Genetic information and biological optimization

by

Michal Hledík

February, 2024

A thesis submitted to the
Graduate School
of the
Institute of Science and Technology Austria
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Committee in charge:

Christoph Lampert, Chair
Nick Barton
Gašper Tkačik
Richard Durbin
Fyodor Kondrashov

Institute of
Science and
Technology
Austria

The thesis of Michal Hledík, titled *Genetic information and biological optimization*, is approved by:

Supervisor: Nick Barton, Institute of Science and Technology Austria, Klosterneuburg, Austria

Signature: ______________________________

Co-supervisor: Gašper Tkačik, Institute of Science and Technology Austria, Klosterneuburg, Austria

Signature: ______________________________

Committee Member: Richard Durbin, Department of Genetics, University of Cambridge, Cambridge, United Kingdom

Signature: ______________________________

Committee Member: Fyodor Kondrashov, Okinawa Institute of Science and Technology, Okinawa, Japan

Signature: ______________________________

Defense Chair: Christoph Lampert, Institute of Science and Technology Austria, Klosterneuburg, Austria

Signature: ______________________________

Signed page is on file

ISTA Thesis, ISSN: 2663-337X

I hereby declare that this thesis is my own work and that it does not contain other people's work without this being so stated; this thesis does not contain my previous work without this being stated, and the bibliography contains all the literature that I used in writing the dissertation.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee, and that this thesis has not been submitted for a higher degree to any other university or institution.

I certify that any republication of materials presented in this thesis has been approved by the relevant publishers and co-authors.

Signature: ______________________________

Michal Hledík
February, 2024

Signed page is on file

Abstract

This thesis consists of four distinct pieces of work within theoretical biology, with two themes in common: the concept of optimization in biological systems, and the use of information-theoretic tools to quantify biological stochasticity and statistical uncertainty.

Chapter 2 develops a statistical framework for studying biological systems which we believe to be optimized for a particular utility function, such as retinal neurons conveying information about visual stimuli. We formalize such beliefs as maximum-entropy Bayesian priors, constrained by the expected utility. We explore how such priors aid inference of system parameters with limited data and enable optimality hypothesis testing: is the utility higher than by chance?

Chapter 3 examines the ultimate biological optimization process: evolution by natural selection. As some individuals survive and reproduce more successfully than others, populations evolve towards fitter genotypes and phenotypes. We formalize this as accumulation of genetic information, and use population genetics theory to study how much such information can be accumulated per generation and maintained in the face of random mutation and genetic drift. We identify the population size and fitness variance as the key quantities that control information accumulation and maintenance.

Chapter 4 reuses the concept of genetic information from Chapter 3, but from a different perspective: we ask how much genetic information organisms actually need, in particular in the context of gene regulation. For example, how much information is needed to bind transcription factors at correct locations within the genome? Population genetics provides us with a refined answer: with an increasing population size, populations achieve higher fitness by maintaining more genetic information. Moreover, regulatory parameters experience selection pressure to optimize the fitness-information trade-off, i.e. minimize the information needed for a given fitness. This provides an evolutionary derivation of the optimization priors introduced in Chapter 2.

Chapter 5 proves an upper bound on mutual information between a signal and a communication channel output (such as neural activity). Mutual information is an important utility measure for biological systems, but its practical use can be difficult due to the large dimensionality of many biological channels. Sometimes, a lower bound on mutual information is computed by replacing the high-dimensional channel outputs with decodes (signal estimates). Our result provides a corresponding upper bound, provided that the decodes are the maximum posterior estimates of the signal.

Acknowledgements

First, I thank my two advisors, Nick Barton and Gašper Tkačik – for the opportunity to learn about several exciting fields of science, their insights, the freedom to pursue my interests, and support in difficult times.

I want to thank Richard Durbin and Fyodor Kondrashov for being on my thesis committee and providing valuable feedback.

I want to thank my collaborators for their help and friendship. Wiktor Młynarski and Tomek Sokolowski contributed directly to the work included here. I learned much from both and enjoyed many conversations together. Katka Boďová inspired and helped me to learn about population genetics. Work and conversations with Jitka Polechová, Ksenia Khudiakova, Réka Borbély and Natália Ružičková helped me develop the ideas included here, and all were also a source of encouragement.

The Barton and Tkačik groups have been great colleagues and friends. Activities like the Barton Group Experience, book club and GGL deserve a mention with their organizers, Anja and Camila among others. In and around the groups, I want to mention Barbora, Himani, Stefanie, Sreyam and Andrea. Among other friends, Yosman and Kasia, and Maťo and Boris at home have been especially important to me.

Many people in ISTA administration make the graduate school a friendly place, and I want to thank May Chan and Sarah Seider in particular.

Chapter 1 and other work was supported by the Scientific Service Units (SSU) of ISTA through resources provided by Scientific Computing (SciComp). I have received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 665385, from the Human Frontiers Science grant no. HFSP RGP0032/2018, and from the European Research Council (101055327 HaplotypeStructure).

About the Author

Michal Hledík completed a Bc. in physics at the Comenius University in Bratislava, and joined ISTA in 2016. His main projects combined population genetics and information theory, with a focus on gene regulation, and excursions into quantitative genetics and computational neuroscience. He published his main results in Neuron and PNAS, and presented at the IEEE Information Theory Workshop, EMBO Workshop: Predicting Evolution and seminars at University of Washington , University of Pennsylvania and University of Vienna. He was involved in Covid-19 a modelling team and helped organize an evolutionary biology competition for high school students, STEB.

List of Collaborators and Publications

Chapter 2 uses, with small modifications (wording and section organization), the full publication

Wiktor Młynarski[12*], Michal Hledík[1*], Thomas R. Sokolowski[13], and Gašper Tkačik[1]. Statistical analysis and optimality of neural systems. *Neuron*, 109(7), 2021.

- W.M. and M.H. performed the research. W.M., M.H., T.R.S. and G.T. designed research and wrote the manuscript. Individual analyses performed by W.M are listed in Chapter 2.

Chapter 3 uses, with small changes to wording, the full publication

Michal Hledík[1], Nick Barton[1], and Gašper Tkačik[1]. Accumulation and maintenance of information in evolution. *Proceedings of the National Academy of Sciences*, 119(36), 2022.

- M.H. performed research; and M.H., N.B., and G.T. designed research and wrote the manuscript.

Chapter 4 contains unpublished results of work with Nick Barton and Gašper Tkačik.

- M.H, N.B., and G.T. designed research; M.H. performed research and wrote the manuscript.

Chapter 5 uses the full publication

Michal Hledík[1], Thomas R. Sokolowski[13], and Gašper Tkačik[1]. A Tight Upper Bound on Mutual Information. In *2019 IEEE Information Theory Workshop (ITW)*, 2019. (© 2019 IEEE)

- M.H. performed research; and M.H, T.R.S., and G.T. designed research and wrote the manuscript.

Additional publication. The following publication is not included in the thesis:

Michal Hledík[1*], Jitka Polechová[4*], Mathias Beiglböck[4], Anna Nele Herdina[5], Robert Strassl[5], and Martin Posch[6]. Analysis of the specificity of a COVID-19 antigen test in the Slovak mass testing program. *PLOS ONE*, 16(7), 2021

- J.P. and M.B. conceptualized the study. M.H. and J.P. curated and investigated the data. M.H. and M.P. conducted the formal analysis and visualized the results. M.H., J.P. and M.P. validated the results. M.H., J.P., M.B., A.N.H., R.S. and M.P. designed the methodology and wrote the article.

[*] *Equal contributions.*

[1] *Institute of Science and Technology Austria (ISTA), Klosterneuburg, Austria.*

[2] *Current address: Faculty of Biology, Ludwig Maximilian University of Munich, Munich, Germany and Bernstein Center for Computational Neuroscience, Munich, Germany.*

[3] *Current address: Frankfurt Institute for Advanced Studies (FIAS), Frankfurt am Main, Germany.*

[4] *Department of Mathematics, University of Vienna, Vienna, Austria.*

[5] *Division of Clinical Virology, Department of Laboratory Medicine, Medical University of Vienna, Vienna, Austria.*

[6] *Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria.*

Table of Contents

List of Figures

List of Abbreviations

GP map Genotype-phenotype map. 57, 59, 60, 65, 68

IC Information content. 58, 61

KL divergence Kullback-Leibler divergence. 4, 37–39, 41, 45, 46, 61, 72

TF Transcription factor. 57, 58, 60, 63, 67–69

CHAPTER 1

Introduction

The chapters of this thesis discuss several different topics within theoretical biology, but there are themes and concepts that they have in common. This introductory chapter briefly discusses these themes, and that way provides general background and motivation for individual chapters.

1.1 Biological optimization

Organisms evolve under natural selection. Across biology, this fact has lead to reasoning that the various organs and behaviors of organisms, as well their genomes, were shaped by selection to help the organisms survive and reproduce, i.e. achieve high fitness (Tinbergen, 1963; Rosen, 2013; Orzack, 2001).

Such arguments can be formulated as normative theories, predicting that among the possible ways that a biological system could be configured, the configuration that will be realized will be one that maximizes fitness, or some proxy – a *utility function* via which that system contributes to organismal fitness. (We leave aside the subtleties of defining fitness and situations where selection does not maximize it.) A classical example are theories of efficient coding in sensory neurons (Barlow, 1961), postulating that the way that sensory neurons respond to stimuli should allow them to represent the stimuli as well as possible, given biophysical constraints, an approach that has found success in predicting response curves and receptive fields (e.g. (Laughlin, 1981; Srinivasan et al., 1982)). Normative theories are not without caveats, though. Not every aspect of every biological system is necessarily optimized in the first place, and even if it is, identifying the utility function for which it is optimized can be difficult (Gould and Lewontin, 1979). Optimization does not need to be perfect (Pérez-Escudero et al., 2009): it operates within the context of biophysical constraints and stochasticity, and being *good enough* may suffice. We may also wish to study optimization in systems where we have some uncertainty over their precise state. The Bayesian statistical framework we introduce in Chapter 2 addresses some of these issues: beliefs about optimization are formalized as maximum-entropy prior distributions over system configurations, with a preference for high utility.

Some instances of biological optimization take place within an individual organism's lifetime, and may be driven by its interactions with its environment (e.g. organisms can learn). Nonetheless, natural selection is the ultimate justification for any normative theory (e.g. because it acts on the capacity to learn), and it is itself an important force for optimization on evolutionary time scales. It is therefore interesting to relate normative theories to population genetics theory, where selection is studied explicitly and quantitatively.

In population genetics, selection is studied in combination with non-adaptive forces such as random genetic drift, mutation and recombination (Fisher, 1930). These non-adaptive forces can sometimes overwhelm selection. Early observations that genome size does not scale with apparent complexity of species (the c-value paradox (Thomas, 1971)) led to the idea that many genomes contain junk DNA (Ohno, 1972), i.e. DNA that is not under selection to perform any function. In humans, the majority of the genome seems to be junk (Ponting and Hardison, 2011; Rands et al., 2014; Doolittle, 2013). (A further complication is that parts of the genome could be selfish, functioning to benefit their own rather than organismal fitness.) Famously, Kimura (Kimura, 1968) used early estimates of rates of molecular evolution to argue that much of evolution must be nearly neutral, i.e. with little to no effect of selection (Kimura, 1983). While the initial arguments needed to be reevaluated, the neutral theory became an important baseline model (Kern and Hahn, 2018; Jensen et al., 2019). Selection causes evolution to depart from this neutral baseline by pushing evolving populations towards fitter genotypes and phenotypes, giving rise to optimization in genetically encoded systems. Optimizing systems with more parameters, or with a greater precision, will require a greater departure from neutrality, but non-adaptive forces impose various limits to selection (Haldane, 1957; Robertson and Waddington, 1960; Barton and Partridge, 2000).

What are the limits to selection with regard to its capacity to optimize heritable phenotypes? We turn to this question in Chapter 3. To formulate such limits in general, we quantify the departure from neutrality using information-theoretic quantities: selection *accumulates information* in the genome. Even though this intuition has been around at least since Kimura's article on this topic in 1961 (Kimura, 1961), and Claude Shannon himself briefly worked in population genetics (Crow, 2001), a general mathematical theory of genetic information has not yet been developed (Maynard Smith, 2000; Griffiths, 2001; Wagner, 2017). In Chapter 3, we propose information measures that bridge three important levels of description (Maynard Smith, 2000): the population (how large is the effect of selection on stochastically evolving populations?), the genotype (how constrained is the DNA sequence?), and the phenotype (how well can selection optimize phenotypes?). We then prove a general bound on how quickly information can be accumulated by selection.

But how much genetic information do organisms actually need? In Chapter 4, we ask this question specifically in the context of gene regulation. Gene regulation is interesting both because of the importance of regulatory sequences (in humans, the majority of conserved sequences are non-coding (Ponting and Hardison, 2011; Rands et al., 2014)) and because gene regulation is particularly amenable to analysis due to its combinatorial nature (sets of genomic regions experience selection for different regulatory phenotypes, but share a genotype-phenotype map). For example, how much information is needed to bind transcription factors in desired genomic locations (Schneider et al., 1986; Wagner, 2017)? The answers to such questions depend not only on the regulatory task, but also on the strength of selection and the effective population size (if selection is weak or the population small, the regulatory task might not be solved) and properties of the genotype-phenotype map (how many genotypes solve the task?).

Regulatory genotype-phenotype maps are an important topic in their own right. The relationships between the DNA sequence and transcription factor binding or transcription are increasingly well described by experiments and predictive models (Yona et al., 2018; Lagator et al., 2022; de Boer et al., 2020; Vaishnav et al., 2022; Fuqua et al., 2020; Galupa et al., 2023), but this opens up questions about why gene regulation is organized the way it is. Chapter 4 combines the concept of genetic information from Chapter 3 with the optimization perspective from Chapter 2. If regulatory genotype-phenotype maps can evolve, what outcome should we expect? While we cannot predict which specific DNA sequences map to which specific phenotype, we can study *how many* sequences map to each phenotype – a property we refer to as the *architecture* of the regulatory phenotypes.

We build on a body of work in population genetics – the equilibrium distributions by Wright (1937), fixation probabilities by Kimura (1962) and the theory of free fitness, developed by independently by Iwasa (1988) and Sella and Hirsh (2005) and extended by Mustonen and Lässig (2010). Free fitness is analogous to free energy in statistical physics, and is maximized as evolution converges to an equilibrium distribution. When written in a suitable form, its entropic term is the genetic information as defined in Chapter 3. By applying free fitness theory to joint systems of regulatory and target loci, we derive an optimization principle for regulatory architectures, which provides an evolutionary justification for the optimization priors introduced in Chapter 2. This result can be used to make predictions about optimal values of some regulatory parameters, and estimate the required genetic information. While individual regulatory sequences evolve to solve specific regulatory tasks, the regulatory architecture evolves to facilitate these solutions across the genome by minimizing the required information. The optimal architecture may require novel regulatory mechanisms, which must also be encoded in the genome – but we find that the information saved by optimization can be so large, that a number of additional genes that implement it can be afforded.

1.2 Information theoretic tools

The systems we consider need a probabilistic description. There are several reasons for this. In Chapter 2, we consider systems with unknown parameter values that we analyze statistically. Chapters 2 as well as Chapter 5 also consider systems that encounter random external stimuli which elicit (possibly noisy) responses. Chapters 3 and 4 consider the evolution of finite populations, i.e. with random genetic drift. All these phenomena are described using probability distributions, and we use several information-theoretic quantities to describe the properties of these distributions.

The Shannon entropy (Cover and Thomas, 2006) of a random variable X,

$$H(X) = -\sum_x P(x) \log_2 P(x), \tag{1.1}$$

here expressed in bits, was originally developed in the context of communication and data compression (Shannon, 1948). If a random source of signals emits signal x with probability $P(x)$, then encoding these signals in bits will take at least $H(X)$ bits per signal on average. Sources where the signals are more difficult to guess ($P(x)$ is more evenly spread among many possible signals) have a higher entropy. Even without the communication context, $H(X)$ can be used simply as a measure of randomness or unpredictability. An example use case are maximum-entropy distributions (Jaynes, 1982, 2003). The optimization priors we introduce in Chapter 2 are an instance of maximum-entropy distributions, and express our partial ignorance about the parameters of the analyzed systems.

The mutual information (Cover and Thomas, 2006) between two random variables X and Y,

$$I(X;Y) = \sum_{x,y} P(X,Y) \log_2 \frac{P(X,Y)}{P(X)P(Y)} \tag{1.2}$$

is a measure of statistical dependence between X and Y. It is zero when X and Y are independent and positive otherwise. It was originally introduced to describe communication via noisy channels. X is the channel input and Y is the channel output, and broadly speaking, the goal is to reconstruct X based on Y. We use mutual information in Chapter 2 as a utility function for biological systems that sense the environment, with X being the state of the environment and Y the response of the system. Chapter 5 examines the relationship between $I(X;Y)$ and the error probabilities when attempting to reconstruct X based on Y.

Finally, the Kullback-Leibler divergence (KL divergence), or relative entropy (Cover and Thomas, 2006) between two distributions P and Q,

$$D_{KL}(P||Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)} \tag{1.3}$$

quantifies how different the two distributions are. It is zero when $P(x) = Q(x)$ for all x and positive otherwise. Originally, it was used to quantify the inefficiency of sub-optimal codes in communication. We use it in Chapter 1 to generalize optimization priors. In situations where our beliefs about a system in absence of any optimization are given by a non-uniform distribution $Q(x)$, optimization priors minimize $D_{KL}(P||Q)$ instead of maximizing entropy $H(X)$. It is also central in Chapters 3 and 4, where we use it to define the information accumulated by selection in evolution. We associate P and Q with evolution with and without selection respectively, such that $D_{KL}(P||Q)$ quantifies how much influence selection has on the variable X.

CHAPTER 2

Optimality and statistical analysis of biological systems

This was a collaboration with Wiktor Młynarski, Thomas R. Sokolowski and Gašper Tkačik (Młynarski et al., 2021). W.M. performed a number of key analyses: the basic toy model analysis in Fig. 2.2, degree of optimality estimation in Fig. 2.4, resolving optimality prediction ambiguities in Fig. 2.5 (including Fig. A.1 and A.2), receptive field analyses in Fig. 2.7 and 2.8 (including Fig. A.3). The remaining analyses were performed or had a major contribution by M.H. Figures mentioned above, as well as Fig. 2.1, 2.3 and 2.9 were also visualized by W.M.

Abstract. Normative theories and statistical inference provide complementary approaches for the study of biological systems. A normative theory postulates that organisms have adapted to efficiently solve essential tasks, and proceeds to mathematically work out testable consequences of such optimality; parameters that maximize the hypothesized organismal function can be derived *ab initio*, without reference to experimental data. In contrast, statistical inference focuses on efficient utilization of data to learn model parameters, without reference to any *a priori* notion of biological function. Traditionally, these two approaches were developed independently and applied separately. Here we unify them in a coherent Bayesian framework that embeds a normative theory into a family of maximum-entropy "optimization priors." This family defines a smooth interpolation between a data-rich inference regime, and a data-limited prediction regime. Using three neuroscience datasets, we demonstrate that our framework allows us to address fundamental challenges relating to inference in high-dimensional, biological problems.

2.1 Introduction

Ideas about optimization are at the core of how we approach biological complexity (Rosen, 2013; Bialek, 2012; Tkačik and Bialek, 2016). Quantitative predictions about biological systems have been successfully derived from first principles in the context of efficient coding (Laughlin, 1981; van Hateren, 1992), metabolic (Kacser and Burns, 1995; Ibarra et al., 2002), reaction (Savir et al., 2010; Tkačik et al., 2008), and transport (Tero et al., 2010) networks, evolution (Orzack, 2001), reinforcement learning (Alexander, 2003), and decision making (Geisler, 2011; Gold and Shadlen, 2007), by postulating that a system has evolved to optimize some utility function under biophysical constraints. Normative theories generate such predictions about living systems *ab initio*, with no (or minimal) appeal to experimental data. Yet as such theories become increasingly high-dimensional and optimal solutions stop being unique, it gets progressively hard to judge whether theoretical predictions are consistent with data (Doi et al., 2012; Bittner et al., 2019), or to define rigorously what that even means (Wang et al., 2016; Park and Pillow, 2017; Eichhorn et al., 2009). Alternatively, data may be “close to” but not “at” optimality, and different instances of the system may show variation “around” optima (Pérez-Escudero et al., 2009; De Martino et al., 2018), but we lack a formal framework to deal with such scenarios. Lastly, normative theories typically make non-trivial predictions only under quantitative constraints which, ultimately, must have an empirical origin, blurring the idealized distinction between a data-free normative prediction and a data-driven statistical inference.

In contrast to normative theories which derive system parameters *ab initio*, the fundamental task of statistical inference is to reliably estimate model parameters from experimental observations. Here, too, biology has presented us with new challenges. While data is becoming increasingly high-dimensional, it is not correspondingly more plentiful; the resulting curse of dimensionality that statistical models face is controlled neither by intrinsic symmetries nor by the simplicity of disorder, as in statistical physics. To combat these issues and simultaneously deal with the noise and variability inherent to the experimental process, modern statistical methods often rely on prior assumptions about system parameters. These priors either act as statistical regularizers to prevent overfitting or to capture low-level regularities such as smoothness, sparseness or locality (Park and Pillow, 2011). Typically, however, their statistical structure is simple and does not reflect the prior knowledge about system function.

Normative theories and inference share a fundamental similarity: they both make statements about parameters of biological systems. While these statements have traditionally been made in opposing “data regimes” (Fig. 2.1), we observe that the two approaches are not exclusive and could in fact be combined with mutual benefit. To this end, we develop a Bayesian statistical framework that combines data likelihood with an “optimization prior” derived from a normative theory; contrary to simple, typically applied priors, optimization priors can induce a complex statistical structure on the space of parameters. This construction allows us to rigorously formulate and answer the following key questions: (1) Can one derive a statistical hypothesis test for the consistency of data with a proposed normative theory? (2) Can one define how close data is to the proposed optimal solution? (3) How can data be used to set the constraints in, and resolve the degeneracies of, a normative theory? (4) To what extent do optimization priors aid inference in high-dimensional statistical models?

The primary focus of this work is to develop conceptual and theoretical links between normative theories and statistical analyses. We illustrate the application of these developments to simple model systems, and demonstrate their relevance to real-world data analysis on three diverse, yet still relatively tractable, examples. Applying similar methodology to large-scale high-dimensional data would necessitate further development of sophisticated computational or approximative schemes. We recognize that as an outstanding and highly relevant challenge for future research.

2.2 Bayesian inference and optimization priors

Given a probabilistic model for a system of interest, $P(x|\theta)$, with parameters θ, and a set of T observations (or data) $\mathcal{D} = \{x_t\}_{t=1}^{T}$, Bayesian inference consists of formulating a (log)

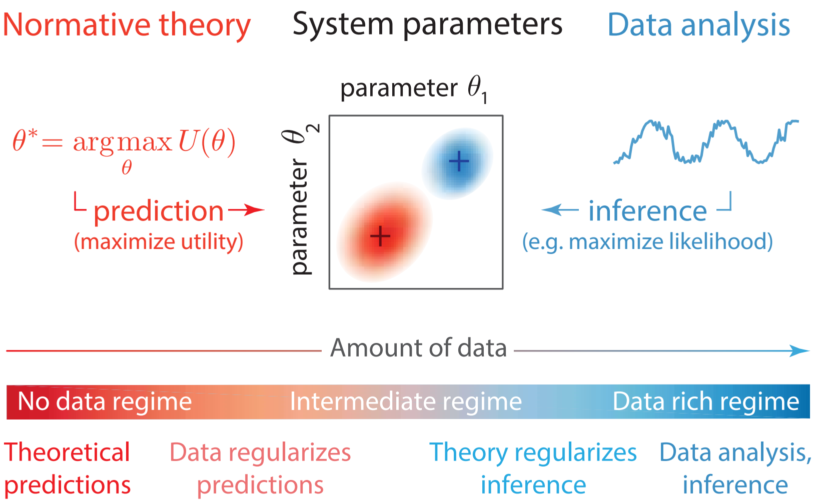

Figure 2.1: **Normative theories and statistical inference.** Both approaches make statements about values of system parameters (middle row; center panel). Normative theories predict which parameters would be of highest utility to the system (middle row in red; left panel) without reference to experimental data. Data analysis infers parameter values from experimental observations (middle row in blue; right panel). Large amounts of data support reliable inference of parameters. We consider a continuum of regimes that are applicable with different amounts of data (bottom row).

posterior over parameters given the data:

$$\log P(\theta|\mathcal{D}) = \log \mathcal{L}(\theta) + \log P(\theta) + \text{const}, \tag{2.1}$$

where the constant term is independent of the parameters, $\mathcal{L}(\theta) = \prod_{t=1}^{T} P(x_t|\theta)$ is the likelihood assuming independent and identically distributed observations x_t, and $P(\theta)$ is the prior, or the postulated distribution over the parameters in absence of any observation. Much work has focused on how the prior should be chosen to permit optimal inference, ranging from uninformative priors (Jeffreys, 1946), priors that regularize the inference and thus help models generalize to unseen data (MacKay, 2003b; Murphy, 2012), or priors that can coarse-grain the model depending on the amount of data samples, T (Machta et al., 2013).

Our key intuition will lead us to a new class of priors fundamentally different from those considered previously. A normative theory for a system of interest with parameters θ can typically be formalized through a notion of a (upper-bounded) utility function, $U(\theta; \xi)$, where ξ are optional parameters which specify the properties of the utility function itself. Optimality then amounts to the assumption that the real system operates at a point in parameter space, θ^*, that maximizes utility, $\theta^*(\xi) = \text{argmax}_\theta \, U(\theta; \xi)$. Viewed in the Bayesian framework, the assertion that the system is optimal thus represents an infinitely strong prior where the parameters are concentrated at θ^*, i.e., $P(\theta|\xi) = \delta(\theta - \theta^*(\xi))$. In this extreme case, no data is needed to determine system parameters: the prior fixes their values and typically no finite amount of data will suffice for the likelihood in Eq (2.1) to move the posterior away from θ^*. This concentrated prior can, however, be interpreted as a limiting case of a softer prior that "prefers" solutions close to the optimum.

Consistent with the maximum entropy principle put forward by Jaynes (Jaynes, 2003), we therefore consider for our priors distributions that are as random and unstructured as possible while attaining a prescribed average utility:

$$P(\theta|\beta, \xi) = \frac{1}{Z(\beta, \xi)} \exp\left[\beta U(\theta; \xi)\right]. \tag{2.2}$$

This is in fact a family of priors, whose strength is parametrized by β: when $\beta = 0$, parameters are distributed uniformly over their domain without any structure and in absence of any optimization; as $\beta \to \infty$, parameter probability localizes at the point $\theta^*(\xi)$ that maximizes the utility to $U_{\max}(\xi)$ (if such a point is unique) irrespective of whether data supports this or not. At finite β, however, the prior is "smeared" around $\theta^*(\xi)$ so that the average utility, $\bar{U}(\beta, \xi) = \int d\theta \; P(\theta|\beta, \xi) U(\theta, \xi) < U_{\max}(\xi)$ increases monotonically with β. For this reason, we refer to β as the "optimization parameter," and to the family of priors in Eq (2.2) as "optimization priors." In evolutionary context, the optimization parameter β is related to the effective population size, see Sec. 2.8.4 and Chapter 4.

The intermediate regime, $0 < \beta < \infty$, in the prior entering Eq (2.1) is interesting from an inference standpoint. It represents the belief that the system may be "close to" optimal with respect to the utility $U(\theta; \xi)$ but this belief is not absolute and can be outweighed by the data: the log likelihood, $\log \mathcal{L}$, grows linearly with the number of observations, T, matching the roughly linear growth of the log prior with β. Varying β thus literally corresponds to the interpolation between an infinitely strong optimization prior and pure theoretical prediction in the "no data regime" and the uniform prior and pure statistical inference in the "data rich regime", as schematized in Fig. 2.1.

Additional parameters of the utility function, ξ, determine its shape in the domain of parameters θ. Parameters ξ can be known and fixed for a specific theory or, if unknown *a priori*, inferred from the data in a Bayesian fashion. When there are no utility parameters ξ to consider, we will suppress them for notational simplicity.

In the following, we apply this framework to a toy model system, a single linear-nonlinear neuron, which is closely related to logistic regression. This example is simple, well-understood across multiple fields, and low-dimensional so that all mathematical quantities can be constructed explicitly; the framework itself is, however, completely general. We then apply our framework to a more complex neuron model and to three experimental data sets. Taken together, these examples demonstrate how the ability to encode the entire shape of the utility measure into the optimization prior opens up a more refined and richer set of optimality-related statistical analyses.

2.2.1 Example: Efficient coding in a simple model neuron

Let us consider a simple probabilistic model of a spiking neuron (Fig. 2.2A), a broadly applied paradigm in sensory neuroscience (Sharpee and Bialek, 2007; Kastner et al., 2015; Paninski et al., 2007; Tkačik et al., 2010; Gjorgjieva et al., 2014). The neuron responds to one-dimensional continuous stimuli x_t either by eliciting a spike ($r_t = 1$), or by remaining silent ($r_t = 0$). The probability of eliciting a spike in response to a particular stimulus value is determined by the nonlinear saturating stimulus-response function. The shape of this function is determined by two parameters: position x_0 and slope k (see Methods).

Parameters $\theta = \{x_0, k\}$ fully determine the function of the neuron, yet remain unknown to the external observer. Statistical inference extracts parameter estimates $\hat{\theta}$ using experimental data $\mathcal{D}$ consisting of stimulus-response pairs (Fig. 2.2B, left panel), by first summarizing the data with the likelihood, $\mathcal{L}(\theta)$ (Fig. 2.2B, right panel), followed either by maximization of the likelihood, $\hat{\theta} = \mathrm{argmax}_\theta \mathcal{L}(\theta)$ in the maximum-likelihood (ML) paradigm, or by deriving $\hat{\theta}$ from the posterior, Eq (2.1), in the Bayesian paradigm.

