathbf{s})$, of the rat's position given the neural responses which exhibited multiple peaks as shown in the inset surface plot. For this particular example, the true position was found close to the maximal point of the surface plot as indicated by the arrows although such was not always the case. Panel **b** depicts the cumulative distribution of the decoding errors of the 30 RNs (violet squares) and 30 HD INs (yellow stars) in the ADn of Mouse 12 in Session 120806. The mean and standard errors of the cumulative distribution of decoding errors of 30 randomly selected ADn neuron ($n = 1000$ realizations) are shown in grey. On the other hand, panel **c** depicts the cumulative decoding error distribution of the 30 RNs (violet squares) and 30 HD INs (yellow stars) in the ADn (crosses) and PoS (circles) of Mouse 28 in Session 140313. The mean and standard errors of the cumulative distribution of decoding errors of 30 randomly selected ADn or PoS neuron ($n = 1000$ realizations) are shown in grey. As the random selection included neurons from the ADn, which contain a pure head directional information and can decode the positions better than the neurons in the PoS, the decoding errors from the 30 randomly selected neurons were, on average, comparable to that of the relevant or head directionally informative PoS neurons. In all the decoding procedures, time points where all the neurons in the ensemble was silent were discarded in the decoding process.

be observed that the top RNs decode the positions better than the ensemble composed solely of grid cells.

Because of the sizable overlap between the top RNs and the top spatial INs, one might argue that much of the spatial information needed for positional decoding is concentrated on the neurons in the overlap (ONs). To address this, we randomly selected 6 neurons among the mEC neurons outside the overlap and, together with the 14 ONs, decoded for the position as done above. If the positional decoding information is contained in the ONs, then we should observe the same decoding efficiency as either the top RNs or top spatial INs. However, we found that the decoding efficiency of the ONs decreased (see Fig. D.7d). We also found that for the decoded positions within 5 cm from the true position, the decoding efficiency of the top RNs were up to 4σ from the mean decoding efficiency of the ONs, as measured by the z-score compared to that of the top spatial INs which was at around 2σ. This indicates that the 6 RNs outside the overlap provide better decodable spatial representation than those of the 6 spatial INs.

Because local variations of the neurons in the mEC correlated with spatial information, we sought to find whether neurons with high local variations (LVNs) also contained decodable spatial representation. We took the top 20 LVNs in the mEC and decoded for the position as done above. We found that the decoding efficiency of top LVNs are much lower compared to top RNs (see Fig. D.3c). This indicates that the repertoire of responses coming from local variations in the interspike intervals of mEC neurons alone can not represent space in freely-behaving rats.

To substantiate the decoding results obtained for neurons in the mEC, we also took the ADn RNs and the HD INs in the of Mouse 12 (Session 120806) in Fig. 5.5 to decode for HD. Mouse 12 was chosen as this animal had the most HD cells recorded among the mice that only had recordings in the ADn [117]. In particular, we looked at the HD decoding at longer time scales (in this case, $\Delta t = 100$ ms), where we could model the neural activity using a Poisson distribution, $p(\mathbf{n}_j|\theta_j)$ similar to that in Eq. (D.11). Bayesian decoding adopts an equation

$$\hat{\theta}_j = \arg\max_{\theta_j} p(\theta_j|\mathbf{n}_j) = \arg\max_{\theta_j} p(\mathbf{n}_j|\theta_j)p(\theta_j) \tag{5.4}$$

similar to Eq. (5.3) to estimate the decoded HD, $\hat{\theta}_j$, where $p(\theta_j)$ is the HD occupation as estimated from the data. We compared the decoding efficiency of the 30 RNs with the 30 HD INs which had 22 neurons that are relevant (see Fig. D.3b). We also compared the decoding efficiencies of the ADn RNs or HD INs with 30 randomly selected ADn neurons ($n = 1000$ realizations). As seen in Fig. 5.8b, the RNs decoded just as well as the neural population composed of HD INs. Furthermore, the decoding efficiency of the RNs were observed to be far better than the decoding efficiency of a random selection of neurons in the ensemble.

We also compared the decoding efficiency of the ADn and PoS neurons from Mouse 28 (Session 140313) which had the most HD cells recorded among the mice that had recordings in both ADn and PoS [117] as in Fig. 5.6. As seen in Fig. 5.8c, neurons in the ADn decoded the HD more efficiently than the neurons in the PoS. These results are consistent with the notion that the ADn contains pure HD modulation which allow for neurons in the ADn to better predict the mouses HD compared to the neurons in the PoS which contain, instead, true spatial information [80, 116]. For the neurons in Mouse 28 (Session 140313), it had to be noted that the 30 ADn RNs also happened to be the 30 ADn HD INs (see Fig. D.7c). On the other hand, among the 30 PoS RNs, 23 were HD INs (also see Fig. D.7c). We observed that the PoS RNs decode just as efficient as the PoS HD INs consistent with the findings for Mouse 12 (Session 120806).

Taken altogether, despite being blind to the rat's position and of the mouse's HD, the MSR is able to capture neurons that can decode the position and HD just as well as the spatial INs and as the HD INs.

5.4 Discussion

In the present work, we introduced a novel, parameter-free and fully featureless method – which we called multiscale relevance (MSR) – to characterize the temporal structure of the activities of neurons within a heterogeneous population. We have shown that the neurons showing persistently broad spike frequency distributions across a wide range of time scales, as measured by the MSR, usually carry information about the external correlates related to the behavior of the observed animal. By analyzing the neurons in the mEC and nearby brain regions and the neurons in the ADn and PoS – areas in the brain that are pertinent to spatial navigation – we showed that the RNs in these regions have firing behaviors that are selective for spatial location and HD. Here, we found that in many cases, the neurons that display broad spike distributions tend to have conjugated representations in that they exhibit high mutual information with multiple behavioral features. These findings are consistent with those observed experimentally in Ref. [91] and statistically in Ref. [82].

The fact that the MSR can be used to select informative neurons as well as neurons that show high decoding performance is consistent with the assumption that the information carried by the activity of a given neuron is encapsulated in the long-ranged statistical patterns of the spike activity. In order to quantify this information, we used the ideas in Refs. [31, 30] to hypothesize that neurons having such non-trivial temporal structures, as manifested by broad distributions of the neural firing behavior, are important to the representations that the brain region encodes. At a given resolution, as defined in Eq. (5.1), we estimate the complexity of the temporal code by the relevance defined in Eq. (5.2). The latter captures the broadness of the spike frequency distribution at that resolution. Since natural and dynamic stimuli and behaviors often operate on multiple time scales, the MSR integrates over different resolution scales, thus allowing us to spot neurons exhibiting persistent non-trivial spike codes across a broad range of time scales.

Broad distributions of spike frequencies, characterized by a high MSR, exhibit a stochastic variablility that requires richer parametric models [31]. In a decoding perspective, these non-trivial distributions afford a higher degree of distinguishability of neural responses to a given stimuli or behavior. Indeed, by decoding for either spatial position or for HD using statistical approaches, we found that the responses of the RNs allow upstream processing units to efficiently decode the external correlates just as well as the neurons whose resulting tuning maps contain information about those external correlates.

Finally, we observed that the population of relevant neurons, as identified by the MSR, is not homogeneous, e.g., the relevant neurons in the mEC data are not composed solely by grid cells and the relevant neurons in the ADn and PoS are not necessarily composed solely of HD cells. Noteworthy, the decoding efficiency of the relevant neurons was observed to be better compared to the ensemble comprising solely the grid cells. When taken altogether, these observations support the idea that population heterogeneity may play a role towards efficient encoding of stimuli [118, 119].

The insistence on broad distributions, on which the MSR relies on, tails with the fact that biological systems such as the one under study, hardly ever generates well-sampled datasets of their complex behavior. The dynamical range which the experimenter can probe is limited by the size of the dataset and its often far from saturating biological dynamical ranges. The MSR takes advantage of this feature and identifies those variables that exhibit a richer variability. This intuition, discussed theoretically in Refs. [31, 30], has also been used to identify biologically and evolutionary relevant amino acid sites in protein sequences [46]. Indeed, in spite of all advances in sequencing techniques, the genomes from which we can learn are only those left to us by evolution. Hence, Ref. [46] show that subsequences that exhibit a wider response in frequency – as measured by Eq. (5.2) – to evolutionary dynamics contain a wealth of biologically relevant information. This same strategy can also be used to identify relevant and hidden (latent) variables in statistical learning (see e.g. Refs. [120, 48]).

In principle, the MSR can be extended to measure the information carried by the spike time series with respect to known external correlates or features, such as spatial location and HD. This requires discretizing the feature space (e.g. space) in bins and computing the number of time with which a neuron fires in each bin. From the distribution of these counts, one can derive a measure of resolution and relevance, as in Eqs. (5.1,5.2), and draw a curve as in Fig. 5.2 upon changing the bin size. Like the MSR discussed here, this MSR would be designed to spot neurons having spike frequencies with non-trivial distributions when projected onto the feature space. In particular, grid cells have been shown to exhibit field-to-field variability [109, 121] which is far from being an artefact of non-uniform spatial sampling. These variabilities offer grid cells an additional channel for transmission of local spatial information [122] and thus, has multiple implications including the capacity of grid cells to contain contextual information contrary to the findings in Refs. [123, 124, 125] as well as the remapping of place cells without the need for changing grid cell phases [121]. Although the application of MSR adapted to space to characterize spatial inhomogeneity of firing behavior is an interesting avenue of further research, here we limited our analysis to temporal binning precisely because it is defined in terms of the sole spike activity – the only information available to upstream neurons – to decode a representation of the feature space.

The fact that the MSR captures functional information from the temporal code is a remarkable feat of this measure. This method can then be used as a pre-processing tool to impose a less stringent criteria compared to those widely used in many studies (e.g., mean vector length, spatial sparsity and grid scores) thereby directing further investigation to interesting neurons. The MSR is expected to be particularly useful in detecting relevant neurons in high-throughput studies where the activity of thousands of neurons are measured and where the function of these neural ensemble are not necessarily known *a priori*.

Whether this measure can also be used to identify functionally relevant neuronal units recorded through calcium imaging or through fMRI is also an exciting direction for future studies. Furthermore, while the current application focuses on multiple-electrode experiments, we also expect MSR to be useful even in single-electrode neural recordings where, under a given task, an experiment is done multiple times. Since the variations as captured by the MSR are a possible signature of a neuron under a particular task or behavior, the MSR can be used to assess whether the electrode was struck on the right brain region or not. Finally, these same principles used to construct MSR can also be used to study neural assemblies [126] and to probe on the importance of correlated firing of neurons in representing external stimuli or behaviors.

6

Conclusions

We live in an age where technological advancements have allowed us to probe the components of many biological systems simultaneously. From the sequencing of an organism's genome to single neuron recordings in different brain regions, these experiments have allowed us to appreciate the complexity of how the natural world around us operates. However, these advancements have also made us realize how tiny a fraction of Nature we can observe. Alongside these technical progress, an explosion of machine learning algorithms have allowed for the reconstruction of the underlying structure in these empirical data. However, the more we have expanded the scope of what we can observe, the more we have realized that the complexity of natural systems grow very fast. Because the size of samples that we can collect is much, much smaller than the dimensionality in which Nature operates, it is crucial to perform appropriate reduction schemes in order to transform the high-dimensional and noisy data into a low-dimensional representation where the statistics are sufficient to make robust conclusions. Guiding principles are then needed to reduce the dimensionality of the problem in ways that minimize the loss of information (in the sense of the data processing inequality).

In this thesis, we work around the principle that the information content of a sample can be quantified by the coding cost, i.e. the number of bits that are needed to code each datum. Some of these bits convey useful information on the generative process, some are just noise. Here, we have shown that an upper bound to the number of useful bits is given by the entropy of the frequency with which different outcomes occur in the sample, that we call relevance. This allows us to define maximally informative samples as those that maximize the trade-off between how detailed we wanted to see our data (i.e., the resolution or coding cost) and how informative the samples are given this detail (i.e., the relevance). As such, maximizing this information content amounts to finding representations that are efficient, i.e., the sample is expressed in terms of the relevant variables.

In order to verify our claims, we then sought to study the properties of samples that are efficiently represented. To this, we resorted to the minimum description length principle to find a distribution which compresses a sample efficiently even in the worst-case scenario when the sample we observe leads to a costly encoding. We show that the codes that achieve optimal compression in MDL are critical in a very precise sense. First, when they are taken as generative models of samples, they generate samples with broad empirical distributions and with an high value of the relevance, defined as the entropy of the empirical frequencies. Second, MDL codes sit precisely at a second order phase transition point where the symmetry between the sampled outcomes is spontaneously broken. The order parameter controlling the phase transition is the coding cost of the samples. The phase transition is a manifestation of the optimality of MDL codes, and it arises because codes that achieve a higher compression do not exist. These results suggest a clear interpretation of the widespread occurrence: *statistical criticality is a consequence of an efficient representation.*

Besides the academic interest of such an interpretation of statistical criticality, the implication of our work on machine learning and data analysis are far reaching because the proposed measure of relevance can then be used as a guiding principle towards the search of optimal dimensional reduction schemes or the identification of relevant variables. Here, we have shown that we can use these principles to characterize how the information on the neural response patterns are represented across multiple temporal resolutions. We found that neurons having low MSR tend to have low mutual information and low firing sparsity across the correlates that are believed to be encoded by the region of the brain where the recordings were made. In addition, neurons with high MSR contain significant information on spatial navigation and allow to decode spatial position or head direction as efficiently as those neurons whose firing activity has high mutual information with the covariate to be decoded.

In all of these discussions, we have assumed that the only information we have is the empirical distribution of the frequencies with which the information content $\hat{H}[k]$ is measured. In practice, however, one may have other prior information about the structure of the data. It is then a particularly interesting direction for future research how the theory discussed in Chapter 3.2 will be modified to incorporate such additional information on the true distribution $p(s)$. Furthermore, the discussions we have presented assume independence between observations which never always holds. Indeed, it would also be of interest how non-stationarity between the data generating process will modify our current understanding of relevance.

Another particularly interesting direction for future investigation is that of finding the maximal trade-off between the resolution $\hat{H}[s]$ and the relevance $\hat{H}[k]$ as a guiding principle towards dimensionality reduction for inference. Such a method will be a drastic improvement from the method in Ref. [46] wherein they fix $\hat{H}[s]$ by fixing the size of the subset of variables which maximizes $\hat{H}[k]$. We believe that this approach will enable us to capture the most relevant variables in a given dataset without the need for specifying the subsample size. The relevant variables can then be studied further to understand the underlying mechanisms that drive the dynamics of the system or explain the variations in the data. However, as it is now, some technical aspects need to be addressed in order to optimize the algorithm for variable selection through relevance maximization and make it more accessible for public use. This provides an exciting avenue for future collaborations in the computer science as well as in the machine learning community.

This page has been intentionally left blank.

Bibliography

[1] Miguel A Muñoz. Colloquium: Criticality and dynamical scaling in living systems. *Reviews of Modern Physics*, 90(3):031001, 2018.

[2] Mark EJ Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[3] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[4] Jayanth R Banavar, Amos Maritan, and Igor Volkov. Applications of the principle of maximum entropy: from physics to ecology. *Journal of Physics: Condensed Matter*, 22(6):063101, 2010.

[5] Jayanth R Banavar, Amos Maritan, and Andrea Rinaldo. Size and form in efficient transportation networks. *Nature*, 399(6732):130, 1999.

[6] Amos Maritan, Andrea Rinaldo, Riccardo Rigon, Achille Giacometti, and Ignacio Rodríguez-Iturbe. Scaling laws for river networks. *Physical review E*, 53(2):1510, 1996.

[7] Filippo Simini, Tommaso Anfodillo, Marco Carrer, Jayanth R Banavar, and Amos Maritan. Self-similarity and scaling in forest communities. *Proceedings of the National Academy of Sciences*, 107(17):7658–7662, 2010.

[8] Hans Ter Steege, Nigel CA Pitman, Daniel Sabatier, Christopher Baraloto, Rafael P Salomão, Juan Ernesto Guevara, Oliver L Phillips, Carolina V Castilho, William E Magnusson, Jean-François Molino, et al. Hyperdominance in the amazonian tree flora. *Science*, 342(6156):1243092, 2013.