To apply our reasoning, we must propose a normative theory for neural function, form the

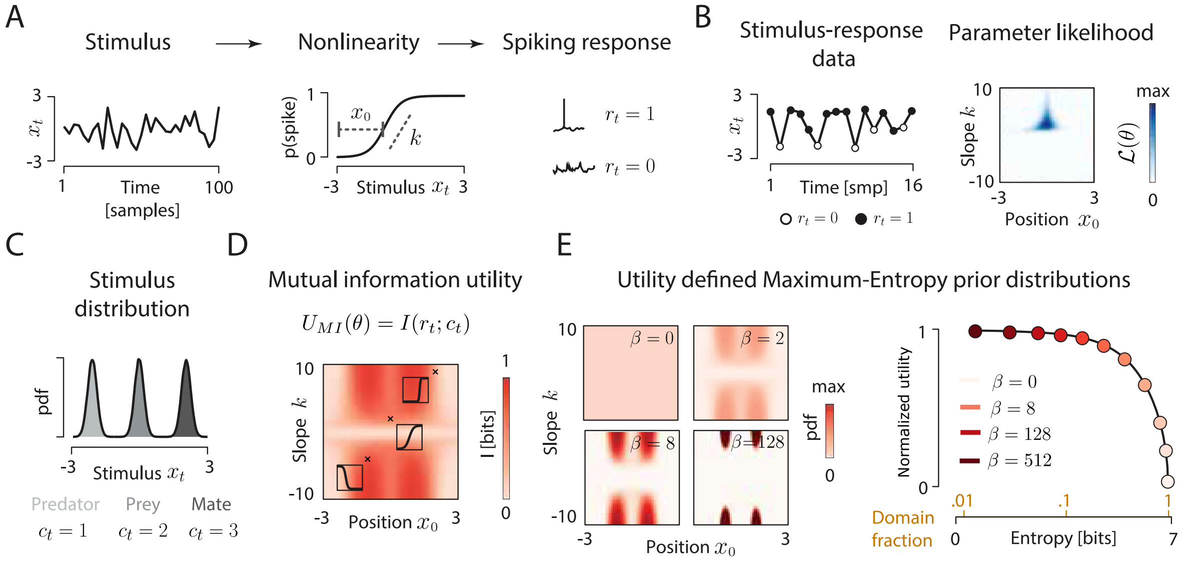

Figure 2.2: **Efficient coding in a toy model neuron and the corresponding optimization prior.** **(A)** Model neuron uses a logistic nonlinearity (middle panel) to map continuous stimuli x_t (left panel) to a discrete spiking response r_t (right panel). The shape of the nonlinearity is described by two parameters: slope k and offset x_0. **(B)** An example dataset (left panel) consisting of stimulus values (black line) and associated spiking responses (empty circles – no spike, full circles – spike). Likelihood function of the nonlinearity parameters defined by the observed data. Dark blue corresponds to most likely parameter values. **(C)** Distribution of natural stimuli to which the neuron might be adapted. In this example, each mode corresponds to a behaviorally relevant state of the environment: presence of a predator, a prey or a mate. **(D)** Efficient coding utility function, here, the mutual information between neural response r_t and the state of the environment, c_t, with stimuli drawn from the distribution in panel C. The amount of information conveyed by the neuron depends on the position and slope of the nonlinearity. Insets depict example nonlinearities corresponding to parameter values marked with black crosses. **(E)** Four maximum-entropy optimization priors over parameters for the neural nonlinearity (left panel). Distributions are specified by the utility of each slope-offset combination. Increasing parameter β constrains the distribution (lowers its entropy) and increases the expected utility of the parameters (right panel). Here we plot the normalized utility $\tilde{U}(\theta)$ - see main text for explanation. Orange numbers on the horizontal axis specify the fraction of the entire domain effectively occupied by parameters at given β.

optimization prior, and combine it with the likelihood in Fig. 2.2B, as prescribed by the Bayes rule in Eq (2.1). An influential theory in neuroscience called "efficient coding" postulates that sensory neurons maximize the amount of information about natural stimuli they encode into spikes given biophysical constraints (Barlow, 1961; van Hateren, 1992; Tkačik et al., 2010; Olshausen and Field, 1996; Smith and Lewicki, 2006; Chalk et al., 2018). This information-theoretic optimization principle (Shannon, 1948) has correctly predicted neural parameters such as receptive field (RF) shapes (Olshausen and Field, 1996; Hyvärinen et al., 2009) and the distribution of tuning curves (Ganguli and Simoncelli, 2014; Wang et al., 2016), as well as other quantitative properties of sensory systems (Laughlin, 1981; Ratliff et al., 2010; Borghuis et al., 2008; Młynarski, 2015; Młynarski and McDermott, 2018; Carlson et al., 2012), *ab initio*, from the distribution of ecologically relevant stimuli (Olshausen and Field, 1996; Bialek, 2012).

To apply efficient coding, we need to specify a distribution from which the stimuli x_t are drawn. In reality, neurons would respond to complex and high-dimensional features of sensory inputs, such as a particular combination of odorants, timbre of a sound or a visual texture, in order to help the animal discriminate between environmental states of very different behavioral relevance (e.g. a presence of a predator, a prey or a mate). To capture this intuition in our simplified setup, we imagine that the stimuli x_t are drawn from a multi-modal distribution, which is a mixture of three different environmental states, labeled by c_t (Fig. 2.2C). Efficient coding then postulates that the neuron maximizes the mutual information, $I(r_t; c_t)$, between the environmental states, c_t, that gave rise to the corresponding stimuli, x_t, and the neural responses, r_t.

Mutual information, which can be evaluated for any choice of parameters k, x_0, provides the utility function, $U_{\mathrm{MI}}(k, x_0) = I(r_t; c_t)$, relevant to our case; in this simple example, the utility function has no extra parameters ξ. Figure 2.2D shows that U_{MI} is bounded between 0 and 1 bit (since the neuron is binary), but does not have a unique maximum. Instead, there are four combinations of parameters that define four degenerate maxima, corresponding to the neuron's nonlinearity being as steep as possible (high positive or negative k) and located in any of the two "valleys" in the stimulus distribution (red peaks in Fig. 2.2D). Moreover, the utility function forms broad ridges on the parameter surface, and small deviations from optimal points result only in weak decreases of utility. Consequently, formulating clear and unambiguous theoretical predictions is difficult, an issue that has been recurring in the analysis of real biological systems (Brinkman et al., 2016; Pitkow and Meister, 2012).

Given the utility function, the construction of the maximum-entropy optimization prior according to Eq (2.2) is straightforward. Explicit examples for different values of β are shown in Fig. 2.2E (left panel). Generally, the average utility of the prior monotonically increases as the prior becomes more localized around the optimal solutions, as measured by the decrease in entropy of the prior (Fig. 2.2E, right panel). This can be interpreted as restricting the system into a smaller part of the parameter domain. If an increase in average utility requires a reduction in entropy by 1 bit, this means that the parameters will be sampled from at most half the available domain.

Before proceeding, we note that our approach depends on several non-trivial choices. First, the fact that system parameterization and the size of the parameter domain can affect Bayesian inferences is well recognized (Gelman, 2004) and we discuss how it relates to our case in Supplemental Information (Sec. 2.8.1, 2.8.3; Fig. 2.10, 2.11). Second, β and the utility function enter the optimization prior of Eq (2.2) as a product, leaving the scale of each quantity arbitrary. For interpretation purposes we therefore define the normalized utility, $\tilde{U} = (\bar{U}(\beta) - \bar{U}(\beta = 0))/(U_{\max} - \bar{U}(\beta = 0))$, which takes on values between 0 and 1 for non-negative β, and is insensitive to linear scaling. We discuss the issue of β scaling in Supplemental Information (Sec. 2.8.4). Third, data and optimality theories could be combined in multiple ways. However combining them via maxent optimization priors enjoys favorable theoretical guarantees that alternative approaches may lack, which we demonstrate in Supplemental Information (Sec. 2.8.5, Fig. 2.13, 2.14 Methods S5, Fig. S4, S5). These considerations complete our setup and allow us to address the four questions posed in the Introduction.

2.3 Question 1: Statistical test for the optimality hypothesis

Given a candidate normative theory and experimental data for a system of interest, a natural question arises: Does the data support the postulated optimality? This question is non-trivial for two reasons. First, optimality theories typically do not specify a sharp boundary between optimal and non-optimal parameters, but rather a smooth utility function $U(\theta)$ (Fig. 2.3A): How should the test for optimality be defined in this case? Second, a finite dataset $\mathcal{D}$ might be insufficient to infer a precise estimate of the parameters θ, but will instead yield a (possibly broad) likelihood surface (Fig. 2.3B): How should the test for optimality be formulated in the presence of such uncertainty?

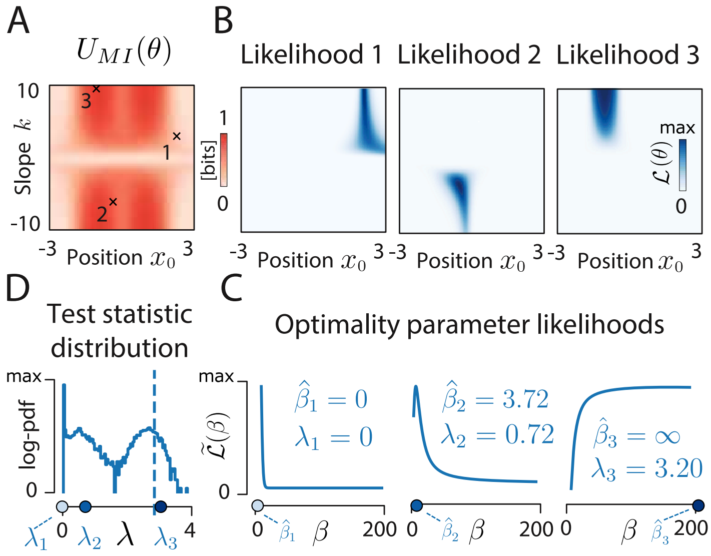

Figure 2.3: **Statistical test of optimality.** **(A)** Utility function $U_{\text{MI}}(k, x_0)$. Crosses and numbers show the locations of ground truth parameters. **(B)** Likelihood of the nonlinearity parameters obtained from 20 stimulus–response (x_i, r_i) pairs. The three examples correspond to three ground truth parameter values (black crosses in A), and are ordered by increasing utility. **(C)** Marginal likelihood of the optimality parameter β, $\tilde{\mathcal{L}}(\beta) = P(\mathcal{D}|\beta)$, corresponding to data in A. Maximum likelihood estimates $\hat{\beta}_{1,2,3}$ (blue circles) indicate that the data would be most probable with no preference for high utility U_{MI} (left panel, $\hat{\beta}_1 = 0$ – note that we do not allow negative $\hat{\beta}$), some preference for high U_{MI} (middle panel, $\hat{\beta}_2 > 0$ finite) and strong preference for high U_{MI} (right panel, $\hat{\beta}_3 \to \infty$; blue circle displayed at $\beta = 200$ for illustration purposes). Likelihood ratio statistic $\lambda_{1,2,3}$ compares the marginal likelihood of β at $\beta = 0$ vs. $\beta = \hat{\beta}_{1,2,3}$ (see Methods). **(D)** Null distribution of the test statistic λ. Point mass at $\lambda = 0$ corresponds to cases where the maximum likelihood optimality parameter is zero, $\hat{\beta} = 0$. High values of λ are evidence against the null hypothesis that $\beta = 0$, and hence support optimality. Dashed vertical line represents $p = 0.05$ significance threshold, blue circles show $\lambda_{1,2,3}$. Only λ_3 crosses the threshold, indicating significant preference for high utility parameters.

Here we devise an approach to address both issues. The basis of our test is a null hypothesis that the system is not optimized, i.e., that its parameters have been generated from a uniform random distribution on the biophysically accessible parameter domain. This distribution is exactly the optimization prior $P(\theta|\beta = 0)$. The alternative hypothesis states that the parameters are drawn from a distribution $P(\theta|\beta)$ with $\beta > 0$. To discriminate between the two hypotheses, we use a likelihood ratio test with the statistic λ, which probes the overlap of high-likelihood and high-utility parameter regions. Specifically, we define the marginal likelihood of β given data, $\tilde{\mathcal{L}}(\beta) = P(\mathcal{D}|\beta) = \int d\theta \mathcal{L}(\theta) P(\theta|\beta)$ (Fig. 2.3C), and then define λ as the log ratio between the maximal marginal likelihood, $\max_{\beta>0} \tilde{\mathcal{L}}(\beta)$, and the marginal likelihood under the null hypothesis, $\tilde{\mathcal{L}}(\beta = 0)$ (see Methods). Here, we assumed for simplicity that the utility function U does not depend on any additional parameters ξ; this simplification is relaxed in the Supplemental Information (Sec. 2.8.2, Fig. 2.12).

The test statistic λ has a null distribution that can be estimated by sampling (Fig. 2.3D), with large λ implying evidence against the null hypothesis; thus, given a significance threshold, we can declare the system to show significant degree of optimization, or to be consistent with no optimization. This is different from asking if the system is "at" an optimum: such a narrow view seems too restrictive for complex biological systems. Evolution, for example, might not have pushed the system all the way to the biophysical optimum (e.g., due to mutational load or because the adaptation is still ongoing), or the system may be optimal under utility function or resource constraints slightly different than those postulated by our theory (De Martino et al., 2018). Instead, the proposed test asks if the system has relatively high utility, compared to the utility distribution in the full parameter space. This approach has been used e.g. in the context of the genetic code, which has been argued to be exceptionally robust withing the space of possible codes (Haig and Hurst, 1991; Freeland and Hurst, 1998).

While principled, this hypothesis test is computationally expensive, since it entails an integration over the whole parameter space to compute the marginal likelihoods, $\tilde{\mathcal{L}}(\beta)$, as well as Monte Carlo sampling to generate the null distribution. The first difficulty can be resolved when the number of observations T is sufficient such that the likelihood of the data, $\mathcal{L}(\theta)$, is sharply localized in the parameter space; in this case the value of the utility function at the peak of the likelihood itself becomes the test statistic and the costly integration can be avoided (see Methods). The second difficulty can be resolved when we can observe many systems and collectively test them for optimality; in this case the distribution of the test statistic approaches the standard χ^2 distribution (see Methods).

2.4 Question 2: Inferring the degree of optimality

Hypothesis testing provides a way to resolve the question whether the data provides evidence for system optimization or not (or to quantify this evidence with a p-value). However, statistical significance does not necessarily imply biological significance: with sufficient data, rigorous hypothesis testing can support the optimality hypothesis even if the associated utility increase is too small to be biologically relevant. Therefore, we formulate a more refined question: How strongly is the system optimized with respect to a given utility, $U(\theta)$?

Methodologically, we are asking about the value of the optimization parameter, β, that is supported by the data $\mathcal{D}$. In the standard Bayesian approach, all parameters of the prior are considered fixed before doing the inference; the prior is then combined with likelihood to generate the posterior (Fig. 2.4A). Our case corresponds to a hierarchical Bayesian scenario, where β is itself unknown and of interest. In the previous section we chose it by maximizing the marginal likelihood, $\tilde{\mathcal{L}}(\beta)$, to devise a yes/no hypothesis test. Here, we consider a fully Bayesian treatment, which is particularly applicable when we observe many instances of the same system. In this case, we interpret different instances (e.g., multiple recorded neurons) as samples from a distribution determined by a single population optimality parameter β (Fig. 2.4B) that is to be estimated. Stimulus-response data from multiple neurons are then used directly to estimate a posterior over β via hierarchical Bayesian inference.

To explore this possibility, we generate parameters θ_n of $n = 1, \dots, N$ model neurons from three different distributions: strongly optimized ($\beta = 12$; Fig. 2.4C, left panel), weakly optimized ($\beta = 4$; Fig. 2.4C, middle panel) and non-optimal (Gaussian distribution of parameters; Fig. 2.4C, right panel). For each of the three examples, we simulate stimulus-response data

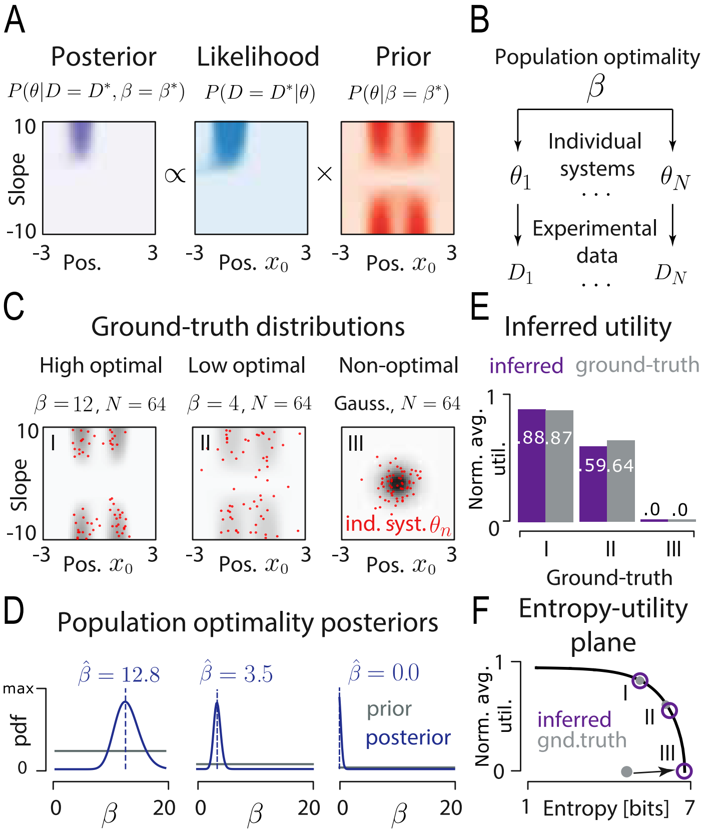

Figure 2.4: **Inference of the degree of optimality.** **(A)** Posterior over nonlinearity parameters, inferred for a single system with a utility-derived prior at fixed optimality parameter, $\beta = \beta^*$. **(B)** A hierarchical model of a population of optimized systems. Population optimality parameter β controls the distribution of parameters for individual systems $(n = 1, \ldots, N)$, θ_n, which give rise to observed data, $\mathcal{D}_n$. **(C)** Nonlinearity parameters (64 red dots per distribution) sampled from three different ground truth distributions (denoted by roman numerals in panels G-J): a strongly optimized population ($\beta = 12$; left), a weakly optimized population ($\beta = 4$; middle), a non-optimal distribution (Gaussian distribution; right). For each model neuron θ_n, data $\mathcal{D}_n$ consists of 100 stimulus-response pairs. **(D)** Results of hierarchical inference. Posteriors over β (purple lines) and MAP estimates, $\hat{\beta}$ (dashed purple lines) were obtained using simulated data from G. Priors (gray lines) were uniform on the $[0, 20]$ interval. **(E)** Normalized utility $\tilde{U}$. Estimated values (purple bars) closely match ground truth (gray bars). **(F)** Entropy and normalized utility of ground truth distributions (gray, filled circles) and inferred distributions parametrized by $\hat{\beta}$ (purple, empty circles).

for all neurons and use these data in a standard hierarchical Bayesian inference to compute posterior distributions over the population optimality parameter, β (Fig. 2.4D; see Methods).

Following hierarchical inference, we can interpret the inferred population optimality parameter $\hat{\beta}$, by mapping it onto normalized utility (cf. Fig. 2.2E). This reports optimality on a $[0, 1]$ scale, with 1 corresponding to the maximum achievable utility $U_{\max}$ and thus a fully optimal system, and 0 corresponding to the average utility under random parameter sampling, $\bar{U}(\beta = 0)$. Normalized utility for the three examples is shown in Fig. 2.4E.

Our framework enables us to draw inferences about optimality which are not possible otherwise. For example, in addition to estimating the normalized utility, we can also quantify how restrictive the optimization needs to be in order to achieve that level of utility. This restriction is measured by the entropy associated with $\hat{\beta}$ (Fig. 2.4F). In example I from Fig. 2.4C-E,

$\hat{\beta} = 12.8$ is associated with a decrease in entropy of about 1.75 bits compared to $\beta = 0$, meaning that nonlinearity parameters are effectively restricted to a fraction about $2^{-1.75} \approx 0.3$ of the parameter domain. Example III with $\hat{\beta} = 0$ is consistent with a high-entropy optimization prior and indicates almost no parameter space restriction. This is despite the fact that the actual parameters were sampled from a Gaussian highly concentrated (i.e., with low entropy) in the parameter space—but not in a region of high utility. This mismatch suggests that such a system could be optimized for a different utility function or shaped by other constraints. The system could also be anti-optimized, i.e. prefer negative values of $\tilde{U}$, which could easily be identified by permitting negative β values during inference. Another clear benefit of the probabilistic framework is the possibility of computing uncertainty estimates of β and the associated utility and entropy.

2.5 Question 3: Data resolves ambiguous theoretical predictions

Predictions derived from optimality theories can be non-unique and ambiguous. This ambiguity can manifest itself in different ways.

The *first kind of ambiguity* results from the existence of multiple maxima of the utility function. Before formulating statistical questions, it is important to pause and clarify the underlying biological context: Could different observed instances of the system freely sample from all utility maxima (as in Fig 2.4C, example I), or is a single maximum relevant, perhaps because it is the only one that nature realized by evolutionary adaptation? In the later case, the first task of statistical analysis is to identify that single maximum. For low-dimensional systems, this ambiguity can be resolved trivially: in our toy model, for example, a few data points suffice to zero in on one of the four degenerate utility maxima (Supplemental Methods Sec. A.2 and Fig. A.1). In contrast, in high-dimensional parameter spaces the task of finding the "closest optimum" is non-trivial (Doi et al., 2012) and could be aided by sampling methods derived from optimization priors, which is a topic for further research.

The *second kind of ambiguity* results from system parameters which enter the utility function, but are unconstrained by the optimization theory in question. Such parameters limit the performance of the whole system, with the utility typically achieving its global maximum when they take on extremal values (e.g., $\pm\infty$, 0, etc.); yet, these extremal values often correspond to physically implausible scenarios (infinite averaging time or energy consumption, zero noise, instantaneous response time, etc.). Optimization theory cannot make a non-trivial prediction about these parameters, so they must either be fixed *a priori* based on known external constraints, or inferred from data simultaneously with the optimization of the remaining parameters. An additional subtlety comes into play when we analyze multiple instances of a system (e.g., neurons): either each individual neuron has its own value of the constraint parameter, to be determined from data (which we address in the following paragraph), or all neurons share a single value of the constraint that needs to be inferred jointly.

In our model, the nonlinearity slope k is unconstrained by optimization: mutual information increases monotonically as $|k| \to \infty$ (Fig. 2.5A). This corresponds to vanishing noise in neural spiking. Since such noise cannot physically vanish, we must change the interpretation of the utility function, $U_{\mathrm{MI}}(\theta)$, and evaluate it only over positions x_0, while treating the slope k as a constraint to be fit from data—which we indicate by writing $U_{\mathrm{MI}}(x_0; k)$. Here, slope k determines the entire shape of the utility function (Fig. 2.5B). Unreliable neurons with a small slope have a unique optimal position $x_0 = 0$, while for neurons with large $|k|$ the utility is bimodal, with optimal positions separating peaks of the stimulus distribution. As before, we can infer both parameters for a "noisy" (Case I) and "precise" (Case II) simulated neuron (Fig. 2.5C); this time, however, the optimization prior acts only on x_0, while the prior over slope k remains uniform. To properly assess optimality, we must normalize the utility by the maximal utility achievable at the estimated value of k: $\tilde{U}(\hat{x}_0; \hat{k}) = (U(\hat{x}_0; \hat{k}) - \bar{U}(\beta = 0; \hat{k}))/((U_{\max}(x_0; \hat{k}) - \bar{U}(\beta = 0; \hat{k}))$. In both cases, the relative utility exceeds 0.9 (Fig. 2.5C). Because theoretical predictions now depend on the biophysical constraint—which itself is a free parameter adjustable separately for each system instance—high values of normalized utility can be achieved by neurons with very different x_0.

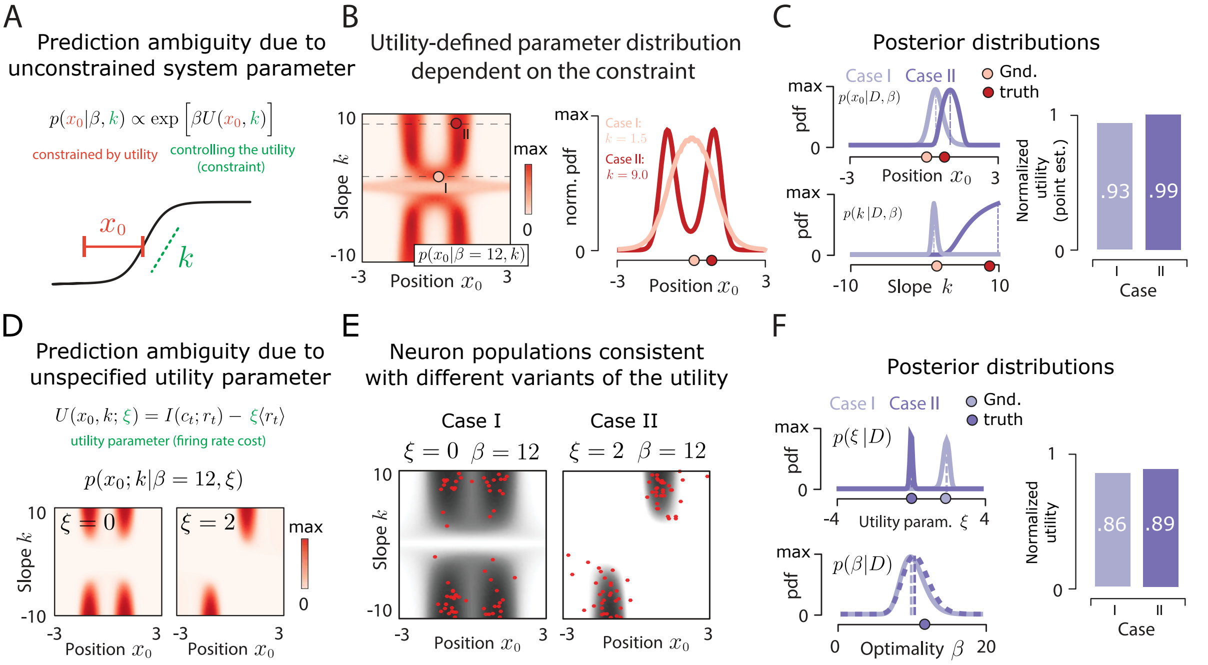

Figure 2.5: **Resolving ambiguities of theoretical predictions.** **(A)** Prediction ambiguity due an unconstrained system parameter. Utility is evaluated over the position parameter x_0 (red), with the slope parameter k (green) interpreted as an externally imposed biophysical constraint. k is inferred from data for each neuron separately; for different k, optimality may predict different optimal positions, x_0. **(B)** Optimization priors for x_0 are conditional maxent distributions over x_0 parametrized by values of k (rows of the matrix), here at fixed $\beta = 12$ (left). Distributions over x_0 for two example values of k (dashed black lines at left) are displayed in the right panel, with optimal x_0 values marked (pink and red circles for cases I and II, respectively). **(C)** Posteriors over the position (x_0, left column, top) and the slope (k, left column, bottom) parameters, estimated for cases I and II (light and dark purple lines, respectively; dashed lines – MAP estimates), by marginalizing the joint posterior. Ground-truth values are marked with circles. Normalized utility of x_0, relative to the maximal utility for k inferred separately for cases I and II. **(D)** Prediction ambiguity due to an unspecified utility function. Utility prefers high mutual information I at a low average firing rate $\langle r \rangle$, with an unknown trade-off parameter ξ. Optimization prior with no firing rate constraint (left, $\xi = 0$) shows four degenerate maxima; the constraint (right, $\xi = 2$) partially lifts the degeneracy. **(E)** Two ground truth distributions (gray) corresponding to different values of the firing rate constraint ξ. Red dots denote $N = 64$ sample neurons. **(F)** Posteriors over the firing rate constraint ξ (left column, top) and the optimality parameter β (left column, bottom), estimated for cases I and II (light and dark purple lines, respectively; dashed lines – MAP estimates), by marginalizing the joint posterior. Ground-truth values are marked with circles. Normalized utilities computed for ξ inferred separately for cases I and II.

The *third kind of ambiguity* arises when the utility function itself depends on additional parameters, ξ. The mutual information utility U_{MI} of our toy model can be extended by considering the cost of neural spiking, resulting in a new compound function, $U(x_0, k; \xi) = U_{\mathrm{MI}}(x_0, k) - \xi\langle r_t\rangle$, with the trade-off parameter ξ. Increasing ξ changes the shape of the new utility function (Fig. 2.5D). Given multiple instances of a biological system (Fig. 2.5E), we can ask about the most likely form of U (i.e., the single value of ξ shared across all instances of the system), together with the most likely value of the optimization parameter, β. Note that such joint determination of β and ξ corresponds to answering Question 2 (" Inferring the degree of optimality"), in the presence of ambiguity. This problem is solved by hyperparameter inference, which generates joint posteriors and MAP estimates of β and ξ (Fig. 2.5F). Here, too, the normalized utilities are defined relative to the inferred value of ξ and can thus be comparable even when the underlying utility functions are substantially different.

The difference between ambiguities of the second and third kind is subtle, yet important. Broadly speaking, the second kind of ambiguity arises if only a subset of system parameters θ depends on the optimality parameter β, while the remaining parameters act as constraints that have to be inferred. In the third kind of ambiguity *all* system parameters θ depend on the optimality parameter β as well as on additional parameters of the utility function ξ. The corresponding differences in parameter dependency patterns are summarized graphically in the Supplemental Information (Sec. A.2, Fig. A.2).

2.6 Question 4: Optimization priors improve inference for high-dimensional problems

Here, we extend our toy model neuron with 2 parameters to a more realistic case with hundreds of parameters. We focus on a Linear-Nonlinear-Poisson (LNP) model (Paninski et al., 2007), whose responses to natural image stimuli are determined by a linear filter (also referred to as a receptive field - RF) - $\phi \in \mathbb{R}^{16\times16}$ (Fig. 2.6A). The purpose of this exercise is to show the tractability of our approach and the power of optimization priors for high-dimensional inference problems. Inference of neural filters, ϕ, from data is a central data analysis challenge in sensory neuroscience, making our example practically relevant.

Experimentally observed filters ϕ in the visual cortex have been suggested to maximize the sparsity of responses s_t to natural stimuli (Olshausen and Field, 1996). A random variable is sparse when most of its mass is concentrated around 0 at fixed variance. These experimental observations have been reflected in the normative model of sparse coding, in which maximization of sparsity has been hypothesized to be beneficial for energy efficiency, flexibility of neural representations, and noise robustness (Hyvärinen et al., 2009; Olshausen and Field, 2004). Filters optimized for sparse utility $U_{\mathrm{SC}}(\phi)$ (see Methods) are oriented and localized in space and frequency (Fig. 2.6B, leftmost panel) and famously resemble RFs of simple cells in the primary visual cortex (V1). A significant fraction of neural RFs, however, differ from optimally sparse filters (Ringach, 2002), perhaps due to the existence of additional constraints. One possible constraint is spatial locality, which leads to suboptimally sparse filters that increasingly resemble localized blobs (Doi and Lewicki, 2014), as shown in Fig. 2.6B.

In our framework, sparse coding utility U_{SC} and locality U_{LO} combine into a single utility function with a parameter ξ that specifies the strength of the locality constraint. We wondered whether an optimization prior based on sparsity, even in the presence of an additional constraint of unknown strength, could successfully regularize the inference of linear filters, ϕ.

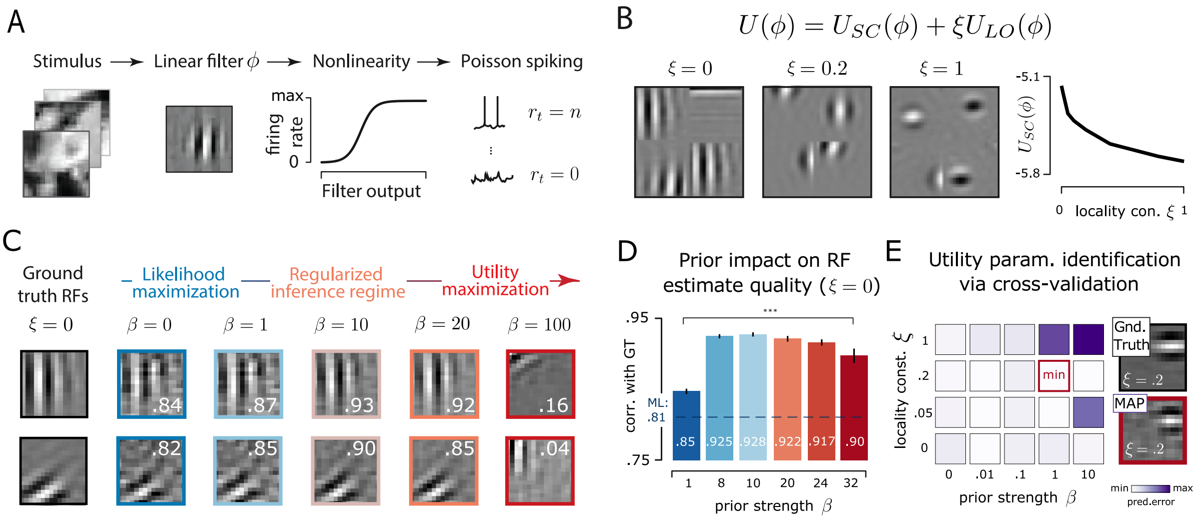

Figure 2.6: **Optimality priors improve inference of high-dimensional receptive fields. (A)** Linear-nonlinear-Poisson (LNP) neuron responding to 16×16 pixel natural image patches, x_t. Stimuli are projected onto a linear filter ϕ, which transforms them via logistic nonlinearity into average firing rate of Poisson spiking, r_t. **(B)** Receptive fields optimized for maximally sparse response to natural stimuli with a locality constraint ξ. First three panels on the left display 2×2 example filters optimized at increasing ξ. Rightmost panel shows the decrease in average sparse utility of filters with increasing ξ. **(C)** MAP estimates of two optimally sparse filters ($\xi = 0$) obtained with optimality prior of increasing strength β. White digits denote correlation with the corresponding ground truth. **(D)** Average correlations of $N = 100$ filter estimates with the ground truth as a function of prior strength β for locality constraint $\xi = 0$. Dashed blue line denotes the average correlation for ML estimates. MAP estimate correlations are significantly higher than ML estimate correlations (t-test; *** denotes $p < 0.001$). Error bars denote standard errors of the mean. **(E)** Identification of prior strength β and locality constraint ξ via cross-validation. Left panel, cross-validation errors in predicting withheld neural responses for a range of β and ξ values (heatmap). Parameter combination resulting in minimal error is marked with a red frame. Top right, a ground truth filter optimized with $\xi = 0.2$. Bottom right, MAP estimate of the filter, obtained with correctly identified values for β and ξ.