[9] Richard Condit, Suzanne Lao, Rolando Pérez, Steven B. Dolins, Robin Foster, and Stephen Hubbell. Barro colorado forest census plot data (version 2012) (https://doi.org/10.5479/data.bci.20130603), 2012.

[10] Felisa A Smith, S Kathleen Lyons, SK Morgan Ernest, Kate E Jones, Dawn M Kaufman, Tamar Dayan, Pablo A Marquet, James H Brown, and John P Haskell. Body mass of late quaternary mammals. *Ecology*, 84(12):3403–3403, 2003.

[11] Andrea Mazzolini, Marco Gherardi, Michele Caselle, Marco Cosentino Lagomarsino, and Matteo Osella. Statistics of shared components in complex component systems. *Physical Review X*, 8(2):021023, 2018.

[12] Laura Bottomley. Epa-http. (http://ita.ee.lbl.gov/html/contrib/epa-http.html), 1995.

[13] Simon A Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, et al. Cosmic: somatic cancer genetics at high-resolution. *Nucleic acids research*, 45(D1):D777–D783, 2016.

[14] Socorro Gama-Castro, Heladia Salgado, Alberto Santos-Zavaleta, Daniela Ledezma-Tejeida, Luis Muñiz-Rascado, Jair Santiago García-Sotelo, Kevin Alquicira-Hernández, Irma Martínez-Flores, Lucia Pannier, Jaime Abraham Castro-Mondragón, et al. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic acids research*, 44(D1):D133–D143, 2015.

[15] Rama Balakrishnan, Julie Park, Kalpana Karra, Benjamin C Hitz, Gail Binkley, Eurie L Hong, Julie Sullivan, Gos Micklem, and J Michael Cherry. Yeastmine – an integrated data warehouse for *Saccharomyces cerevisiae* data as a multipurpose tool-kit. *Database*, 2012, 2012.

[16] Thierry Mora and William Bialek. Are biological systems poised at criticality? *Journal of Statistical Physics*, 144(2):268–302, 2011.

[17] Edward D Lee, Chase P Broedersz, and William Bialek. Statistical mechanics of the US Supreme Court. *Journal of Statistical Physics*, 160(2):275–301, 2015.

[18] Thierry Mora, Aleksandra M Walczak, William Bialek, and Curtis G Callan. Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences*, 107(12):5405–5410, 2010.

[19] William Bialek, Andrea Cavagna, Irene Giardina, Thierry Mora, Edmondo Silvestri, Massimiliano Viale, and Aleksandra M Walczak. Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences*, 2012.

[20] Gašper Tkačik, Thierry Mora, Olivier Marre, Dario Amodei, Stephanie E Palmer, Michael J Berry, and William Bialek. Thermodynamics and signatures of criticality in a network of neurons. *Proceedings of the National Academy of Sciences*, 112(37):11508–11513, 2015.

[21] Per Bak. *How Nature Works: The Science of Self-Organized Criticality, Copernicus*. Springer Science+Business Media New York, 1996.

[22] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96, 2012.

[23] Jorge Hidalgo, Jacopo Grilli, Samir Suweis, Miguel A Muñoz, Jayanth R Banavar, and Amos Maritan. Information-based fitness and the emergence of criticality in living systems. *Proceedings of the National Academy of Sciences*, 111(28):10095–10100, 2014.

[24] Woodrow L Shew, Wesley P Clawson, Jeff Pobst, Yahya Karimipanah, Nathaniel C Wright, and Ralf Wessel. Adaptation to sensory input tunes visual cortex to criticality. *Nature Physics*, 11(8):659, 2015.

[25] Iacopo Mastromatteo and Matteo Marsili. On the criticality of inferred models. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(10):P10012, 2011.

[26] David J Schwab, Ilya Nemenman, and Pankaj Mehta. Zipf's law and criticality in multivariate data without fine-tuning. *Physical Review Letters*, 113(6):068102, 2014.

[27] Laurence Aitchison, Nicola Corradi, and Peter E Latham. Zipf's law arises naturally in structured, high-dimensional data. *arXiv preprint arXiv:1407.7135*, 2014.

[28] Laurence Aitchison, Nicola Corradi, and Peter E Latham. Zipf?s law arises naturally when there are underlying, unobserved variables. *PLoS Computational Biology*, 12(12):e1005110, 2016.

[29] Marcel Nonnenmacher, Christian Behrens, Philipp Berens, Matthias Bethge, and Jakob H Macke. Signatures of criticality arise from random subsampling in simple population models. *PLoS Computational Biology*, 13(10):e1005718, 2017.

[30] Matteo Marsili, Iacopo Mastromatteo, and Yasser Roudi. On sampling and modeling complex systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(09):P09003, 2013.

[31] Ariel Haimovici and Matteo Marsili. Criticality of mostly informative samples: a Bayesian model selection approach. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(10):P10013, 2015.

[32] Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

[33] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[34] Edwin T Jaynes. Information theory and statistical mechanics. ii. *Physical review*, 108(2):171, 1957.

[35] George Kingsley Zipf. *Selected studies of the principle of relative frequency in language*. Harvard university press, 1932.

[36] Henrik Jeldtoft Jensen. *Self-organized criticality: emergent complex behavior in physical and biological systems*, volume 10. Cambridge university press, 1998.

[37] Jakob H Macke, Manfred Opper, and Matthias Bethge. Common input explains higher-order correlations and entropy in a simple model of neural population activity. *Physical Review Letters*, 106(20):208102, 2011.

[38] Xavier Gabaix. Zipf's law for cities: an explanation. *The Quarterly Journal of Economics*, 114(3):739–767, 1999.

[39] Charlotte James, Sandro Azaele, Amos Maritan, and Filippo Simini. Zipf's and taylor's laws. *Physical Review E*, 98(3):032408, 2018.

[40] Chris G Langton. Computation at the edge of chaos: phase transitions and emergent computation. *Physica D: Nonlinear Phenomena*, 42(1-3):12–37, 1990.

[41] Nils Bertschinger and Thomas Natschläger. Real-time computation at the edge of chaos in recurrent neural networks. *Neural computation*, 16(7):1413–1436, 2004.

[42] Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A*, 222(594-604):309–368, 1922.

[43] Javier D Burgos and Pedro Moreno-Tovar. Zipf-scaling behavior in the immune system. *Biosystems*, 39(3):227–232, 1996.

[44] Daniel L Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. In *Advances in Neural Information Processing Systems*, pages 551–558, 1994.

[45] Matteo Marsili et al. Dissecting financial markets: sectors and states. *Quantitative Finance*, 2(4):297–302, 2002.

[46] Silvia Grigolon, Silvio Franz, and Matteo Marsili. Identifying relevant positions in proteins by Critical Variable Selection. *Molecular BioSystems*, 12(7):2147–2158, 2016.

[47] Ryan John Cubero, Matteo Marsili, and Yasser Roudi. Finding informative neurons in the brain using Multi-Scale Relevance. *arXiv preprint arXiv:1802.10354*, 2018.

[48] Juyong Song, Matteo Marsili, and Junghyo Jo. Resolution and relevance tradeoffs in Deep Learning. *arXiv preprint arXiv:1710.11324*, 2017.

[49] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[50] Yann LeCun, Corinna Cortes, and Christopher JC Burges. The MNIST database of handwritten digits. (https://yann.lecun.com/exdb/mnist/), 1998.

[51] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[52] Erich Leo Lehmann and Henry Scheffé. Completeness, similar regions, and unbiased estimation: Part i. *Sankhya*, 10(4):305–340, 1950.

[53] Abram M Kagan and Yaakov Malinovsky. On the structure of umvues. *Sankhya*, 78(1):124–132, 2016.

[54] Susanne Still, William Bialek, and Léon Bottou. Geometric clustering using the information bottleneck method. In *Advances in Neural Information Processing Systems*, pages 1165–1172, 2004.

[55] DJ Strouse and David J Schwab. The deterministic information bottleneck. *Neural Computation*, 29(6):1611–1630, 2017.

[56] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[57] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[58] Jose M Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 113–147, 1979.

[59] Per Bak, Chao Tang, and Kurt Wiesenfeld. Self-organized criticality: and explanation of $1/f$ noise. *Physical Review Letters*, 59:381–384, 1987.

[60] Ryan John Cubero, Junghyo Jo, Matteo Marsili, Yasser Roudi, and Juyong Song. Minimally sufficient representations, maximally informative samples and Zipf's law. *arXiv preprint arXiv:1808.00249*, 2018.

[61] Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.

[62] Ming Li and Paul Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer Heidelberg, 1997.

[63] Alan M Turing. On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London mathematical society*, 2(1):230–265, 1937.

[64] Jorma Rissanen. *Information and complexity in statistical modeling*. Springer Science & Business Media, 2007.

[65] Peter D Grünwald. A tutorial introduction to the minimum description length principle. *arXiv preprint arxiv:math/0406077*, 2004.

[66] Yuri M Shtarkov. Universal sequential coding of single messages. *(*translated from*) Problems of Information Transmission*, 23:175–186, 1987.

[67] Jorma J Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, 1996.

[68] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In Boris Nikolaevich Petrov and Frigyes Cski, editors, *2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971*, pages 267–281, 1973.

[69] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[70] Vijay Balasubramanian. MDL, Bayesian inference, and the geometry of the space of probability distributions. *Advances in minimum description length: Theory and applications*, pages 81–98, 2005.

[71] Alberto Beretta, Claudia Battistin, Clélia de Mulatier, Iacopo Mastromatteo, and Matteo Marsili. The stochastic complexity of spin models: Are pairwise models really simple? *Entropy*, 20(10):739, 2018.

[72] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[73] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[74] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[75] Mario Filiasi, Giacomo Livan, Matteo Marsili, Maria Peressi, Erik Vesselli, and Elia Zarinelli. On the concentration of large deviations for fat tailed distributions, with application to financial data. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(9):P09030, 2014.

[76] Marco Zannetti, Federico Corberi, and Giuseppe Gonnella. Condensation of fluctuations in and out of equilibrium. *Physical Review E*, 90(1):012143, 2014.

[77] Federico Corberi. Large deviations, condensation and giant response in a statistical system. *Journal of Physics A: Mathematical and Theoretical*, 48(46):465003, 2015.

[78] Simon B Laughlin. Energy as a constraint on the coding and processing of sensory information. *Current opinion in neurobiology*, 11(4):475–480, 2001.

[79] Hanne Stensola, Tor Stensola, Trygve Solstad, Kristian Frøland, May-Britt Moser, and Edvard I Moser. The entorhinal grid map is discretized. *Nature*, 492(7427):72–78, 2012.

[80] Adrien Peyrache, Marie M Lacroix, Peter C Petersen, and György Buzsáki. Internally organized mechanisms of the head direction sense. *Nature Neuroscience*, 18(4):569–575, 2015.

[81] Stefano Panzeri, Nicolas Brunel, Nikos K Logothetis, and Christoph Kayser. Sensory neural codes using multiplexed temporal scales. *Trends in Neurosciences*, 33(3):111–120, 2010.

[82] Kiah Hardcastle, Niru Maheswaranathan, Surya Ganguli, and Lisa M Giocomo. A multiplexed, heterogeneous, and adaptive code for navigation in medial entorhinal cortex. *Neuron*, 94(2):375–387, 2017.

[83] Rebecca A Mease, Thomas Kuner, Adrienne L Fairhall, and Alexander Groh. Multiplexed spike coding and adaptation in the thalamus. *Cell Reports*, 19(6):1130–1140, 2017.

[84] Shigeru Shinomoto, Keiji Miura, and Shinsuke Koyama. A measure of local variation of inter-spike intervals. *Biosystems*, 79(1-3):67–72, 2005.

[85] Shigeru Shinomoto, Keisetsu Shima, and Jun Tanji. Differences in spiking patterns among cortical neurons. *Neural Computation*, 15(12):2823–2842, 2003.

[86] David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148(3):574–591, 1959.

[87] Michael M Merzenich, Paul L Knight, and G Linn Roth. Representation of cochlea within primary auditory cortex in the cat. *Journal of Neurophysiology*, 38(2):231–249, 1975.

[88] Jeffrey S Taube, Robert U Muller, and James B Ranck. Head-direction cells recorded from the postsubiculum in freely moving rats. i. description and quantitative analysis. *Journal of Neuroscience*, 10(2):420–435, 1990.

[89] Jeffrey S Taube. Head direction cells recorded in the anterior thalamic nuclei of freely moving rats. *Journal of Neuroscience*, 15(1):70–86, 1995.

[90] John O'Keefe and Jonathan Dostrovsky. The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34:171–175, 1971.

[91] Francesca Sargolini, Marianne Fyhn, Torkel Hafting, Bruce L McNaughton, Menno P Witter, May-Britt Moser, and Edvard I Moser. Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science*, 312(5774):758–762, 2006.

[92] Emilio Kropff, James E Carmichael, May-Britt Moser, and Edvard I Moser. Speed cells in the medial entorhinal cortex. *Nature*, 523(7561):419–424, 2015.

[93] Tor Stensola, Hanne Stensola, May-Britt Moser, and Edvard I Moser. Shearing-induced asymmetry in entorhinal grid cells. *Nature*, 518(7538):207–212, 2015.

[94] Julija Krupic, Marius Bauza, Stephen Burton, Caswell Barry, and John O?Keefe. Grid cell symmetry is shaped by environmental geometry. *Nature*, 518(7538):232?235, 2015.

[95] Richard B Stein. The information capacity of nerve cells using a frequency code. *Biophysical Journal*, 7(6):797–826, 1967.

[96] F Rieke, D Warland, and W Bialek. Coding efficiency and information rates in sensory neurons. *EPL (Europhysics Letters)*, 22(2):151, 1993.

[97] Richard B Stein, E Roderich Gossen, and Kelvin E Jones. Neuronal variability: noise or part of the signal? *Nature Reviews Neuroscience*, 6(5):389, 2005.

[98] Shigeru Shinomoto, Hideaki Kim, Takeaki Shimokawa, Nanae Matsuno, Shintaro Funahashi, Keisetsu Shima, Ichiro Fujita, Hiroshi Tamura, Taijiro Doi, Kenji Kawano, et al. Relating neuronal firing patterns to functional differentiation of cerebral cortex. *PLoS Computational Biology*, 5(7):e1000433, 2009.

[99] Patricia E Sharp and Catherine Green. Spatial correlates of firing patterns of single cells in the subiculum of the freely moving rat. *Journal of Neuroscience*, 14(4):2339–2356, 1994.

[100] Patrick Latuske, Oana Toader, and Kevin Allen. Interspike intervals reveal functionally distinct cell populations in the medial entorhinal cortex. *Journal of Neuroscience*, 35(31):10963–10976, 2015.

[101] Christian Laut Ebbesen, Eric Torsten Reifenstein, Qiusong Tang, Andrea Burgalossi, Saikat Ray, Susanne Schreiber, Richard Kempter, and Michael Brecht. Cell type-specific differences in spike timing and spike shape in the rat parasubiculum and superficial medial entorhinal cortex. *Cell Reports*, 16(4):1005–1015, 2016.

[102] Alessandro Treves and Stefano Panzeri. The upward bias in measures of information derived from limited data samples. *Neural Computation*, 7(2):399–407, 1995.

[103] Steven P Strong, Roland Koberle, Rob R de Ruyter van Steveninck, and William Bialek. Entropy and information in neural spike trains. *Physical Review Letters*, 80(1):197, 1998.

[104] K-I Goh and A-L Barabási. Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, 81(4):48002, 2008.

[105] Jonathan J Couey, Aree Witoelar, Sheng-Jia Zhang, Kang Zheng, Jing Ye, Benjamin Dunn, Rafal Czajkowski, May-Britt Moser, Edvard I Moser, Yasser Roudi, et al. Recurrent inhibitory circuitry as a mechanism for grid formation. *Nature Neuroscience*, 16(3):318–324, 2013.

[106] Hugh Pastoll, Lukas Solanka, Mark CW van Rossum, and Matthew F Nolan. Feedback inhibition enables theta-nested gamma oscillations and grid firing fields. *Neuron*, 77(1):141–154, 2013.

[107] Yasser Roudi and Edvard I Moser. Grid cells in an inhibitory network. *Nature Neuroscience*, 17(5):639–641, 2014.

[108] Benjamin Dunn, Maria Mørreaunet, and Yasser Roudi. Correlations and functional connections in a population of grid cells. *PLoS Computational Biology*, 11(2):e1004052, 2015.

[109] Benjamin Dunn, Daniel Wennberg, Ziwei Huang, and Yasser Roudi. Grid cells show field-to-field variability and this explains the aperiodic response of inhibitory interneurons. *arXiv preprint arXiv:1701.04893*, 2017.

[110] William E Skaggs, Bruce L McNaughton, and Katalin M Gothard. An information-theoretic approach to deciphering the hippocampal code. In *Advances in Neural Information Processing Systems*, pages 1030–1037, 1993.

[111] William E Skaggs, Bruce L McNaughton, Matthew A Wilson, and Carol A Barnes. Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus*, 6(2):149–172, 1996.

[112] Christina Buetfering, Kevin Allen, and Hannah Monyer. Parvalbumin interneurons provide grid cell–driven recurrent inhibition in the medial entorhinal cortex. *Nature Neuroscience*, 17(5):710–718, 2014.

[113] Trygve Solstad, Charlotte N Boccara, Emilio Kropff, May-Britt Moser, and Edvard I Moser. Representation of geometric borders in the entorhinal cortex. *Science*, 322(5909):1865–1868, 2008.

[114] Rosamund F Langston, James A Ainge, Jonathan J Couey, Cathrin B Canto, Tale L Bjerknes, Menno P Witter, Edvard I Moser, and May-Britt Moser. Development of the spatial representation system in the rat. *Science*, 328(5985):1576–1580, 2010.

[115] Bruce L McNaughton, Francesco P Battaglia, Ole Jensen, Edvard I Moser, and May-Britt Moser. Path integration and the neural basis of the 'cognitive map'. *Nature Reviews Neuroscience*, 7(8):663–678, 2006.

[116] Adrien Peyrache, Natalie Schieferstein, and Gyorgy Buzsáki. Transformation of the head-direction signal into a spatial code. *Nature Communications*, 8(1):1752, 2017.

[117] Adrien Peyrache and György Buzsáki. Extracellular recordings from multi-site silicon probes in the anterior thalamus and subicular formation of freely moving mice (http://dx.doi.org/10.6080/k0g15xs1), 2015.

[118] Mircea I Chelaru and Valentin Dragoi. Efficient coding in heterogeneous neuronal populations. *Proceedings of the National Academy of Sciences*, 105(42):16344–16349, 2008.

[119] Leenoy Meshulam, Jeffrey L Gauthier, Carlos D Brody, David W Tank, and William Bialek. Collective behavior of place and non-place neurons in the hippocampal network. *Neuron*, 96(5):1178–1191, 2017.

[120] Claudia Battistin, Benjamin Dunn, and Yasser Roudi. Learning with unknowns: analyzing biological data in the presence of hidden variables. *Current Opinion in Systems Biology*, 1:122–128, 2017.

[121] Benjamin R Kanter, Christine M Lykken, Daniel Avesar, Aldis Weible, Jasmine Dickinson, Benjamin Dunn, Nils Z Borgesius, Yasser Roudi, and Clifford G Kentros. A novel mechanism for the grid-to-place cell transformation revealed by transgenic depolarization of medial entorhinal cortex layer ii. *Neuron*, 93(6):1480–1492, 2017.

[122] Revekka Ismakov, Omri Barak, Kate Jeffery, and Dori Derdikman. Grid cells encode local positional information. *Current Biology*, 27(15):2337–2343, 2017.

[123] Federico Stella, Erika Cerasti, Bailu Si, Karel Jezek, and Alessandro Treves. Self-organization of multiple spatial and context memories in the hippocampus. *Neuroscience & Biobehavioral Reviews*, 36(7):1609–1625, 2012.

[124] Alexander Mathis, Andreas VM Herz, and Martin Stemmler. Optimal population codes for space: grid cells outperform place cells. *Neural Computation*, 24(9):2280–2317, 2012.

[125] Ila R Fiete, Yoram Burak, and Ted Brookings. What grid cells convey about rat location. *Journal of Neuroscience*, 28(27):6858–6871, 2008.

[126] Eleonora Russo and Daniel Durstewitz. Cell assemblies at multiple time scales with arbitrary lag constellations. *eLife*, 6:e19428, 2017.

[127] Boris V Gnedenko. *Theory of probability*. CRC Press, 1998.

[128] Jayanta K Ghosh, Mohan Delampady, and Tapas Samanta. *An introduction to Bayesian analysis: theory and methods*. Springer Science & Business Media, 2007.

[129] Kechen Zhang, Iris Ginzburg, Bruce L McNaughton, and Terrence J Sejnowski. Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. *Journal of Neurophysiology*, 79(2):1017–1044, 1998.

A

Basic Notions

In this chapter, we shall introduce several basic concepts of information theory, coding theory and large deviations theory that was needed in the preceding chapters. From the sampling process defined in Chapter 2.1, we shall define *entropy* – a cope concept throughout the dissertation. This definition will allow us to build other information theoretic quantities like the Kullback-Leibler divergence, mutual information and the data processing inequality.

From the sampling process defined in the previous section, we shall define entropy – a core concept throughout the dissertation. This definition will allow us to build other information theoretic quantities like Kullback-Leibler divergence, mutual information and data processing inequality. We shall also discuss the concept of sufficiency for parametric models. This will allow us to finally close the section with the maximum entropy principle.

A.1 Information theory

A.1.1 Entropy

Consider now the generative process described above which results in drawing the sample $\hat{s} = (s^{(1)}, \dots, s^{(N)})$ of N independent observations. In terms of the frequency, we have $\{k_1, \dots, k_S\}$ such that

$$\sum_{s=1}^{S} k_s = N. \tag{A.1}$$

Given $\{k_1, \dots, k_S\}$, we want to know how "uncertain" we are about the sample $\hat{s}$. Note that because the sampling procedure is independent, one can have S^N possible outcomes. Hence, one can ask, out of the S^N possible outcomes, in how many ways can we realize the sample $\hat{s}$ compatible with $\{k_1, \dots, k_S\}$. The answer to this is the multinomial coeffficient

$$W = \frac{N!}{\prod_{s \in \chi} k_s!}. \tag{A.2}$$

Notice that when W is large, there are many outcomes of the sampling procedure that realizes the set of frequencies $\{k_1, \dots, k_S\}$. And thus, intuitively, such samples are more "uncertain". Notice also that taking a monotonically increasing function of W will not affect our intuition of "uncertainty". Thus, we take $\log W$ so that as $N \to \infty$, we immediately have Stirling's approximation, $\log N! \simeq N \log N - N$[1], and one finds that

$$\log W \simeq - \sum_{s \in \chi} \frac{k_s}{N} \log \frac{k_s}{N} = \hat{H}[s]. \tag{A.3}$$

In the limit when $N \to \infty$ such that $k_s/N \to p(s)$, then this "uncertainty" is given by

$$H[s] = - \sum_{s \in \chi} p(s) \log p(s) \tag{A.4}$$

and $H[s]$ is what we shall call the *entropy*.

[1] As it turns out, the choice of the logarithm is intuitive as it corresponds to the minimum number of true or false questions one can ask to reduce the uncertainty of the sample $\hat{s}$. Indeed, we shall see in Appendix A.2 that the entropy $H[s]$ provides a lower bound to the expected length of the codeword which efficiently compresses the sample $hats$.

As it turns out, the result in Eq. (A.4) is the only function that is consistent with Shannon's suggestion towards measuring the "amount of uncertainty" represented by a discrete probability distribution. In particular, Eq. (A.4) satisfies the following conditions that reflect our intuition about what a reasonable measure of uncertainty would be:

1. $H[s]$ is a continuous function of the $p(s)$, i.e., we do not want any infinitesimal changes in $p(s)$ to produce drastic changes in the amount of uncertainty in the distribution[2].

2. If all possible limiting distributions $p(s)$ are equal, i.e., $p(s) = 1/N$, $\forall s \in \chi$, then $H[s]$ is at a maximum. Furthermore, in this case, $H[s]$ is a monotonically increasing function of N, i.e., if we increase the number of observations, it is intuitively acceptable that our uncertainty also increases. And finally,

3. The amount of uncertainty must be independent of the steps by which certainty may be achieved. That is, instead of giving the limiting distributions directly to each of the outcomes $s \in \chi$, we can group together the outcomes $x_1 \in \xi_1 = \{1, \ldots, n\}$ and $x_2 \in \xi_2 = \{n+1, \ldots, S\}$ with limiting distributions $p(x_1) = \sum_{s=1}^{n} p(s)$ and $p(x_2) = \sum_{s=n+1}^{S} p(s)$. Given this composition, we require that the uncertainty in s should be a weighted sum of the uncertainty in x i.e., $H[s] = H[x] + p(x_1)H[s|x = 1] + p(x_2)H[s|x_2]$ where $H[s|x_i]$ is the *conditional entropy* given by

$$H[s|x_i] = -\sum_{s\in\xi_i} p(s|x_i) \log p(s|x_i) \qquad \text{(A.5)}$$

with $p(s|x_i) = p(s)/p(x_i)$ being the conditional distribution.

[2] The continuity condition also implies that when $p(s') = 0$ then $p(s') \log p(s') = 0$.

A.1.2 Kullback-Leibler divergence

When one instead approximates the limiting distributions $p(s)$ by another distribution, say $q(s)$, one can then ask how "far" is $p(s)$ from $q(s)$. This is answered by computing for the *Kullback-Leibler divergence*, $D_{KL}(p(s)\|q(s))$ given by

$$D_{KL}(p(s)\|q(s)) = \sum_{s\in\chi} p(s)\log\frac{p(s)}{q(s)}. \tag{A.6}$$

Note that the Kullback-Leibler divergence is not a proper distance metric since it is not symmetric, i.e., $D_{KL}(p(s)\|q(s)) \neq D_{KL}(q(s)\|p(s))$ and it does not satisfy the triangle inequality, i.e., $D_{KL}(p(s)\|q(s)) \nleq D_{KL}(p(s)\|r(s)) + D_{KL}(r(s)\|q(s))$ for some distribution $r(s)$ defined over $s \in \chi$. The only properties that it shares with distance is that $D_{KL}(p(s)|q(s)) \geq 0$ and $D_{KL}(p(s)|q(s)) = 0$ if $p(s) = q(s), \forall s \in \chi$. Furthermore, the Kullback-Leibler divergence can be applied to any distributions $p(s)$ and $q(s)$ provided that these distributions are defined over the same finite space.

A.1.3 Mutual information

Suppose now that one draws an observation $s \in \chi$ from $p(s)$ and another observation $v \in \mathcal{V}$ from $p(v)$ with which the limiting *joint distribution* $p(s,v)$ can be constructed[3]. One can then ask whether observing v sheds light into s and vice versa, i.e., if v contains some *information* about s. This is answered by the *mutual information*, $I(s,v)$ between v and s given by

$$I(s,v) = -\sum_{s\in\chi, v\in\mathcal{V}} p(s,v)\log\frac{p(s,v)}{p(s)p(v)}. \tag{A.7}$$

Note that the mutual information $I(s,v)$ is symmetric, i.e., $I(s,v) = I(v,s)$ indicating that $I(s,v)$ quantifies how certain we are about v if we know s and vice versa. Notice as well that when v does not contain any information about s, then $p(s,v) = p(s)p(v)$

[3] Note that, from probability theory, the joint distribution $p(s,v)$ can be expressed in terms of conditional distributions, i.e., $p(s,v) = p(s|v)p(v) = p(v|s)p(s)$. From here, one can define Bayes' theorem $p(s|v) = p(v|s)p(s)/p(v)$.

indicating that the two observations are independent. Furthermore, notice that the mutual information $I(s,v)$ takes the form of the Kullback-Leibler divergence in Eq. (A.6). Hence, the mutual information $I(s,v)$ quantifies how far from independence are the observations s and v.

A.1.4 Data processing inequality

Let us consider three draws of observations: first, draw $x \in \mathcal{X}$ from $p(x)$, second, using x, draw $y \in \mathcal{Y}$ drawn from $p(y|x)$ and finally, using y, draw $z \in \mathcal{Z}$ from $p(z|y)$. This process forms a Markov chain

$$x \rightarrow y \rightarrow z \tag{A.8}$$

such that their joint probability distribution can be written as

$$p(x,y,z) = p(x)p(y|x)p(z|y). \tag{A.9}$$

This implies the following consequences: *(i)* x, y, and z form a Markov chain $x \rightarrow y \rightarrow z$ if and only if x and z are conditionally independent given y. Indeed, Markovianity implies the conditional independence since

$$p(x,z|y) = \frac{p(x,y,z)}{p(y)} = \frac{p(x,y)p(z|y)}{p(y)} = p(x|y)p(z|y). \tag{A.10}$$

(ii) This said, $x \rightarrow y \rightarrow z$ implies as well that $z \rightarrow y \rightarrow x$. Hence, in some formulations (e.g., the information bottleneck [51]), the Markov chain is written down as $x \leftrightarrow y \leftrightarrow z$. And finally, *(iii)* if $z = g(y)$ for any function g, then $x \rightarrow y \rightarrow z$ holds.

Now, given the observations x, y and z, then one can express the mutual information (as in Section A.7) in two different ways:

$$I(x,y,z) = I(x,z) + I(x,y|z) = I(x,y) + I(x,z|y). \tag{A.11}$$

However, because x and z are conditionally independent on y, then $I(x, z|y) = 0$. Hence, the *data processing inequality* follows which states that if $x \to y \to z$, then

$$I(x, y) \geq I(x, z) \tag{A.12}$$

where $I(x, y)$ and $I(x, z)$ are the mutual information between x and y, and x and z as in Eq. (A.7) respectively. This statement still stands if $z = g(y)$, i.e., a function of the data y cannot increase the information about x. Consequently, the data processing inequality states that no clever manipulation of data can normally improve any inference that can be performed on data.

A.2 Coding theory

A.2.1 Prefix-free codes

In this section, we shall discuss a particular type of code – *prefix-free codes* – which allow for efficient compression of messages. Here, we treat the message $\hat{s}$ as a sequence of characters $s^{(i)}$ where, under a given coding strategy, a character $s \in \chi$ is mapped onto a codeword $C(s)$ with an alphabel of size D and a length E_s in bits. When comparing two codewords, C_1 and C_2, we shall index the codeword lengths, $E_s^{(C_1)}$ and $E_s^{(C_2)}$, to denote the coding cost of s under C_1 and C_2 respectively. We shall show that for the *prefix-free codes* to be uniquely decodable, it needs to satisfy a property called the *Kraft inequality*. We shall then use this property to show that the coding cost, E_s, is related to the probability distribution $p(s)$ as

$$E_s = -\log p(s) \tag{A.13}$$

and thus, with $E = \sum_{s \in \hat{s}} E_s$, gives the coding cost of the sample $\hat{s}$ in Eq. (4.1).

Prefix-free codes are a set of codewords for which there is no codeword that is the prefix of another codeword[4]. In the context of MDL, prefix-free codes are good can-

[4] In the literature, prefix-free codes are instead called prefix codes. However, we find that "prefix

didate codes for two reasons: *(i)* prefix-free codes are *non-singular*, i.e. the mapping between a character s and its codeword $C(s)$ is one-to-one, and *(ii)* prefix-free codes are uniquely decodable, i.e., the code $C(\hat{s}) = C(s^{(1)})C(s^{(2)})C(s^{(3)})\dots C(s^{(N)})$ is non-singular. These properties allow prefix-free codes for lossless decoding.

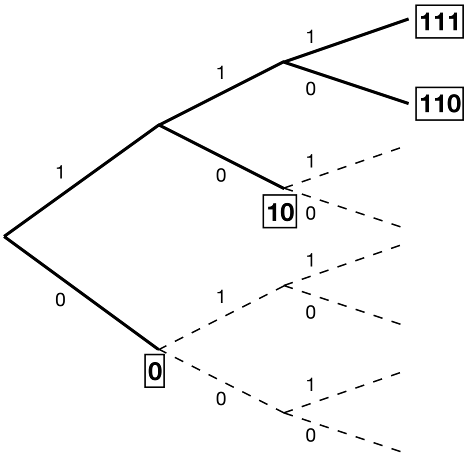

FIGURE A.1. **An illustration of the construction of prefix-free codes from a 2-ary tree.** Each codeword is assigned to a node in the binary tree in such a way that no one of the codewords is the ancestor of another.