We first consider a scenario where the locality constraint is known *a priori* to equal zero. We simulate spike trains of 100 model neurons optimized under sparse utility U_{SC} responding to a sequence of 2000 natural image patches (see Methods for details). Using these simulated data we infer the filter estimates, $\hat{\phi}$, using Spike Triggered Average (STA) (Sharpee, 2013; Park and Pillow, 2017), which under our assumptions are equivalent to the maximum likelihood (ML) estimates (Paninski et al., 2007) (see Methods). STAs computed from limited data recover noisy estimates of neural filters (Fig. 2.6C; column second from the left).

Can sparse coding provide a powerful prior to aid inference of high-dimensional filters? Using our sparse coding utility, $U_{\text{SC}}(\phi)$, we formulate optimization priors for various values of β and compute maximum-a-posteriori (MAP) filter estimates $\hat{\phi}(\beta)$ from simulated data (Fig. 2.6C; four rightmost columns; see Methods for details). Increasing values of β interpolate between pure data-driven ML estimation (Fig. 2.6C, second column from the left) that ignores the utility, and pure utility maximization (Fig. 2.6C, right column) at very high $\beta = 10^2$ where the predicted filters become almost completely decoupled from data; these two regimes seem to be separated by a sharp transition. For intermediate $\beta = 1, 10, 20$, MAP filter estimates show a significant improvement in estimation performance relative to the ML estimate (Fig. 2.6D).

We next consider a scenario where the locality constraint is not known *a priori*, but can be identified together with the prior strength β using cross-validation (Kass et al., 2014), as described in Question 3. To this end, we simulate responses of a single neuron whose filter was optimized with the locality constraint $\xi = 0.2$ (Fig. 2.6E, "Gnd. Truth"). We then use a subset of 1800 out of 2000 stimulus-response pairs to compute the MAP estimate of the filter using a range of β and ξ values. Each MAP estimate of the filter is used to compute the prediction error for neural responses over withheld portion of the data. Cross-validation correctly identifies the true ξ and the optimal β values which minimize the prediction error (Fig. 2.6E); the resulting filter estimate (Fig. 2.6E, "MAP") closely resembles the ground truth.

Optimization priors achieve a boost in performance because they quantitatively encode many characteristics we ascribe to the observed receptive fields (localization in space and bandwidth, orientation), which the typical regularizing priors (e.g., L2 or L1 regularization of ϕ components) will fail to do. While hand-crafted priors designed for receptive field estimation can capture some of these characteristics (Park and Pillow, 2017; Savin and Tkacik, 2016), optimization priors grounded in the relevant normative theory represent the most succinct and complete way of summarizing our prior beliefs. For flexible optimization priors whose strength and additional parameters are set by cross-validation, one might expect that the postulated optimality theory need not be exactly correct to aid inference, so long as it captures *some* of the statistical regularities in the data.

2.7 Applications

2.7.1 Application 1: Receptive fields in the visual cortex

Here we analyze receptive fields of neurons in the primary visual cortex (V1) of the Macaque monkey (Ringach, 2002) (Fig. 2.7A). This system is a good test case, for which multiple candidate optimality theories were developed and tested against data (Olshausen and Field, 1996; Wiskott and Sejnowski, 2002; Hyvärinen et al., 2009; Van Hateren and van der Schaaf, 1998). As in the example of Fig. 2.6, we focus on sparse coding using utility U_{SC}, which prioritizes RFs localized in space and frequency (Fig. 2.7B; see Methods). An alternative utility prioritizing slow features is presented in Supplemental Information (Sec. A.3, Fig. A.3).

We first ask whether RFs of individual neurons support the optimality hypothesis, as in Question 1. Given the high-quality of RFs estimates, costly marginalization of the likelihood can be avoided and the utility of estimated RFs can be used directly as a test statistic. To construct the null distribution for the test, we sample 10^6 random filters consistent with optimization prior $P(\phi|\beta = 0)$, and declare the 95th percentile to be the optimality threshold (Fig. 2.7C). As expected, a large majority (204 neurons, green dots / example frame in Fig. 2.7C) of V1 neurons pass the optimality threshold, with 46 neurons failing the test (orange dots / example frame in Fig. 2.7C).

We next ask whether all RFs can be used together to quantify the degree of population optimality, as in Question 2. We estimate approximate posteriors over parameter β via rejection sampling (see Methods), using all RFs in the population (Fig. 2.7D, purple line). For comparison, we also compute posteriors using 250 utility-maximizing and 250 utility-minimizing filters (Fig. 2.7D, red and gray lines, respectively). MAP estimates of β obtained with simulated maximal and minimal utility RFs provide a reference for the interpretation of β estimated from real data. This estimate, $\hat{\beta}_{\mathrm{V1}}$, is very close to the parameter value of the optimally

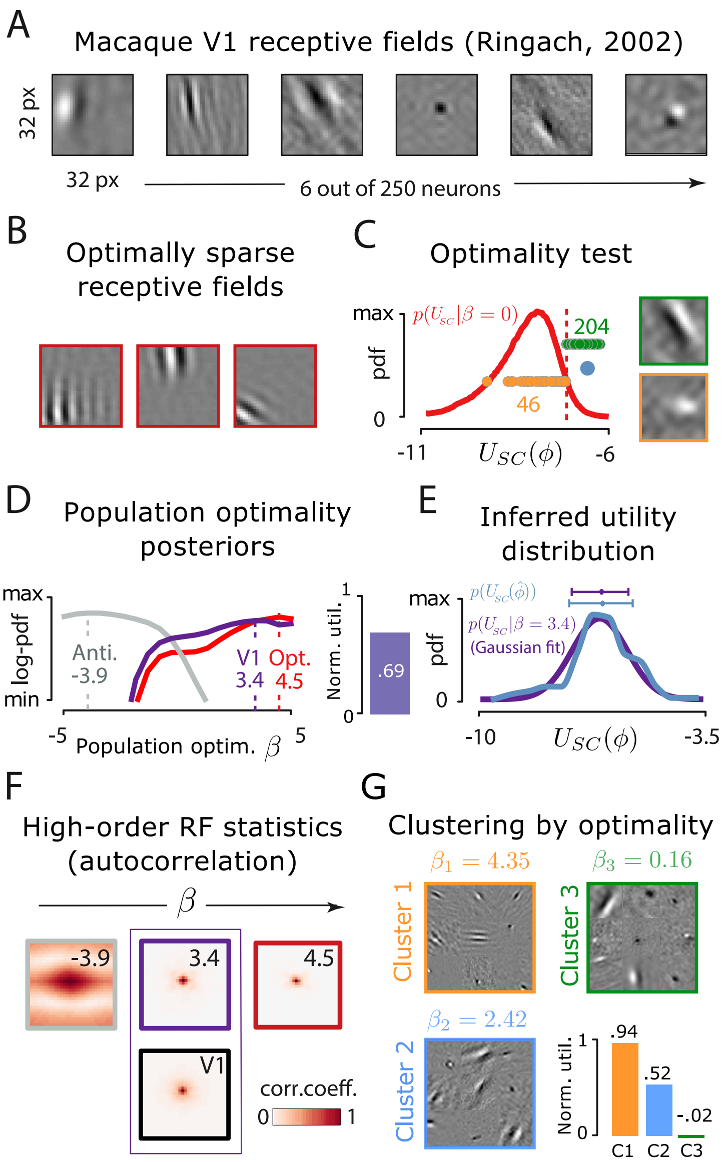

Figure 2.7: **Optimality of V1 receptive fields.** **(A)** Six example receptive fields (RFs) from Macaque visual cortex (courtesy of Dario Ringach (Ringach, 2002)). **(B)** Example simulated RFs optimized for sparsity. **(C)** Null distribution of utility values used to test for optimality under sparse utility and the 95-percentile significance threshold (red dashed line). Significant (green) and non-significant (orange) receptive fields denoted with dots (x axis is truncated for visualization purposes); example RFs are shown in frames of matching colors. Blue dot shows the average RF utility (99.6$^{\text{th}}$ percentile of the null distribution). **(D)** Approximate log-posteriors over population optimality parameter β derived from 250 RFs estimates (purple line), 250 maximum-utility filters (red line) and 250 minimal-utility filters (gray line). Dashed lines mark MAP estimates. **(E)** Empirical distribution of RF utilities (blue line) compared with utility distribution consistent with the inferred $\hat{\beta}_{\text{V1}}$ (purple line). **(F)** Spatial autocorrelation of RFs predicted for different β values (reported in top-right corner of each panel, cf. inferred values in D). Note a good match between data-derived RF autocorrelation (black frame) and the predicted autocorrelation at the inferred $\hat{\beta}_{\text{V1}}$ (purple frame). **(G)** Three clusters with different β, learned with a MaxEnt mixture-model. For each cluster, 3×3 sample receptive fields are displayed, together with the corresponding normalized utility values in the bottom-right panel.

sparse filters, implying high degree of optimization. The normalized utility is 0.69, implying a significant, yet not complete, degree of optimization.

Since population optimality β parametrizes the entire distribution of receptive fields, inferring β allows us to make predictions inaccessible by other means. For example, given the inferred degree of optimality, we predict the entire distribution of utility values (not only its mean) across neurons. In principle, the predicted distribution (or its higher-order moments, e.g., variance) could deviate from the empirically observed distribution, if the real system were adapted to a different utility or set of constraints. For V1 neurons, the predicted and empirical sparse utility distributions are very similar (Fig. 2.7E).

Another prediction concerns the correlation between system parameters, in our case, RF shapes. Different values of β predict very different spatial autocorrelation functions of RFs (Fig. 2.7F), with the prediction at inferred β resembling the data-derived autocorrelation better than the alternative or extremal β values. These examples demonstrate that once the single parameter β is inferred, the optimality framework makes quantitative, rigorous, and parameter-free predictions of non-trivial statistics that can be directly tested against data.

Our framework can also be used to dissect sources of deviation from optimality. We fit a mixture model, where each mixture component was parametrized by a separate value of β (Fig. 2.7G; see Methods). This procedure clusters the RFs into three groups spanning a broad range of utility values. The largest cluster (135 RFs) achieves a nearly maximal normalized utility of 0.94; neurons in this cluster all passed the significance test in Fig. 2.7C. The existence of second- and third-largest clusters (95 RFs, normalized utility of 0.52; 20 RFs, normalized utility ~ 0, respectively) suggests that these cells might be a subject to additional unknown constraints or might be optimizing a different utility. We emphasize that we analyze the optimality of *individual* neurons, whereas the optimization of complete populations could yield a more diverse set of RFs that are individually suboptimally sparse (Olshausen and Field, 1996; Zylberberg et al., 2011; Hyvärinen et al., 2009), accounting for the deviations we observe. Our analysis is intended as a demonstration of the applicability of our framework, rather than a definitive optimality claim about V1 neurons. Population-level analysis of optimality is a subject of future work.

2.7.2 Application 2: Receptive fields in the retina

Here we analyze temporal receptive fields of 117 retinal ganglion cells (RGCs) in the rat retina (Deny et al., 2017). Temporal RFs have a characteristic bimodal shape (Fig. 2.8A, left) which can be captured well by a simple filter model with three parameters (Sun et al., 2017). Two parameters (c_1, c_2) describe the amplitudes of both modes, while the third (a) determines the temporal scale of the filter (Fig. 2.8A, right panel). In what follows, we focus on the optimality of filter shapes in the space of these three parameters.

RGC receptive fields long have been hypothesized to instantiate predictive coding (PC) – a canonical example of a normative theory in sensory neuroscience (Srinivasan et al., 1982). Temporal PC postulates that, instead of tracking the exact stimulus value directly in their responses, neurons encode a difference between the stimulus and its linear prediction computed using past stimuli. Such a strategy has many potential benefits: it reduces the dynamic range of signals, minimizes use of metabolic resources, and can lead to efficient coding in the low noise limit, by performing stimulus decorrelation and response whitening (Srinivasan et al., 1982; Bialek, 2012; Dong and Atick, 1995; Van Hateren and van der Schaaf, 1998; Chalk et al., 2019).

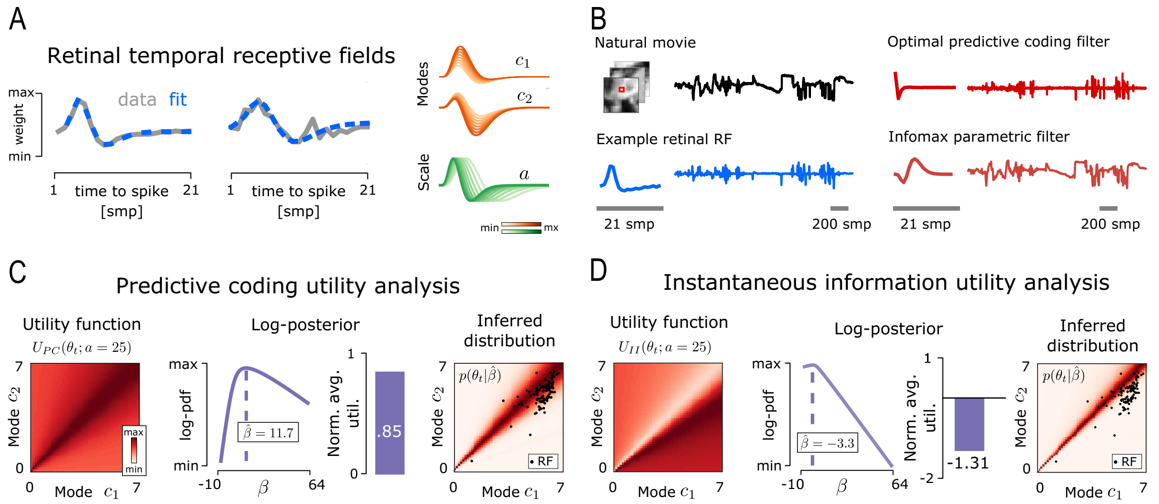

Figure 2.8: **Optimality of retinal receptive fields.** **(A)** Two example temporal receptive fields of rat retinal ganglion cells. Gray lines show RF estimates (courtesy of Olivier Marre (Deny et al., 2017)), dashed blue lines show parametric fits. Fit parameters correspond to amplitudes of filter modes (parameters c_1, c_2, orange) and scale (parameter a, green). **(B)** Example natural stimulus: light intensity of a single pixel of a natural movie (top-left, black). Representative retinal RF and its linear response to the natural stimulus (bottom-left, blue line). Optimal predictive coding filter and its response to the same stimulus (top-right, dark red line). Optimal instantaneous information transmission filter and its response (bottom-right, pink line). **(C)** Analysis of temporal RFs with the generalized predictive coding utility, U_{PC}. First panel: Utility function of filter modes c_1, c_2 constrained by timescale $a = 25$. Second panel: Log-posterior (solid purple line) over population optimality parameter β (dashed vertical line – MAP estimate). Third panel: Normalized utility of the RF population. Fourth panel: Optimization prior distribution over (c_1, c_2) at the inferred $\hat{\beta}$, marginalized over all values of the timescale parameter a (black dots – data-derived RFs). **(D)** Analysis of temporal RFs with the instantaneous information utility, U_{II}, analogous to C.

An optimal predictive coding filter must be adapted to the statistics of stimuli it encodes (Srinivasan et al., 1982). We optimize PC filters using natural light intensity time-courses (see Methods). Optimal PC filter responses qualitatively resemble the responses of a representative retinal filter convolved with the same natural stimulus (Fig. 2.8C). Both filters generate strong, spike-like transients to sudden changes in the stimulus mean, while their output remains close to 0 when the stimulus is not changing. This pattern is different from the response of a parametric bimodal filter (with parameters a, c_1, c_2) optimized to track the stimulus, obtained by maximizing instantaneous information transmission in a low-noise regime (U_{II}, see Methods). Importantly, predicted responses can be very distinct despite the qualitative similarity between retinal, PC, and instantaneous information filters.

To evaluate the optimality of retinal RFs, we propose a new utility, U_{PC}, that mathematically generalizes the canonical formulation of predictive coding (Srinivasan et al., 1982). This utility prioritizes filters which minimize power in their output, given a fixed filter norm, while allowing the filters to operate on timescales distinct from the stimulus frame rate (see Methods for details). We evaluate $U_{\mathrm{PC}}(\theta; a)$ as a function of the two filter mode parameters, $\theta = (c_1, c_2)$, but consider the timescale a to be an external constraint to be inferred from data for each neuron separately, as in Question 3. Parameter a is a constraint because, much like k in the toy neuron example of Fig. 2.2, its value is not set by optimality (which prefers $a \to 0$) but by biophysical constraints or by the temporal horizon at which prediction is of highest use to the organism. For a broad range of a values, U_{PC} is highest close to the diagonal of the (c_1, c_2) plane, representing nearly balanced filters, as shown in Fig. 2.8C (left).

We use all retinal RFs jointly to compute the posterior over the optimality parameter β (Fig. 2.8C, second panel). The inferred $\hat{\beta} \approx 11.7$ yields a normalized utility of 0.85, implying strong optimization for PC (Fig. 2.8C, third panel from the left); even relative to the non-parametric optimal PC filter with no timescale constraint, the utility of retinal filters remains as high as 0.74. The high degree of optimization is visually evident in the (c_1, c_2) plane, where individual neurons fall onto high utility regions of the maximum entropy distribution given inferred $\hat{\beta}$ and marginalized over timescale a (Fig. 2.8C, right). An analogous analysis performed using maximization of instantaneous information U_{II} (see Methods) reveals a negative β estimate and thus anti-optimization for this alternative utility, with real neurons avoiding high-utility regions of the maximum entropy distribution.

2.7.3 Application 3: Neural wiring in *C. elegans*

Here we analyze neural wiring in *C. elegans*, which has been the subject of several normative studies (Chen et al., 2006; Chklovskii, 2004; Pérez-Escudero et al., 2009; Pérez-Escudero and de Polavieja, 2007). Relative positions of neurons could be partially predicted by minimizing the total wiring cost under the constraint that muscles and sensors need to be properly connected (Chen et al., 2006; Pérez-Escudero and de Polavieja, 2007). Instead of trying to predict individual neuron positions, we ask a different question: Are the measured neuron positions optimized to minimize the wiring cost to muscles and sensors?

For each neuron i, the wiring cost is determined by the number of muscles it connects to, the distance between the neuron's position, x_i, and positions of muscles, $m_{i,j}$, and the number of synapses formed by each connection, $n_{i,j}$ (Fig. 2.9A). The resulting utility function for each neuron can be written as $U_{\mathrm{WC}}(x_i; m_i, n_i, \xi) = -\sum_{j=1}^{N_i} n_{i,j} |x_i - m_{i,j}|^{\xi}$, where N_i is the number of muscles the neuron i connects with, and ξ is an exponent determining the form of the utility as a function of distance (Chen et al., 2006) (Fig. 2.9B). The precise value of ξ is not specified by the theory and thus needs to be inferred from data, following the *ambiguity of the third kind* scenario (cf. Question 3).

Our analysis shows that a large proportion of 126 neurons that form connections with muscles align closely with the maxima of the utility function (Fig. 2.9B, left panel). We estimate the joint posterior distribution over the optimality parameter β and the connection exponent ξ, for neuron-muscle and neuron-sensor connections separately (Figs. 2.9B,C, middle panels). In both cases, the normalized utility exceeds 90 $\%$, implying strong optimization. Interestingly, the estimates for the exponent ξ are relatively high: 1.6 for neuron-muscle connections and 1.9 for neuron-sensor connections, suggesting that neurons are penalized only weakly for small deviations from optimal positions but much stronger for large deviations. This is in contrast to previously published analysis that focused instead on neuron-neuron connections (Pérez-Escudero et al., 2009), where the authors find (and we confirm) $\xi \approx 0.5$. This implies a cost that is considerable for short connections and grows only slowly with distance, a pattern consistent with the clustering of neurons within ganglia and the nerve ring. The lower exponent for neuron-neuron connections could be related to their anatomy. In *C. elegans*, input and output synapses can be located on the same neurite (Donato et al., 2019)). This might reduce the cost associated with connections between distant neurons, e.g. because some computation

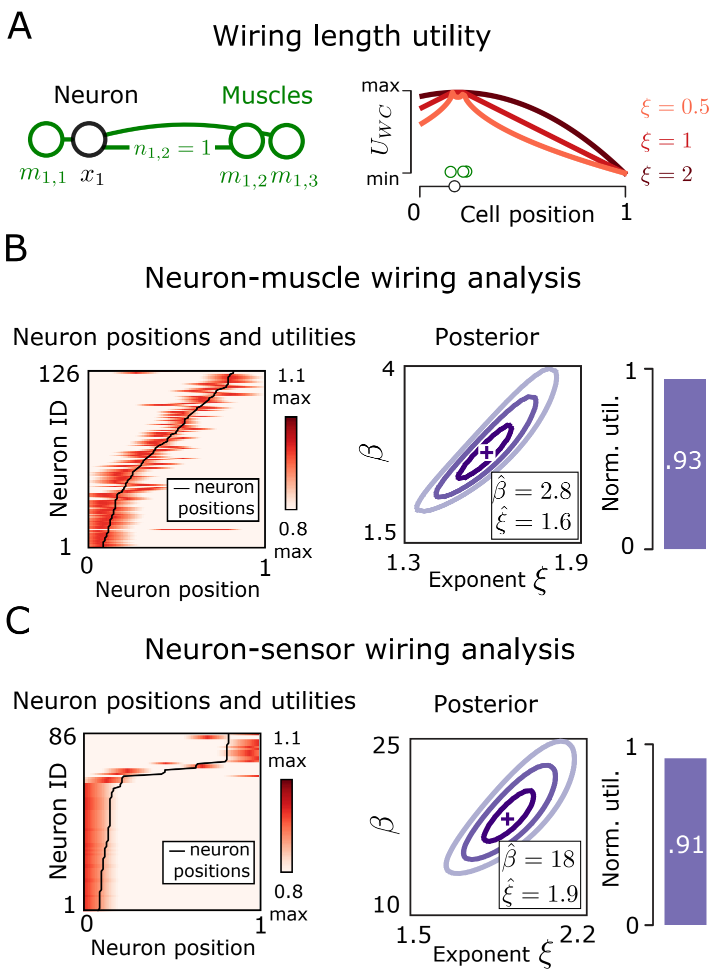

Figure 2.9: **Optimality of neural wiring in *C. elegans*.** **(A)** Left panel: Connection schematic between example neuron at position x_1 (black circle) and three muscles at positions $m_{1,1}, m_{1,2}, m_{1,3}$ (green circles). Number of synapses between neuron x_1 and muscle $m_{1,j}$ is denoted $n_{1,j}$. The example neuron forms monosynaptic connections (green lines) only with the three muscles. Right panel: wiring cost utility, $U_{\text{WC}}(x_1; \xi)$, as a function of position x_1, corresponding to the scenario depicted at left. Position axis spans the entire *C. elegans* body length. Utility functions are shown for three exponent values ξ. **(B)** Neuron-muscle connection analysis. Left panel: Utility $U_{\text{WC}}(x; \xi = 2)$ (red, scaled to $[0, 1]$ for each neuron) for all 126 neurons (rows), as a function of neuron positions $x \in [0, 1]$. Black line denotes positions of real neurons. Middle panel: joint posterior over optimality parameter β and the exponent ξ (cross denotes MAP estimates reported in the legend). Right panel: normalized utility of neuron-muscle connectivity. **(C)** Neuron-sensor connection analysis, analogous to B.

could be performed locally near clusters of synapses away from the cell bodies Ruach et al. (2023).

2.8 Properties and extensions of the framework

Here we discuss some advantages and limitations of using optimization priors, as well as extensions that address them.

2.8.1 Effects of parametrization

Related to Sec. Bayesian inference and optimization priors.

Biological systems can often be described using different sets of parameters. One set of parameters can be converted into another using a mathematical transformation, such as the log-transform (replacing the value of slope k by $\log k$) or the reciprocal (replacing k by $1/k$). The choice of parametrization is largely a matter of convenience or convention. However, when working with probability distribution functions such as the optimization priors (Eq. 2), special care needs to be taken, since the functional form of the priors is intertwined with the choice of parametrization. In this section we highlight the subtleties involved. In 2.8.3 we introduce a generalized form of the optimization priors, that allows us to clearly separate the choice of the parametrization from the choice of the normative prior distribution.

Maximum entropy priors depend on parametrization. Suppose that a researcher, Alice, defines the family of optimization priors as introduced in Eq. 2,

$$P_A(\theta|\beta) \propto e^{\beta U(\theta)}. \tag{2.3}$$

Suppose that Alice then decides to use a different set of parameters, $\phi = f(\theta)$, to perform some computational task. The prior Eq. (2.3), as function of ϕ, can be obtained by the standard method of changing variables,

$$P_A(\phi|\beta) \propto e^{\beta U(f^{-1}(\phi))} |\det J(\phi)|, \tag{2.4}$$

where $J_{ij}(\theta) = \frac{\partial f^{-1}(\phi)_i}{\partial \phi_j}$ is the Jacobian. Some parameter regions may appear expanded or shrunk in the new parametrization; the Jacobian corrects for this and makes sure that the underlying probability distribution does not change. A change of variables like this can be done whenever it is convenient.

Notice, however, that the distribution $P_A(\phi|\beta)$ in Eq. (2.4) is different from what another researcher, Bob, who has been using the parametrization ϕ from the beginning, has obtained from Eq. 2 directly,

$$P_B(\phi|\beta) \propto e^{\beta U(f^{-1}(\phi))}. \tag{2.5}$$

Namely, Bob's distribution $P_B(\phi|\beta)$ does not include the Jacobian that is present in $P_A(\phi|\beta)$, since Bob did not perform a change of variables from θ. This means that Alice and Bob are using different optimization priors and will get different results downstream in the analysis.

This is particularly clear for $\beta = 0$, i.e. with zero optimization. Bob's prior $P_B(\phi|\beta)$ is then uniform in ϕ; but Alice's prior $P_A(\phi|\beta) \propto |\det J(\phi)|$ is uniform in θ, but non-uniform in ϕ (unless θ and ϕ are related by a linear transformation, in which case the Jacobian $|\det J(\phi)|$ is constant).

Illustration with the toy model. Fig. 2.10 shows three different parametrizations of our toy model,

$$\text{Alice:} \quad \theta = (x_0, k)\,, \tag{2.6}$$

$$\text{Bob:} \quad \phi = \left(x_0^{1/3}, k^{1/3}\right), \tag{2.7}$$

$$\text{Carol:} \quad \psi = \left(\frac{1}{2}(1+\mathrm{erf}(\frac{x_0}{\sqrt{2}\sigma_{x_0}})), \frac{1}{2}(1+\mathrm{erf}(\frac{k}{\sqrt{2}\sigma_k}))\right). \tag{2.8}$$

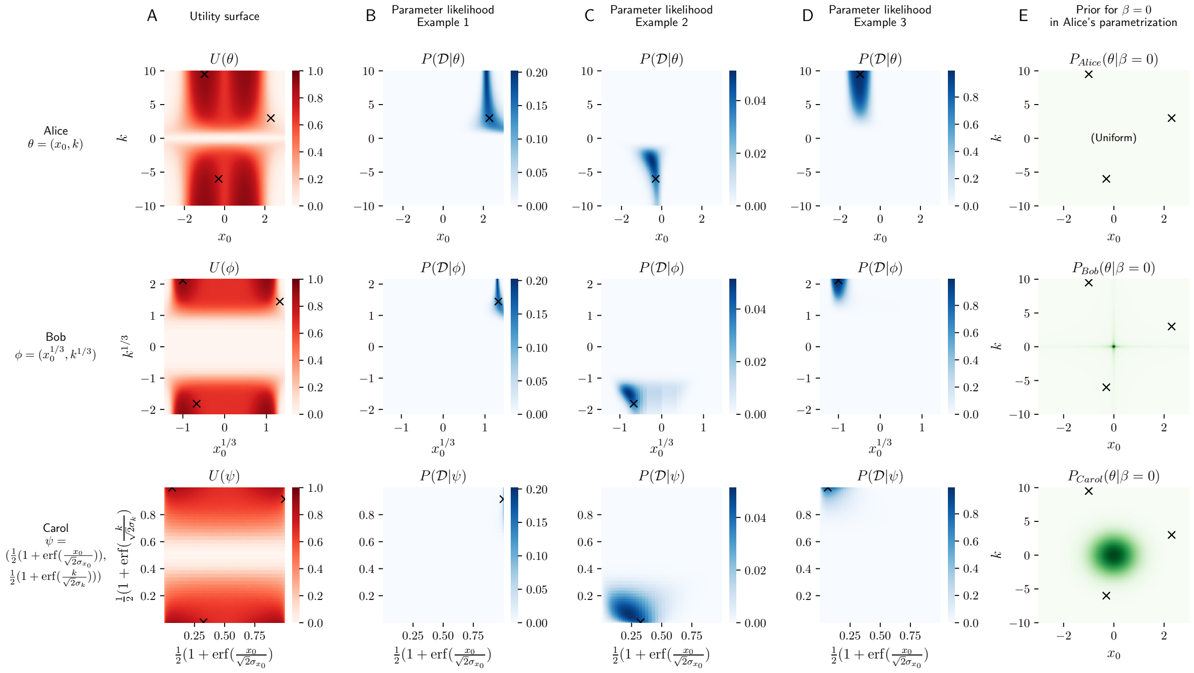

Figure 2.10: **An illustration of three different parametrizations of the toy model,** related to Sec. Bayesian inference and optimization priors and Fig. 2B,D. Alice, Bob and Carol use θ, ϕ and ψ – shown in top, middle and bottom row respectively. Column **A**: Utility plotted in the different parametrizations. Columns **B**-**D**: Likelihood from the three examples used in the main text Fig. 3B, here plotted in 3 different parametrizations. Column **E**: Optimization priors under $\beta = 0$ that Alice, Bob and Carol are effectively using when they applied Eq. 2 using their choice of parametrization. For comparison, all are plotted in Alice's parametrization θ.

The ψ parametrization corresponds to the Gaussian distribution that we used as a non-optimised example in the paper, Fig. 3G (III).

In Fig. 2.10, column A shows the utility surface in the three different parametrizations. Compared to Alice's parametrization θ (top row; this is the parametrization also used throughout the paper), the parameter regions around $(x_0, k) = (0, 0)$ are inflated in Bob's and Carol's parametrizations ϕ and ψ (middle and bottom rows). Columns B-D show the likelihood surfaces, which show the same distortion between the parametrizations. Column E in Fig. 2.10 shows the optimization priors of Alice, Bob and Carol under no optimization, $\beta = 0$, transformed into Alice's parameters θ. Alice has defined her prior family according to Eq. 2 using θ and it is therefore uniform for $\beta = 0$. Bob and Carol have defined their priors to be uniform ϕ and ψ, and they are therefore not uniform in θ. Specifically, the region around $(x_0, k) = (0, 0)$ which was inflated in ϕ and ψ has higher probability density when "shrunk" back to θ. We later refer to the $\beta = 0$ optimization prior, $P(\theta|\beta = 0)$, as the *null model.*

Since Alice, Bob and Carol are effectively using different priors, they obtain different results in downstream analyses. To demonstrate this, Fig. 2.11 shows the likelihood of β for the three example systems (subplots) and three parametrizations (blue, orange, green). The differences between parametrizations are mostly qualitative for examples 1 and 3, which are on the two extremes – non-optimised and highly optimised ($\hat{\beta}_{ML} = 0$ and ∞, regardless of parametrization).