For example, if we take a code C_1 (with an alphabel $\{0, 1\}$ so that $D = 2$) defined by $C_1(a) = 0$, $C_1(b) = 10$ and $C_1(c) = 01$, then sending the message $\hat{s}_1 = acbaab$ corresponds to sending the codeword 001100010 which has $E^{(C_1)} = 9$ bits. However, the receiver only sees one bit at a time following the sequence. Hence, upon seeing the 2nd bit, the receiver decodes the message as aa. However, upon seeing the 3rd and 4th bit, he finds that he has no codeword for 11 and thus, he realizes that he made a mistake somewhere. This is an example of a lossy decoding and is brought by the fact that $C_1(c)$ is prefixed by $C_1(a)$. Instead, if we take a code C_2 defined by $C_2(a) = 0$, $C_2(b) = 10$ and $C_2(c) = 11$, then the same message will have the codeword $\hat{s}_2 = 011100010$ which also has $E^{(C_2)} = 9$ bits. However, in this case, the receiver is able to losslessly decode the message because the code is a prefix code.

code" is a bit misleading as such codes are supposed not to have prefixes in the code.

A.2.2 Kraft's inequality

Prefix-free codes can be constructed efficiently which entails that a fundamental constraint on the codeword lengths exists. In particular, the *Kraft inequality* states that the lengths E_s of prefix-free codes must satisfy the inequality

$$\sum_{s=1}^{S} D^{-E_s} \leq 1 \tag{A.14}$$

where D is the alphabel of the code C. Conversely, given a set of codeword lengths E_s which satisfy this inequality, there exists a prefix-free code with these codeword lengths.

To show this, we consider the codewords $\{C(1), C(2), \dots, C(S)\}$ with lengths $\{E_1, E_2, \dots, E_S\}$ of a prefix-free code which are ordered such that

$$E_1 \leq E_2 \leq \ldots \leq E_S. \tag{A.15}$$

Now, we can try to construct a prefix-free code in an increasing order as in Eq. (A.15) and in Fig. A.1. There exists a prefix-free code if and only if at each step s, there is at least one codeword $C(s)$ to choose that does not contain any of the previous $s-1$ codewords as a prefix, i.e., the codeword $C(s)$ must not contain any of the previous codewords $\{C(1), C(2), \dots, C(s-1)\}$ as a prefix. For the codeword s, there are D^{E_s} combinations of codewords possible if there was no prefix constraints. However, the prefix constraints tell us that $D^{E_s - E_{s-1}}$ codewords are forbidden. Therefore, the total number of forbidden codewords at s is

$$\sum_{i=1}^{s-1} D^{E_s - E_i}. \tag{A.16}$$

There exists a prefix code if and only if we have a codeword to choose at every $s \in \{1, \dots, S\}$, i.e.,

$$D^{E_s} > \sum_{i=1}^{s-1} D^{E_s - E_i}, \forall s = 2, 3, \dots, S. \tag{A.17}$$

Since every term in the sum above is an integer, then it is equivalent to

$$D^{E_s} \geq \sum_{i=1}^{s-1} D^{E_s-E_i} + 1 = \sum_{i=1}^{s} D^{E_s-E_i}, \forall s = 1, 2, \ldots, S. \tag{A.18}$$

And thus, dividing both sides by D^{E_s}, we find

$$\sum_{s=1}^{S} D^{-E_s} \leq 1.$$

Notice that every argument in the proof goes both ways. And thus, this proves that the Kraft inequality is a necessary and sufficient condition for the existence of a prefix-free code. Interestingly, it turns out that the Kraft inequality must also hold for all uniquely decodable codes (*McMillian inequality*).

A.2.3 Correspondence between the coding cost and the probability distribution

With Kraft inequality, we now have a fundamental constraint that must be fulfilled to construct a prefix-free code with an average codeword length given by

$$\langle E_s \rangle = \sum_{s=1}^{S} E_s p(s). \tag{A.19}$$

With this, we are now in a position to to find the optimal codeword length which involves solving the optimization problem given by

$$\text{minimize}_{E_s} \quad \sum_{s=1}^{S} E_s p(s) \tag{A.20}$$

$$\text{subject to} \quad \sum_{s=1}^{S} 2^{-E_s} \leq 1 \tag{A.21}$$

Using the method of Lagrange multipliers, the solution reads as

$$E_s^* \geq -\log p(s) \tag{A.22}$$

And thus, the expected codelength is lower bounded by the entropy

$$\langle E_s^* \rangle = \sum_{s=1}^{S} E_s^* p(s) \geq -\sum_{s=1}^{S} p(s) \log p(s) = H[s] \tag{A.23}$$

where $H[s]$ is the entropy as defined in Eq. (A.4). Notice that when the *optimal codes* are used to compress the data, we saturate the inequality in Eq. (A.22) and we find the correspondence between coding cost and the probability distribution.

A.3 Large deviations

In this section, we consider the typical properties of the samples that correspond to large deviations in the fluctuations of the empirical average

$$\hat{G}[s] = \frac{1}{N} \sum_{i=1}^{N} g(s^{(i)}) \tag{A.24}$$

where $g(s^{(i)})$ is a function of the observation $s^{(i)}$ such that a typical value is obtained when the size of the sample is large enough. This means that the law of large numbers [127] given by

$$\lim_{N \to \infty} P\left\{ \left| \frac{1}{N} \sum_{i=1}^{N} g(s^{(i)}) - \langle g \rangle_P \right| > \epsilon \right\} = 0 \tag{A.25}$$

where $\langle g \rangle_P = \sum_{s \in \chi} sp(s)$. Here, we shall be interested in typical properties of samples that correspond to large deviations in $\hat{G}[s]$, i.e., when $\hat{G}[s] = E$ different from the typical value. In particular, we want to compute the probability $P\{\hat{G}[s] = E\}$ of this event. Because the observations $s^{(i)}$ in the sample $\hat{s}$ take value in a finite set χ, we can invoke Sanov's theorem [56] which states that the probability $P\{\hat{H}[s] = E\}$ is asymptotically given by

$$P\{\hat{G}[s] = E\} \simeq e^{-N D_{KL}(P_\beta(\hat{s}) \| P(\hat{s}))} \tag{A.26}$$

where $D_{KL}(Q(\hat{s}) \| P(\hat{s})) = \sum_{s \in \chi} Q(\hat{s}) \log \left[Q(\hat{s}) / P(\hat{s}) \right]$ is the Kullback-Leibler divergence from the sample distribution $P(\hat{s})$ to some distribution $Q(\hat{s})$ and $P_\beta(\hat{s})$ is the distribution that minimizes $D_{KL}(Q(\hat{s})\|P(\hat{s}))$ on all $Q(\hat{s})$ that satisfy $\hat{G}[s] = E$. This constrained optimization has the solution given by a "tilted" distribution

$$P_\beta(\hat{s}) = \frac{P(\hat{s})e^{\beta N \hat{G}[s]}}{Z(\beta)} \tag{A.27}$$

where

$$Z(\beta) = \sum_{\hat{s}} P(\hat{s})e^{\beta N \hat{G}[s]} \tag{A.28}$$

is a normalization constant and the parameter $\beta = \beta(E)$ is adjusted such that $E = \sum_{s\in\chi} P_\beta(\hat{s})\hat{G}[s]$. Notice that when $\beta = 0$, then $P_\beta(\hat{s}) = P(\hat{s})$ is the "untilted" sample distribution. Hence, adjusting β allows one to "explore" rare events with large fluctuations of $\hat{G}[s]$.

In limit of infinitesimal δE, one can invoke Gärtner-Ellis theorem [74] with which the probability $P\{\hat{G}[s] = E\}$ can be expressed in terms of the distribution of E as $P\{\hat{G}[s] = E\} \approx e^{-NI(E)}\delta E$, where the rate function $I(E)$ (also called the Cramér's function) is given by

$$I(E) = -\lim_{N\to\infty} \frac{1}{N} \log P\{\hat{G}[s] = E\} = D_{KL}(P_\beta(\hat{s})\|P(\hat{s})). \tag{A.29}$$

Hence, with Eq. (A.27), the probability $P\{\hat{G}[s] = E\}$ can be computed as

$$\lim_{N\to\infty} \frac{1}{N} \log P\{\hat{G}[s] = E\} = -\beta E + \phi(\beta) \tag{A.30}$$

where $\phi(\beta)$ has the form of the free energy with

$$\frac{\partial}{\partial\beta}\phi(\beta) = \langle \hat{G}[s] \rangle_\beta = E. \tag{A.31}$$

This page has been intentionally left blank.

B

Efficient representations exhibit statistical criticality

B.1 Derivation of Equation (3.24)

In this section, we shall show the calculation of the Kullback-Leibler divergence between the posterior distribution $p(\theta|\hat{s})$ and the prior distribution $p_0(\theta)$ as in Eq. (3.27) of Chapter 3.4.

Consider the Kullback-Leibler divergence

$$D_{KL}(p(\theta|\hat{s})\|p_0(\theta)) = \int d\theta p(\theta|\hat{s}) \log \frac{p(\theta|\hat{s})}{p_0(\theta)}. \tag{B.1}$$

Because of Bayes' theorem, this integral can be cast as

$$\begin{aligned} D_{KL}(p(\theta|\hat{s})\|p_0(\theta)) &= \int d\theta p(\theta|\hat{s}) \log \frac{p(\hat{s}|\theta)}{\int d\theta' p(\hat{s}|\theta')p_0(\theta')} && \text{(B.2)} \\ &= \int d\theta p(\theta|\hat{s}) \log p(\hat{s}|\theta) - \int d\theta p(\theta|\hat{s}) \log\left[\int d\theta' p(\hat{s}|\theta')p_0(\theta')\right]. && \text{(B.3)} \end{aligned}$$

Notice that the argument inside the logarithm in the second term in Eq. (B.3) does not depend on the parameter θ. Hence, we start by considering the integral $\int d\theta p(\hat{s}|\theta)p_0(\theta)$. For $N \gg 1$, the integral is dominated by the point $\theta = \hat{\theta}(\hat{s})$ that maximizes the log-likelihood $\log p(\hat{s}|\theta)$ and it can be computed by a Laplace approximation (or saddle point method). Performing a Taylor expansion around the maximum likelihood parameters, $\hat{\theta}(\hat{s})$, one finds (up to leading orders in N)

$$\log p(\hat{s}|\theta) \simeq \log p(\hat{s}|\hat{\theta}(\hat{s})) - \frac{1}{2}\sum_{i,j} N L_{i,j}(\hat{\theta})(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j) + \mathcal{O}((\theta - \hat{\theta})^3). \tag{B.4}$$

where, for exponential families, the Hessian of the log-likelihood

$$L_{i,j}(\theta) = -\sum_{s\in\chi} \frac{k_s}{N}\frac{\partial^2 \log p(\hat{s}|\theta)}{\partial\theta_i\partial\theta_j} \tag{B.5}$$

$$= -\sum_{s\in\chi} p(s|\theta)\frac{\partial^2 \log p(\hat{s}|\theta)}{\partial\theta_i\partial\theta_j} \tag{B.6}$$

is independent of the data and it coincides with the Fisher Information matrix [70]. The integral can then be computed by Gaussian integration, as

$$\int d\theta p(\hat{s}|\theta)p_0(\theta) \simeq p(\hat{s}|\hat{\theta}(\hat{s}))p_0(\hat{\theta}(\hat{s})) \int d\theta e^{-\frac{N}{2}\sum_{i,j}\hat{L}_{ij}(\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)} \tag{B.7}$$

$$= p(\hat{s}|\hat{\theta}(\hat{s}))p_0(\hat{\theta}(\hat{s}))\left(\frac{2\pi}{N}\right)^{\frac{k}{2}}\frac{1}{\sqrt{\det L(\hat{\theta})}}. \tag{B.8}$$

where k is the number of parameters.

Notice that the likelihood $p(\hat{s}|\theta)$ is normally distributed with a mean centered at the maximum likelihood estimator $\hat{\theta}(\hat{s})$ and variance $L(\hat{\theta})^{-1}/N$. Also, the likelihood distribution scales with the sample size while the prior distribution does not. Thus, when we have a reasonable amount of data, i.e., $N \gg 1$, the likelihood distribution becomes more and more significant and consequently, will outweigh the prior distribution. Hence, the posterior distribution $p(\hat{s}|\theta)$ will be normally distributed as the likelihood distribution. Using a well-known result about the exponent of a multivariate normal distribution that (see also page 127 of Ref. [128])

$$\int d\theta \left[-\frac{1}{2} \sum_{i,j} N L_{i,j}(\hat{\theta})(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j) \right] p(\theta|\hat{s}) = -\frac{k}{2}, \tag{B.9}$$

we find that the first term in Eq. (B.3) can be evaluated as

$$\int d\theta p(\theta|\hat{s}) \log p(\hat{s}|\theta) \simeq \log p(\hat{s}|\hat{\theta}(\hat{s})) - \frac{k}{2}. \tag{B.10}$$

Thus, putting everything together, we find that

$$D_{KL}\left(p(\theta|\hat{s}) \| p_0(\theta)\right) \simeq \frac{k}{2} \log \frac{N}{2\pi} - \log p_0(\hat{\theta}) + \frac{1}{2} \log \det \hat{L}(\hat{\theta}) + \mathcal{O}(1/N)$$

as shown in Eq. (3.27) of Chapter 3.4. If we take the prior distribution $p_0(\theta)$ to be the Jeffreys prior, i.e., $p_0(\theta) = \mathcal{N}\sqrt{\det L(\theta)}$ where $\mathcal{N}$ is a normalization constant, then it can easily be seen that, for $N \gg 1$, the Kullback-Leibler divergence $D_{KL}\left(p(\theta|\hat{s}) \| p_0(\theta)\right)$ attains a maximum value.

This page has been intentionally left blank.

C

Distribution that minimize description lengths are critical

C.1 Derivation for the parametric complexity

In order to compute the parametric complexity, given in Eq. (4.4), let us consider the integral $\int d\theta f(\hat{s}|\theta)g(\theta)$ for a generic function $g(\theta)$. For $N \gg 1$, the integral is dominated by the point $\theta = \hat{\theta}(\hat{s})$ that maximizes $\log f(\hat{s}|\theta)$, and it can be computed by the saddle point method. Performing a Taylor expansion around the maximum likelihood parameters, $\hat{\theta}(\hat{s})$, one finds (up to leading orders in N)

$$\log f(\hat{s}|\theta) = \log f(\hat{s}|\hat{\theta}(\hat{s})) - \frac{1}{2}\sum_{i,j} N(\theta_i - \hat{\theta}_i)L_{i,j}(\hat{\theta})(\theta_j - \hat{\theta}_j) + \mathcal{O}((\theta - \hat{\theta})^3). \tag{C.1}$$

where

$$L_{i,j}(\hat{\theta}) = -\frac{1}{N}\frac{\partial^2 \log f(\hat{s}|\theta)}{\partial\theta_i\partial\theta_j} \tag{C.2}$$

$$= -\sum_{s\in\chi}\frac{k_s}{N}\frac{\partial^2 \log f(s|\theta)}{\partial\theta_i\partial\theta_j}. \tag{C.3}$$

Note that for exponential families, the Hessian of the log-likelihood is independent of the data, and hence it coincides with the Fisher Information matrix [70]

$$L_{i,j}(\theta) = -\sum_{s\in\chi} f(s|\theta)\frac{\partial^2 \log f(s|\theta)}{\partial\theta_i\partial\theta_j}. \tag{C.4}$$

The integral can then be computed by Gaussian integration, as

$$\int d\theta f(\hat{s}|\theta)g(\theta) \simeq f(\hat{s}|\hat{\theta}(\hat{s}))g(\hat{\theta}(\hat{s}))\int d\theta e^{-\frac{N}{2}\sum_{i,j}(\theta_i-\hat{\theta}_i)L_{ij}(\hat{\theta})(\theta_j-\hat{\theta}_j)} \tag{C.5}$$

$$= f(\hat{s}|\hat{\theta}(\hat{s}))g(\hat{\theta}(\hat{s}))\left(\frac{2\pi}{N}\right)^{\frac{k}{2}}\frac{1}{\sqrt{\det L(\hat{\theta})}}. \tag{C.6}$$

where k is the number of free parameters. If we choose $g(\theta)$ to be

$$g(\theta) = \left(\frac{N}{2\pi}\right)^{\frac{k}{2}}\sqrt{\det L(\theta)} \tag{C.7}$$

and take a sum over all samples $\hat{s}$ on both sides of Eq. (C.5), Eq. (C.6) becomes

$$\sum_{\hat{s}} f(\hat{s}|\hat{\theta}(\hat{s})) \simeq \sum_{\hat{s}}\left(\frac{N}{2\pi}\right)^{\frac{k}{2}}\int d\theta f(\hat{s}|\theta)\sqrt{\det L(\theta)} \tag{C.8}$$

$$= \left(\frac{N}{2\pi}\right)^{\frac{k}{2}}\int\sqrt{\det L(\theta)}d\theta. \tag{C.9}$$

Hence, the parametric complexity, $\bar{\mathcal{R}} = \log\sum_{\hat{s}} f(\hat{s}|\hat{\theta}(\hat{s}))$, is asymptotically given by Eq. (4.5) when $N \gg 1$.