In the intermediate example 2, the ML estimates vary based on parametrization. For example,

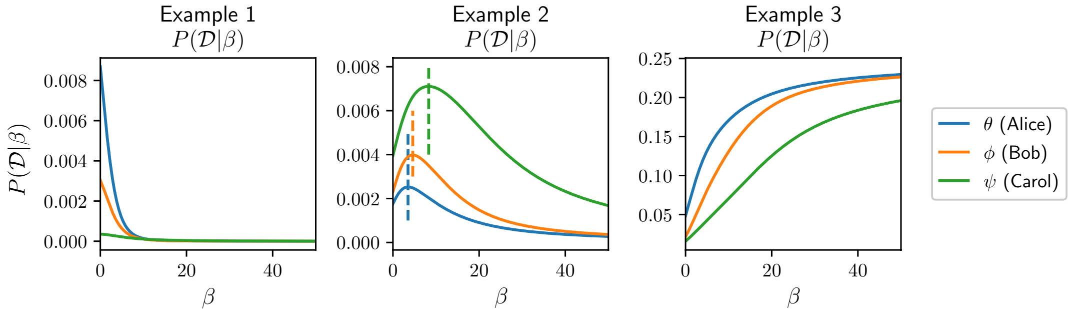

Figure 2.11: **Likelihood of β for three example systems and three parametrizations θ, ϕ and ψ,** related to Sec. Bayesian inference and optimization priors and Fig. 3C. **Example 1** (see likelihood in Fig. 2.10B) is not optimised and ML β is always 0, but the curve differs between parametrizations. Similarly, **Example 3** (likelihood in Fig. 2.10D) is strongly optimised and ML β is always ∞. The intermediate **Example 2** (likelihood in Fig. 2.10C) has finite ML estimate of β, which depends on parametrization (dashed vertical lines). The values are $\hat{\beta}_{ML,\theta} = 3.5$ $\hat{\beta}_{ML,\phi} = 4.6$ and $\hat{\beta}_{ML,\psi} = 8.3$.

ψ squeezes the utility peaks into the corners (Fig. 2.10A, bottom). This leads to $\hat{\beta}_{ML,\psi} = 8.3$, higher than $\hat{\beta}_{ML,\theta} = 3.5$ for the original parametrization, where the peaks are more spread out.

These differences raise the question of which parametrization leads to correct results. First, we stress that the answer is problem-specific and cannot be addressed in general. One should consider whether a prior uniform in the chosen parameters under zero optimization ($\beta = 0$) is appropriate.

Second, whenever this is desirable, the choice of parameters can be decoupled from the prior distribution. A change in variables can be performed after the optimization priors are defined. Alternatively, a null distribution under zero optimization can be specified explicitly – this is discussed in the following section.

2.8.2 Statistical test of optimality with additional utility parameters

Related to Question 1: Statistical test for the optimality hypothesis and Fig. 3A-D

Suppose we have a utility function $U(\theta; \xi)$ and a likelihood function $\mathcal{L}(\theta) = P(\mathcal{D}|\theta)$, which is evaluated using some experimental dataset $\mathcal{D}$. The optimality theory, and the associated question of evidence for it, may be stated in one of two forms:

(a) The system parameters have been optimized for high $U(\theta; \xi)$ for a specific value of ξ^* which we assume to know. Can we reject the optimization prior $P(\theta|\beta = 0, \xi = \xi^*)$ with in favour of $P(\theta|\beta, \xi = \xi^*)$ with some positive $\beta > 0$?

In this case, ξ is fixed and the framework described in the main text applies. Note that the null hypothesis does not in fact depend on ξ^*, since utility plays no role under $\beta = 0$, and we can write $P(\theta|\beta = 0, \xi = \xi^*) = P(\theta|\beta = 0)$.

(b) The system parameters have been optimized for high $U(\theta; \xi)$ for *some unknown* value of ξ. Can we reject the optimization prior $P(\theta|\beta = 0)$ in favour of $P(\theta|\beta, \xi)$ with some positive $\beta > 0$ and some value of ξ?

> This can be approached as a multiple hypothesis testing problem. We perform the likelihood ratio test for a range of values of ξ, obtaining a p-value for each, $p(\xi)$. This allows us to detect and compare evidence for optimality under different assumptions about the utility function, parametrized by ξ. A multiple testing correction is necessary to manage the risk of false positives.

Example. We illustrate how this can be done within our toy model. The utility $U(x_0, k; \xi) = U_{MI}(x_0, k) - \xi \langle r \rangle$ combines the mutual information utility with the cost of spiking, parametrized by ξ – as introduced in the main text, Question 3. The utility is plotted for three values of ξ in Fig. 2.12A.

Fig. 2.12B-E shows the statistical test of optimality for two example systems. Likelihood functions of the neuron parameters $P(\mathcal{D}|\theta)$ are displayed in Fig. 2.12B. For system I (Fig. 2.12B, top panel), the likelihood is concentrated near one of the mutual information peaks, and for system II (Fig. 2.12B, bottom panel) it is concentrated in the areas corresponding to low spiking rate. Fig. 2.12C displays the corresponding marginal likelihood $P(\mathcal{D}|\beta, \xi) = \int d\theta P(\mathcal{D}|\theta) P(\theta|\beta, \xi)$. Given a sample of stimulus-response pairs, system I is likely to be generated under a zero or moderate cost of spiking ξ. System II is likely under high ξ that prioritizes low spiking rate.

The likelihood ratio test can be performed for any value of ξ; we show the results for $\xi = 0$ in Fig. 2.12D (which is equivalent to the scenario described in the main text, Question 1). According to the test, and under the 0.05 significance threshold, system I is significantly optimal ($p = 0.013$). System II, however, is not ($p = 1$). The null distribution of the likelihood ratio statistic λ is based on 10,000 data sets sampled according to $P(\mathcal{D}|\beta = 0)$. The likelihood ratio test is performed across a range of $\xi \in (-4, 4)$, each value yielding a p-value $p(\xi)$. Note that the null distribution of likelihood ratios under each ξ can be estimated using the same set of samples.

Performing multiple hypothesis tests needs to be balanced by using a more stringent p-value threshold, if we are to maintain a low probability of falsely claiming optimality. We aim to keep the probability of one or more false positives – the family-wise error rate (FWER) below $\alpha = 0.05$. Common generic methods such as the Bonferroni correction, which sets the p-value threshold at α/n with n being the number of tests, can be too conservative, since tests performed for similar values of ξ are correlated – especially if ξ takes values from a dense grid.

To choose the appropriate p-value threshold, we compute the minimal p-value across ξ, $p_{min} = \min_\xi p(\xi)$. FWER is the probability that p_{min} sampled under $\beta = 0$ falls below the threshold, which means that the threshold can be chosen as the α-quantile of the null distribution of p_{min}. This distribution, computed from 10,000 samples, and the threshold corresponding to $\alpha = 0.05$, is shown in Fig. 2.12E (the transformation $-\log_{10} p_{min}$ is used for visualization purposes). A system with p_{min} below this threshold is declared optimal at significance level α. Fig. 2.12E shows that both example systems pass this test at $\alpha = 0.05$, despite being optimal for different variants of the utility function corresponding to different values of the spiking cost parameter ξ.

2.8.3 Optimization priors with general null models

Related to Sec. Bayesian inference and optimization priors.

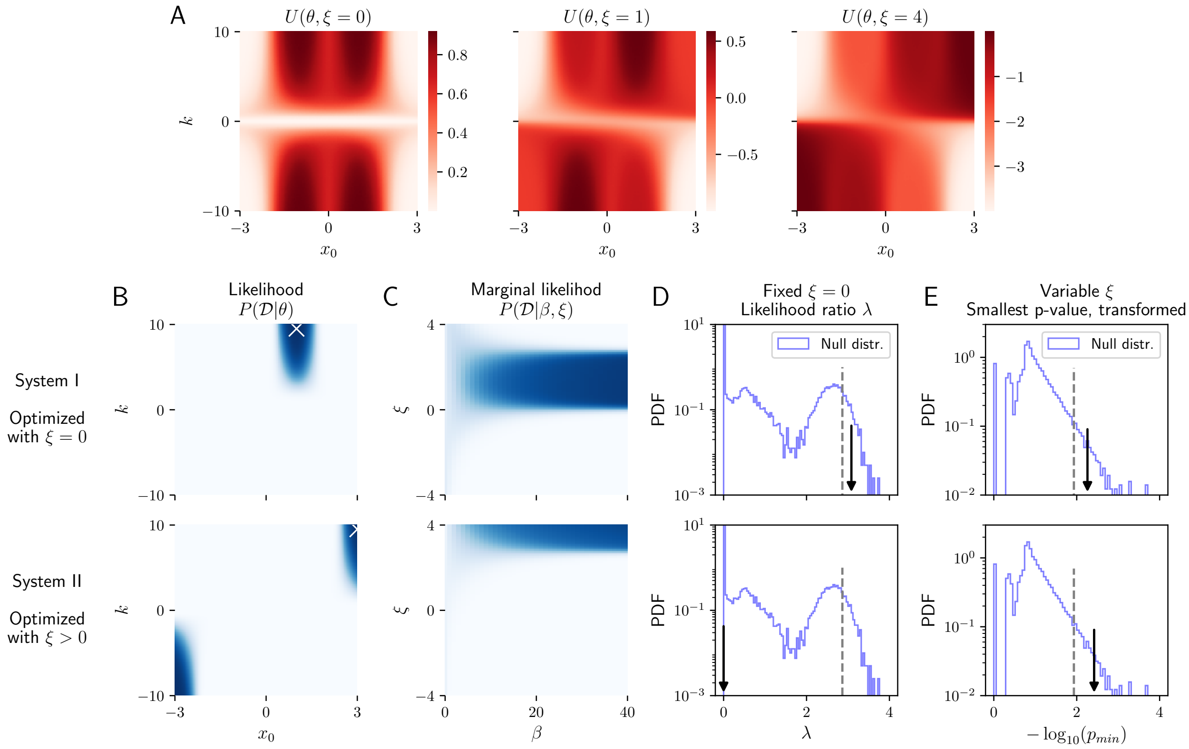

Figure 2.12: **Statistical test of optimality with unknown utility parameter ξ,** related to Question 1: Statistical test for the optimality hypothesis and Fig. 3A-D. **A**: Toy model utility function $U(x_0, k; \xi) = U_{MI}(x_0, k) - \xi\langle r \rangle$ of position x_0 and slope k for three different values of spiking cost – parameter ξ. **B**: Likelihood of the nonlinearity parameters obtained from 20 stimulus–response pairs. The ground truth parameters shown as white crosses. **C**: Marginal likelihood of the optimality parameter β and utility parameter ξ. **D**: Likelihood ratio test of optimality for a fixed $\xi = 0$. The null distribution is based on 10,000 samples of $\theta = (x_0, k)$ from the uniform distribution $P(\theta|\beta = 0)$ and dataset $\mathcal{D}$ according to $P(\mathcal{D}|\theta)$. The black arrows indicate the likelihood ratio statistic for the two example systems. Only the system I passes the test at significance level 0.05 (to the right of dashed vertical line). **E**: Test of optimality under variable ξ. The likelihood ratio test was performed across a range of $\xi \in (-4, 4)$; plotted on the horizontal axis is the smallest p-value $p_{min} = \min_\xi p(\xi)$, transformed as $-\log_{10} p_{min}$ for better visualization at low p_{min}. Dashed vertical line shows the significance threshold $\alpha = 0.05$, which lies at $p_{min} \approx 0.012$ – this serves as the multiple testing correction guaranteed to keep FWER below α. Both systems pass this threshold (black arrows) but they achieve this with different values of utility parameter ξ (for system I, p_{min} is achieved at $\xi \approx 0.7$; for system II at $\xi \approx 4$).

The maximum entropy optimization priors from Eq. 2 are uniform for $\beta = 0$ (we refer to the $P(\theta|\beta = 0)$ as the *null model*). However, finding a parametrization where this is appropriate (see 2.8.1) can be difficult. For example, how to parametrize receptive fields such that a reasonable null model is uniform in some domain? The choice of parameters can also be dictated by convenience or convention. For such cases and in general, we can decouple the choice of parametrization from the choice of the null model.

Let the null model – the parameter distribution without optimization – be $q(\theta)$. Then the normative prior family

$$P(\theta|\beta) = \frac{1}{Z(\beta)} q(\theta)\, e^{\beta U(\theta)} \tag{2.9}$$

is maximum entropy in the sense that is solves the optimization problem

$$\begin{aligned} P(\theta|\beta) &= \arg\max_{p(\theta)} \Big(-D_{KL}(p||q) + \beta\, \mathbb{E}_{\theta\sim p(\theta)}\, U(\theta) \Big) && (2.10) \\ &= \arg\max_{p(\theta)} \int p(\theta) \left(-\log \frac{p(\theta)}{q(\theta)} + \beta\, U(\theta) \right) d\theta, && (2.11) \end{aligned}$$

where D_{KL} is the Kullback-Leibler divergence or relative entropy. Intuitively, $P(\theta|\beta)$ is as similar to $q(\theta)$ as possible, while constraining average utility. If $q(\theta)$ is uniform in some domain, we recover the maximum entropy solution from Eq. 2.

When changing parametrization, q will transform accordingly. E.g. Alice and Bob might choose two null models $q_A(\theta)$ and $q_B(\phi)$, each using their preferred parametrization. They are equivalent (Alice and Bob will get the same results) if

$$q_A(\theta)\, |\det J(\theta)| = q_B(f(\theta)). \tag{2.12}$$

How to choose the right parametrization/null model? In general, the choice is problem-specific and beyond the scope of our paper. Note that even when $q(\theta)$ is uniform on some domain, the domain needs to be chosen and affects the results.

In some biological systems, an established null model may be available. Distributions of the form Eq. (2.9) have a history in population genetics (Wright, 1937; Barton and Coe, 2009) of describing the equilibrium distributions of allele frequencies; the functional form emerges from a stochastic model of evolution. The null model $q(\theta)$ is then the equilibrium distribution under mutation and random drift; natural selection enters through the factor $e^{\beta U(\theta)}$. The parameter β quantifies the strength of natural selection in favour of U relative to random drift.

In other cases, the null model may simply reflect our limited knowledge about the system. As a half-joke example, a simple null model for positions of neurons along the AP axis in *C. elegans* might be uniform, ignoring the intricacies of its body shape. The situation is different for an animal like the brontosaurus, since "All brontosauruses are thin at one end, much, much thicker in the middle, and then thin again at the far end" (Elk, 1972). More elaborate null models might take into account cell lineages, or assume that neurons are allowed to permute their positions, but not occupy new places (making θ discrete).

In the example of visual receptive fields, it may be worth considering if properties such as smoothness should be included in the null model q or in the utility part of the prior $e^{\beta U(\theta)}$. If smoothness is present without optimization, then for the purposes of testing and quantifying optimality, it should be included in q. For Bayesian inference this might not matter, since performance is more important than the justification of the prior/null model.

2.8.4 Interpreting the magnitude of optimization parameter β

Related to Question 2: Inferring the degree of optimality.

The optimization parameter β enters the optimization priors in a product with utility, $e^{\beta U(\theta)}$. This means that the magnitude of β cannot be directly compared between different biological systems, or different utility functions considered in the same system. However, there are several ways to interpret the size of β.

Probability of observing a parameter combination. As follows from the functional form of the optimization priors, each additional unit of utility U makes a parameter combination θ more probable by a factor e^{β}. One can ask how much of a probability increase is then conferred by a biologically significant increase in U.

Normalized utility. Each value of β has an associated average utility

$$\bar{U}(\beta) = \int P(\theta|\beta)U(\theta)d\theta \tag{2.13}$$

achieved by the optimization prior. Normalized utility,

$$\tilde{U} = \frac{\bar{U}(\beta) - \bar{U}(\beta = 0)}{U_{\max} - \bar{U}(\beta = 0)} \tag{2.14}$$

is plotted throughout the paper. This converts β to the expected, biologically meaningful increase in utility, and where it falls on the scale from not optimized at all (at $\beta = 0$) and maximum possible utility ($U_{\max}$). Negative values for normalized utility are possible for cases of "anti-optimization", when the inferred $\beta < 0$.

Note that normalized utility is a distributional property, since it is defined as an average value of utility over an optimization prior at an inferred value of β. It is also possible to evaluate and normalize the point estimate of utility at a particular value of (e.g. inferred) parameters θ, $\frac{U(\hat{\theta})-\bar{U}(\beta=0)}{U_{\max}-\bar{U}(\beta=0)}$. This may be more appropriate if we only deal with data for a single instance of a system, which may not permit a reliable estimate of β.

The mapping between β and $\bar{U}(\beta)$ is not trivial and depends on the distribution of U under no optimization. A direct calculation shows that the average utility grows with its variance,

$$\frac{d\bar{U}(\beta)}{d\beta} = \overline{U^2}(\beta) - \bar{U}^2(\beta) = \mathrm{Var}\, U(\beta). \tag{2.15}$$

In particular, close to $\beta = 0$, $\bar{U}(\beta)$ grows linearly with slope $\mathrm{Var}\, U(\beta = 0)$. This can be expected intuitively – if the available parameter combinations provide a large variety of utility values, even small β (weak selection, see below) can induce a large change in $\bar{U}$.

If the distribution of utility under $\beta = 0$ is Gaussian, this linear growth continues indefinitely, otherwise the growth is nonlinear and depends on higher moments of that distribution.

Alternative normalization of utility functions. Based on the above arguments, it would be also possible to "standardize" the raw utility values by subtracting the average utility at $\beta = 0$ and dividing by $\mathrm{Std}\, U(\beta = 0)$. Such standardized utility would then enter the optimization prior and inference. In this representation, inferred β values would be directly interpretable: an increase of 1 for the value of β would lead to an approximate increase of 1 in the standardized utility; while zero, as before, would correspond to the expected utility with no optimization. The advantage of this method is that $U_{\max}$ need not be known in advance (which could be intractable to find exactly in high-dimensional spaces); the drawback is that one needs to estimate $\bar{U}(\beta = 0)$ and $\operatorname{Std} U(\beta = 0)$ by Monte Carlo sampling to standardize the utility *before* doing any inference. We did not use this alternative normalization in the paper.

Strength of natural selection. A factor analogous to $e^{\beta U(\theta)}$ appears in the equilibrium distribution of allele frequencies (Wright, 1937; Barton and Coe, 2009) in population genetics, if log fitness if proportional to U. The parameter β corresponds to the log fitness advantage per unit utility, multiplied by the effective population size. Fitness is the expected number of offspring in the next generation; population size enters because selection is more efficient in larger populations, where stochastic effects (random drift) are weaker. This connection is developed in more detail in Chapter 4.

2.8.5 Trade-offs between utility and likelihood

Related to Sec. Bayesian inference and optimization priors.

The normative priors defined in Eq. (2), and the optimization parameter β, are used in a Bayesian framework to interpolate between predictions from normative theories and inferences from data. Along this interpolation path, various trade-offs between theory and data are made. Here we focus on the maximum posterior parameter estimates, and show that they achieve the best possible utility-likelihood trade off. This provides an additional justification for the maximum entropy form of the optimization priors, Eq. (2).

Consider the the maximum posterior (MP) estimates of θ, as a function of the optimization parameter β,

$$\hat{\theta}_{MP}(\beta) = \arg\max_{\theta} \frac{e^{\beta U(\theta)} P(\mathcal{D}|\theta)}{Z(\beta)}. \tag{2.16}$$

As β increases from 0 to infinity, $\hat{\theta}_{MP}(\beta)$ goes from the maximum likelihood estimate $\hat{\theta}_{ML}$ to the maximum utility prediction, θ^{*}.

The maximum posterior trajectory $\hat{\theta}_{MP}(\beta)$, parametrized by β, achieves the best possible trade-off between the normative theory and data in the following sense: it finds the parameter combinations with the highest utility given each value of likelihood. Or conversely, it finds the most likely parameter combination, for a given each value of utility. This can be shown by rewriting the MP formula[1] as

$$\hat{\theta}_{MP}(\beta) = \arg\max_{\theta} \left(\log P(\mathcal{D}|\theta) + \beta U(\theta)\right) \tag{2.17}$$

$$= \arg\max_{\theta} \left(U(\theta) + \frac{1}{\beta}\log P(\mathcal{D}|\theta)\right). \tag{2.18}$$

Here, β takes the role of a Lagrange multiplier constraining $U(\theta)$ when maximizing $\log P(\mathcal{D}|\theta)$; or equivalently, $1/\beta$ constrains $\log P(\mathcal{D}|\theta)$ when maximizing $U(\theta)$.

[1] We could also include a non-uniform null distribution $q(\theta)$ as discussed in 2.8.3, and the MP formula would be $\hat{\theta}_{MP}(\beta) = \arg\max_{\theta} \left(\log q(\theta) + \log P(\mathcal{D}|\theta) + \beta U(\theta)\right)$

We compare the utility-likelihood trade-off achieved achieved by maximum posterior estimates based on the maximum entropy optimization priors with two alternative methods of interpolation between theory and data.

Alternative interpolation method: Linear interpolation between maximum likelihood and maximum utility parameters In this approach we reduce both the normative theory and the data to points in the parameter space, maximum utility θ^* and maximum likelihood $\hat{\theta}_{ML}$, and interpolate between them, using

$$\theta(\gamma) = \gamma\theta^* + (1-\gamma)\hat{\theta}_{ML} \tag{2.19}$$

with higher values of parameter $\gamma \in (0,1)$ giving more importance to the normative theory. This path can be compared to the maximum posterior trajectory Eq. (2.16). While the linear interpolation follows a straight line (cyan), the MP estimates follow a curved trajectory (purple). While on many situations the two trajectories can be similar, sometimes there is a substantial difference in the trade-off, see Fig. 2.13A-C. In addition to the better trade-off, maximum entropy optimization priors also yield posterior distributions over parameters, Fig. 2.13D, which can be used for more refined analyses.

Alternative interpolation method: Interpolation using a Gaussian prior. Keeping the Bayesian approach, one could choose a family of optimization priors other than maximum entropy (MaxEnt). Here we consider Gaussian priors around a utility peak, with two diffferent levels of detail. The comparison with MaxEnt is in Fig. 2.14.

The simplest option seems to be a Gaussian prior centered at the utility peak,

$$P_{G1}(\theta|\beta) \propto e^{-\frac{\beta}{2}(\theta-\theta^*)^T C^{-1}(\theta-\theta^*)}. \tag{2.20}$$

As an example using our toy model, we chose the covariance matrix C naively such that the Gaussian is roughly symmetric in the plots, corresponding to a visual "distance from the optimum".

A more elaborate approach is to take into account the shape of the peak – the rate at which utility decreases in different directions. This is similar to the approach of Pérez-Escudero et al. (Pérez-Escudero et al., 2009). We can fit the utility function around the peak with a quadratic function, and take the exponential to obtain a Gaussian,

$$P_{G2}(\theta|\beta) \propto e^{\beta\tilde{U}(\theta)}; \qquad \tilde{U}(\theta) = a + \sum_i b_i(\theta-\theta^*)_i + \sum_{ij} c_{ij}(\theta-\theta^*)_i(\theta-\theta^*)_j. \tag{2.21}$$

The linear coefficients b_i can be set to zero if the peak is inside the parameter domain – in our case the peak lies at the boundary, and utility has nonzero gradient there, and hence b_i will be nonzero. The coefficients a, b, c can be obtained by Taylor expansion if a formula for $U(\theta)$ is available; in our case U is computed numerically – so we fit a, b, c by minimizing mean square error in the vicinity of the peak.

Both types of Gaussian priors are parametrized by β, analogously to MaxEnt. Example priors for $\beta = 10$ are shown in Fig. 2.14A-C. To compare them in terms of the utility-likelihood trade-off, we compute the MP trajectory $\hat{\theta}_{MP}(\beta)$ for increasing β as in the previous section. The trajectories are shown on top of each prior and on top of the likelihood surface, Fig. 2.14D. The panel 2.14E shows the utility-likelihood trade-off.

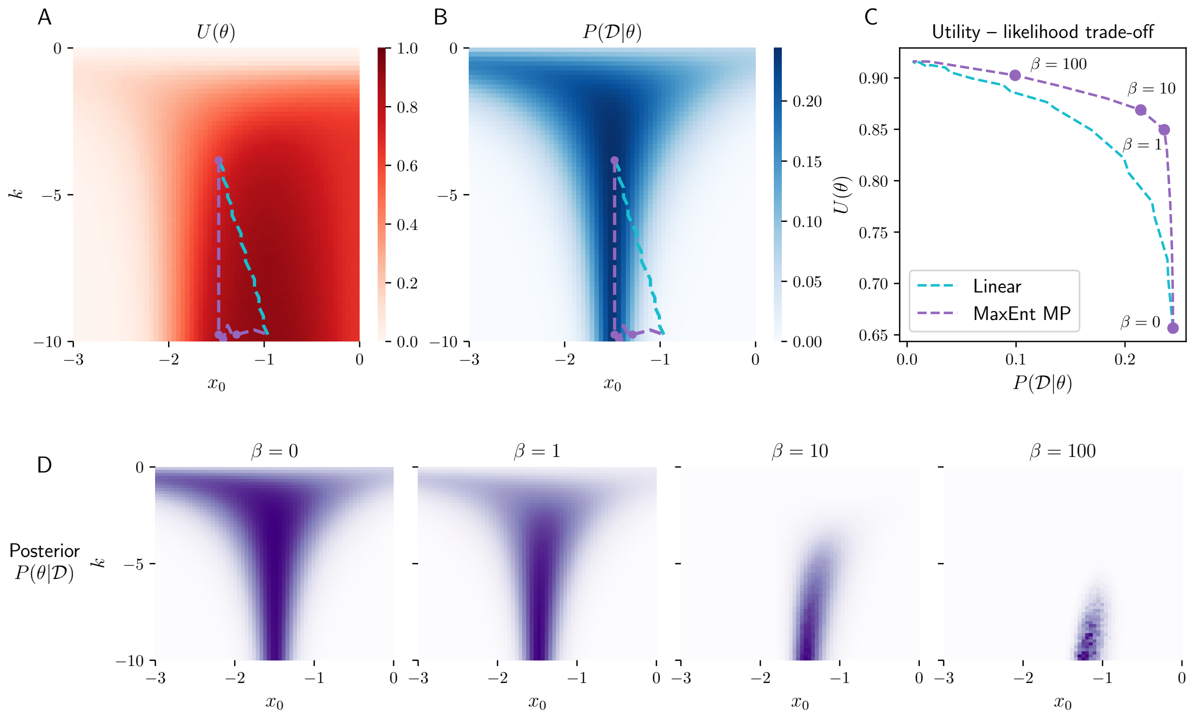

Figure 2.13: **Comparison with linear interpolation between data and a normative theory in the parameter space,** related to Sec. Bayesian inference and optimization priors. Linear interpolation takes a straight trajectory from maximum likelihood $\hat{\theta}_{ML}$ to maximum utility θ^*, cyan dashed lines. Maximum posterior (MP) interpolates between the same two, but takes a curved trajectory, purple dashed lines. **A**-**B**: Interpolation trajectories are shown on top of a utility heatmap (A) and a likelihood heatmap (B). **C**: MP achieves a better trade-off between likelihood and utility along the way. **D**: Posterior heatmaps, for the 4 values of β highlighted as purple points in the upper plots. Notes: The trajectories are bumpy, because parameters are rounded to the nearest vertex in a 128 by 128 grid, and the utility values are computed with noise (mutual information estimates from 200,000 samples). Data was chosen to get a marked difference between linear interpolation and maximum posterior trajectories. Likelihood is based on stimuli $x = \{-1.5, -1.49, 0, 2\}$ and responses $r = \{0, 1, 0, 0\}$. For randomly generated data, the two trajectories are often similar in the utility-likelihood plot (even if they differ visibly in the x_0, k plot).

The trade-off is poor for the naive Gaussian $P_{G1}(\theta|\beta)$. The fitted Gaussian prior $P_{G2}(\theta|\beta)$ and the full MaxEnt prior achieve similar trade-offs for high β (i.e. near the peak), but $P_{G2}(\theta|\beta)$ underperforms MaxEnt for lower β. As argued in the previous section, the trade-off is optimal for MaxEnt. However, the fitted Gaussian prior may often serve as a convenient approximation to reduce computational costs, especially if the utility has a unique maximum.

Discussion. This optimal trade-off achieved by the maximum entropy optimization priors is due to its sensitivity to the full shape of the likelihood and utility function – not only their peaks. In addition to the specific shape of the trajectory, this means that if the utility has multiple peaks, they are naturally taken into account.

The linear and Gaussian require specifying the “correct” peak, which can be difficult in high-dimensional systems. Consider the RF inference in Fig. 6C in the main text. Small values of β improve inference by pulling towards the nearest local maximum and gradually increasing utility. Too large β can lead the inference towards a different maximum with higher utility,

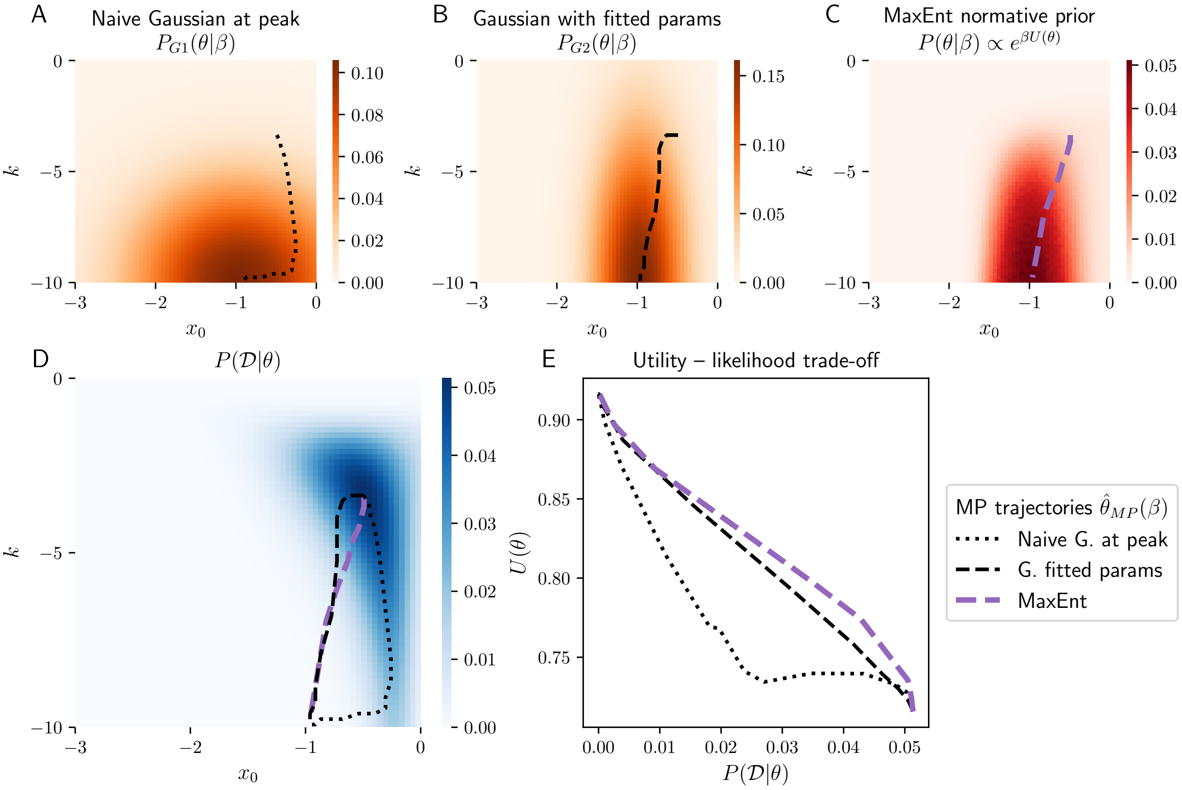

Figure 2.14: **Comparison with interpolation using Gaussian priors,** related to Sec. Bayesian inference and optimization priors. **A-C:** Alternative normative priors with $\beta = 10$. A: Naive Gaussian $P_{G1}(\theta|\beta)$, see Eq. Eq. (2.20). Covariance is $C = ((10^2, 0), (0, 3^2))$ such that the Gaussian is visually symmetric. B: Fitted Gaussian $P_{G2}(\theta|\beta)$, see Eq. Eq. (2.21). Parameters are $a = 0.916$, $b = (-0.021, 0.212)$, $c = ((0.003, -0.030), (-0.030, -0.603))$ obtained from a least squares fit around the peak. C: MaxEnt prior $P(\theta|\beta) \propto e^{\beta U(\theta)}$. **D**: Likelihood surface used for further analysis. Generated from an intermediately optimised ground truth parameters and 20 spikes; same as in main text Fig. 3 likelihood 2. The trajectories (also shown in top panels) correspond to the maximum posterior $\hat{\theta}_{MP}(\beta)$ with varying β in the three priors above. **E**: Utility-likelihood trade off. MaxEnt performs best, and $P_{G2}(\theta|\beta)$ is an approximation to it that performs similarly for high β (high utility region of the plot). The naive Gaussian $P_{G2}(\theta|\beta)$ performs poorly, since it does not correspond to the shape of the utility peak.

which is however inconsistent with data. Linear interpolation towards this "wrong" maximum would yield meaningless results.