Notice also that $\bar{P}(\hat{s})$ induces a distribution over the space of parameters θ. With the choice

$$g(\theta) = \left(\frac{N}{2\pi}\right)^{\frac{k}{2}} \sqrt{\det L(\theta)}\delta(\theta - \theta_0), \tag{C.10}$$

the same procedure as above shows that

$$\begin{aligned}
\sum_{\hat{s}} \bar{P}(\hat{s})\delta\left(\hat{\theta}(\hat{s}) - \theta_0\right) &= e^{-\bar{\mathcal{R}}} \sum_{\hat{s}} f(\hat{s}|\hat{\theta}(\hat{s}))\delta\left(\hat{\theta}(\hat{s}) - \theta_0\right) && \text{(C.11)}\\
&= e^{-\bar{\mathcal{R}}} \left(\frac{N}{2\pi}\right)^{\frac{k}{2}} \sqrt{\det I(\theta_0)} && \text{(C.12)}\\
&= \frac{\sqrt{\det L(\theta_0)}}{\int d\theta \sqrt{\det L(\theta)}} && \text{(C.13)}
\end{aligned}$$

which is the Jeffreys prior.

The choice of the Jeffreys prior can be motivated in the following manner. We have noted earlier that the maximum likelihood distribution, $f(\hat{s}|\hat{\theta}(\hat{s}))$, depends on the sample, $\hat{s}$. Hence, each sample represents a point in the space of probability distributions parametrized by θ. When the number of observations, N, is finite, one cannot rule out the possibility that two distributions, $f(\hat{s}|\hat{\theta}(\hat{s}))$ and $f(\hat{s}|\hat{\theta}'(\hat{s}))$, can be distinguished from one another. By distinguishability, we mean that the Kullback-Leibler divergence, $D_{KL}(\theta||\theta') = \langle \log \frac{f(\hat{s}|\theta)}{f(\hat{s}|\theta')} \rangle_\theta$ is less than a given threshold i.e., $D_{KL}(\theta||\theta') \leq -\frac{\log \epsilon}{N}$. This condition implies that one can count the number of distributions that are *indistinguishable* from $f(\hat{s}|\theta)$ given a finite number of observations, N, and was found to be [70]

$$V_{\epsilon,N}(\theta) = \left(-\frac{2\pi \log \epsilon}{N}\right)^{\frac{k}{2}} \frac{1}{\Gamma\left(\frac{k}{2}+1\right)\sqrt{\det L(\theta)}} \tag{C.14}$$

where, as before, k is the dimensionality of the model, $f(\hat{s}|\theta)$, and $I(\theta)$ is the Fisher information matrix. In effect, the space of probability distributions parametrized by θ can be partitioned into n regions wherein the Fisher information, $L(\theta)$ remains a constant. Thus, the prior distribution over the i^{th} partition with a volume U_i will be

$$\rho_i = \frac{\frac{U_i}{V(\theta_i)}}{\sum_{j=1}^n \frac{U_j}{V(\theta_j)}} = \frac{U_i\sqrt{\det L(\theta_i)}}{\sum_{i=1}^n U_j\sqrt{\det L(\theta_j)}}. \tag{C.15}$$

Hence, in the continuum limit where $N \to infty$ such that $n \to \infty$, then we have

$$\rho(\theta) = \frac{d\theta\sqrt{\det L(\theta)}}{\int d\theta\sqrt{\det L(\theta)}} \tag{C.16}$$

which is the Jeffreys prior.

C.2 Calculating the parametric complexity

In this section, we calculate the parametric complexity for the Dirichlet model for $\rho \gg 1$ and the paramagnetic Ising model.

C.2.1 Dirichlet model

In the regime where $\rho \gg 1$ and k large such that we can employ Stirling's approximation, $k! = \sqrt{2\pi k}k^k e^{-k}$, the normalization can be calculated as

$$\sum_{k=0}^{\infty} \frac{k^k e^{-k} e^{-z^*(\rho)k}}{k!} \approx \sum_{k=0}^{\infty} \frac{e^{-z^*(\rho)k}}{\sqrt{2\pi k}} \tag{C.17}$$

$$= \int_0^\infty \frac{e^{-z^*(\rho)k}dk}{\sqrt{2\pi k}} \tag{C.18}$$

$$= \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\pi}{z^*(\rho)}} \tag{C.19}$$

$$= \frac{1}{\sqrt{2z^*(\rho)}}. \tag{C.20}$$

Similarly, we can also calculate

$$\sum_{k=0}^{\infty} \frac{k^{k+1} e^{-k} e^{-z^*(\rho)k}}{k!} \approx \sum_{k=0}^{\infty} \frac{k e^{-z^*(\rho)k}}{\sqrt{2\pi k}} \tag{C.21}$$

$$= \int_0^{\infty} \sqrt{\frac{k}{2\pi}} e^{-z^*(\rho)k} dk \tag{C.22}$$

$$= \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\pi}}{2(z^*(\rho))^{\frac{3}{2}}} \tag{C.23}$$

$$= \frac{1}{(2z^*(\rho))^{\frac{3}{2}}} \tag{C.24}$$

and thus, the saddle point value z^* can now be evaluated as

$$z^*(\rho) \simeq \frac{1}{2\rho}. \tag{C.25}$$

Also, the variance $\langle k^2 \rangle_{z^*} - \langle k \rangle_{z^*}^2$ can be calculated along the similar lines where

$$\sum_{k=0}^{\infty} \frac{k^{k+2} e^{-k} e^{-z^*(\rho)k}}{k!} \approx \sum_{k=0}^{\infty} \frac{k^2 e^{-z^*(\rho)k}}{\sqrt{2\pi k}} \tag{C.26}$$

$$= \int_0^{\infty} \sqrt{\frac{k^3}{2\pi}} e^{-z^*(\rho)k} dk \tag{C.27}$$

$$= \frac{3}{\sqrt{2\pi}} \frac{\sqrt{\pi}}{2(z^*(\rho))^{\frac{5}{2}}} \tag{C.28}$$

$$= \frac{3}{(2z^*(\rho))^{\frac{5}{2}}} \tag{C.29}$$

and thus, one finds that

$$\langle k^2 \rangle_{z^*} - \langle k \rangle_{z^*}^2 = 2\rho^2. \tag{C.30}$$

In the same regime, given the determinant $\det I(\theta)$ of the Fisher information matrix for the Dirichlet model,

$$\det L(\theta) = \prod_{s \in \chi} \frac{1}{\theta_s}, \tag{C.31}$$

the parametric complexity can be approximated as

$$
\begin{aligned}
e^{\bar{\mathcal{R}}} &\simeq \left(\frac{N}{2\pi}\right)^{(S-1)/2} \int d\theta \sqrt{\det L(\theta)} && \text{(C.32)} \\
&= \left(\frac{N}{2\pi}\right)^{(S-1)/2} \frac{\Gamma(\frac{1}{2})^S}{\Gamma(\frac{S}{2})} && \text{(C.33)} \\
&\simeq \frac{e^{\frac{S}{2}(1+\log\rho)}}{\sqrt{2\rho}} && \text{(C.34)}
\end{aligned}
$$

which, together with Eq. (4.16) and Eq. (C.30), implies that $\Phi(z^*(\rho)) = \frac{1}{2}(1+\log\rho)$.

C.2.2 Paramagnet model

The parametric complexity for the paramagnetic Ising model, given $\bar{P}(m)$ in Eq. (4.30), is given by

$$
e^{\bar{\mathcal{R}}} = \sum_{M=-N}^{N} \binom{N}{\frac{N-M}{2}} e^{\left[M \tanh^{-1}(M/N) - N \log\left(2\cosh\left(\tanh^{-1}(M/N)\right)\right)\right]}. \qquad \text{(C.35)}
$$

where $M = -N, -N+2, \ldots, N-2, N$ runs on $N+1$ values. When $N \gg 1$, the magnetization, $m = M/N$, can be treated as a continuous variable and consequently, the sum can be approximated as an integral: $\sum_M \cdots \simeq \frac{N}{2}\int_{-1}^{1} dm \ldots$. Hence, by using the identities $\tanh^{-1}(m) = \frac{1}{2}\log\frac{1+m}{1-m}$, $\cosh\left(\tanh^{-1}(m)\right) = \frac{1}{\sqrt{1-m^2}}$ and $K! \simeq K^K e{-}K\sqrt{2\pi K}$, one finds that

$$
\begin{aligned}
e^{\bar{\mathcal{R}}} &\simeq \frac{N}{2}\int_{-1}^{1} \frac{1}{\sqrt{2\pi N(1-m^2)}} && \text{(C.36)} \\
&= \sqrt{\frac{\pi N}{2}}. && \text{(C.37)}
\end{aligned}
$$

C.3 Simulation details

C.3.1 Sampling universal codes through Markov chain Monte Carlo

Unlike the Dirichlet model and the independent spin model, analytic calculations for the Sherrington-Kirkpatrick (SK) model and the restricted Boltzmann machine (RBM) are generally not possible, because the partition function Z, and consequently, the UC partition function $e^{\bar{\mathscr{R}}}$, is computationally intractable. In order to sample the NML for these graphical models, we turn to a Markov chain Monte Carlo (MCMC) approach in which the transition probability, $P(\hat{s} \to \hat{s}')$, can be built using the following heuristics:

1. Starting from the sample, $\hat{s}$, we calculate the maximum likelihood estimates, $\hat{\boldsymbol{\theta}}(\hat{s})$, of the parameters of the model, $p(\hat{s}|\theta)$ by either solving Eq. (4.37) for the SK model or by Contrastive Divergence (CD$_\kappa$) [72, 73] for the RBM (see Appendix C.3.2).

2. We generate a new sample, $\hat{s}'$ from $\hat{s}$ by flipping a spin in randomly selected r points $s^{(i)}$ of the sample. The number of selected spins, r, must be chosen carefully such that r must be large enough to ensure faster mixing but small enough so the new inferred model, $p(\hat{s}'|\theta)$, is not too far from the starting model, $p(\hat{s}|\theta)$.

3. The maximum likelihood estimators, $\hat{\boldsymbol{\theta}}(\hat{s}')$ for the new sample are calculated as in Step 1.

4. Compute

$$\Delta E = \log p(\hat{s}'|\hat{\boldsymbol{\theta}}(\hat{s}')) - \log p(\hat{s}|\hat{\boldsymbol{\theta}}(\hat{s})) \tag{C.38}$$

   and accept the move $\hat{s} \to \hat{s}'$ with probability $\min\left(e^{N\Delta E}, 1\right)$.

C.3.2 Estimating RBM parameters through Contrastive Divergence

Given a sample, $\hat{\boldsymbol{v}} = (\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(N)})$, of N observations, the log-likelihood for the restricted Boltzmann machine (RBM) is given by

$$\log \mathscr{L}(\boldsymbol{\theta}) = \sum_{k=1}^{N} \log \sum_{\boldsymbol{h}} P(\boldsymbol{v}^{(k)}, \boldsymbol{h}|\boldsymbol{\theta}). \tag{C.39}$$

The inference of the parameters, $\boldsymbol{\theta}$, proceeds by updating $\boldsymbol{\theta}$ such that the log-likelihood, $\log \mathcal{L}(\theta)$, is maximized. This updating formulation for the parameters is given by

$$\Delta\theta = \frac{\epsilon}{N}\frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \theta} \tag{C.40}$$

where ϵ is the learning rate parameter. The corresponding gradients for the parameters, $\boldsymbol{w}$, $\boldsymbol{a}$ and $\boldsymbol{b}$ can then be written down respectively as

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial w_{ij}} = \sum_{k=1}^{N}\left[\sum_{\boldsymbol{h}} v_i^{(k)} h_j P(\boldsymbol{v}^{(k)}, \boldsymbol{h}|\boldsymbol{\theta}) - \sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}} v_i h_j P(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})\right] \tag{C.41}$$

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial a_{i}} = \sum_{k=1}^{N}\left[\sum_{\boldsymbol{h}} v_i^{(k)} P(\boldsymbol{v}^{(k)}, \boldsymbol{h}|\boldsymbol{\theta}) - \sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}} v_i P(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})\right] \tag{C.42}$$

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial b_{j}} = \sum_{k=1}^{N}\left[\sum_{\boldsymbol{h}} h_j P(\boldsymbol{v}^{(k)}, \boldsymbol{h}|\boldsymbol{\theta}) - \sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}} h_j P(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})\right] \tag{C.43}$$

where the first terms denote averages over the data distribution while the second terms denote averages over the model distribution.

Here, we use the contrastive divergence (CD) approach which is a variation of the steepest gradient descent of $\mathcal{L}(\boldsymbol{\theta})$. Instead of performing the integration over the model distribution, CD approximates the partition function by averaging over distribution obtained after taking κ Gibbs sampling steps away from the data distribution.

To do this, we exploit the factorizability of the conditional distributions of the RBM. In particular, the conditional probability for the forward propagation (i.e., sampling the hidden variables given the visible variables) from $\boldsymbol{v}$ to h_j reads as

$$P(h_j = 1|\boldsymbol{v}, \boldsymbol{\theta}) = \frac{1}{1 + \exp\left(-b_j - \sum_i v_i w_{ij}\right)}. \tag{C.44}$$

Similarly, the conditional probability for the backward propagation (i.e., sampling the visible variables from the hidden variables) from $\boldsymbol{h}$ to v_i reads as

$$P(v_i = 1|\boldsymbol{h}, \boldsymbol{\theta}) = \frac{1}{1 + \exp\left(-a_i - \sum_j h_j w_{ij}\right)}. \tag{C.45}$$

The Gibbs sampling is done by propagating a sample, $\boldsymbol{v}^{(k)} = \boldsymbol{v}^{(k)}(0)$, forward and backward κ times: $\boldsymbol{v}^{(k)}(0) \to \boldsymbol{h}^{(k)}(0) \to \boldsymbol{v}^{(k)}(1) \to \ldots \to \boldsymbol{h}^{(k)}(\kappa - 1) \to \boldsymbol{v}^{(k)}(\kappa) \to \boldsymbol{h}^{(k)}(\kappa)$. And thus, the Gibbs sampling approximates the gradient in Eq. (C.41) as

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial w_{ij}} = \sum_{k=1}^{N} \left[v_i^{(k)}(0) h_j^{(k)}(0) - v_i^{(k)}(\kappa) h_j^{(k)}(\kappa) \right]. \tag{C.46}$$

In the CD approach, each parameter update for a batch is called an epoch. While larger κ approximates well the partition function, it also induces an additional computational cost. To find the global minimum more efficiently, we randomly divided the samples into groups of mini-batches. This approach introduces stochasticity and consequently reduce the likelihood of the learning algorithm to be confined in a local minima. However, a mini-batch approach can result to data-biased sampling. To circumvent this issue, we adopted the Persistent CD (PCD) algorithm where the Gibbs sampling extends to several epochs, each using different mini-batches. In the PCD approach, the initial visible variable configuration, $\boldsymbol{v}^{(k)}(0)$, was set to random for the first mini-batch, but the final configurations, $(\boldsymbol{v}^{(k)}(\kappa), \boldsymbol{h}^{(k)}(\kappa))$, of the current batches become the initial configuration for the next mini-batches. In this paper, we performed Gibbs sampling at $\kappa = 10$ steps where we update the parameters, $\boldsymbol{\theta}$, are updated at 2500 epochs at a rate $\epsilon = 0.01$ with 200 mini-batches per epochs. For other details regarding inference of parameters of the RBM, we refer the reader to Refs. [72, 73].