Except in situations when the data points to a point near a unique utility maximum, we need to take into account the full shape of the utility function (and of the likelihood function).

2.9 Discussion

Despite their theoretical appeal, the application of optimization principles to biological systems has been hindered by statistical issues that grow more pressing as the complexity and dimensionality of the models increases. These issues are not new. Instead of developing an *ad hoc* solution whenever called for by a particular application, we decided to tackle these issues head on and flesh them out with simple examples. For instance, the issue of an unconstrained optimization parameter or a trade-off with unknown strength is well-known to the practitioners, but is often solved “by hand”: one manually adjusts the constraint until the optimality predictions are (visually) consistent with data. Such manual “fine-tuning” of constraints is clearly problematic from the statistical viewpoint, as it could easily amount to (over-)fitting that is not controlled for. In contrast, our framework performs inference and optimization jointly and provides a full posterior over constrained and unconstrained parameters alike. Another problematic issue arises from degenerate maxima of the utility functions. A frequent solution has been to postulate further constraints within the theory itself, which disambiguate the predictions (Doi et al., 2012). Our framework proposes a complementary mechanism: using a small amount of data to localize the theoretical predictions to the relevant optimum, against which further statistical tests can be carried out. As a last example, when fitting complex (e.g., nonlinear dynamical systems) models one typically restricts parameters by hand to a domain that is thought to be “biologically relevant.” In contrast, optimization priors automatically suppress vast swaths of parameter space that lead to non-functioning systems, even if these systems are not fully optimized for the postulated utility. In this way, the statistical power of the data can be used with maximum effect in the parameter regime that is of actual biological relevance, without sacrificing statistical rigor.

The ability to exclude biologically irrelevant regions of the parameter space highlights a general advantage of optimality priors over simple, unstructured distributions. Frequently applied “regularization priors” which penalize the norm of parameter values (e.g. Laplace or Gaussian (Park and Pillow, 2017; Sharpee, 2013)) assign highest probability when all parameters are equal to 0. Moreover, these priors are isotropic – they act with the same strength on each parameter and do not take into account interactions between them – which is an essential (and nontrivial) property of real systems. Together, these two requirements enforced by the prior are often contradictory to the notion of a functioning biological system. For example, penalizing parameter magnitudes while inferring the shape of nonlinearity in our toy-model neuron would bias the inferences towards completely non-functional solutions (slope and offset equal to 0). Intuitively, the robustness against overfitting afforded by the regularization prior thus comes at a cost of biasing inferences away from functional solutions. Our approach, in contrast, attempts to avoid such a disastrous trade-off by incorporating knowledge about biological function directly into the structure of the prior.

While our framework provides a principled way to navigate a number of statistical issues in complex biological systems, important questions remain. A key challenge is to identify the relevant optimization criterion for a biological system, and to express it in terms of experimentally measurable quantities. A candidate utility function which embodies an optimality criterion of interest could be selected from a possible discrete set of such functions (Wang et al., 2016; Młynarski and Hermundstad, 2018; Chalk et al., 2019), or by inferring utility function parameters. Because we leverage the well-understood machinery of Bayesian inference, one could perform model selection for the utility function that best explains the data. Such an approach could be used, for example, to rigorously verify whether entire neural populations in the visual cortex are jointly optimized for sparsity or a different utility, such as slowness (Wiskott and Sejnowski, 2002). An important caveat is that the more flexible our choice of the utility function becomes, the easier it is to claim an optimality for a system of interest. In principle, one could postulate a utility function with a fully unconstrained shape: in this limit, our framework would automatically recover the utility function shape from data (if these were sufficient) assuming the observed system is optimal, in a way reminiscent of inverse reinforcement learning (Chalk et al., 2019). This connection is an interesting topic for further research. In this paper, however, we focused on optimization theories where the number of adjustable utility parameters is smaller than the number of system parameters being predicted.

Our framework dovetails with other approaches which address the issues of ambiguity of theoretical predictions and model identifiability given limited data in biology. "Sloppy-modelling" (O'Leary et al., 2015; Gutenkunst et al., 2007), grounded in dynamical systems theory, characterizes the dimensions of the parameter space which yield qualitatively similar behavior of the system. In our framework, these dimensions correspond to regions of the parameter space of equal or similar utility. Another important conceptual advance grounded in statistical inference has been the usage of limited data to coarse-grain probabilistic models (Bialek et al., 1996; Chen et al., 2018; Machta et al., 2013). In our framework, a related coarse-graining occurs when, instead of inferring all system parameters from data directly, optimization sets the values of most of these parameters, leaving only the unconstrained subset to be fitted. The resulting dimensionality reduction could be sizable (e.g., with optimization predicting high-dimensional RF shapes given inferred firing rate, locality, or neural noise constraints), and could efficiently parametrize neuronal heterogeneity in terms of a small number of constraints that vary from neuron to neuron or between neural populations. Another point of connection with recent work concerns the ability to instantiate high-dimensional maximum entropy distributions over parameters with complicated dependency structures (De Martino et al., 2018; Bittner et al., 2019; Lueckmann et al., 2017). Such computational innovations will be essential for statistical analyses of optimality that require sampling from maximum-entropy optimization priors.

Acknowledgements The authors thank Dario Ringach for providing V1 receptive fields and Olivier Marre for providing retinal receptive fields. WM was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 754411. MH was funded in part by Human Frontiers Science Grant HFSP RGP0032/2018.

CHAPTER 3

Accumulation and maintenance of genetic information

Abstract. Selection accumulates information in the genome — it guides stochastically evolving populations towards states (genotype frequencies) that would be unlikely under neutrality. This can be quantified as the Kullback-Leibler divergence (KL divergence) between the actual distribution of genotype frequencies and the corresponding neutral distribution. First, we show that this population-level information sets an upper bound on the information at the level of genotype and phenotype, limiting how precisely they can be specified by selection. Next, we study how the accumulation and maintenance of information is limited by the cost of selection, measured as the genetic load or the relative fitness variance, both of which we connect to the control-theoretic KL cost of control. The information accumulation rate is upper bounded by the population size times the cost of selection. This bound is very general, and applies across models (Wright-Fisher, Moran, diffusion) and to arbitrary forms of selection, mutation and recombination. Finally, the cost of maintaining information depends on how it is encoded: specifying a single allele out of two is expensive, but one bit encoded among many weakly specified loci (as in a polygenic trait) is cheap.

3.1 Introduction

Throughout evolution, selection accumulates information in the genome. It guides evolving populations towards fitter phenotypes, genotypes and genotype frequencies, which would be highly unlikely to arise by chance. This information – the degree to which selection can control the stochastic process of evolution – has been a long-standing subject of research (Kimura, 1961; Eigen, 1971; Worden, 1995; MacKay, 2003a; Watkins, 2008; Peck and Waxman, 2010; Barton, 2017), and relates to basic questions in evolutionary biology and genetics.

How well can selection specify the genotype and the phenotype? The degree to which within- and between-species genetic variation are shaped by selection has been the subject of the neutralist-selectionist debate (Kimura, 1968; Hey, 1999; Kern and Hahn, 2018; Jensen et al., 2019). Today, we know that much of the human genome is involved in various biochemical processes (The ENCODE Project Consortium et al., 2012; Kellis et al., 2014a), but this does not mean that it is strongly shaped by selection (Doolittle, 2013; Graur et al., 2013; Brunet and Doolittle, 2014). Here we ask a related question in information-theoretic terms: how much information can selection accumulate and maintain in the genome? Much of the sequence is to some degree random, and given its size $l \approx 3 \times 10^9$ base pairs, it likely contains far less information than the maximum conceivable 6×10^9 bits of information. A similar question has been raised in the context of origin of life: given high mutation rates, how much information could be maintained in the genome of early organisms (Eigen, 1971)?

Analogous questions can be asked about the phenotype. How many traits can selection optimize? It is easy to list a large number of potentially relevant traits: take the expression of all genes in all cell types and conditions, or regulatory interactions between pairs of genes. For a fit organism, these traits need to be specified with some precision, and this precision is likely limited (even if it is to some degree facilitated by correlations among traits). For example, a study of selective constraint on human gene expression (Glassberg et al., 2019) gave evidence of constraint, but overall, this seems weak. Given the large number of possibly important phenotypes, how precisely can selection specify them?

Quantifying genetic information. An established method in bioinformatics quantifies the information content of a short genomic motif, such as a binding site, by comparing an alignment of its instances across the genome to the genomic background (Schneider et al., 1986; Wasserman and Sandelin, 2004). Our definition of genetic information is mathematically similar, but aims to apply more generally (to large regions without multiple instances available). It is therefore based in theoretical population genetics rather than sequence data analysis. A key related concept is the repeatability of evolution (Lobkovsky and Koonin, 2012; Lässig et al., 2017). Evolution is stochastic due to genetic drift and mutation, but selection can reduce the space of possible outcomes. For example, suppose that in a sequence of length l, n sites are under strong selection for specific nucleotides. By fixing those nucleotides, selection will accumulate $2n$ bits of information. Meanwhile, the remaining $l-n$ sites will be occupied by random nucleotides, and if a replicate population evolves under identical conditions, the $l-n$ nucleotides will likely be different. Therefore our concept of information in a sequence is inversely related to how differently it could have evolved under identical conditions.

In general, however, the information content of the genome cannot be quantified by simply counting the sites that are under selection. A single bit of information can be spread across many loci under weak selection – a phenomenon particularly relevant when selection acts on polygenic traits, long recognized in quantitative genetics and described by the infinitesimal model (Fisher, 1918; Barton et al., 2017). Polygenicity and weak selection also resolve the apparent contradiction between the variety of phenotypes, or biochemical processes involving the DNA, and the lack of strong selective constraint on all of them. Selection might act on a small number of high level traits, which are influenced by large numbers of loci spread across the genome (described by the omnigenic model (Boyle et al., 2017)), which experience only weak selection individually.

In Section 3.2, we define information on three levels – the population state (genotype frequencies), the genotype, and the phenotype. There are simple inequalities between the three levels. This means that the upper bound on information accumulation rate, which we prove at the population level, also implies a bound at the genotype and phenotype levels. We use the KL divergence, a central quantity in information theory (Cover and Thomas, 2006) to quantify the difference between their actual distribution and their corresponding neutral distribution.

Notably, the neutral phenotype distribution corresponds approximately to the phenotype distribution among random DNA sequences. Recent work with random mutant libraries suggests that for some phenotypes, this distribution is accessible experimentally (gene expression driven by random promoters (de Boer et al., 2020; Vaishnav et al., 2022; Lagator et al., 2022) or enhancers (Fuqua et al., 2020)). Any departure from this neutral distribution amounts to accumulation of information.

Cost of information. After defining what genetic information means, we ask how quickly it can accumulate and how much of it can be maintained. We look for answers in terms of the cost of selection – the amount of relative fitness variation in a population. This cost, traditionally measured as the relative fitness variance or the genetic load, is itself limited. In a population with constant size, relative fitness is proportional to the expected number of offspring, and the number of offspring can only vary between zero and the reproductive capacity of the organism.

We rely on an information-theoretic measure of cost of selection, which is itself upper bounded by the relative fitness variance and genetic load, but has favorable mathematical properties. It relates the cost of selection to the KL cost of control (Todorov, 2006; Theodorou, 2015; Nourmohammad and Eksin, 2021), or the thermodynamic power (Pavlichin et al., 2019).

The relationship between information accumulation rate and the cost of selection has been studied by Kimura (Kimura, 1961) and later Worden (Worden, 1995), MacKay (MacKay, 2003a) and Barton (Barton, 2017). In Sec. 3.3 we discuss these works in more detail and present a new, more general bound. The problem of maintenance has been studied by Eigen (Eigen, 1971), Watkins (Watkins, 2008) and Peck and Waxman (Peck and Waxman, 2010). We discuss these in Sec. 3.4 and present example calculations that suggest general trends in the amount of information that can be maintained per unit cost.

3.2 Quantifying genetic information

The measures of information studied in this paper are based on comparisons between the distributions of various variables under selection versus neutrality. The focus on probability distributions accounts for the stochasticity of evolution, and the difference between the distributions with and without selection corresponds to the control that selection exerts on evolution. We quantify this difference in bits, using the KL divergence (Cover and Thomas, 2006)

$$D(U) = \sum_u \psi^U(u) \log_2 \frac{\psi^U(u)}{\varphi^U(u)} \tag{3.1}$$

where U is a variable that takes values u with probabilities $\psi^U(u)$ with selection and $\varphi^U(u)$ under neutrality. Below we focus on three variables – genotype frequencies (which describe population states), genotypes and phenotypes.

For a pair of variables U, V, statistical dependencies are reflected in their joint and conditional KL divergence, $D(U, V)$ and $D(U|V)$ (see SI Sec. B.1 for the definitions). Both are nonnegative quantities, and they follow the chain rule

$$D(U, V) = D(U) + D(V|U) = D(V) + D(U|V). \tag{3.2}$$

The chain rule allows a comparison of the effects of selection on different variables, as well as on the same variable at different times.

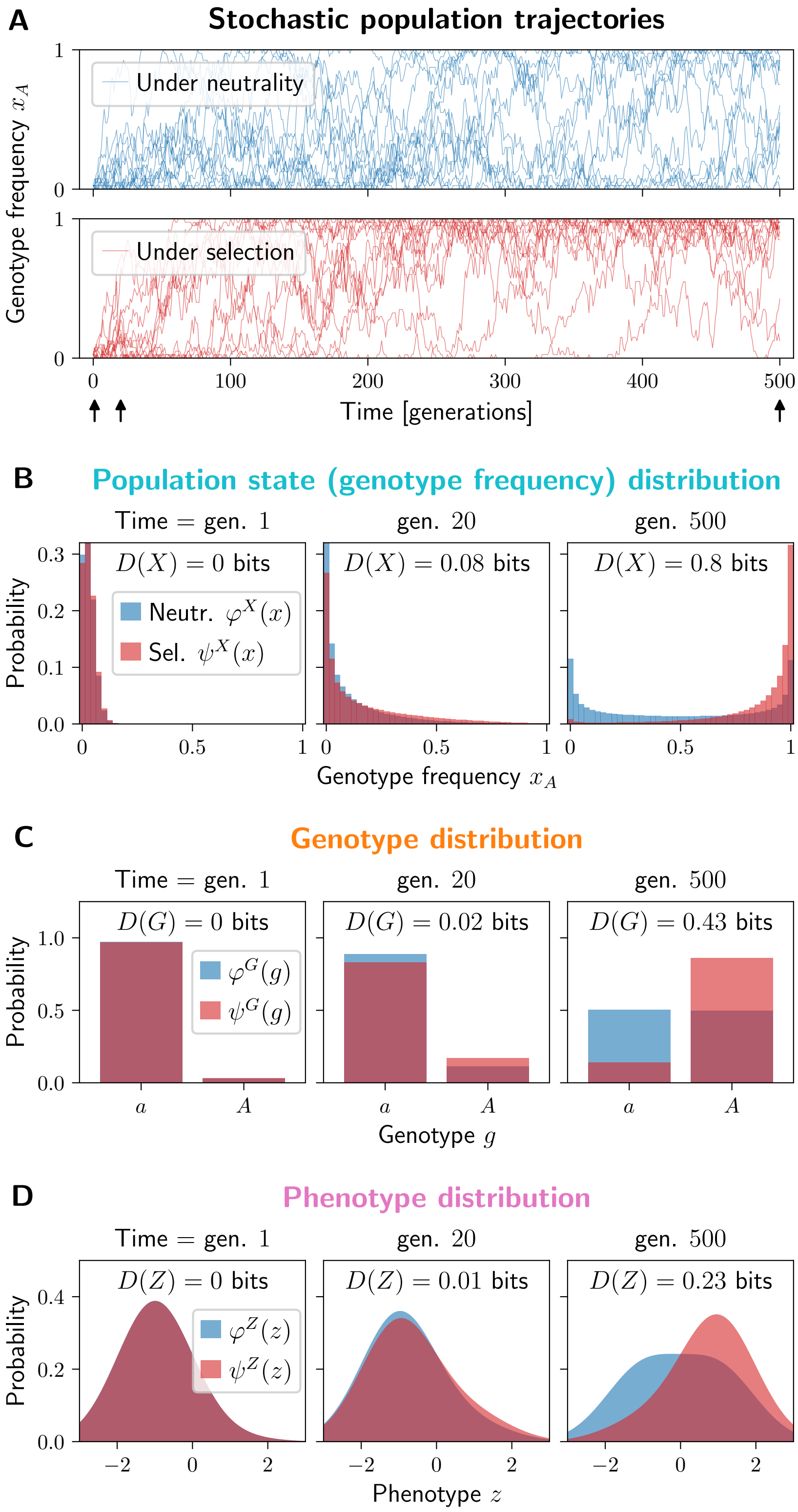

Figure 3.1: Selection controls the evolution of a single locus, two allele system and drives the distribution of the population states, genotype and phenotype away from neutrality. (A) Stochastic trajectories of the frequency x_A of the beneficial allele A, under neutrality and under selection (blue and red). The allele A starts at a single copy, and under selection it tends to increase in frequency. Black arrows indicate the times when the distributions are plotted in (B-D). At time $= 500$ generations, the system is approximately stationary. (Caption continues on the next page.)

Figure 3.1: (Continued from previous page.) (B-D) The probability distributions of the genotype frequency x_A (B), genotype g (C) and a noisy phenotype z (D) under neutrality (blue) and under selection (red) after a varying number of generations of evolution. The associated measures of information $D(X)$, $D(G)$ and $D(Z)$ are indicated in each panel. (B) The neutral distribution φ^X converges to a symmetric U shape, while the distribution under selection is biased towards high frequencies of the beneficial allele A. The information $D(X)$ increases over time. (C) The neutral genotype distribution φ^G converges to a uniform distribution, due to symmetry between alleles a and A. Under selection (ψ^G), the beneficial allele A has a higher probability, but it does not dominate completely, so the genotype-level information $D(G)$ is less than the maximum 1 bit. $D(G)$ is also upper bounded by $D(X)$. (D) A phenotype with different means and a Gaussian noise for each allele, $\zeta(z|g) = \mathcal{N}(z; \mu_g, \sigma)$ with $\mu_a = -1$, $\mu_A = +1$ and $\sigma = 1$. The information $D(Z)$ is upper bounded by $D(G)$, with a gap due to the partially overlapping distributions $\zeta(z|a)$ and $\zeta(z|A)$. Generated using a haploid Wright-Fisher model (SI Sec. B.4) with population size $N = 40$, mutation rate $\mu = 0.005$ and fitness 1 (allele a) and 1.05 (allele A).

3.2.1 Population-level information

Evolution is a stochastic process happening to populations, and genotype frequencies form the state space. We use X to denote the genotype frequencies as a random variable, with each value x being a vector with an element x_g for each genotype g, normalized as $\sum_g x_g = 1$. As an example, Fig. 3.1A shows a common evolutionary scenario where a single locus, two allele system starts from a single copy of a beneficial allele A, and later the frequency evolves stochastically.

X takes values x with probabilities $\psi^X(x)$ under selection and $\varphi^X(x)$ under neutrality. Fig. 3.1B shows examples of these distributions for the single locus system at three different times. In general, these distributions are shaped by various evolutionary forces – mutation, drift, recombination, selection (ψ^X only), and others. We refer to $D(X)$, the KL divergence between ψ^X and φ^X, as the population-level information.

The example in Fig. 3.1 illustrates two important phenomena we discuss in the rest of this chapter. The first phenomenon is the accumulation of information. A population evolves from an initial distribution (in the simplest case, initially $\psi^X = \varphi^X$ and $D(X) = 0$, but this is not necessary). For example, the initial state x may be completely specified as in Fig. 3.1A, or both ψ^X and φ^X may start at the neutral stationary distribution. Over time, selection causes ψ^X to diverge from φ^X and the information $D(X)$ accumulates (Fig. 3.1B). We study this in detail in Sec. 3.3. The second phenomenon is the maintenance of information, and it takes place when both $\psi^X(x)$ and $\varphi^X(x)$ are stationary, and the information $D(X)$ is constant. In Sec. 3.4 we study how much information can be maintained at a given cost of selection.

The population-level information $D(X)$ has been studied under different names and in different roles (Barton and de Vladar, 2009; Mustonen and Lässig, 2010; Bod'ová et al., 2016; Barton, 2017). It captures any departure of the genotype frequency distribution ψ^X from its neutral counterpart φ^X – notably, selection can favor not only high frequencies of fit genotypes, but also higher or (more typically) lower amounts of genetic variation within populations. Note that $D(X)$ refers to the effects of selection on the genotype frequencies, rather than allele frequencies. It therefore includes effects of selection on correlations between loci (linkage disequilibrium), which are generated by physical linkage, by chance in finite populations, or due to functional interactions (epistasis) – see also SI Sec. B.2.

Notably, $D(X)$ (or $D(G)$ introduced below) appears as a term in free fitness – a quantity analogous to free energy which, under some assumptions, increases over time (Iwasa, 1988; Sella and Hirsh, 2005; Mustonen and Lässig, 2010). This implies that evolution maximizes the expected log-fitness while constraining $D(X)$ – see SI Sec. B.8.

3.2.2 Genotype-level information

If we sample a random genotype from a population in a given state x, we find the genotype g with a probability given simply by its frequency $\psi^{G|X}(g|x) = \varphi^{G|X}(g|x) = x_g$. Taking into account evolutionary stochasticity, we average over all population states x with their probabilities $\varphi^X(x)$ or $\psi^X(x)$,

$$\varphi^G(g) = \sum_x \varphi^X(x)\, x_g, \qquad \psi^G(g) = \sum_x \psi^X(x)\, x_g. \tag{3.3}$$

Under symmetric point mutations, the neutral distribution φ^G converges to a uniform distribution over all genotypes, while selection typically concentrates ψ^G among a smaller number of fit genotypes. This is also the case for the single locus system in Fig. 3.1C. The divergence between ψ^G and φ^G is the genotype-level information $D(G)$.

If selection precisely specifies n out of l nucleotides in the genome – i.e. $\psi^G(g)$ is uniform over a fraction $1/4^n$ out of 4^l possible genotypes – this implies $D(G) = 2n$ bits. This corresponds to the intuition of $2n$ bits of information encoded in the genome. More typically, selection will specify many sites only weakly (biasing the probability towards some alleles, see also Fig. 3.1C), and may contribute to $D(G)$ through LD – correlations between linked or epistatically interacting sites. Without linkage or epistasis, $D(G)$ is approximately additive across loci (see Fig. B.1).

$D(G)$ generalizes some previous definitions of genetic information (Kimura, 1961; Worden, 1995; Peck and Waxman, 2010) which focused on strong selection or uniform distributions, and coincides with others in important special cases (Watkins, 2008; MacKay, 2003a).

3.2.3 Phenotype-level information

Finally, selection controls evolution on the level of the phenotype Z. Z could be a categorical trait such as the presence/absence of a disease or the correct/incorrect protein fold, a quantitative trait, a comprehensive characterization of an individual, or its fitness. Given a genotype g, the probability of the phenotype z will be given by the possibly noisy genotype-phenotype relationship $\psi^{Z|G}(z|g) = \varphi^{Z|G}(z|g) = \zeta(z|g)$. When there are no environmental effects or intrinsic noise, $\zeta(z|g)$ will be concentrated at a single value z for each genotype g. Taking into account the variation within populations, as well as the evolutionary stochasticity, the marginal probability of z is

$$\psi^Z(z) = \sum_g \psi^G(g)\, \zeta(z|g), \qquad \varphi^Z(z) = \sum_g \varphi^G(g)\, \zeta(z|g). \tag{3.4}$$

We show the distributions ψ^Z, φ^Z for the single locus system in Fig. 3.1D, where the trait has a genotype-dependent mean and a Gaussian noise. While under neutrality, φ^Z tends to spread out over time, selection causes ψ^Z to be more concentrated. The divergence between ψ^Z and φ^Z is the phenotype-level information $D(Z)$.

If we can take the genotype distribution φ^G to be uniform over all possible DNA sequences of some length, then φ^Z is the phenotype distribution among such random sequences. Examples

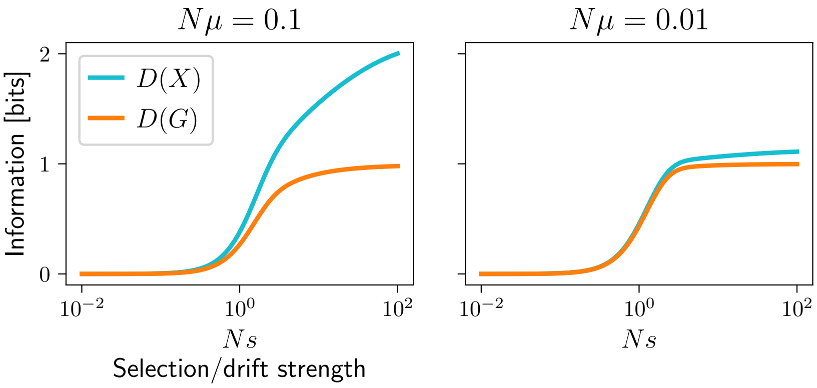

Figure 3.2: Illustration of $D(X)$ (cyan) and $D(G)$ (orange) for a single locus, two allele system at stationary distributions ψ^X, φ^X as a function of selection strength Ns for two different mutation strengths $N\mu$. The genotype-level information $D(G)$ grows with Ns, from 0 up to 1 bit one out of the two alleles dominates, with the steepest increase around $Ns = 1$. The population-level information $D(X)$ can be much greater than $D(G)$ when mutation is strong, and generates diversity within the population that selection can shape (or suppress). When mutation is weak, $D(X)$ and $D(G)$ are similar, since the population state can be specified by the allele that is currently fixed and $D(X|G) = 0$. Computed using a Wright-Fisher model as in Fig. 3.1, with population size $N = 100$.

of this distribution have recently been measured experimentally, for gene expression generated by random promoter sequences in *S. cerevisiae* and *E. coli* (de Boer et al., 2020; Lagator et al., 2022). If a healthy cell requires the gene expression to be in some narrow range, this translates to a requirement on the phenotype-level information $D(Z)$, and this requirement will increase if the expression needs to be specified across cell states.

3.2.4 The relationship between the three levels

The definitions above, combined with the chain rule (Eq. (3.2)) lead to a hierarchy among the three levels,

$$D(X) \geq D(G) \geq D(Z). \tag{3.5}$$

This inequality can be observed across columns of panels in Fig. 3.1B-D.

Intuitively, the phenotype-level information $D(Z)$ is bounded by the genotype-level information $D(G)$, since the information about the phenotype has to be encoded in the genome. A special case of this relationship has been noted by Worden (Worden, 1995), who however, worked in a deterministic setting (see SI Sec. B.3). The difference between the two, $D(G) - D(Z) = D(G|Z)$, can have two sources. First, the phenotype distribution $\zeta(z|g)$ may overlap between genotypes, causing the phenotype to be specified less precisely than the genotype (as in Fig. 3.1D). Second, selection may favor genotypes based on criteria other than the phenotype Z, such as other phenotypes or robustness.

Similarly, $D(G)$ can only be as large as the population-level information $D(X)$. To increase the probability of a genotype g, selection must increase the probability of population states with a high frequency of g. However, selection can also shape the patterns of genetic diversity in populations, without impacting the average genotype frequencies, therefore contributing to the difference $D(X) - D(G) = D(X|G)$. In populations with weak mutation, which tend to have little diversity, this difference is small – see Fig. 3.2.

We rely on the inequalities in Eq. (3.5) in two ways. First, an upper bound on the population-level information $D(X)$ which we prove in Sec. 3.3 also implies an upper bound on the genotype and phenotype-level information $D(G)$ and $D(Z)$. In other words, selection can only fine-tune the phenotype to a degree to which it can control the population state.

Second, $D(X)$ and $D(G)$ can be difficult to estimate directly for systems with multiple loci, due to the high dimensionality (see Fig. B.1). In such situations, $D(Z)$ for fitness or a low-dimensional phenotype Z can serve as a lower bound on $D(G)$ and $D(X)$. If Z is the trait under selection, or fitness itself, this lower bound can be tight. This approach is applicable even for complex or essentially black box genotype-phenotype models, such as models of gene regulation or protein folding.

3.3 Accumulation of information

In this section we show how the rate at which $D(X)$, the population-level information, increases over time, is limited by the population size and the variation in fitness. We start by pointing out a connection between population genetics and control theory.

3.3.1 Accumulation of information and the cost of control

We consider a population evolving over time, with a trajectory $X^0, X^1, \ldots, X^T$ forming a Markov chain between generations 0 and T (such as in Fig. 3.1A). The divergence of the trajectories' distribution from neutrality, $D(X^0, X^1, \ldots, X^T)$, has been proposed as a measure of predictability of evolution (Lässig et al., 2017). Using the chain rule (Eq. (3.2)), we can decompose it in two ways,

$$\begin{aligned} &D(X^0, X^1, \ldots, X^T) \\ &= \underbrace{D(X^0)}_{\text{Initial information}} + \underbrace{\sum_{t=0}^{T-1} D(X^{t+1}|X^t)}_{\text{KL cost of control}} \qquad (3.6) \\ &= \underbrace{D(X^T)}_{\text{Final information}} + \underbrace{\sum_{t=0}^{T-1} D(X^t|X^{t+1})}_{\text{Effect of selection on trajectories reaching } X^T} \ . \qquad (3.7) \end{aligned}$$

In Eq. (3.6), we distinguish between the divergence of the initial states X^0 and the additional conditional divergence in each generation, $D(X^{t+1}|X^t)$. The latter can be recognized as the KL cost of control, averaged over the initial states x^t (Todorov, 2006; Theodorou, 2015). In the context of population genetics, selection takes the role of control.

Eq. (3.7) makes the distinction between the distribution of endpoints X^T, and the conditional distribution of the states that precede those endpoints. Selection can shape the full trajectories, but only the effects on X^T constitute the final population-level information.

Together, Eq. (3.6) and (3.7) imply a bound on the information accumulated between times 0 and T in terms of the KL cost of control,

$$D(X^T) - D(X^0) \leq \sum_{t=0}^{T-1} D(X^{t+1}|X^t). \qquad (3.8)$$

Specifically, the information accumulated over a single generation, $\Delta D(X^t) = D(X^{t+1}) - D(X^t)$, is upper bounded as

$$\Delta D(X^t) \leq D(X^{t+1}|X^t). \tag{3.9}$$

Analogous bounds for continuous time Markov chains and the diffusion approximation are in SI Sec. B.6 and B.7.

Note that control theory is concerned with computing optimal control policies, which maximize an imposed objective while minimizing the cost $\sum_{t=0}^{T-1} D(X^{t+1}|X^t)$. This is analogous to computing the optimal artificial selection – in fact, the KL divergence control theory framework has recently been used to study artificial selection on quantitative traits (Nourmohammad and Eksin, 2021).

In contrast, natural selection is typically given by the biological or ecological circumstances, and not necessarily optimized in this sense. Still, the KL cost of control provides bounds on the rate at which selection accumulates information (Eq. (3.8,3.9)) and it has a meaning in population genetics, which we discuss in the next section.

We also note that Eq. (3.9) is related to the proof that free fitness increases over time (Iwasa, 1988; Sella and Hirsh, 2005), see SI Sec. B.8.