C.3.3 Source codes

All the calculations in this manuscript were done using personalized scripts written in Python 3. The source codes are accessible online[1].

[1] https://github.com/rcubero/UniversalCodes

This page has been intentionally left blank.

D

Finding relevant neurons in the brain using MultiScale Relevance (MSR)

D.1 Materials and methods

D.1.1 Data Collection

The data used in this study are recordings from rodents with multisite tetrode implants. These neurons are of particular interest because they are involved in spatial navigation.

Data from medial entorhinal cortex (mEC)

The spike times of 65 neurons recorded across the mEC area of a male Long Evans rat (Rat 14147) were taken in Ref. [79]. The rat was allowed to freely explore a box of dimension 150×150 cm^2 for a duration of around 20 mins. The positions were tracked using a platform attached to the head with red and green diodes fixed at both ends. Additional details about the data acquisition can be found in the paper in Ref. [79].

Data from the anterodorsal thalamic nucleus (ADn) and post-subiculum (PoS)

The spike times of 746 neurons recorded from multiple areas in the ADn and PoS across multiple sessions in six free moving mice (Mouse 12, Mouse 17, Mouse 20, Mouse 24, Mouse 25 and Mouse 28) while they freely foraged for food across an open environment with dimensions 53×46 cm^2 and in their home cages during sleep were taken from [117]. Mouse 12, Mouse 17 and Mouse 20 only had recordings in the ADn while Mouse 24, Mouse 25 and Mouse 28 had simultaneous recordings from ADn and PoS. The positions were tracked using a platform attached to the heads of the mice with red and blue diodes fixed at both ends. Only the recorded spike times during awake sessions and the neural units with at least 100 observed spikes were considered in this study. Additional information regarding the data acquisition can be found in the paper in Ref. [80] and the CRCNS[1] database entry in [117].

D.1.2 Position and speed filtering

The position time series for the mEC data were smoothed to reduce jitter using a low-pass Hann window FIR filter with cutoff frequency of 2.0 Hz and kernel support of 13 taps (approximately 0.5 s) and were then renormalized to fill missing bins within the kernel duration as done in Ref. [109]. The rat's position was taken to be the average of the recorded and filtered positions of the two tracked diodes. The head direction was calculated as the angle of the perpendicular bisector of the line connecting the two diodes using the filtered positions. The speed at each time point was computed by dividing the trajectory length with the elapsed time within a 13-time point window. When calculating for spatial firing rate maps and spatial information (see below), only time points where the rat was running faster than 5 cm/s were considered. No speed filters were imposed when calculating for head directional tuning curves and head directional information. On the other hand, no position smoothing nor speed filtering were performed when calculating for the spatial firing rate maps and spatial information for the ADn and PoS data.

[1] https://crcns.org

D.1.3 Rate maps

The spike location, $\xi_j^{(i)}$, of neuron i at a spike time $t_j^{(i)}$ was calculated by linearly interpolating the filtered position time series at the spike time. As done in Ref. [109], the spatial firing rate map at position $\mathbf{x} = (x, y)$ was calculated as the ratio of the kernel density estimates of the spatial spike frequency and the spatial occupancy, both binned using 3 cm square bins, as

$$f(\mathbf{x}) = \frac{\sum_{j=1}^{M} K(\mathbf{x}|\xi_j)}{\sum_{j=1}^{M} \Delta t_j K(\mathbf{x}|\mathbf{x}_j)} \tag{D.1}$$

where a triweight kernel

$$K(\mathbf{x}|\xi) = \frac{4}{9\pi\sigma_K^2}\left[1 - \frac{\|\mathbf{x}-\xi\|^2}{9\sigma_K^2}\right]^3, \|\mathbf{x}-\xi\| < 3\sigma_K \tag{D.2}$$

with bandwidth $\sigma_K = 4.2$ cm was used. In place of a triweight kernel, a Gaussian smoothing kernel with $\sigma_G = 4.0$ truncated at $4\sigma_G$ was also used to estimate the rate maps which gave qualitatively similar results. For better visualization, a Gaussian smoothing kernel with $\sigma_G = 8.0$ was used to filter the spatial firing rate map.

On the other hand, for head direction tuning curves, the angles were binned using 9° bins. The tuning curve was then calculated as the ratio of the head direction spike frequency and the head direction occupancy without any smoothing kernels as the head direction bins are sampled well-enough. For better visualization, a Gaussian kernel with smoothing window of 20° was used to filter the tuning curves.

D.1.4 Information, Sparsity and other Scores

Given a feature, ϕ (e.g., spatial position, $\mathbf{x}$, head direction, θ or speed, v), the information between the neural spiking s and the feature can be calculated á la Skaggs-McNaughton [110]. In particular, under the assumption of a non-homogeneous Poisson process with feature dependent rates, $\lambda(\phi)$, under small time intervals Δt, the amount of information, in bits per second, that can be decoded from the rate maps is given by

$$I(s,\phi) = \sum_{\phi} p(\phi)\frac{\lambda(\phi)}{\bar{\lambda}} \log\frac{\lambda(\phi)}{\bar{\lambda}} \tag{D.3}$$

where $\lambda(\phi)$ is the firing rate at ϕ, $p(\phi)$ is the probability of occupying ϕ and

$$\bar{\lambda} \equiv \sum_{\phi} \lambda(\phi)p(\phi) \tag{D.4}$$

is the average firing rate. To account for the bias due to finite samples, the information of a randomized spike frequency was calculated using a bootstrapping procedure. To this end, the spikes were randomly shuffled 1000 times and the information for each shuffled spikes was calculated. The average randomized information was then subtracted from the non-randomized information.

Apart from the information, one of the measures that are used to quantify selectivity of neural firing to a given feature is the firing sparsity [112] which can be calculated using

$$sp_{\phi} = 1 - \frac{\left(\sum_{\phi}\lambda(\phi)p(\phi)\right)^2}{\sum_{\phi}\lambda(\phi)^2 p(\phi)}. \tag{D.5}$$

Apart from the measures of information and sparsity, we also calculated the grid scores, g, for the neurons in the mEC data. The grid score is designed to quantify the hexagonality of the spatial firing rate maps through the spatial autocorrelation maps (or autocorrelograms) and was first used in Ref. [91] to identify putative grid cells. In brief, the grid score is computed from the spatial autocorrelogram where each element ρ_{ij} is the Pearson's correlation of overlapping regions between the spatial firing rate map shifted i bins in the horizontal axis and j bins in the vertical axis and the unshifted rate map. The angular Pearson autocorrelation, $\mathrm{acorr}(u)$, of the spatial autocorrelogram was then calculated using spatial bins within a radius u from the center at lags (or rotations) of $30°$, $60°$, $90°$, $120°$ and $150°$, as well as the $\pm 3°$ and $\pm 6°$ offsets from these angles to account for sheared grid fields [93]. As done in

Ref. [109], the grid score, $g(u)$, for a fixed radius of u, is computed as

$$\begin{aligned} g(u) = \frac{1}{2}&\left[\max\{\mathrm{acorr}(u) \text{ at } 60^\circ \pm (0^\circ, 3^\circ, 6^\circ) + \right. \\ &\left. \max\{\mathrm{acorr}(u) \text{ at } 120^\circ \pm (0^\circ, 3^\circ, 6^\circ)\right] \\ -\frac{1}{3}&\left[\min\{\mathrm{acorr}(u) \text{ at } 30^\circ \pm (0^\circ, 3^\circ, 6^\circ) + \right. \\ &\min\{\mathrm{acorr}(u) \text{ at } 90^\circ \pm (0^\circ, 3^\circ, 6^\circ) + \\ &\left. \min\{\mathrm{acorr}(u) \text{ at } 150^\circ \pm (0^\circ, 3^\circ, 6^\circ)\right]. \end{aligned} \tag{D.6}$$

The final grid score, g, is then taken as the maximal grid score, $g(u)$, within the interval $u \in [12\text{ cm}, 75\text{ cm}]$ in intervals of 3 cm.

Another quantity that was calculated in this paper is the Rayleigh mean vector length, R. Given the angles $\{\theta_1, \ldots, \theta_M\}$ where a neuronal spike was recorded, the mean vector length can be calculated as

$$R = \sqrt{\left(\frac{1}{M}\sum_{i=1}^{M}\cos\theta_i\right)^2 + \left(\frac{1}{M}\sum_{i=1}^{M}\sin\theta_i\right)^2}. \tag{D.7}$$

Note that for head direction cells where the neuron fires at a specific head direction, the angles will be mostly concentrated along the preferred head direction, θ_c, and hence, $R \approx 1$ whereas for neurons with no preferred direction, $R \approx 0$.

D.1.5 Resampling the firing rate map

The calculated rate maps and the real animal trajectory were used to resample the neural activity assuming non-homogeneous Poisson spiking statistics with rates taken from the rate maps. To this end, the real trajectory of the rat was divided into $\Delta t = 1$ ms bins. The position and head direction were linearly interpolated from the filtered positions described above. The target firing rate, f_j in bin j was then calculated by evaluating the tuning profile at the interpolated position or head direction. Whenever the target firing rate was modulated by both the position and head direction, we assumed that the contribution due to each feature was multiplicative and thus, f_j is calculated as the product of the tuning profiles at the interpolated position and the interpolated head direction. A Bernoulli trial was then performed in each bin with a success probability given by $f_j \Delta t$.

D.1.6 Statistical decoding

For positional decoding, we divided the space in a grid of 20×20 cells of 7.5 cm $\times$ 7.5 cm spatial resolution, which was comparable to the rat's body length. Time was also discretized into 20 ms bins which ensured that for most of the time (i.e. in 92% of the cases), the rat was located within a single spatial cell. Under these time scales, the responses of a neuron can be regarded as being drawn from a binomial distribution, i.e., either the neuron i is active ($s_j^{(i)} = 1$) or not ($s_j^{(i)} = 0$) between $(j-1)\Delta t$ and $j\Delta t$. The likelihood of the neural responses, $\mathbf{s}_j = (s_j^{(1)}, \ldots, s_j^{(N)})$ of N independent neurons at a given time conditioned on the position, $\mathbf{x}_j$ is then given by

$$p(\mathbf{s}_j|\mathbf{x}_j) = \prod_{i=1}^{N} (\lambda^{(i)}(\mathbf{x}_j)\Delta t)^{s_j^{(i)}} (1 - \lambda^{(i)}(\mathbf{x}_j)\Delta t)^{1-s_j^{(i)}} \tag{D.8}$$

where $\lambda^{(i)}(\mathbf{x}_j)$ is the firing rate of neuron i at $\mathbf{x}_j$. Given the prior distribution on the position, $p(\mathbf{x}_j)$, which is estimated from the data, the posterior distribution of the position, $\mathbf{x}_j$, given the neural responses, $\mathbf{s}_j$ at time t is given by

$$p(\mathbf{x}_j|\mathbf{s}_j) = \frac{p(\mathbf{s}_j|\mathbf{x}_j)p(\mathbf{x}_j)}{p(\mathbf{s}_j)}. \tag{D.9}$$

The decoded position, as in the Bayesian 1-step decoding in Ref. [129], was calculated as

$$\hat{\mathbf{x}}_j = \arg\max_{\mathbf{x}_j} p(\mathbf{s}_j|\mathbf{x}_j)p(\mathbf{x}_j). \tag{D.10}$$

For head directional decoding, on the other hand, we divided the angles, $\theta \in [0, 2\pi)$ in 9° bins. For this case, time was instead discretized into 100 ms bins. Under these time scales, the neurons could not be regarded simply as either active or not. Hence, it was natural to switch towards the analysis of population vectors, $\mathbf{n}_j$, a vector which represents the number of spikes, $n_j^{(i)}$, recorded from each neuron within the j^{th} time bin, to decode for the head direction. In this case, the number of spikes, $n_j^{(i)}$, that neuron i discharges between $(j-1)\Delta t$ and $j\Delta t$ can be modeled as a non-homogeneous Poisson distribution

$$p(n_j^{(i)}|\theta_j) = \frac{\lambda^{(i)}(\theta_j)^{n_j^{(i)}}}{n_j^{(i)}!} \exp(-\lambda^{(i)}(\theta_j)) \tag{D.11}$$

and thus, under the independent neuron assumption, $p(\mathbf{n}_j|\theta_j) = \prod_{i=1}^{N} p(n_j^{(i)}|\theta_j)$. The decoded head direction can then be calculated as

$$\hat{\theta}_j = \arg\max_{\theta_j} p(\mathbf{n}_j|\theta_j)p(\theta_j). \tag{D.12}$$

where $p(\theta_j)$ is the head directional prior distribution which is estimated from the data. Note that in all of the decoding procedures, we only decoded for time points with which at least one neuron was active.

D.1.7 Source codes

All the calculations in this manuscript were done using personalized scripts written in Python 3. The source codes for calculating multiscale relevance (which is also compatible with Python 2) and for reproducing the figures in the main text are accessible online[2].

D.2 Relation between MSR and other measures of temporal structure

Characterizing the neural spiking can be done by studying the distribution of the time intervals between two succeeding spikes, known in literature as the interspike interval (ISI) distribution which allows us to see whether a neuron fires in bursts [101, 99].

[2] https://github.com/rcubero/MSR

Note that given the time stamps of neural activity $\{t_1, \ldots, t_M\}$, the interspike interval is given by $\{\tau_1, \ldots, \tau_{M-1}\}$ where $\tau_i = t_{i+1} - t_i$. Because the multiscale relevance (MSR) is built to separate relevant neurons from the irrelevant ones through their temporal structures in the neural spiking, we wanted to assess how the proposed measure scales with the characteristics that give structure to temporal events. In the context of the temporal activity of a neuron, a feature of the relevance measure, $H[K]$ is that highly regular, equally-spaced ISI are attributed with a low measure. On the other hand, ISI that follow broad, non-trivial distributions are attributed with a high relevance measure. Hence, we expected that the relevance measure, and therefore the MSR, captures non-trivial bursty patterns of neurons.

To study how MSR behaves with respect to the characteristics of ISI, we considered a stretched exponential distribution

$$P_{SE}^{u}(\tau) = \frac{u}{\tau_0} \left[\frac{\tau}{\tau_0}\right]^{u-1} \exp\left[-\left(\frac{\tau}{\tau_0}\right)^{u}\right] \tag{D.13}$$

with which the parameter u allows us to define the broadness of underlying distribution and τ_0 is the characteristic time constant of the random event. For Poisson processes, the ISI follow an exponential distribution corresponding to $u = 1$ in Eq. (D.13). For $u < 1$, the ISI distribution becomes broad and tends to a power law distribution with an exponent of -1 in the limit when $u \to 0$. On the other hand, for $u > 1$, the distribution becomes narrower and tends to a Dirac delta function in the limit when $u \to \infty$.

Upon fixing the parameters u and τ_0 which fixes the stretched exponential distribution in Eq. (D.13), random ISI, τ_i, could then be sampled independently from Eq. (D.13) so as to generate a time series of 100,000 time units. The MSRs of each time series could then be calculated using the methods described in the main text (Section 2).

To characterize the temporal structures of both the simulated data and neural data, we adapted the measures of bursty-ness and memory of Goh and Barabasi [104].