3.3.2 Variation in fitness as cost of control

To compute $D(X^{t+1}|X^t)$ in population genetics, we need to specify a model. We analyze multiple general model classes in the SI: Wright-Fisher and discrete Moran models in Sec. B.5, continuous time Moran model in Sec. B.6 and the diffusion approximation in Sec. B.7. In summary, the bound in Eq. (3.9) always takes the form

$$\Delta D(X^t) \leq kN \sum_{x^t} \psi^{X^t}(x^t)\, C(x^t) = kN\langle C\rangle_t, \tag{3.10}$$

where N is the population size, kN is the number of individuals that are sampled with selection in each generation ($k = 1$ under asexual reproduction or $k = 2$ under sexual reproduction when 2 parents are sampled with selection for each individual). Note that this does not reflect ploidy – we consider g to be the full genotype of an organism, including polyploidy or any extrachromosomal DNA. Such factors might enable more complex patterns of selection and inheritance and influence the accumulation of information in any particular case, but they do not enter the upper bound. Rather, k reflects sampling 2 parents per individual in sexually reproducing species provides more opportunity for selection to influence which genes get passed on as the fitness of both parents is considered. This does not necessarily mean that information will accumulate faster with sex, as some of the extra information can be lost due to random recombination and segregation.

$C(x^t)$ is the cost of selection at the population state x^t (see below), and $\langle C\rangle_t$ is the expected cost at time t. To upper bound information accumulated over multiple generations, we need to sum over them,

$$D(X^T) - D(X^0) \leq kN \sum_{t=0}^{T-1} \sum_{x^t} \psi^{X^t}(x^t)\, C(x^t) = kN\langle C\rangle_{0,T}. \tag{3.11}$$

The cost $C(x)$ is a measure of fitness variation in a population in the state x,

$$C(x) = \sum_g x_g \hat{w}_g(x) \log_2 \hat{w}_g(x). \tag{3.12}$$

where $\hat{w}_g(x)$ is the (frequency dependent) relative fitness of genotype g. When sampling genotypes as parents for the next generation, g is picked with probability x_g under neutrality and $x_g \hat{w}_g(x)$ under selection – $C(x)$ is the KL divergence between these two distributions.

Under a constant population size N, the strongest possible selection regime would entail a single individual producing all N offspring in the current generation, i.e. $x_i = 1/N$, $\hat{w}_g(x) = N$ for their genotype and $\hat{w}_g(x) = 0$ for all others. This yields an upper bound on $C(x)$ in terms of N, $C(x) \leq \log_2 N$.

$C(x)$ is also related to two more established measures of cost in population genetics – the relative fitness variance $V(x)$ and the genetic load $L(x)$, which have been studied under a number of circumstances – e.g. mutation-selection balance (Haldane, 1937), genetic drift (Kimura et al., 1963; Kondrashov, 1995), certain types of epistasis and the evolution of sex (Kimura and Maruyama, 1966; Kondrashov, 1988), ongoing substitutions (Haldane, 1957; Kimura, 1995; Ewens, 1970) or stabilizing selection on quantitative traits (Lande, 1980). They are defined as

$$V(x) = \sum_g x_g \left(\hat{w}_g(x) - 1\right)^2 \tag{3.13}$$

$$L(x) = 1 - \frac{1}{\hat{w}_{\max}(x)}, \tag{3.14}$$

where $\hat{w}_{\max}(x)$ is the maximum relative fitness present in the population x, $\hat{w}_{\max}(x) = \max_{g;\, x_g > 0} \hat{w}_g(x)$. We derive the relationships between $C(x)$, $V(x)$ and $L(x)$ in the SI Sec. B.9. $V(x)$ and $L(x)$ satisfy the inequality $V(x) \leq \frac{L(x)}{1-L(x)}$ (see also (Shnol et al., 2011)) and both provide an upper bound on $C(x)$,

$$C(x) \leq \frac{V(x)}{\ln 2}, \qquad C(x) \leq \log_2 \frac{1}{1 - L(x)}. \tag{3.15}$$

In addition, under weak selection and in the diffusion approximation, $C(x) = V(x)/(2 \log 2)$. The bounds in Eq. (3.10,3.11) can therefore also be rewritten in terms of $V(x)$ or $L(x)$ using Eq. (3.15).

Assuming constant population size, relative fitness is proportional to the expected number of offspring, and therefore limited by the species' reproductive capacity. The quantities $\hat{w}_{\max}(x)$, $L(x)$, $V(x)$ and $C(x)$ and as a consequence $\Delta D(X)$, are therefore all limited in realistic settings (SI Sec. B.9). In particular, $C(x) \leq \log_2 R(x)$ where $R(x)$ is the expected number of offspring of the fittest genotype contained in x.

In the context of artificial selection or genetic algorithms, an alternative measure of cost is the population size N, which is the number of cultivated plants or animals, or fitness function evaluations (Robertson, 1970; Barton and Paixão, 2013). We note that according to the bounds in Eq. (3.10,3.11), the maximal accumulation rate is also proportional to N. Furthermore, increasing the strength of selection (and therefore $C(x)$) beyond an optimal value may increase the immediate response to selection, but reduces the long term response, due to loss of genetic diversity (Robertson, 1970; Barton and Paixão, 2013). Therefore in practice, $C(x)$ will be limited even in this context.

3.3.3 Example 1: the fates of a beneficial allele

The bounds in Eq. (3.10,3.11) hold in genetically diverse populations with clonal interference or recombination. Still, it is interesting to consider the case of sequential fixation/loss of mutations, as was done previously (Haldane, 1957; Kimura, 1961; Barton, 2017).

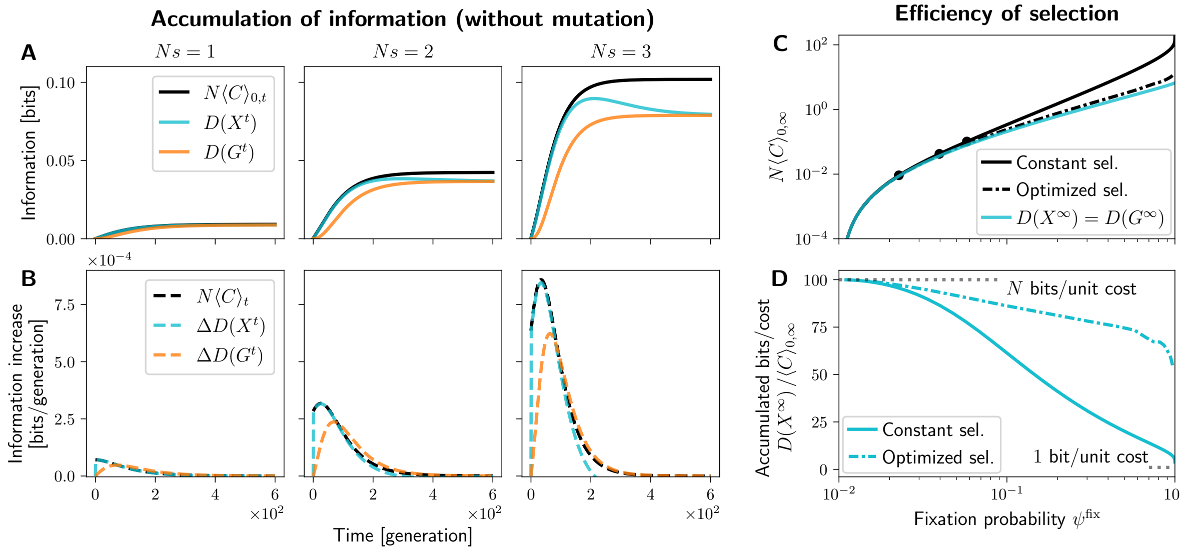

Figure 3.3: Information accumulation associated with the fixation or loss of a beneficial allele in a haploid single locus, two allele system. The beneficial allele starts at a single copy and evolves under drift and selection, but no mutation. (A) The population-level information ($D(X^t)$, cyan) and genotype-level information ($D(G^t)$, orange) over time, for three different strengths of selection (Ns). Both $D(X^t)$ and $D(G^t)$ start at zero, accumulate over time as selection tends to increase the frequency of the beneficial allele, and saturate as the allele is fixed or lost. The black line is the upper bound according to Eq. (3.11) with $k = 1$. (B) The increments in $D(X^t)$ and $D(G^t)$ per generation (cyan and orange dashed), and the upper bound according to Eq. (3.10) with $k = 1$ (black dashed). (C) The cyan line shows the total information accumulated, $D(X^\infty) = D(G^\infty)$, as function of the fixation probability ψ^{fix}. $D(X^\infty)$ serves as a lower bound on N times the total cost of selection, plotted in black, regardless of the form selection takes. The full black line corresponds to constant selection coefficient, with black points showing the three cases in (A,B). The dash-dotted black line shows frequency dependent selection that maximizes ψ^{fix} (and therefore also $D(X^\infty)$) while constraining $N\langle C\rangle_{0,\infty}$. (D) Same data as in (C), but the vertical axis now shows the ratio of the information $D(X^\infty)$ and the total cost of selection $\langle C\rangle_{0,\infty}$ for constant selection (full black) and optimized frequency dependent selection (dash-dotted black). At most N bits can be accumulated per unit cost, and this is achieved at weak selection. At strong selection, this reduces to as low as 1 bit per unit cost. Figure computed using the Wright-Fisher model as in Fig. 3.1, with population size $N = 100$.

Suppose that a beneficial allele A appears in one copy at time $t = 0$, and is guaranteed to be fixed or lost before another mutation appears that could interfere with it. The population and genotype-level information, $D(X^t)$ and $D(G^t)$, start at 0 and accumulate over time as selection tends to increase the frequency of A (Fig. 3.3A). The cumulative cost of selection $N\langle C\rangle_{0,t}$ serves as the upper bound on both $D(X^t)$ and $D(G^t)$.

Note that under relatively strong selection ($Ns = 3$, Fig. 3.3A right), A increases in frequency considerably faster than under neutrality, leading to high $D(X^t)$. But some of these gains are later lost as A is fixed or lost. This is an example of how only the probabilities of endpoints, and not the shape of the trajectories, matters for the information that is ultimately accumulated (the two terms in Eq. (3.7)).

The increments in $D(X^t)$ and $D(G^t)$ in each generation are plotted in Fig. 3.3B, along with the bound by $N\langle C\rangle_t$, Eq. (3.10). The bound on $\Delta D(X^t)$ is relatively tight. $\Delta D(G^t)$ can temporarily exceed $N\langle C\rangle_t$, since the accumulation bound in Eq. (3.10) does not directly apply to the genotype level, but this is only a transient phenomenon due to the inequality between the cumulative genotype- and population-level information $D(G^t) \leq D(X^t)$.

Both $D(X^t)$ and $D(G^t)$ saturate at the same value $D(X^\infty) = D(G^\infty)$, since the ultimate fate of the population is given simply by whether the allele A is fixed or lost. The fixation probability is $1/N$ under neutrality and $\psi^{\text{fix}} = \psi^{X^\infty}(1) = \psi^{G^\infty}(A)$ under selection, and the accumulated information is a function of this probability,

$$D(X^\infty) = D(G^\infty) = \tag{3.16}$$

$$= \psi^{\text{fix}} \log_2(N\psi^{\text{fix}}) + (1-\psi^{\text{fix}}) \log_2 \frac{N(1-\psi^{\text{fix}})}{N-1} \tag{3.17}$$

This function is plotted in cyan in Fig. 3.3C. According to Eq. (3.11), it provides a lower bound on the total cost, $N\langle C\rangle_{0,\infty} \geq D(X^\infty)$, given a fixation probability. This holds when the allele A has a constant, frequency independent selective advantage, as in the three examples in Fig. 3.3A,B (full black line and black points in Fig. 3.3C). By computing a suitable frequency dependent selection, which optimizes the fixation probability while constraining the total cost $N\langle C\rangle_{0,\infty}$, we can reduce the cost considerably (dash-dotted black line in Fig. 3.3C, see Sec. B.11 and Fig. B.4 for details). This is achieved by making selection weaker at high frequencies, where the risk of losing A is low. Still, the cost stays above $D(X^\infty)$, as it has to under arbitrary frequency and time-dependent selection.

Under both forms of selection, the bound is only tight when selection is weak. To emphasize this, we plot the information accumulated per unit cost, $D(X^\infty)/\langle C\rangle_{0,\infty}$, as function of the fixation probability ψ^{fix} in Fig. 3.3D. At weak selection, ψ^{fix} is only perturbed a little from its neutral value $1/N$, but up to N bits can be accumulated per unit cost. A special case of this was shown by Barton (Barton, 2017). Similar scaling with N was also found in a different setting by Kimura (Kimura, 1995).

Stronger selection accumulates more information, but at a disproportionately higher cost, since a large part of it is spent on shaping trajectories rather than outcomes. In the extreme case, to achieve $\psi^{\text{fix}} = 1$, only individuals carrying the A allele can be allowed to reproduce and A gets fixed in only one generation – a highly unlikely way to fixation under neutrality. In this case, selection has the same effect on each genotype sampled as a parent in the first generation, as on the allele that is ultimately fixed (both are A with probability $1/N$ under neutrality and 1 under selection). As a result, the cost is equal to the accumulated information, $\langle C\rangle_{0,\infty} = D(G^\infty) = D(X^\infty)$ and only 1 bit per unit cost is accumulated (Fig. 3.3D). This is why previous results derived in deterministic settings (Kimura, 1961; Worden, 1995) claimed much more stringent limits on accumulation of information.

3.3.4 Example 2: accumulation of information under mutation

Unlike the example above, real systems experience ongoing mutation. On the one hand, mutation is necessary to supply beneficial alleles for adaptation, but on the other hand, mutation can disrupt existing adaptation. In this section, we assume that the single locus, two allele system starts at the neutral stationary distribution with $D(X^0) = D(G^0) = 0$, and then selection is turned on. Adaptation exploits copies of the allele A that either segregate in the population by chance at time 0, or arise later by mutation.

Fig. 3.4A shows the information $D(X^t)$ and $D(G^t)$ over time. Accumulation take place on the time scale of $1/\mu$. Note that the bound Eq. (3.11) is not very tight. This is even more

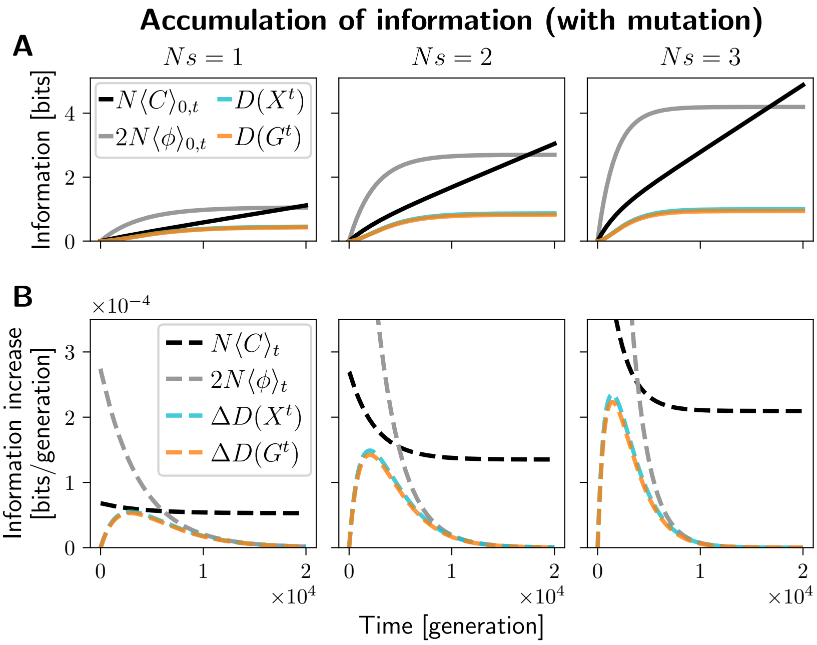

Figure 3.4: Information accumulation in a single locus, two allele system and the associated upper bounds. The system starts from a neutral stationary distribution over allele frequencies, where $D(X^0) = D(G^0) = 0$. Then it evolves under selection with varying strengths (Ns, left to right) for 2×10^4 generations. (A) The cumulative information at the population level ($D(X^0)$, cyan) and genotype level ($D(G^0)$, orange) over time. Due to the weak mutation $N\mu = 0.01$, the two measures of information are similar. The black and gray lines show upper bounds by the cumulative cost of selection and the cumulative fitness flux. (B) The increments in information per generation, $\Delta D(X^t)$ (cyan dashed) and $\Delta D(G^t)$ (orange dashed) and the upper bounds on these increments in terms of the cost of selection $kN\langle C\rangle_t$ (black, in this case $k = 1$) and the expected fitness flux $2N\langle\psi\rangle_t$ (gray). Note that the cost of selection bound is briefly nearly tight under weak selection ($Ns = 1$, left), and the fitness flux bound is tight near stationarity, when both the accumulation rate and the fitness flux approach 0. Figure computed using the Wright-Fisher model as in Fig. 3.1. The population size is fixed at $N = 100$. For technical reasons, the expected fitness flux curves were computed using an equivalent Moran model, see Sec. B.10 and Fig. B.2.

apparent in Fig. 3.4B, where the average cost per generation $N\langle C\rangle_t$ remains positive even after the system has reached the new stationary state, while the increments in $D(X^t)$ and $D(G^t)$ are zero. This corresponds to the cost of maintaining information, which we discuss in Sec. 3.4.

In summary, the accumulation of information is upper bounded by the KL cost of control, which in turn corresponds to the population size times the variation in fitness. However, if selection changes not only the probabilities of the final states, but also the paths that lead there (because it is strong, because adaptation is maintained for a long time, or because adaptation is reversed by time-dependent selection), then the information accumulated is less than the total cost.

3.3.5 Comparison with the fitness flux bound

The fitness flux theorem (Mustonen and Lässig, 2010) implies another upper bound on information accumulation rate, $\Delta D(X^t) \leq 2N\langle\phi\rangle_t$, where $\langle\phi\rangle_t$ is the expected fitness flux per generation. It is plotted in gray in Fig. 3.4. It differs from the cost of selection bound both quantitatively and in terms of interpretation.

Quantitatively, neither bound is tighter in general. In Fig. 3.4B, the cost of selection bound is tighter in early stages of adaptation, and the fitness flux bound is tighter in the late stages. This is consistent with the interpretation of fitness flux as the rate of ongoing adaptation, or the rate of ascent in the mean fitness landscape/seascape (Mustonen and Lässig, 2010). This rate is high in the early stages of adaptation, when the population is far from the fitness peak and tends to climb up quickly. Later, when the population approaches a stationary distribution, there is no more adaptation on average, and $2N\langle\phi\rangle_t$ as well as $\Delta D(X^t)$ vanish. Meanwhile, the cost of selection bound $kN\langle C\rangle_t$ is tighter in the earlier stages when most of the cost is spent on new adaptation, but it remains positive under stationarity, due to maintenance costs.

Technically, the fitness flux theorem was originally derived in (Mustonen and Lässig, 2010) under the diffusion approximation, and requires an additional assumption that the neutral process is at a stationary distribution with detailed balance. We derive and discuss the technical aspects of the fitness flux bound in Sec. B.10 and Fig. B.2 and B.3.

3.4 Maintenance of information

In this section, we ask how much information can be maintained in the genome for a given cost of selection. A general bound analogous to Eq. (3.10) seems to be out of reach for now, but we can study how the information maintained depends on key evolutionary parameters. We start by analyzing the single locus, two allele system, and then proceed to systems with large numbers of loci.

3.4.1 Single locus: weak selection is most efficient

Fig. 3.5A shows the information, $D(X)$ and $D(G)$, maintained by the single locus, two allele system at the stationary state under various strengths of selection. Stronger selection maintains more information – up to 1 bit at the genotype level, and more on the population level. However, it comes with a higher cost of selection $\langle C\rangle$, Fig. 3.5B. Notably, the cost increases faster then the maintained information. As a result, the amount of information maintained per unit cost decreases with selection strength, Fig. 3.5C.

There are two important asymptotic regimes. When selection is very strong, $Ns \gg 1$, deleterious mutations are purged as soon as they arise, and $D(G) \approx 1$ bit. Mutations arise with a probability $N\mu$ per generation, and purging each costs $C \approx 1/(N\ln(2))$ (assuming truncation selection with $\alpha = 1 - 1/N$, see Sec. B.9). In this regime,

$$\text{Strong selection:} \qquad \frac{D(G)}{\langle C\rangle} \approx \frac{\ln 2}{\mu}, \tag{3.18}$$

bits can be maintained per unit cost (see Fig. 3.5C). Similar arguments apply when $N\mu > 1$. The inverse scaling with μ is expected based on the deterministic mutation load (Haldane, 1937) or Eigen's error catastrophe (Eigen, 1971) which occurs when selection cannot maintain sequences without error, and it was also derived by Watkins (Watkins, 2008).

Selection is much more efficient when it is weak, $Ns \ll 1$. Both the cost and the maintained information can be calculated under the diffusion approximation (see SI Sec. B.4.2 for details). If mutation is also weak, $N\mu \ll 1$, the amount of genetic variation (pairwise diversity)

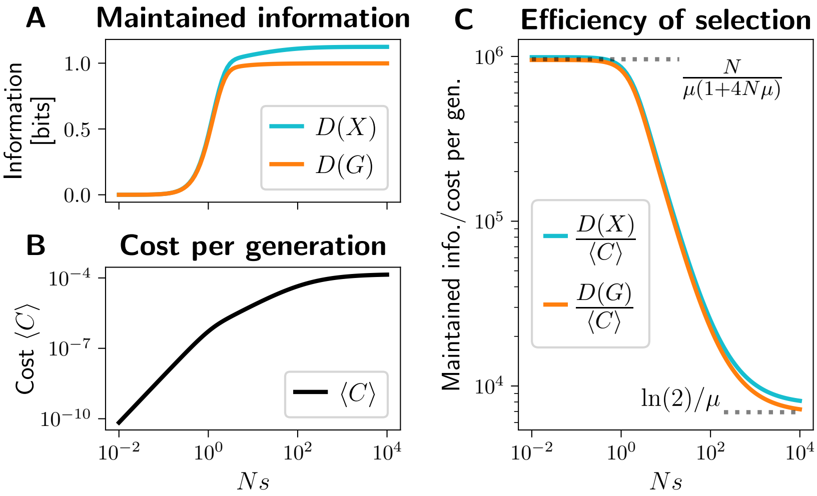

Figure 3.5: Maintenance of information in the single locus, two allele system. (A) The main plot shows the stationary values of information, $D(X)$ (cyan) and $D(G)$ (orange), as function of selection strength Ns. Stronger selection keeps the beneficial allele at higher frequencies, but this is associated with higher average cost of selection $\langle C\rangle$, shown in (B). Note that much of the time, one of the alleles is fixed and the cost C is zero. $\langle C\rangle$ is the average cost per generation over the stationary distribution of allele frequencies. (C) The ratio of the maintained information and the average cost of selection, $D(X)/\langle C\rangle$ (cyan) and $D(G)/\langle C\rangle$ (orange). Selection is most efficient when it is relatively weak ($Ns \ll 1$), maintaining up to $\frac{N}{\mu(1+4N\mu)}$ bits per unit cost at the genotype level, and inefficient when strong ($Ns \gg 1$), maintaining only about $\ln(2)/\mu$ bits per unit cost (dotted horizontal lines). The population size is $N = 100$ and mutation rate $\mu = 10^{-4}$.

scales with $2N\mu$, and the cost (variation in fitness) is approximately $\langle C\rangle \approx N\mu s^2/(2\ln 2)$. Meanwhile, selection shifts the mean frequency of A away from $1/2$ by about $Ns/2$, and this is associated with genotype-level information $D(G) \approx N^2s^2/(2\ln 2)$ bits. In this regime, up to N/μ bits per unit cost are maintained. When mutation $N\mu$ is not negligible, a more accurate result is

$$\text{Weak selection:} \qquad \frac{D(G)}{\langle C\rangle} \approx \frac{N}{\mu(1+4N\mu)}, \tag{3.19}$$

see SI Sec. B.4. This limit is also highlighted in Fig. 3.5C. The special case when $N\mu \gg 1$, $\frac{D(G)}{\langle C\rangle} \approx 1/(4\mu^2)$ was previously derived by Watkins (Watkins, 2008).

By itself, a single locus under weak selection cannot contribute much to biological function. However, selection can act on a polygenic trait influenced by many loci. If they are unlinked, we expect both the maintained information and the cost of selection to be approximately additive, and the ratio $D(G)/\langle C\rangle$ to scale according to Eq. (3.19). To confirm this, we next study a polygenic system.

3.4.2 Information stored among many loci

We use an individual-based model to study a population of N haploids with $l = 1000$ biallelic loci, mutation and free recombination. Offspring are produced by sampling pairs of parents with selection, shuffling their genomes (at each locus, the allele from either parent is inherited with probability $1/2$), and flipping each allele with probability μ. Selection acts on a fully heritable, additive trait with equal effects, $z_g =$ (the number of A alleles in g), with fitness being $w_g = (1+s)^{z_g}$.

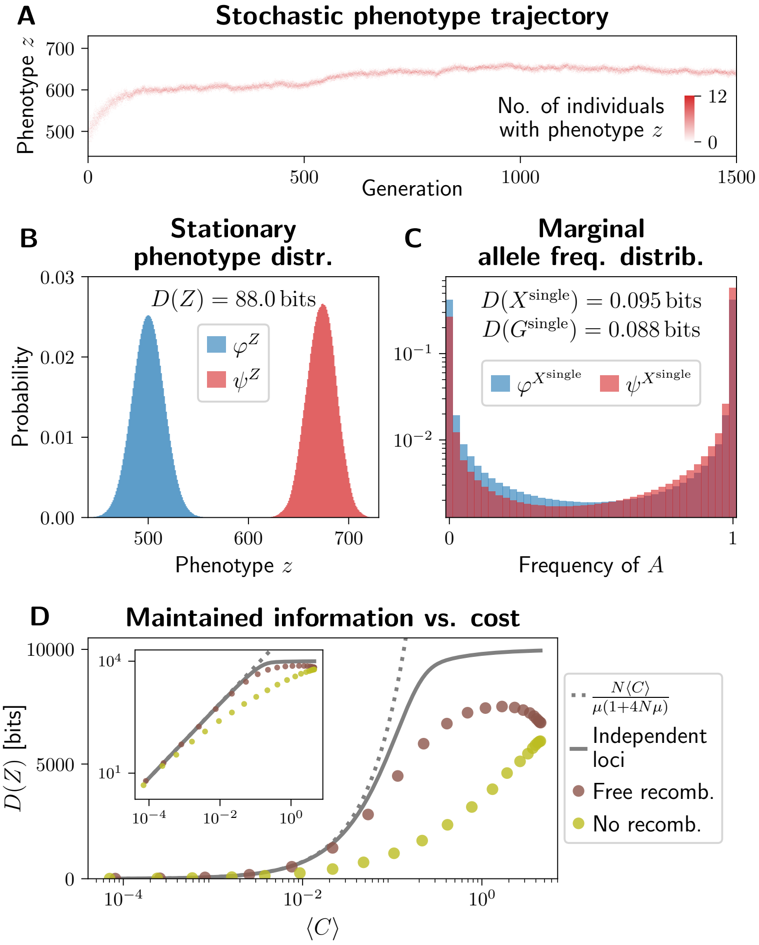

Figure 3.6: Maintenance of information in a system with $l = 1000$ biallelic loci. Selection is directional on an additive trait Z (= the number of beneficial alleles). (A) A heatmap showing the number of individuals in a population occupying each value of the phenotype z at each generation. The population is initialized as a collection of random genomes, each containing the beneficial allele at around $l/2 = 500$ loci. Over time, this number stochastically increases due to selection. Only the first 1500 generations of the trajectory are shown, the full trajectory was 5×10^3 generations of burn-in and 2×10^5 to estimate the stationary distributions in (B,C). (B) The stationary distribution over the phenotype Z, under neutrality (φ^Z, blue) and selection (ψ^Z, red), along with the phenotype-level information $D(Z)$. Due to symmetry between loci and alleles, $\varphi^Z(z) = \mathrm{Binom}(z; l, 0.5)$ is binomial. Under selection, ψ^Z is obtained as the histogram over individuals and over 2×10^5 generations at stationarity. (C) The marginal distribution over allele frequencies at individual loci, under neutrality ($\varphi^{X^{\mathrm{single}}}$, blue, computed using a transition matrix for the single locus system) and under selection ($\psi^{X^{\mathrm{single}}}$, red, computed as a histogram over all loci and 2×10^5 generations at stationarity). The associated $D(X^{\mathrm{single}})$ and $D(G^{\mathrm{single}})$ correspond to information maintained at one locus, and because the loci are approximately independent, the total information is about $l = 1000$ times more. The population size is $N = 40$, mutation strength $N\mu = 0.02$ and selection strength $Ns = 0.4$. (D) The relationship between the maintained information $D(Z)$ and the cost of selection $\langle C \rangle$, with recombination (brown points), without recombination (olive points). This is compared with predictions under the assumption of independent loci (gray line; computed using single locus diffusion approximation and multiplying both information and cost by the number of loci) and the linear scaling with $\langle C \rangle$ based on Eq. (3.19) (dotted gray). Computed for a system with $l = 10^4$ loci, population size $N = 40$, mutation strength $N\mu = 0.02$ and variable Ns. Distributions estimated from a stochastic trajectory over 5×10^4 generations, after 5×10^3 generations of burn-in. The inset shows identical data with a log vertical scale.

The results are shown in Fig. 3.6. The panel 3.6A shows an example of a stochastic population trajectory, indicating the phenotypes present in the population over time. The system is initialized with random genomes that contain the beneficial allele at each locus with probability $1/2$, with z taking values around $l/2 = 500$ with binomial noise. Selection with $s = 0.01$ makes the beneficial alleles more frequent over time. The stationary distribution over phenotypes is in Fig. 3.6B. Under neutrality, $\varphi^Z = \mathrm{Binom}(l, 1/2)$ by symmetry. The distribution ψ^Z under selection is shifted relatively far from φ^Z, leading to $D(Z) = 88.0$ bits of information on the phenotype level.

The population state distribution and the genotype distribution are inaccessible due to their dimensionality (see Fig. B.1). However, we know that they are lower bounded by $D(Z)$, which is easy to compute, and $D(Z) \approx D(G)$ since Z is the only trait under selection. Since the loci are unlinked and have equal effects, the information $D(Z)$ can be divided evenly among them. The marginal distribution over allele frequencies is only slightly different from neutrality (Fig. 3.6C), by about $D(X^{\mathrm{single}}) = 0.095$ bits in terms of allele frequency distribution and $D(G^{\mathrm{single}}) = 0.088$ in terms of allele probabilities. The 1000 loci, however, combine to produce a large shift in the phenotype distribution, $D(Z) \approx 1000 D(G^{\mathrm{single}})$.

This information is maintained at a very low cost of selection, $\langle C \rangle = 0.0012$ bits per generation, or relative fitness variance $\langle V \rangle = 0.0017$. This amounts to $D(Z)/\langle C \rangle = 7.1 \times 10^4$ bits per unit cost, only a little below the single locus limit $N/\mu/(1 + 4N\mu) = 7.4 \times 10^4$ under weak selection.

3.4.3 Interference between loci

In practice, the selection on different loci might interfere, and this can hinder the maintenance of information. The interaction may be due to Hill-Robertson interference, linkage, or epistasis.

In Fig. 3.6D we vary the selection coefficient s on individual alleles in a $l = 10^4$ locus system, and plot the maintained $D(Z)$ against the cost $\langle C \rangle$. We use the individual-based model to compute these with free recombination (as in 3.6A-C) and with zero recombination (offspring genotypes are identical to those of single parents, up to mutation). We compare the results with the weak selection scaling according to Eq. (3.19), and results for 10^4 loci that evolve independently (cost and information are summed over 10^4 single locus systems).

With free recombination, weak selection maintains about as much information as if the loci were independent (brown points and gray line in Fig. 3.6D, inset), approximately according to Eq. (3.19) (gray dotted line). However, when selection is strong ($\langle C \rangle \approx 0.1$ or more), individual alleles experience additional fluctuations in frequency due to random associations with alleles at other loci in a finite population (Hill and Robertson, 1966; Barton, 1995), reducing the efficiency of selection. As a result, the freely recombining loci maintain less information than if they were independent. This is in addition to the fact that under strong selection, maintenance is more costly even for independent loci (full gray line departs from dotted gray line, Fig. 3.6D). Extremely strong selection, which removes potentially adaptive variation at other loci, maintains even less information than more moderate selection, and it makes recombination ineffective (brown points at high $\langle C \rangle$ in Fig. 3.6D).