While the bursty-ness coefficient, b defined as

$$b = \frac{\sigma_\tau - \mu_\tau}{\sigma_\tau + \mu_\tau}, \tag{D.14}$$

measures the broadness of the underlying ISI distribution with μ_τ and σ_τ as the mean and standard deviations of the ISI respectively, the memory coefficient, m defined as

$$m = \frac{1}{M-2}\sum_{j=1}^{M-2}\frac{(\tau_j - \mu_\tau)(\tau_{j+1} - \mu_\tau)}{\sigma_\tau^2}, \tag{D.15}$$

measures the short-time correlation between events.

For the stretched exponential distribution in Eq. (D.13), the mean and standard deviations could be computed as

$$\mu_\tau = \tau_0 \Gamma\left(\frac{u+1}{u}\right) \tag{D.16}$$

and

$$\sigma_\tau = \tau_0 \sqrt{\Gamma\left(\frac{u+2}{u}\right) - \Gamma\left(\frac{u+1}{u}\right)^2} \tag{D.17}$$

where $\Gamma(x) \equiv (x-1)!$ is the gamma function. With these closed-form relationships, we could now study the limiting properties of the burstiness and memory coefficients. For Poisson processes, the mean, μ_τ, and standard deviation, σ_τ, coincide, i.e. $\mu_\tau = \sigma_\tau = \tau_0$, and thus with Eq. (D.14), give $b = 0$. For broad distributions, $u < 1$ in Eq. (D.13), $\sigma_\tau > \mu_\tau$ which gives $b > 0$ and tends to approach $b \to 1$ in the limit $u \to 0$. On the other hand, for narrow distributions, $u > 1$ in Eq. (D.13), $\sigma_\tau < \mu_\tau$ resulting to $b < 0$ and tends to $b \to -1$ in the limit $u \to \infty$. Hence, the bursty-ness parameter, b, is a bounded parameter, i.e., $b \in [-1, 1]$.

For the synthetic datasets, note that fixing the parameter u automatically fixes the bursty-ness coefficient, b. However, because the synthetic ISI are sampled independently, the memory coefficient, m, is approximately zero. Short-term memory can then be introduced by first sorting the ISI in decreasing (or increasing) order which results to $m \approx 1$. Randomly shuffling a subset of the ordered ISI (100 events at a time in this case) results to a monotonic decrease of m. In the limit of infinite data, the memory coefficient is bounded by $[-1, 1]$. These bounds may no longer hold in the case of limited data. Despite this, a positive memory coefficient indicates that a short (long) ISI between events tends to be followed by another short (long) interval and a negative memory coefficient indicates that a short (long) ISI between events tends to be followed by a long (short) interval.

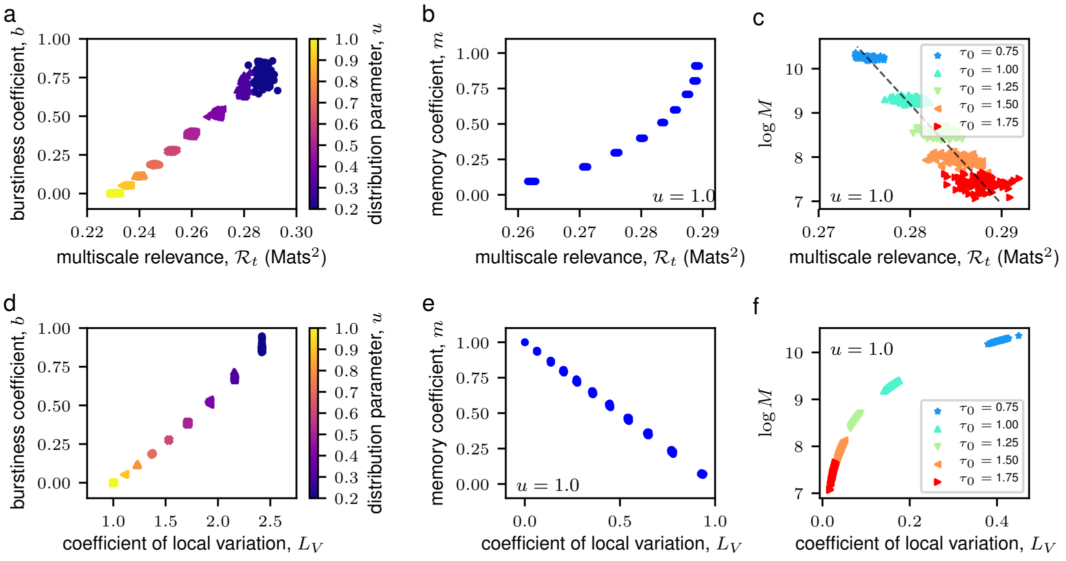

FIGURE D.1. **Synthetically-generated neural data reveals relationship of the MSR and of the coefficient of local variation, L_V, with the bursty-ness and memory coefficients.** Interevent times were drawn from a stretched exponential distribution to simulate random events up to 100,000 time units where short-term memory effects were introduced through a shuffling procedure and the number of random events, M, were varied by modifying the characteristic time constant, τ_0. Scatter plots show how the multiscale relevance (MSR) scales with the bursty-ness coefficient, b (panel **a**), the memory coefficient, m (panel **b**), and $\log M$ (panel **c**). In panel **b**, random events were drawn from a stretched exponential distribution with $u = 1.0$ while in panel **c**, the parameter u was set to 0.3. Panels **d**, **e** and **f**, on the other hand, show the relationship between L_V and bursty-ness coefficient, memory coefficient and $\log M$ respectively. The results for 100 realizations of such random events are shown. Notice, in **c** and **f**, that both the MSR and the L_V are sensitive to the number of spiking events.

With this, we found that the MSR increased with bursty-ness and memory for the synthetically generated dataset as seen in Fig. D.1a and b. We also sought to characterize the relationship between the number of events, M, with the MSR which can be addressed by changing the characteristic time constant, τ_0, in Eq. (D.13) wherein decreasing τ_0 leads to more events and thus, increased $\log M$. We found that MSR decreased with $\log M$ as seen in Fig. D.1c. This result is indicative that MSR of randomly generated events can be explained by $\log M$.

Since the MSR is constructed as a measure of dynamical variablity, we compared our results on synthetically generated datasets with the coefficient of local variation, L_V, [84, 85, 98] defined as

$$L_V = \frac{1}{M-1} \sum_{j=1}^{M-1} \frac{3\left(\tau_j - \tau_{j+1}\right)^2}{\left(\tau_j + \tau_{j+1}\right)^2} \tag{D.18}$$

where the factor 3 in the summand was taken such that, for a Poisson process, $L_V = 1$. With this, we found that the L_V increases with increasing bursty-ness coefficient, b, indicating that power law ISI distributions lead to highly locally variating spiking events. Also, we found that the L_V decreases with increasing short-term memory, m. Finally, like the MSR, we also found a dependence of the L_V with the $\log M$.

Following the results on synthetic data, we also analyzed temporal characteristics in real neural dataset. In the case of neurons in the mEC data, we also found that MSR decreased with the logarithm of the number of observed spikes, $\log M$, as shown in Fig. D.1b. To determine how much of the calculated MSRs can be explained by the number of observed spikes, M, we linearly regressed MSR with $\log M$ shown as the dashed line in Fig. D.2b. Residuals were then calculated as the deviation of the calculated MSR from the regression line and thus, captures the amount of MSR that cannot be explained by $\log M$ alone. We showed in Fig. D.2c and d that the MSR for real dataset still contained information going beyond $\log M$ as the residual MSRs (with respect to $\log M$) still retained the dependence with spatial and HD information as already observed in the main text (Fig. 2). We also observed a positive correlation between MSR and L_V. However, through residual analysis, we found that the residual MSRs (with respect to L_V) still contained spatial and HD information as seen in Fig. D.2f and g.

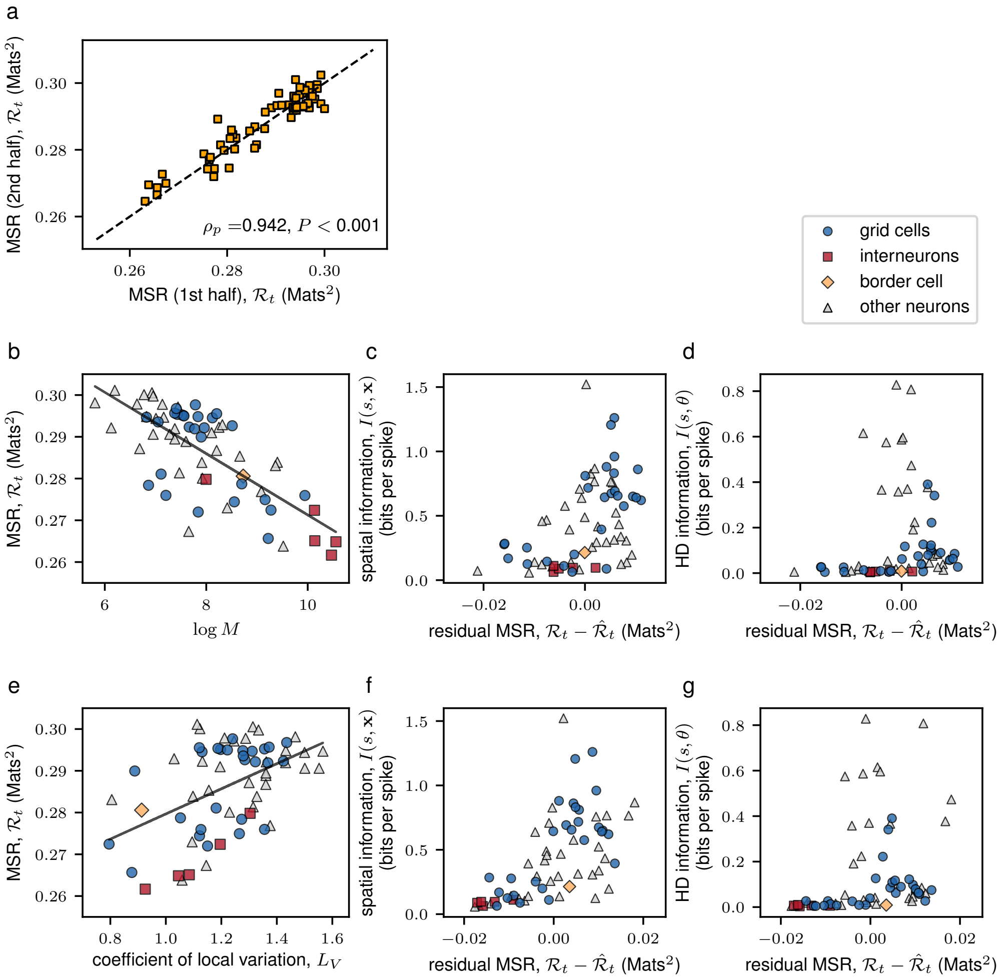

FIGURE D.2. **The MSR is a robust measure and contains information beyond what the number of spikes and local variations can explain.** For each neuron, the MSR was calculated using only the first half and only the second half of the data (panel **a**). The scatter plot reports the two results. The linearity of the relationship between the two sets of partial data is quantified by the Pearson correlation ρ_p along with its P-value. The black dashed line indicates the linear fit. For the neurons in the mEC dataset, the MSR was linearly regressed with $\log M$ (panel **b**). The residual MSR, defined as the deviation of the MSR from the black regression line, were then correlated against spatial (panel **c**) and HD (panel **d**) information. The MSR was also linearly regressed with the coefficient of local variation, L_V (panel **e**). The residual MSRs were then correlated against spatial (panel **f**) and HD (panel **g**) information.

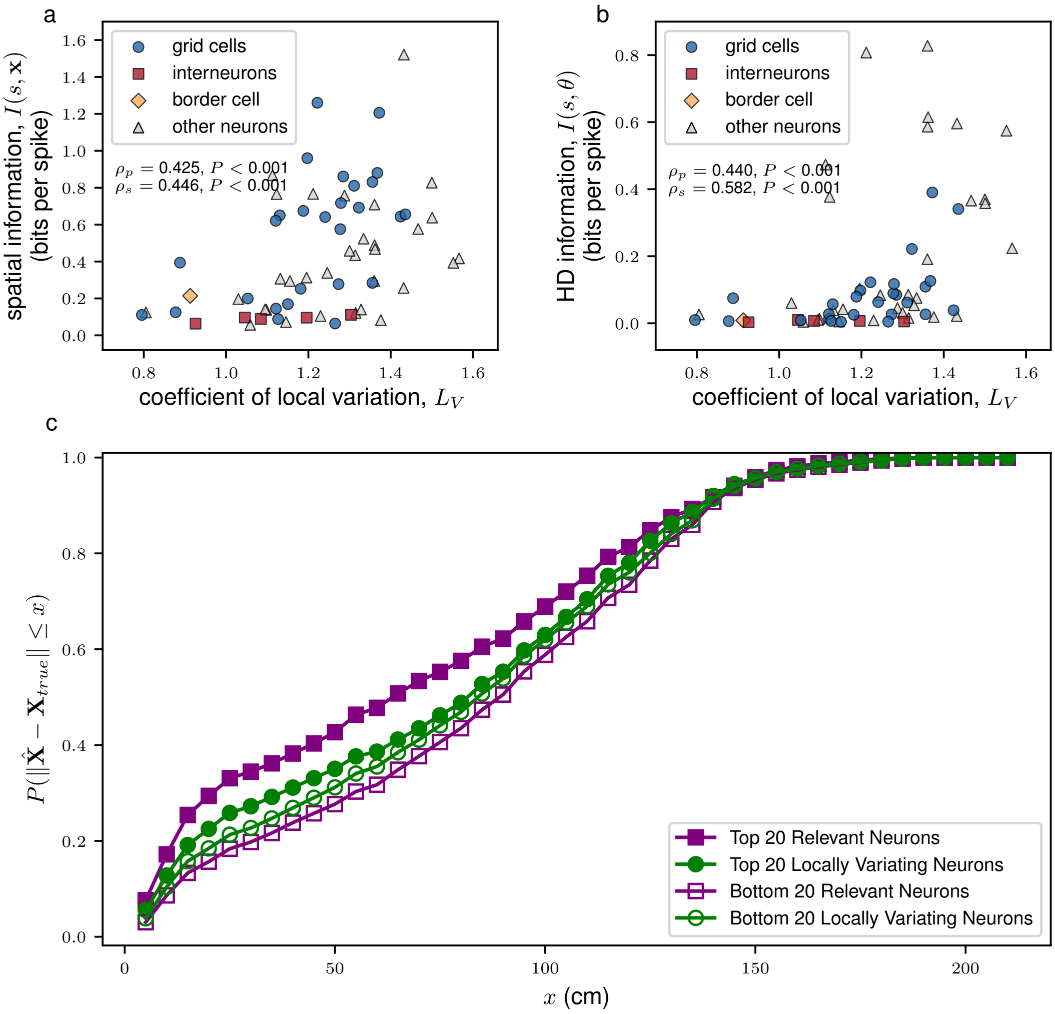

FIGURE D.3. **Local variations in the interspike intervals can capture spatial and HD information but not decodable spatial information.** A scatter plot of the coefficient of local variation L_V vs. the spatial (HD) information is shown in **a** (**b**). The shapes of the scatter points indicate the identity of the neuron according to Ref. [79]. The linearity and monotonicity of the multiscale relevance and the information measures were assessed by the Pearson's correlation, ρ_p, and the Spearman's correlation, ρ_s, respectively. The 20 top and bottom locally variating neurons (LVNs) were then used to decode position (See Main Text Section 5.6). Panel **c** shows the cumulative distribution of the decoding error, $\|\hat{\mathbf{X}} - \mathbf{X}_{true}\|$, for the RNs (solid violet squares) and LVNs (solid green circles) neurons as well as for the non-RNs (dashed violet squares) and non-LVNs (dashed green circles). In all the decoding procedures, time points where all the neurons in the ensemble was silent were discarded in the decoding process.