Without recombination, less information is maintained at any given cost (olive points in Fig. 3.6D). In fact, Watkins (Watkins, 2008) has shown that due to clonal interference, organisms with no recombination cannot maintain more than the order of $\ln(N)/\mu$ bits of information even if the cost is unlimited, making Haldane's and Eigen's results (Haldane, 1937; Eigen, 1971) pertinent to asexual populations.

The advantage of recombination has also been recognized in a similar context by MacKay (MacKay, 2003a) and Peck and Waxman (Peck and Waxman, 2010), and relates to the evolution of sex and epistasis. Recombination is advantageous when facing unconditionally deleterious or beneficial alleles (Kondrashov, 1988). However, recombination can be disadvantageous when adaptation depends on beneficial combinations of alleles (Kondrashov and Kondrashov, 2001). For example, under stabilizing selection on a quantitative trait, fitness is highest near an intermediate, optimal trait value, but recombination generates variation around it. This raises questions about how information scales with relevant parameters (such as recombination rate) in such scenarios. For now, it is not clear if any form of selection can maintain more information at a given cost than $\frac{N}{\mu(1+4N\mu)}$ achieved by weak directional selection with strong recombination.

3.5 Discussion

Selection exerts control on evolving populations, but its capacity is limited. The limits to selection have been approached from various angles. Here we build upon previous work that had developed the idea that selection accumulates and maintains information in the genome (Kimura, 1961; Eigen, 1971), and that this is associated with a cost in terms of variation in fitness, such as genetic load or fitness variance (Haldane, 1937, 1957). The early work has suggested remarkably simple limits to selection: that the maximal rate of accumulation is bounded by the cost itself (Kimura, 1961; Worden, 1995), and that maintenance is limited to about $1/\mu$ functional sites in the genome (Haldane, 1937; Eigen, 1971).

Later work has pointed out that both accumulation (Barton, 2017; MacKay, 2003a) and maintenance (Peck and Waxman, 2010; Watkins, 2008) can exceed these limits, notably when recombination is involved. However, the general bounds remained unclear, possibly in part due to the difficulty of defining genetic information in general.

The measures of information that we have introduced in Sec. 3.2 coincide with or generalize previous definitions, and offer two advantages. First, they facilitate connections between different levels – e.g. between the abstract population-level information that has been studied theoretically in different contexts (Mustonen and Lässig, 2010; Barton and de Vladar, 2009; Bod'ová et al., 2016) and the effect that selection has on the distribution of phenotypes.

Second, the generality of our definition allows proving a general bound on information accumulation rate. This turns out to be a factor N faster than the early bounds, but depends on selection on individual loci being weak. The bound relies on a measure of cost of selection that connects the genetic load and fitness variance (Shnol et al., 2011) with the KL cost in control theory (Todorov, 2006; Theodorou, 2015), recently used in the context of artificial selection (Nourmohammad and Eksin, 2021).

How much information can be maintained in the genome at a given cost remains an open problem, but we have discussed how this might scale with the population size and the mutation rate. The scaling in Eq. (3.19) generalizes a result by Watkins (Watkins, 2008) to realistic populations with $N\mu < 1$. Still, more work is needed to make claims about the information content of any real organism's genome. Typical populations have N_e/μ much greater than the genome size, suggesting that the genome size or other factors are more limiting than Eq. (3.19). The maintenance can be made more difficult by linkage or epistasis, and parts of the genome are likely under strong selection which is more costly. Still, Eq. (3.19) suggests that in theory, the genome could contain a substantial amounts of information among weakly selected loci, e.g. coding for polygenic traits. This is consistent with recent work (Galeota-Sprung et al., 2020) pointing out that mutation load does not pose severe limitations to the functional fraction of the human genome.

Similarly, the bound on accumulation rate in Eq. (3.10) hypothetically allows accumulation of information amounting to 10% of the human genome in about 10^6 generations (6×10^8 bits, assuming effective population size $N_e \approx 10^4$, $k = 2$ and meager cost $\langle C \rangle \approx 0.03$ or relative fitness variance $\langle V \rangle \approx 0.018$ devoted to accumulation). But this is unlikely to have happened. Some selection was likely strong and more costly, and selection could have fluctuated, reversing previous adaptation. However, under the right conditions, information can accumulate very fast.

Our findings are complementary to the point raised by Kondrashov (Kondrashov, 1995), that the survival of populations could be threatened by large numbers of weakly deleterious mutations ($Ns < 1$). While selection cannot purge them, it can perturb the allele frequency distribution of each by a small amount, and thus shift the distribution of higher level traits very far from neutrality. This is similar to the response by Charlesworth (Charlesworth, 2013). In fact, information accumulation and maintenance are most cost-efficient in this regime. This does not mean that a genomic architecture, where most mutations operate at $Ns < 1$ and information is encoded among many weakly specified sites, would evolve as an adaptation to maximise information gain. Nevertheless, such an architecture might arise in multicellular organisms as a side effect of their small effective population sizes and long genomes (Lynch and Conery, 2003; Lynch, 2007).

Focus on the information content of genomes, rather than their fraction under selection, could help better frame the controversy sparked by some publications from the ENCODE project (The ENCODE Project Consortium et al., 2012; Doolittle, 2013; Graur et al., 2013; Brunet and Doolittle, 2014; Kellis et al., 2014a; Graur, 2017; Galeota-Sprung et al., 2020). On the one hand, genomic regions under detectable selection (less than 15% in humans (Rands et al., 2014)) likely contain less than 2 bits per base pair, because their current function could be achieved by a number of alternative sequences (e.g. due to synonymous mutations in coding regions, or flexibility of transcription factor binding site sequence and location). On the other hand, regions without detectable selection could contain a considerable amount of information in the aggregate, at a low cost, encoding polygenic traits.

In bioinformatics, there already is a measure of information content applicable to short regulatory motifs (Schneider et al., 1986; Wasserman and Sandelin, 2004). Future work could examine the precise relationship between this measure and our theoretical definitions. The generality of our framework also opens new directions for future research. One is to predict the maximal amount of information that can be maintained in genomes and populations with realistic parameters. Another is to study the information content of genomic elements with well-described genotype-phenotype maps (e.g. promoters (de Boer et al., 2020; Vaishnav et al., 2022)), under different hypotheses about selection on the phenotype.

Acknowledgements. We thank Ksenia Khudiakova, Wiktor Młynarski, Sean Stankowski and two anonymous reviewers for discussions and comments on the manuscript. G.T. and M.H. acknowledge funding from the Human Frontier Science Program (HFSP RGP0032/2018). N.B. acknowledges funding from ERC grant 250152 "Information and Evolution".

CHAPTER 4

Evolution and information content of regulatory sequences

Abstract. The evolution of regulatory regions is shaped by the genetic architecture of regulatory phenotypes, i.e. properties or the regulatory Genotype-phenotype map (GP map). This architecture depends broadly on how gene regulation is organized (e.g. the availability of regulatory mechanisms such as nucleosome binding or RNA interference), as well as on the particular parameter values associated with such mechanisms – both of which can evolve. We argue that selection on regulatory phenotypes across the genome induces selection on the architecture, which in turn evolves to facilitate adaptation across the genome. Using population genetics theory, we derive an optimality principle, implying that GP maps associated with the evolved architecture tend to maximize the number of genotypes associated with the fit combinations of regulatory phenotypes. Mathematically, this builds on analogies between population genetics and statistical physics, as well as optimal coding in information theory. Optimal values of some regulatory parameters can be derived even without a complete knowledge of the molecular mechanisms involved. We illustrate the theory using a simplified model of Transcription factor (TF) binding to promoters: the fraction of possible promoter sequences that bind the TF evolves to match the fraction of promoters under selection to bind it. We discuss the subtleties and limitations of the theory, associated with biophysical constraints and the need for mutational robustness.

4.1 Introduction

Evolving genomes experience selection that is shaped by underlying GP maps. For example, regulatory regions such as promoters may be selected to bind specific TFs. But the GP maps themselves can evolve: changes to the TF concentration or DNA-binding domain will affect which promoter sequences bind it via changes to the size and sequence of potential binding sites. Mutations that affect the map may have highly pleiotropic effects, because a single TF can interact with sites across the entire genome, and therefore experience strong selection associated with phenotypic changes at the target loci. If regulatory GP maps evolve under this type of selection, what kind of maps should we expect as a result?

A number of specific questions about gene regulation remains unanswered, and might benefit from a suitable evolutionary framework. It is not clear why TFs have such short motifs in eukaryotes (Wunderlich and Mirny, 2009) and rely on other mechanisms for promoter search (Field et al., 2008; Brodsky et al., 2020), why so much of human DNA is transcribed (Mattick et al., 2010), or why regulatory networks are so densely interconnected by weak links (Biggin, 2011). A number of recent studies that characterize the genotype-expression maps of promoter and enhancer regions in *E. coli* (Yona et al., 2018; Lagator et al., 2022), *S. cerevisiae* (de Boer et al., 2020; Vaishnav et al., 2022) and Drosophila (Fuqua et al., 2020; Galupa et al., 2023) found that a large fraction of random DNA sequences can act as a promoter or enhancer, and even those that do not are only a few point mutations away. Should we be surprised by this finding?

In this paper, we study the evolution of regulatory parameters, such as those associated with TFs, which parametrize the genetic architecture of regulatory phenotypes. More specifically, we focus on how the space of possible DNA sequence at a locus is divided among the relevant regulatory phenotypes. For example, what fraction of possible promoter-sized DNA sequences bind a TF?

Our key finding can be summarized in a simple principle: evolved parameters associate the fit phenotypes with a large number of genotypes. This provides a bridge between population genetics theory and optimality theories: a way to derive optimal parameter values that is rooted in population genetics theory.

Our approach builds on an analogy between population genetics and statistical physics (Sella and Hirsh, 2005; de Vladar and Barton, 2011), in particular the theory of free fitness developed by Iwasa (Iwasa, 1988) and Sella and Hirsh (Sella and Hirsh, 2005), who in turn built on Wright's equilibrium distributions (Wright, 1937) and Kimura's fixation probabilities (Kimura, 1962). Free fitness is a quantity that is maximized as evolution approaches equilibrium. We use this result to express a trade-off between fitness and genetic information, a measure of evolutionary constraint on the genotype (the difference between the genotype distribution under selection vs. under neutrality). For a given genotype-phenotype-fitness map, there is a minimal information required for any given expected fitness, with the population size controlling how much of each is maintained. Regulatory parameters are under selection to maximize free fitness.

This type of information is similar to the bioinformatics concept of the Information content (IC) of binding motifs (Schneider et al., 1986; Wasserman and Sandelin, 2004). IC quantifies, for a collection of TF binding sites, how different they are statistically from the rest of the genome (a comparison similar to that between distributions under selection vs. under neutrality, if selection drove the evolution of the binding sites). The requirement to localize TF binding to a precise site, i.e. to make that site unique within the genome, implies a necessary amount of IC. But a similar question can be asked about any regulatory task. How much information does a promoter need to contain so that it interacts with a given set of TFs, or expresses the gene in a given set of environments (Wagner, 2017)? We find that if selection is strong, the evolved architectures resemble optimal codes in information theory, reflecting the statistics of "messages", i.e. phenotypes selected for across the genome.

We find that the difference in free fitness between optimal and sub-optimal architectures can be dramatic, amounting to millions of bits across the whole genome for plausible parameters, due to the sheer number of TF-DNA interactions affected. The magnitude of such savings is critical. Optimization may require additional molecular mechanisms, which must also be encoded in the genome, and require an information overhead which must not exceed the savings. Savings of millions of bits imply that a number of additional regulatory genes (thousands of bits each) can be easily afforded. However, the comparisons of various architectures in terms of information can be done even if the precise molecular mechanisms involved are not known.

Moreover, quantification of the necessary information can be interesting in the context of debates about the functionality within genomes. Based on comparisons with related species, up to about 15% of the human DNA is under detectable selection (Ponting and Hardison, 2011; Rands et al., 2014), most of which is non-coding and must affect fitness via some regulatory processes. But a much larger fraction (up to about 80%) of the genome seems to have the capacity, or shows signs of involvement in such processes (Pheasant and Mattick, 2007; Mattick et al., 2010; The ENCODE Project Consortium et al., 2012; Kellis et al., 2014a). This has led to debates about technical problems and underappreciated evidence (Graur et al., 2013; Mattick and Dinger, 2013), the meaning of *function* (Doolittle, 2013; Brunet and Doolittle, 2014; Kellis et al., 2014b; Linquist, 2022) and insights from population genetics theory (Graur, 2017; Galeota-Sprung et al., 2020). One source of difficulties are the unknown and possibly complicated genotype-phenotype relationship of regulatory sequences. Long regions of DNA could perform simple functions that only weakly constrain their evolution. Quantification of the information necessary for some regulatory tasks, as we propose here, avoids the dichotomy between functional and non-functional, and can be done even without a complete knowledge of the GP maps involved: instead we ask what the optimal map would be.

We develop the theory in Sec. 4.2. Starting from basic population genetics assumptions, we derive an optimization principle for the regulatory parameters, and connect it to optimal coding in information theory. The assumptions are then questioned in Sec. 4.3, where we discuss how biophysical constraints on gene regulation affect the information requirements, and how the presence of strong mutation (neglected in Sec. 4.2) adds robustness as an additional factor in the optimization.

4.2 Theory: evolution of regulatory parameters

The theory contains three key ingredients. First, a class of population genetics models that is very general but still tractable at stationarity, introduced in Sec. 4.2.1. Second, an analysis analogous to energy-entropy trade-off in statistical physics, see Sec. 4.2.2. And third, an interpretation of this trade-off as an optimization principle relevant for regulatory systems, see Sec. 4.2.3. In Sec. 4.2.4 we point out a connection to optimal coding in information theory, which provides intuition about what the optimal architectures look like.

4.2.1 Population genetics setting

We track a population with effective size N that evolves under selection, mutation and drift. We assume the successive fixations regime, which requires a low enough mutation rate, $NL\mu \ll 1$, where L is the size of the genome considered. This limits the applicability of this theory, but we show in Sec. 4.3.3 that similar results hold also in genetically diverse populations.

The population state is given by the most recently fixed genotype (labeled $i, j, \dots$), and evolutionary stochasticity is captured by the distribution over these genotypes, ψ_i. Mutations are proposed at a rate μ_{ij} (from i to j), and accepted with the fixation probability (Kimura, 1962; Sella and Hirsh, 2005)

$$p_{ij}^{\text{fix}} = \frac{1 - e^{-2(w_j - w_i)}}{1 - e^{-2N(w_j - w_i)}}, \tag{4.1}$$

where w_i and w_j are log-fitness values. Together, μ_{ij} and p_{ij}^{fix} determine the substitution rates $\alpha_{ij} = \mu_{ij} p_{ij}^{\mathrm{fix}}$, which can be used to define a continuous time Markov chain,

$$\frac{d\psi_i}{dt} = \sum_{j \neq i} \left(\psi_j \alpha_{ij} - \psi_i \alpha_{ij} \right). \tag{4.2}$$

We make two further assumptions. First, the effective population size N and log-fitness landscape w_i are constant over time. Second, the mutation probabilities μ_{ij} satisfy detailed balance,

$$\frac{\mu_{ij}}{\mu_{ji}} = \frac{\varphi_j}{\varphi_i}, \tag{4.3}$$

where φ_i is the neutral stationary distribution over i, i.e. stationary solution to Eq. (4.2) if $w_i =$ const. for all genotypes i (analogous to the density of states in physics). Under the simplest model of only single nucleotide replacements and no mutation bias, φ_i will be uniform over all sequences of a given length.

Under these assumptions, ψ will converge to an equilibrium distribution (Wright, 1937; Berg et al., 2004; Sella and Hirsh, 2005; de Vladar and Barton, 2011),

$$\psi_i^{eq} = \frac{\varphi_i e^{2Nw_i}}{Z}, \tag{4.4}$$

with the partition sum $Z = \sum_i \varphi_i e^{2Nw_i}$. Selection increases the probability for genotypes with high fitness, and is more efficient in larger populations. This is analogous to the Boltzmann distribution in physics, with fitness instead of (negative) energy and $2N$ instead of inverse temperature.

This model, or its more specific instances, have been used in similar contexts of evolution of gene regulation (Berg et al., 2004; Sella and Hirsh, 2005; Lynch and Hagner, 2015). We introduce an extension to genetically diverse populations in Sec. 4.3.3 and discuss limitations of equilibrium-based theory in Sec. 4.4.

Particular systems at equilibrium are fully specified by the fitness landscape w_i (in this paper, motivated by gene regulation), the neutral distribution φ_i, and the population size N. Often, we can simplify the description by grouping all genotypes by their fitness. Log-fitness w occurs with probability $\varphi(w) = \sum_{i;\, w_i = w} \varphi_i$ under neutrality and $\varphi(w) e^{2Nw}/Z$ under selection.

In Fig. 4.1A, we introduce a simple binary model of TF binding. It consists of n promoter regions, some of which (nf) are under selection to bind a TF and the rest $(n(1-f))$ to avoid it. TF binding is treated as a binary phenotype, and each promoter with the wrong phenotype (failure to bind where wanted or ectopic binding where unwanted) incurs a log-fitness penalty s, such that the log-fitness is $w_i = -(n_i^{\mathrm{fail}} + n_i^{\mathrm{ect}})s$, where n_i^{fail} and n_i^{ect} are the counts of both kinds of error in genotype i. (Extensions with different s for positive and negative selection, or different for each promoter, are possible but not done here for simplicity.) We assume that all promoters have the same GP map, and that under neutrality, a single promoter binds the TF with a probability q (e.g. because a fraction q of all possible promoter sequences bind the TF). Therefore q parametrizes the regulatory architecture in this system. Under neutrality, n^{fail} and n^{ect} will both be binomial, $n_i^{\mathrm{fail}} \sim \mathrm{Bin}(nf, 1-q)$ and $n_i^{\mathrm{ect}} \sim \mathrm{Bin}(n(1-f), q)$. This allows us to compute the neutral distribution of log-fitness $\varphi(w)$, shown in gray in Fig. 4.1B. Under selection, the distribution $\psi(w)$ is pushed towards higher w, depending on the population size (red in Fig. 4.1B).

In summary, the example system consists of an architecture parametrized by the regulatory parameter q, and a phenotype-fitness map parametrized by s and f. This system is intentionally simple for illustration purposes. Later we introduce continuously varying binding probabilities, multiple TFs and non-promoter regions. Note, however, that this analysis does not depend on which promoter sequences bind the TF, or the mechanism of TF binding. Details such as promoter length, TF binding motif or possible cooperativity between binding sites are all absorbed into a single parameter of the architecture, q.

4.2.2 Fitness-information trade-off

As the population size N increases, selection becomes more efficient, and the population achieves a higher expected log-fitness $\langle w \rangle = \sum_i \psi_i w_i$. For this to happen, the distribution ψ must concentrate more and more probability at genotypes with high rather than low fitness. We can quantify this as *genetic information* D, using the Kullback-Leibler divergence (KL divergence) (Cover and Thomas, 2006)

$$D = \sum_i \psi_i \ln \frac{\psi_i}{\varphi_i}. \tag{4.5}$$

This quantifies how the distribution ψ differs from φ. If ψ and φ are equal, e.g. in absence of selection or before it had time to act, D is zero. Whenever ψ and φ differ, D is positive. In the important special case of uniform neutral distribution $\varphi_i =$ const., D is equal to the reduction in Shannon entropy of ψ compared to φ, $D = H(\varphi) - H(\psi)$, see Sec. 4.5.1.

Both $\langle w \rangle$ and D are plotted as function of population size in Fig. 4.1C for the TF binding example. The plots show the information in bits, $D/\ln 2$. Before developing the equilibrium theory, we point out a few more properties of the information D and its connection to broader literature.

The measure D is similar to the IC of binding motifs in bioinformatics (Schneider et al., 1986; Wasserman and Sandelin, 2004). IC focuses on the particular problem of localizing TF binding. It is mathematically similar to Eq. (4.5), but it is computed from nucleotide frequencies within experimentally discovered binding sites (e.g. by ChiP-seq) vs. across the genome, rather than probabilities under selection vs. neutrality. If the discovered binding sites were all under identical selection (to bind a TF, for example), and genome-wide nucleotide frequencies corresponded to neutrality, then IC would be a special case of D. IC has been used to study how much information is needed in a binding motif to localize it precisely within the genome (Schneider et al., 1986; Wunderlich and Mirny, 2009). The information D generalizes this approach: in any particular system, how much information is needed to achieve a given expected log-fitness?

A similar question was asked by Wagner (Wagner, 2017), who focused on binary phenotypes and defined an information measure as the log-ratio of all genotypes to the subset of those with a required phenotype. This corresponds to D if φ is uniform across all and ψ across the subset of genotypes. We also use binary examples for illustration purposes (except in Sec. 4.3.1), but D applies to any fitness landscape, notably when selection is finite and acts on continuously varying phenotypes.

Finally, we point out that D can also be computed or lower bounded using distributions over phenotypes. If genotype i has a conditional probability $\zeta_{z|i}$ of developing a phenotype z, then $\psi^Z_z = \sum_i \psi_i \zeta_{z|i}$ and $\varphi^Z_z = \sum_i \varphi_i \zeta_{z|i}$ are the distributions of z under selection and under neutrality. We then find the inequality

$$D \geq D^Z = \sum_z \psi_z^Z \ln \frac{\psi_z^Z}{\varphi_z^Z}. \tag{4.6}$$

In particular, if z is fully determined by the genotype ($z = z(i)$ and $\zeta_{z|i} = \delta_{z,z(i)}$), the bound is tight $D = D^Z$. This allows us to compute $D = D^W$ from fitness distributions such as in Fig. 4.1B. Further connections to literature, and an analysis of how selection accumulates and maintains information is in Chapter 3 (ref. (Hledik et al., 2022)).

We show the relationship between $\langle w \rangle$ and D in Fig. 4.1D. With increasing population size, the parametric curve takes us from relatively low fitness at zero information to high fitness and higher information. Crucially, the region below this curve is inaccessible: the equilibrium distribution in Eq. (4.4) has the lowest information D possible in this system at any given expected log-fitness $\langle w \rangle$, or conversely, the highest $\langle w \rangle$ at a given D.

This can be easily proved by showing that the equilibrium distribution maximizes a quantity called free fitness F,

$$\psi^{eq} = \underset{\psi}{\operatorname{argmax}}\, F(\psi), \qquad \text{where} \qquad F(\psi) = \langle w \rangle - \frac{1}{2N} D. \tag{4.7}$$

In the maximization problem, $1/(2N)$ can be seen as a Lagrange multiplier, constraining D while maximizing $\langle w \rangle$. Free fitness F (developed independently by Iwasa (Iwasa, 1988) and Sella and Hirsh (Sella and Hirsh, 2005)) is analogous to free energy in physics where a similar trade-off occurs between energy and entropy. In fact, F is not only maximized at equilibrium, but is a non-decreasing function of time before equilibrium is reached. We include a short proof in Sec. 4.5.2, and other proofs for a variety of other models are given in references (Iwasa, 1988; Sella and Hirsh, 2005; Mustonen and Lässig, 2010; Hledik et al., 2022). A related body of work is the maximum entropy approximation (Barton and de Vladar, 2009; Bod'ová et al., 2016) which approximates the evolution of quantitative traits by assuming quasi-equilibrium form of allele frequency distribution at all times.

From now on, all distributions ψ will be assumed to be at equilibrium, and we drop the superscript from ψ^{eq}. We note that at equilibrium, the free fitness can be expressed using the partition function, $F = \frac{\ln Z}{2N}$.

4.2.3 Optimization of regulatory parameters

The equilibrium theory implies a minimal D for any $\langle w \rangle$ for any specific system, but this minimal value will change if the regulatory architecture changes. In the TF binding toy model, its only parameter is q, the fraction of promoter sequences that bind the TF. We plot the fitness-information trade-off curve for several different values of q in Fig. 4.1E. Like other parameters of regulatory systems, q, is encoded elsewhere in the genome, and can evolve. For example, q could evolve via changes to the TF's own regulatory sequence (making the TF concentration higher or lower) or to the TF DNA-binding domain (making it bind more promiscuously or specifically). What can we say about evolution of q in our model?

Depending on the population size N, some values of q seem more advantageous than others. For example, in small populations, selection within promoters is ineffective, marked by D close to zero. It seems best to choose q such that random sequences have a high expected fitness $\langle w \rangle$. Large populations are practically certain to evolve the desired binding phenotype at each

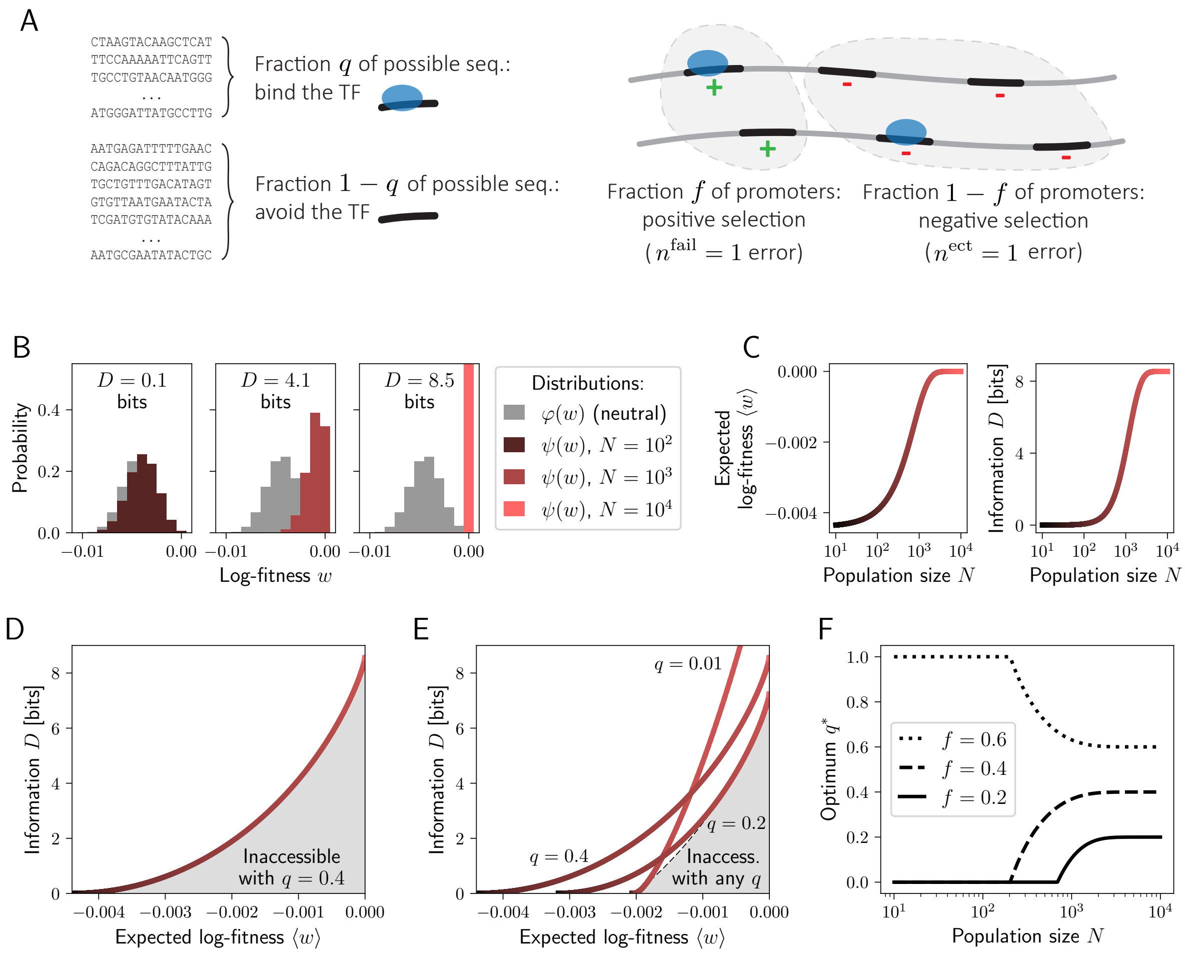

Figure 4.1: Illustration of the key concepts. (A) An example binary model of TF binding. Promoters (black lines) either bind or avoid a TF, depending on their sequence. Among all possible promoter sequences, a fraction q binds and $1-q$ avoids the TF. The genotype consists of n (here $n=6$) promoters, of which a fraction f is under positive selection (bind) and $1-f$ under negative selection (avoid). Each error reduces log-fitness by s. Here, one promoter under positive selection fails to bind the TF ($n^{\text{fail}} = 1$), and one promoter under negative selection binds the TF ectopically ($n^{\text{ect}} = 1$), implying $w = -2s$. (B) Distributions of log-fitness under neutrality and under selection at different effective population sizes. (C) Expected log-fitness and information both increase with increasing population size. (D) Relationship between the expected log-fitness $\langle w \rangle$ and information D. Equilibrium distributions have the lowest possible D at a given $\langle w \rangle$. (E) If the genotype-phenotype map changes (via parameter q), new combinations of $\langle w \rangle$ and D are unlocked. (E) Values of q^* that maximize free fitness, i.e. achieve best fitness-information trade-off, as function of N. The optimum depends on the form of selection, here parametrized by the fraction f of promoters under positive selection.

promoter. It seems best to choose q such that this can be achieved with as little information D as possible, i.e. to maximize the number of fit genotypes. If we can choose any q, a larger region becomes accessible in the $\langle w \rangle$, D diagram. The values of q that achieve the best trade-offs can be obtained by maximizing the free fitness F for each population size, and the optima are shown in Fig. 4.1F. In small populations, the best choice for q is to accommodate the most prevalent requirement (weighted, if needed, by selection strength). Since the fraction f of promoters is under selection for binding the TF and $1-f$ to avoid the TF, the optimum at low N is $q^* = 0$ for $f < 0.5$ and $q^* = 1$ for $f > 0.5$. At higher N, both phenotypes need to be represented in the sequence space, so that each promoter can adapt.

To answer this, we need to include the loci controlling q in the model. We will refer to them as *regulator loci* and label their genotype r, while the other, *target loci* keep the label i. Their joint equilibrium distribution is

$$\psi_{r,i} = \frac{\varphi_{r,i} e^{2Nw_{r,i}}}{Z} \tag{4.8}$$

We assume that the two sets of loci do not overlap, such that under neutrality, r and i are independent and $\varphi_{r,i} = \varphi_r \varphi_i$. We are especially interested in the evolution of the regulatory loci. The marginal distribution over r is

$$\psi_r \propto \sum_i \varphi_r \varphi_i e^{2Nw_{r,i}} = \varphi_r Z_r = \varphi_r e^{2NF_r} \tag{4.9}$$

where $Z_r = \sum_i \varphi_i e^{2Nw_{r,i}}$ is the partition function for the target loci conditional on a regulatory genotype r, and $F_r = \ln Z_r/(2N) = \langle w_r \rangle - D_r/(2N)$ is the corresponding free fitness. Therefore, while the equilibrium distribution of entire systems has log-fitness in the exponent (Eq. (4.4)), the marginal distribution of any subset of loci that interact with others has free fitness in the exponent. Regulator genotypes do not even have fitness by themselves, but only in combination with genotypes at the target loci. Advantage goes to regulator genotypes that achieve high fitness with a large number of target genotypes, and the interplay of fitness and numbers is captured by free fitness (cf. (Barton and Coe, 2009)).

Often, the interaction between the regulatory loci and their targets is only via a collection of regulatory parameters λ (such as $\lambda = \{q\}$), and we might not know or care how these are encoded in the genome. In that case, we can write log-fitness as $w_i(\lambda)$, conditional free fitness as $F(\lambda)$ and summarize the possibly complicated genetic architecture of λ as $\varphi(\lambda) = \sum_{r;\,\lambda(r)=\lambda} \varphi_r$. The equilibrium distribution of λ will be

$$\psi(\lambda) \propto \varphi(\lambda)\, e^{2NF(\lambda)}. \tag{4.10}$$

$\varphi(\lambda)$ expresses how common the value λ would be under neutrality, i.e. among random DNA sequences. But in large populations, or with many target loci, the factor $e^{2NF(\lambda)}$ will overwhelm it and λ will be strongly optimized for the free fitness of its targets.