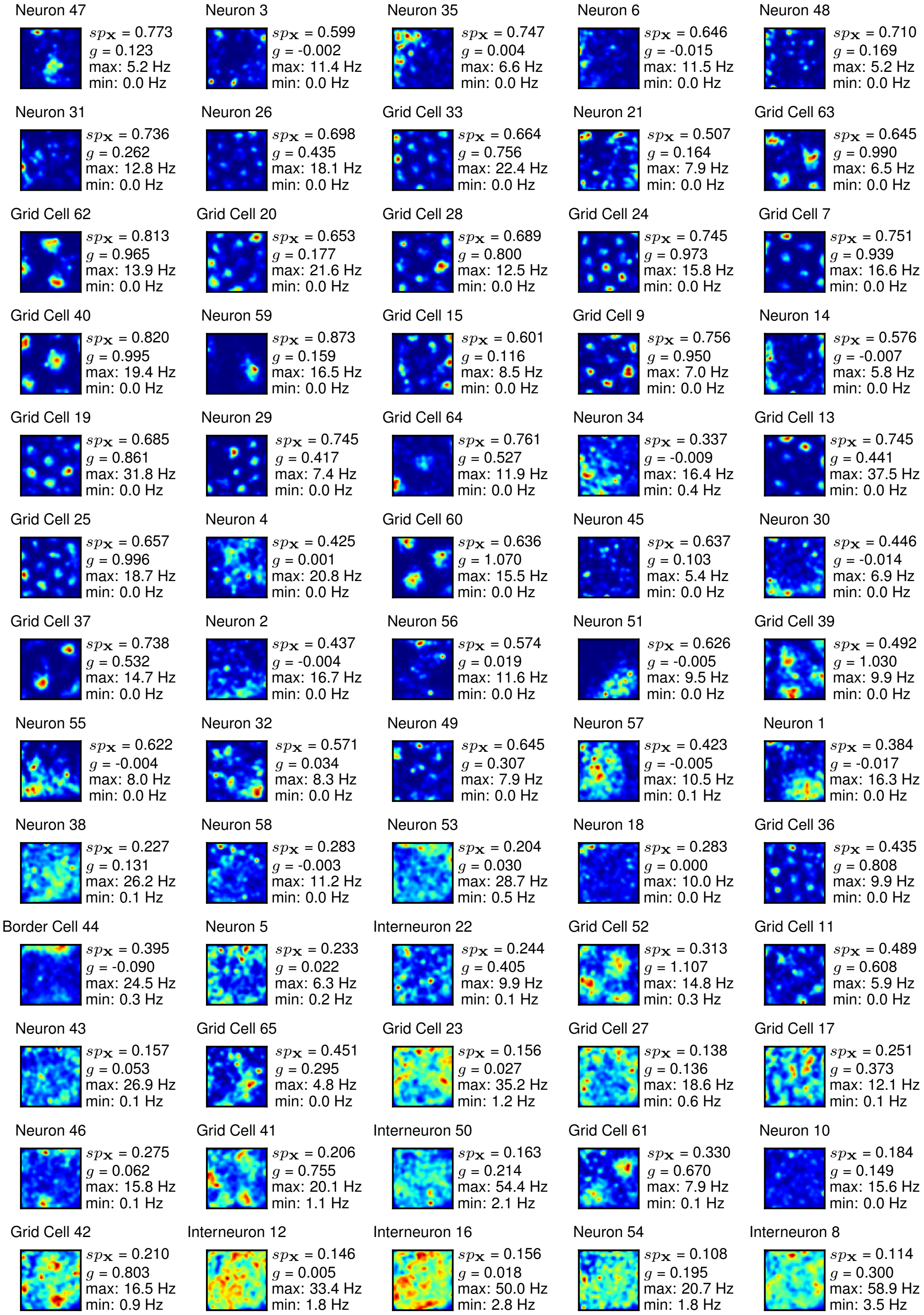

FIGURE D.4. **RNs in the mEC exhibit spatially selective firing compared to non-RNs.** The spatial firing rate maps of the 65 neurons in the mEC data, sorted according to their MSR scores, are shown together with the calculated spatial sparsity, $sp_{\mathbf{x}}$, the grid score, g, and the maximum and minimum firing.

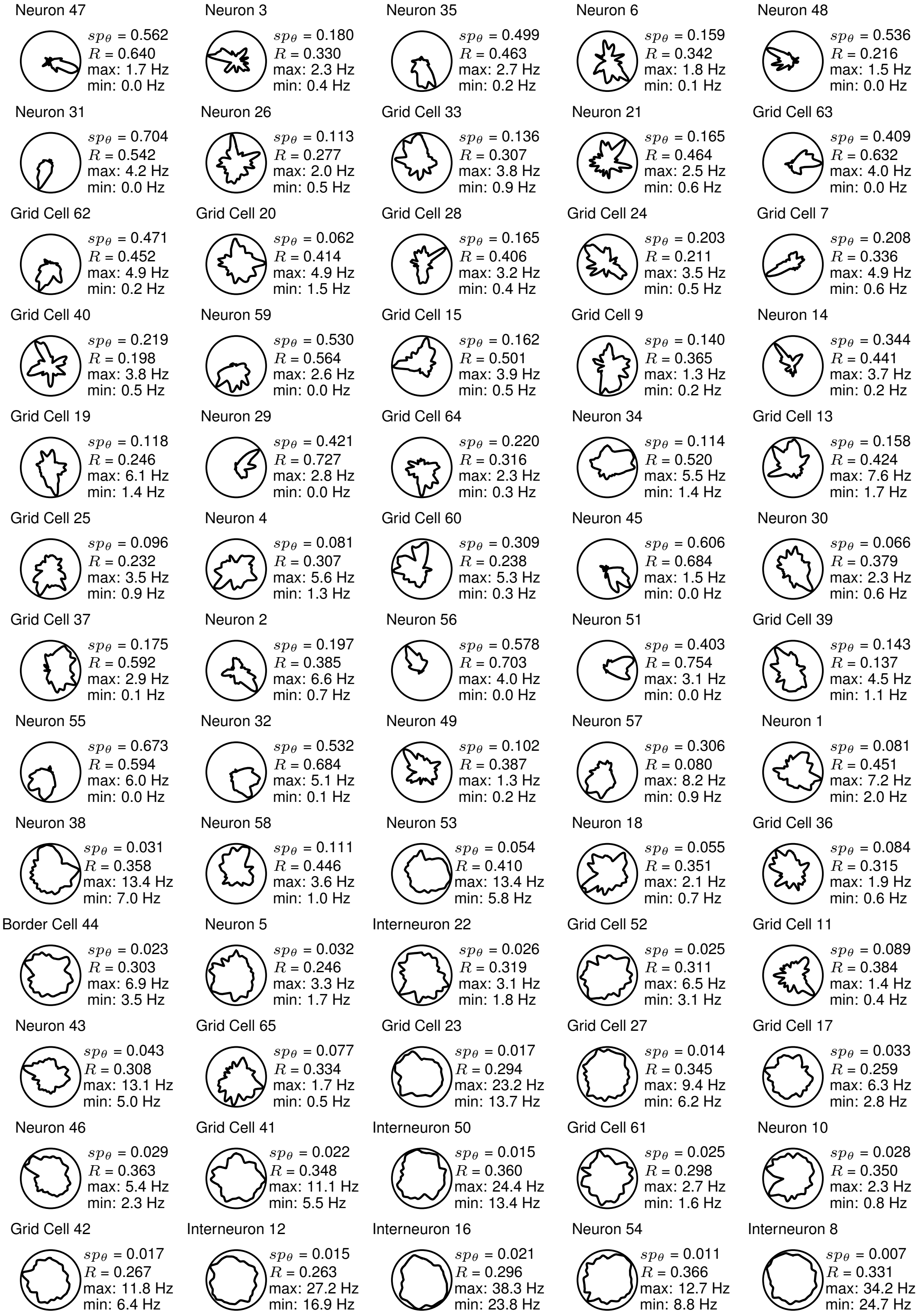

FIGURE D.5. **RNs in the mEC exhibit HD selective firing compared to non-RNs.** The HD tuning curves of the 65 neurons in the mEC data, sorted according to their MSR scores, are shown together with the calculated HD sparsity, sp_θ, the Rayleigh mean vector length, R, and the maximum and minimum firing.

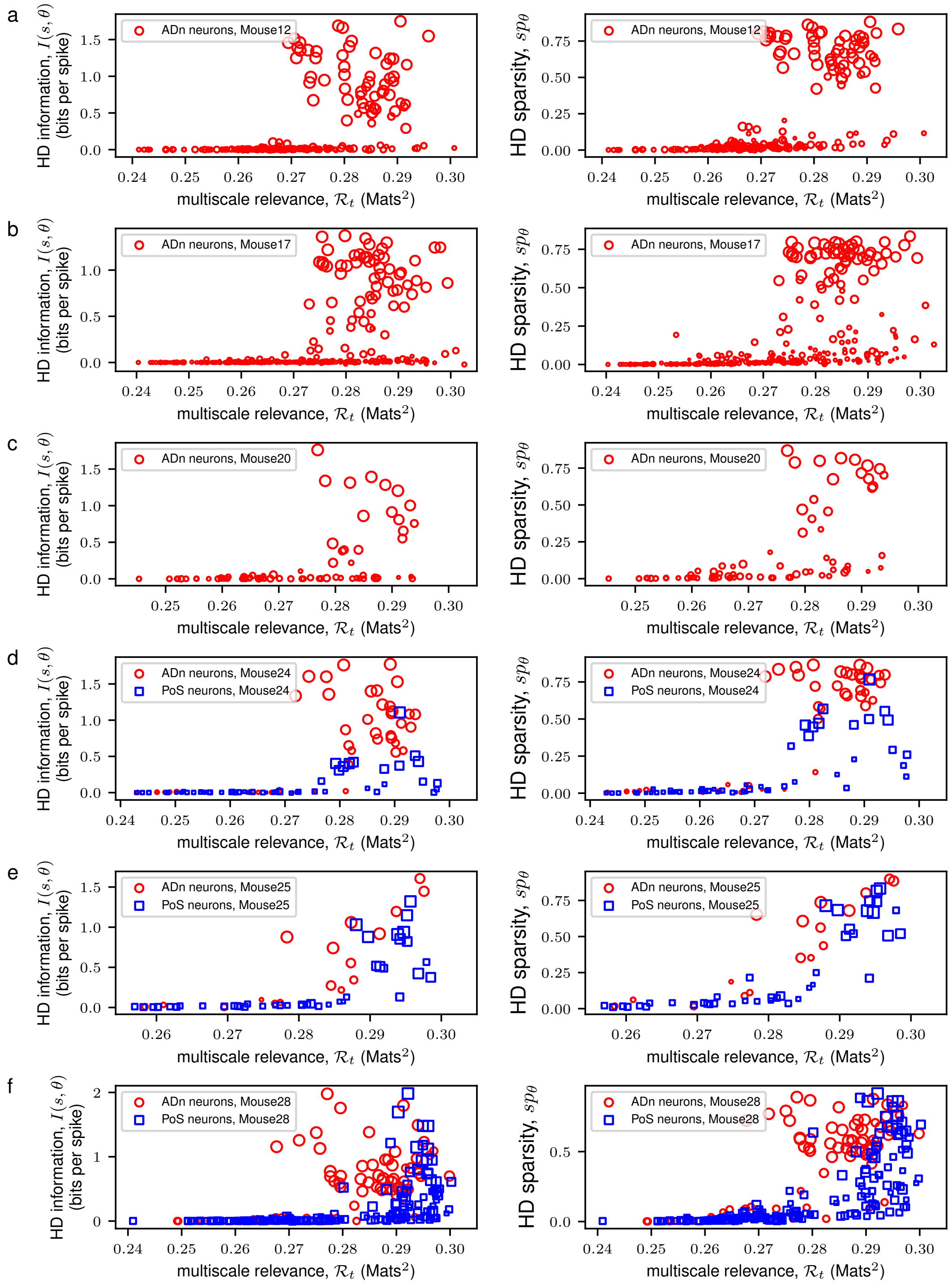

FIGURE D.6. **MSR of neurons from the anterodorsal thalamic nucleus (ADn) and post-subicular (PoS) regions of 6 freely-behaving mice pooled from multiple recording sessions**. For each mice, the MSR of the recorded neurons which had more than 100 recorded spikes in a session were calculated. The corresponding the HD information and sparsity (in bits per spike, see Main Text Section 5.4: Information, Sparsity and other Scores) were also calculated. ADn neurons are depicted in red circles while PoS neurons in blue squares. The size of each point reflect the mean vector lengths of the neurons wherein larger points indicate a unimodal distribution in the calculated HD tuning curves.

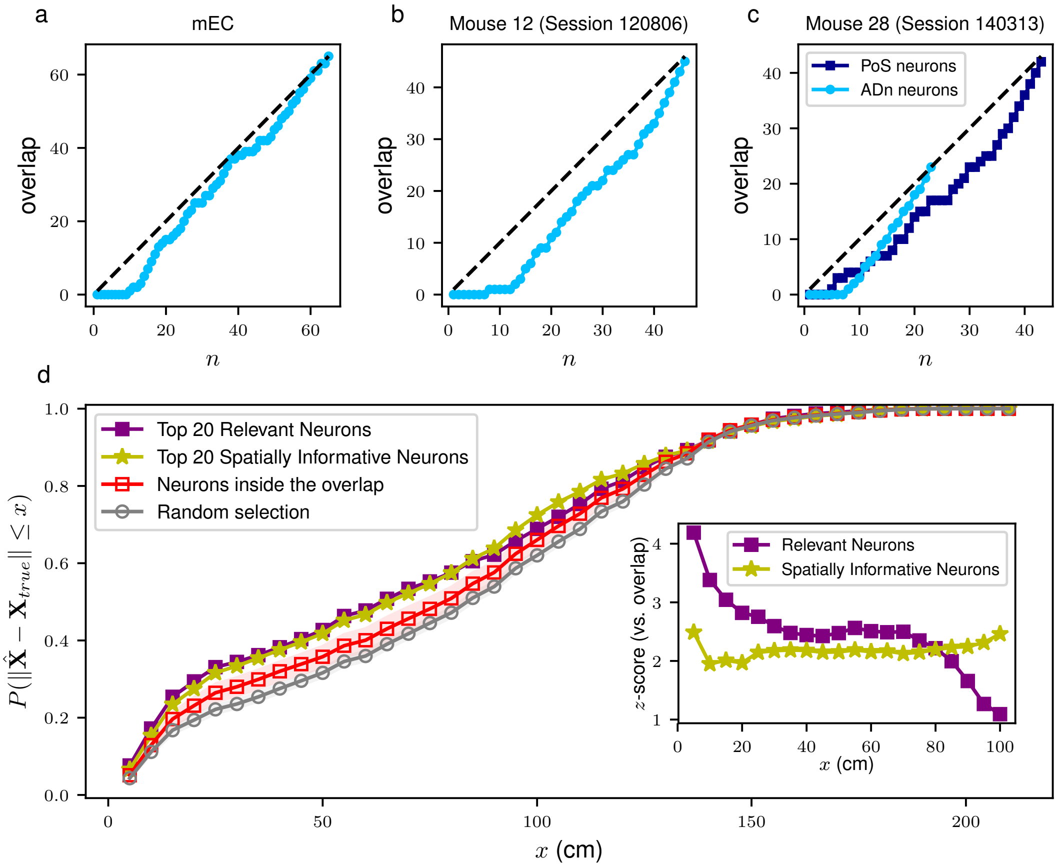

FIGURE D.7. **The mEC neurons that were both RNs and INs do not contain the bulk of the decodable spatial information.** The overlap between the set of RNs and INs as a function of the size, n, of each set for the mEC (**a**), for the ADn of Mouse 12 (Session 120806) (**b**) and for the ADn and PoS of Mouse 28 (Session 140313) (**c**). To show how much decodable spatial information there is in the overlap between the RNs and spatial INs in the mEC at $n = 20$, we took the 14 overlapping neurons (ONs) and randomly chose 6 neurons outside of this overlap and performed a Bayesian positional decoding (see Main Text Section 5.6). The mean and standard errors of the cumulative distribution of decoding errors, $\|\hat{\mathbf{X}} - \mathbf{X}_{true}\|$, of the 14 ONs + 6 random neurons (n = 100 realizations) are shown in grey (**D**) together with the cumulative distribution of decoding errors of the RNs (violet squares) and spatial INs (yellow stars). For a given position error, x, a z-score can be calculated by measuring how many standard errors from the mean of the decoding errors for ONs is the decoding error of the RNs or of the spatial INs. These z-scores are shown in the inset of panel **d**.