This is illustrated for the toy model in Fig. 4.2. The panel A shows three examples of a possible neutral distribution $\varphi(q)$. In panel B, we increase the population size and observe how under selection on 100 target promoters, the distribution $\psi(q)$ converges to the same form. Note that as observed in Fig. 4.1F, the optimal q also changes with N. Beyond around $N = 10000$, further increases in N do not make a difference: all promoters are adapted to the required phenotype, and the distribution $\psi(q)$ is dominated simply by what fraction of possible genotypes realize the required combination of phenotypes.

Fig. 4.2CD shows the effect of increasing the number of target promoters, n. The conditional free fitness $F(\lambda)$ is additive across target promoters, and therefore the term $e^{2NF(\lambda)}$ not only overwhelms $\varphi(q)$ with large n, but also becomes increasingly peaked around an optimum. Therefore, unless this is physically impossible, we would expect that regulatory parameters interacting with many target loci are very strongly optimized at equilibrium.

Boltzmann-like types distributions have been used before to model biological optimization. For example, Berg and von Hippel (1987) derived it for TF binding site sequences as a maximum-entropy distribution with a constraint on the free energy of TF binding. A more general formulation of such approaches beyond genetics is in Chapter 2 (ref. (Młynarski et al., 2021)). On the other hand, equilibrium distributions such as Eq. (4.4) has been derived from population genetics theory by Wright (1937) (or later Sella and Hirsh (2005) in the low mutation regime studied here), providing a mechanistic explanation for this form. We argue that these approaches can be bridged: the distribution in Eq. (4.10) implies an optimization principle for λ, but is derived from population genetics.

If regulatory parameters are indeed optimized, this raises some questions: what form do the optimal architectures tend to have? And after optimization, how much information is still needed for adaptation?

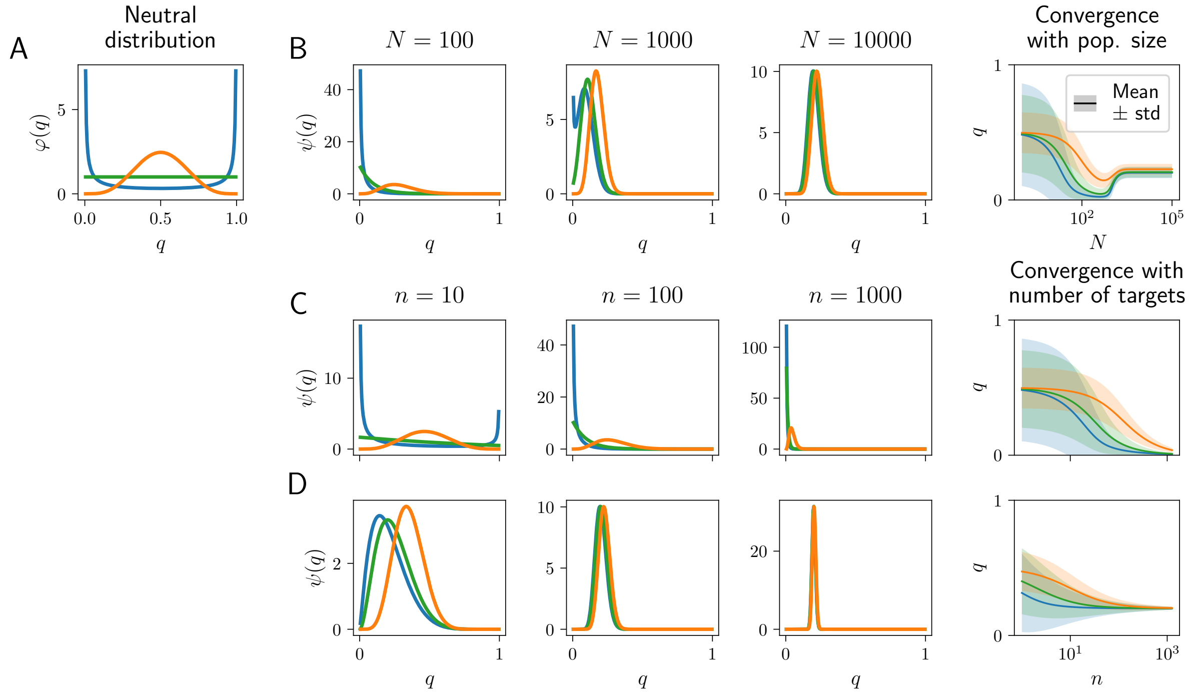

Figure 4.2: If a regulator has many targets, it will be strongly optimized. (A) Several neutral distributions of q, expressing the availability of q among all possible DNA sequences. (B) Distribution $\psi(q)$ under increasing population size. Other parameters: $n = 100, f = 0.2, s = 0.001$ (C,D) Distribution $\psi(q)$ under increasing number of target promoters for $N = 100$ (C) and $N = 10000$ (D). Other parameters: $f = 0.2, s = 0.001$.

4.2.4 Optimal architectures and the associated information

To develop intuition about the emergent architectures, we focus on a simple best-case scenario. We focus on the regime of large population size, where maximization of free fitness $F(\lambda) = \langle w(\lambda)\rangle - D(\lambda)/(2N)$ actually means minimization of $D(\lambda)$, since the expected log-fitness $\langle w(\lambda)\rangle$ is anyway at its maximum.

We assume discrete phenotypes z (such as binding/avoiding a TF), and n genomic regions, such as promoters or non-promoter windows. All of these share the same GP map where the phenotype z is realized with probability q_z under neutrality, i.e. by a fraction q_z of random sequences (parametrizing the genetic architecture). We allow all q_z values to be optimized, up to the constraint $\sum_z q_z = 1$ such that they do not overlap in the sequence space. Of the n regions, fractions f_z are forced by strong selection to evolve the phenotype z. How should be sequence space be divided among phenotypes, so that the required phenotypes are realized by the largest number of genotypes, i.e. with the lowest $D(\{q_z\})$?

Given q and f, the information is $D(\{q_z\}) = \sum_z nf_z \ln \frac{1}{q_z}$, since nf_z regions must realize z with probability 1, while under neutrality it would only be with probability q_z. The optimal partitions q_z^* are given by

$$\{q_z^*\} = \underset{\{q_z\};\, \sum_z q_z = 1}{\operatorname{argmin}} \sum_z nf_z \ln \frac{1}{q_z}. \tag{4.11}$$

This minimization problem is identical to the classical problem of optimal codes in information theory dating back to Shannon (Shannon, 1948). A source generates messages z with probabilities f_z, and each is assigned a code word that is l_z symbols long. We want to choose code words such that the average length, $\sum_z f_z l_z$ is minimized. In Eq. (4.11), $\ln(1/q_z)$ corresponds to the lengths l_z and the condition $\sum_z q_z = 1$ corresponds to the Kraft inequality (Cover and Thomas, 2006), a requirement for unique decodability of the original messages. (Kraft inequality is $\sum_z q_z \leq 1$ but equality is achieved at the optimum. Also, in information theory, e and the natural $\ln$ are usually replaced by 2 and $\log_2$ to express code length in bits). The solution is

$$q_z^* = f_z, \qquad \text{with minimal information} \qquad D(\{q_z^*\}) = n \sum_z f_z \ln \frac{1}{f_z} = nH(\{f_z\}), \tag{4.12}$$

where $H(\{f_z\})$ is the Shannon entropy of probabilities f_z. In other words, the fraction of sequence space devoted to each phenotype should match its desired frequency among the relevant genomic regions – we refer to this as *frequency matching*. And the minimal information to encode these phenotypes is given by the entropy of these frequencies. The DNA sequence of each region must encode a message about which phenotype should be realized there, and the information needed is proportional to the length of code needed to transmit such messages. In engineered systems we use information theory to devise codes that approach this minimum (Cover and Thomas, 2006), but apparently selection is similarly driving organisms to efficiently solve an analogous problem, by evolving novel regulatory mechanisms that approach a mathematically identical bound.

Rare messages, i.e. phenotypes such as strong TF binding that is required only in a minority of genomic regions, require a considerable amount of information to encode, $\ln \frac{1}{f_z} > 1$ per genomic region. In contrast, phenotypes that are common (e.g. do not bind here) take little information, $f_z \approx 1$ and $\ln \frac{1}{f_z} \approx 0$. Even across the genome, the most common phenotype require less information to encode than rare ones. However, it is the need to encode the common phenotype that drives the optimization towards devoting most sequence space to them, leaving less for the rare but nonetheless important phenotypes.

An example of frequency matching is in Fig. 4.1F, where at large N, optimal q matches the fraction f of promoters that must bind the TF. A real-world example is TF binding in prokaryotes. Each TF binds only in a small number of locations, and binding requires a correspondingly specific sequence – with the associated information depend on the size of the genome where the TF does not bind (Schneider et al., 1986; Wunderlich and Mirny, 2009). In contrast, nucleosomes cover the majority of eukaryotic genomes, and their localization relies on specific sequences they avoid rather than bind (Field et al., 2008). TFs in eukaryotes tend to have low information content compared to prokaryotes, likely because their binding is restricted by the chromatin landscape and multiple binding sites can be required for transcriptional regulation (Wunderlich and Mirny, 2009). Nonetheless, the information needed to arrange these more complex mechanisms for a desired outcome might follow the same coding principles.

4.3 Approximating theoretically optimal maps with biological mechanisms

The idealized scenario leading to the coding analogy, as well as the population genetics assumptions needed to derive the free fitness theory, depends on strong simplifications. In this section we address three important aspects of real systems. First, phenotypes are often not discrete, but quantitative or probabilistic (e.g., TF molecules bind DNA with some sequence and parameter-dependent probability). In Sec. 4.3.1, we show that this leads to *overspecification*: additional information is needed to guarantee or avoid binding with high probability. Second, if the phenotype z is highly multidimensional (e.g. if z is the binding status of many TFs), it might not be possible to independently tune all the probabilities q_z and achieve the optimum in Eq. (4.12). Nonetheless, suitable regulatory mechanisms can enable organisms to get closer to it, as we discuss in Sec. 4.3.2. Finally, we question a key assumption made in Sec. 4.2.1 that evolution proceeds by successive fixations in otherwise monomorphic populations. We outline how the results change in genetically diverse populations.

4.3.1 Biophysical constraints and overspecification

Real TF binding is not binary: a given genomic region can be bound with a probability that depends on the target sequence, TF properties, and its concentration. We extend our toy model by assuming that this probability p_B is a function of binding free energy ΔG,

$$p_B = \frac{1}{1+e^{\Delta G}}, \tag{4.13}$$

i.e. the system is at thermodynamic equilibrium. Under neutrality, we assume that a random promoter has ΔG drawn from a normal distribution with mean μ and standard deviation σ, e.g. because it is the sum of independent free energy contributions from a number of nucleotides (with a continuous approximation for simplicity). This model is illustrated in Fig. 4.3A. The transition from binding to not binding takes place approximately between $\Delta G = -5$ and 5, and depending on μ and σ, sequences may exist on either side as well as inside the transition range.

Selection still acts on the binding phenotype; now the log-fitness penalty is $p_B s$ for spurious binding and $(1 - p_B)s$ for failing to bind when needed. Either type of selection on individual promoters will restrict their distribution of ΔG to only sufficiently high and sufficiently low ΔG, see red and blue distributions in Fig. 4.3A.

Note that the sequences inside the transition range are eliminated by positive and negative selection alike. Intermediate binding such as $p_B = 0.5$ is never desirable in this model, and an optimal frequency-matching code would not assign it any genotypes. But with a unimodal distribution over ΔG, it is impossible to avoid intermediate binding without also eliminating either of the desirable phenotypes ($p_B \approx 1$ and $p_B \approx 0$). The next best solution would be to make σ very large, so that the vast majority of genotypes is far on either side of the transition, but σ is likely limited, for instance by the strength of molecular recognition interactions or the binding site length.

These biophysical constraints mean that achieving the same expected log-fitness now takes more information, Fig. 4.3B. This is an instance of *overspecification* (von Hippel and Berg, 1986) – more information is required than might be supposed from a simple combinatorial analysis. The optimal μ (Fig. 4.3C) behaves similarly to the binary model. In small populations, promoters cannot adapt anyway and the best choice is to accommodate the majority of promoters (which require no binding) and choose a very large μ. In large populations, intermediate values of μ can approximate frequency matching.

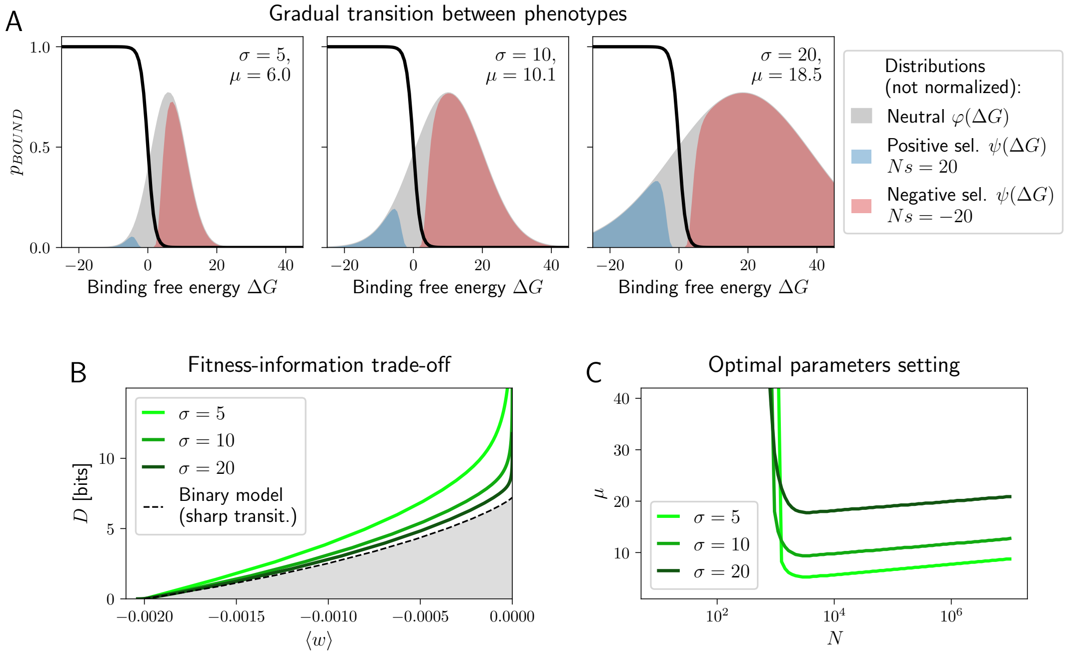

Figure 4.3: Biophysical constraints and overspecification. (A) TF occupancy depends on the binding free energy ΔG. This is normally distributed. (B) Fitness-information trade-offs achieved by optimal μ with various σ. For large σ, we approach the binary model, as fewer sequences have intermediate occupancies. (C) Optimal μ as function of N.

4.3.2 Availability of regulatory mechanisms

The optimal frequency-matching code may also be impossible because the available mechanisms of gene regulation may not have enough flexibility to individually adjust the probability q_z of each phenotype z. To illustrate this, we return to the binary model of TF binding, but consider a large number T of TFs, as well as a set of m non-promoter genomic regions (in addition to n promoters).

Each TF must bind some fraction f of promoters, and avoid the rest (subject to a penalty s per error). Non-promoter regions have the same GP map, but are selected to avoid all TF (penalty also s for simplicity). Under neutrality, any single TF binds to any promoter or non-promoter region with probability q, independently of other TFs. A schematic figure is in Fig. 4.4A.

Under neutrality, the chances for a promoter to bind the correct set of fT TFs and avoid the other $(1-f)T$ are $q^{fT}(1-q)^{(1-f)T}$. A non-promoter region successfully avoids all TFs with probability $(1-q)^T$. For a given value of q, we can again compute $\langle w \rangle$ and D as function of population size, black curves in Fig. 4.4B. The fitness-information trade-off is shown in Fig. 4.4C.

Note that other verbal descriptions lead to the same mathematical model. For example, instead of binding a single TF molecule, we can require several molecules binding a cluster binding sites, perhaps because they are needed to cooperatively initiate transcription. Instead of genomic regions binding different TFs, we could talk about these regions being transcribed under different conditions. In all these cases, we obtain a list of binary phenotypes, each realized by a fractions q of possible sequences, each required at a list of genomic windows parametrized by n, m and f, and a fitness penalty s for errors.

What is the optimal architecture in this situation? At low N, q^* is zero as before, Fig. 4.4D. At high N, the ideal solution would be frequency matching. Perfect frequency matching would divide the sequence space among the at most $n+1$ different phenotypes actually needed – n binding profile (at most one unique profile per promoter) and 1 *avoid all* phenotype for all the non-promoter regions. But this is certainly impossible to realize by varying the single parameter q. Even more complex but realistic models of gene regulation would not be able to prioritize the required set of TF combinations out of 2^T possible ones – these must be ultimately specified by the promoter sequences.

In our model, varying q can only lead to more or less binding overall, the optimal q is therefore $q^* = nf/(n+m)$, matching the fraction of all regions that need to bind any particular TF, and the necessary information is

$$D(q^*) = T(n+m)H\left(\frac{nf}{n+m}\right), \tag{4.14}$$

proportional to the number of TFs. This arrangement is very inefficient from the perspective of the non-promoter regions, which must avoid each TF separately. This does not necessarily mean that the total information in non-promoter regions is high. Instead, the need to avoid them means that the optimal q^* is low, leading to a high information requirements for binding TFs inside promoters. For example, with $q^* = 0.01$, about 82% of $D(q^*)$ is inside promoters (87% with $q^* = 0.001$ or 71% with $q^* = 0.1$).

Can organisms reduce the required information by evolving new regulatory mechanisms? This is possible by introducing an additional layer of regulation – a mechanism that leaves promoters open to TF binding, while non-promoters are prevented from binding all TFs in a single step. We refer to these two states as open and closed chromatin, but other molecular mechanisms may be involved instead or regulate the chromatin state – for example, a special class of pioneering TFs or DNA methylation (Bird and Wolffe, 1999; Deaton and Bird, 2011; Héberlé and Bardet, 2019).

Under neutrality, there is a probability q_{op} of being open and $1-q_{\mathrm{op}}$ of being closed. Therefore, promoters achieve the required phenotype with the probability $q_{\mathrm{op}}q^{fT}(1-q)^{(1-f)T}$. Non-promoter regions with probability $1-q_{\mathrm{op}}+q_{\mathrm{op}}(1-q)^T \approx 1-q_{\mathrm{op}}$ – either by being closed, or by being open but avoiding all TFs; but the latter option is negligible with an appreciable number of TFs. This leads to a considerable improvement in the fitness-information trade-off, visible in Fig. 4.4C. At high N, the optimal values are $q^*_{\mathrm{op}} = n/(n+m)$, matching the frequency of promoters that need to be open to binding, and $q^* = f$, matching the fraction of promoters that any TF needs to bind (Fig. 4.4D). The necessary information is

$$D_{\mathrm{chr}}(q^*_{\mathrm{op}}, q^*) = (n+m)H\left(\frac{n}{n+m}\right) + TnH(f). \tag{4.15}$$

The first term term corresponds to the information necessary to specify which regions are promoters, and this does not grow with the number of TFs. The second quantifies the specification of which promoters must be bound by each of the T TFs, and this is independent of the number of non-promoter regions.

The addition of a chromatin-like mechanism can lead to a dramatic reduction in the required information compared to the single-layer regulation in Eq. (4.14). We illustrate the difference in Fig. 4.4E for parameters roughly corresponding to the human and yeast genomes. We assume that the genome contains $n = 20000$ promoters (for humans; or $n = 6300$ for yeast), each $\ell = 2000$bp $(300$bp$)$ long. The rest of the genome consists of $L/\ell - n$ non-promoter regions, with genome size $L = 3.2 \times 10^9$bp $(1.2 \times 10^7$bp$)$. We assumed $T = 1500$ (300) TFs each binding a fraction $f = 0.1$ (0.1) of promoters.

In Fig. 4.4E, we varied the genome size L to show how the amount of non-promoter regions affects the information calculations. If TFs bind completely independently, the information grows with L, because binding must be more specific to avoid the aditional non-promoter regions. With chromatin, this growth practically disappears, because the first term in Eq. (4.15), corresponding to promoter/non-promoter specification, is negligible. At the actual values of L, the amount of information saved by chromatin is very large.

These estimates might be optimistic if some non-promoter regions need to be open in practice, reducing the advantage of chromatin. Also, implementing the additional level of regulation itself requires information: genes involved in chromatin regulation, and their own regulatory elements, must be encoded in the genome. Nonetheless, even if 10^4 bits were needed to encode each such additional gene, several can be afforded in yeast, and many in humans, from the available savings. In other words, random walks in the sequence space are more likely to stumble upon several entire chromatin-regulating genes and efficiently encoded TF binding, than inefficiently encoded TF binding. This suggests an economy of scale principle in gene regulation: with many genes and TFs, even a small efficiency improvement in encoding all their interactions can pay for encoding a new gene that implements it.

4.3.3 Robustness and theory extension to strong mutation regime

A major assumption of the theory in Sec. 4.2 is that evolution proceeds by successive fixations in otherwise monomorphic populations. This assumption is justified for short genomic elements, such that $NL\mu \ll 1$, where L is the size of the genomic region considered, and μ is the mutation rate per base pair. This may be sufficient for individual regulatory elements in small populations, but not, for example, when considering a large number of promoters as well as loci controlling their TF regulators. The low-mutation theory can also be applied separately to independently evolving loci – assuming no linkage and no epistasis between them. However, our key results about optimization depend on epistasis between a set of regulatory loci and a large number of target loci.

The strong mutation regime is more technically difficult to study, as the level of description must move from individual genotypes to genotype frequencies in diverse populations. Nonetheless, equilibrium distributions and free fitness were both originally developed on the population-level (Wright, 1937; Iwasa, 1988). The derivation for systems consisting of many multi-allelic loci at linkage equilibrium with arbitrary epistasis are clearly laid out in ref. (Mustonen and Lässig, 2010). The necessary assumption is that individual each locus is either in the low mutation regime or has a particularly simple form of mutation. We summarize the relevant results in Sec. 4.5.3.

In the weak mutation regime, it was sufficient to consider fixations of individual genotypes; as shown by Eq. (4.4), those with higher fitness were more likely to evolve. In contrast, under strong mutation, the success of any one genotype also depends on the success of its mutational neighbors, which occur in many of its descendants and ancestors. This causes a departure

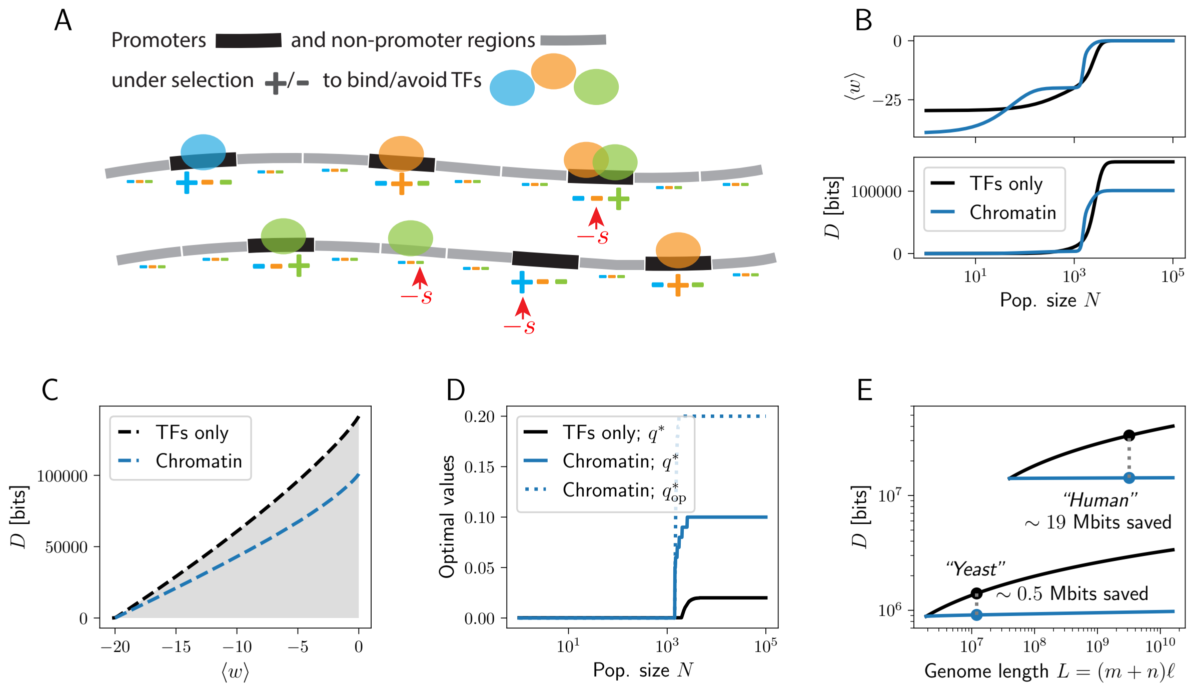

Figure 4.4: Novel regulatory mechanisms can improve coding efficiency. (A) An extended model with T TFs, n promoters and m non-promoter regions (here, $T = 3$, $n = 6$ and $m = 14$). Each promoter is under positive selection for a fraction f of promoters and negative selection for $1 - f$ (here $f = 1/3$), while all non-promoter regions are under negative selection only. Ectopic binding events, and failures to bind where it is required, reduce log-fitness by s (red arrows). (B) Expected log-fitness and information as function of N. Black curves are based on the model with TF binding only, and blue curves are based on the model which also includes open/closed chromatin states. Parameter values are $n = 2000$, $m = 8000$, $T = 100$, $f = 0.1$, $s = 10^{-3}$, $q = 0.01$ (TFs only) or $q = 0.1$ and $q_{\mathrm{op}} = 0.2$ (with chromatin). (C) Compared to a model with TFs only (black), chromatin regulation (blue) enables better trade-offs between fitness and information. Parameter values as in (B), but q and q_{op} use optimal values as shown in (D). (D) Parameter optima for the models with and without chromatin. Parameter values as in (B,C). (E) Comparison of information needed at strong selection, with and without chromatin. Parameter values inspired by the human and yeast genome (see text).

from Eq. (4.4) such that an additional advantage goes to fit genotypes with fit mutational neighbors, i.e. are mutationally robust van Nimwegen et al. (1999); Rao and Leibler (2022). Below we discuss this phenomenon from the perspective of equilibrium and free fitness theory.

On the population level, an equilibrium distribution $\tilde{\psi}(x)$ can be derived over x, a vector of allele frequencies. It has a Boltzmann-like form

$$\tilde{\psi}(x) \propto \tilde{\varphi}(x) e^{2N \ln \bar{\omega}(x)}, \tag{4.16}$$

where $\tilde{\varphi}(x)$ is the equilibrium in absence of selection, and $\bar{\omega}(x) = \sum_i x_i \omega_i$ is the mean fitness within within a population with frequencies x_i of genotypes x_i, and genotypic fitness values ω_i. This distribution maximizes a population-level free fitness,

$$\tilde{F} = \langle \ln \bar{\omega} \rangle - \frac{1}{2N} \tilde{D} \tag{4.17}$$

The term $\tilde{D}$ is again a type of genetic information, but on the population level – it quantifies how different the distribution over population states differs under selection, $\psi(x)$, and under neutrality, $\varphi(x)$. We can relate it to the information D defined in Eq. (4.5) for distributions over genotypes by computing the probability that a randomly sampled genotype from a randomly sampled population is i,

$$\psi_i = \int \psi(x) x_i dx \qquad \text{and} \qquad \varphi_i = \int \varphi(x) x_i dx, \tag{4.18}$$

or in other words, the expected genotype frequencies. It can be shown that the population-level information is larger than the genotype-level information,

$$\tilde{D} = D + \sum_i \psi_i \int \psi(x|i) \ln \frac{\psi(x|i)}{\varphi(x|i)} = D + D(X|G) \geq D \tag{4.19}$$

where $\psi(x|i) = \psi(x)x_i/\psi_i$ is the conditional distribution over population states, if a randomly sampled genotype was i. If $\psi(x)$ is taken as a Bayesian prior, then $\psi(x|i)$ is the posterior after looking at a randomly sampled genotype. The second term, $D(X|G)$, is a non-negative conditional KL divergence (Cover and Thomas, 2006), implying the inequality $\tilde{D} \geq D$.

The difference $D(X|G)$ vanishes when mutation is weak and the population is mostly monomorphic, because after we sample a single genotype i, it is practically certain that it is fixed in the population, regardless of selection ($\psi(x|i) = \varphi(x|i)$ both peaked at $x_i = 1$). In this regime, we can also replace the log mean-fitness $\ln \bar{\omega}(x)$ by log-fitness w_i of the fixed genotype i, recovering the mean fitness in Eq. (4.7). In contrast, when mutation is strong, $\psi(x|i)$ and $\varphi(x|i)$ can differ because populations are genetically diverse and selection can change the patterns of diversity (e.g. reduce it when acting against phenotypic variance). In such cases, $D(X|G)$ will be positive.

As in Sec. 4.2.3, we can consider a joint system of regulatory loci and their targets, and compute the marginal distribution over the allele frequencies x^R at the regulatory loci,

$$\psi(x^R) \propto \varphi(x^R) e^{2N\tilde{F}(x^R)}, \tag{4.20}$$

$\tilde{F}(x^R)$ is the free fitness of the target loci conditional on the regulator loci having allele frequencies x^R. Therefore, even in diverse populations, selection acts to optimize regulatory parameters by maximizing free fitness, but there are two differences.

First, Eq. (4.20) does not automatically translate to a distribution over regulatory parameter values as in Eq. (4.10), because the λ itself can have diverse values in the population, depending on the allele frequencies x^R at regulatory loci. Optimization still takes place, but in the space of population states, each with some variation in λ. If λ has a sharply peaked optimum (as in Fig. 4.2 with large N and many target loci), states x with too much diversity in λ will be suppressed. This will cause departures from Eq. (4.10) by favoring values of λ that are more robust to changes by mutation, or where such changes cause smaller reduction in free fitness.

Second, mutational robustness is also important for the target loci. Under successive fixations, free fitness expressed the need to achieve high fitness with a large number of genotypes. With stronger mutation, the focus shifts from genotypes to populations. Populations need to be kept at high log-*mean* fitness $\ln \bar{\omega}$, while minimizing the KL divergence from the neutral distribution over population states, $\tilde{D}$, which contains an additional term $D(X|G)$ reflecting the effects of selection on genetic variation, and depends on the robustness of occupied genotypes. If the genotype ψ_i is robust, mutations to it have little effect on fitness, the population is likely to have the same levels of diversity under selection as under neutrality, and the contribution $\int \psi(x|i) \ln \frac{\psi(x|i)}{\varphi(x|i)}$ will be low. On the other hand, if i sits on top of a sharp fitness peak or near a steep cliff, the inferior mutational neighbors are likely to be missing from the population under selection, even though they would have been present under neutrality. Minimization of $D(X|G)$ therefore favors robust genotypes, and the marginal genotype distribution ψ_i departs from the low-mutation equilibrium in Eq. (4.4). It also means that D is no longer minimized at a given $\langle w \rangle$, a phenomenon that could be seen as a case of robustness-associated overspecification.

Mutational robustness is a known factor in evolution in general (van Nimwegen et al., 1999; Wilke et al., 2001; Rao and Leibler, 2022) as well as in the context of gene regulation (Payne and Wagner, 2014, 2015). It is interesting to contrast it with the minimization of D. Mutational robustness is a local phenomenon, associated with the preference for genotypes with fit mutational neighbors (Rao and Leibler, 2022). In contrast, low information D is closely related to genotypic redundancy – a large number of genotypes with high fitness (Láruson et al., 2020), and it does not depend on *which* genotypes these are. Mutational robustness may emerge indirectly as a by-product of the minimization of D, if the genotypes with high fitness happen to be neighbors (Payne and Wagner, 2015), but if mutation is high, robustness becomes important in itself.

In summary, if mutation is high, the trade-off between fitness and information changes to a triple trade-off between fitness, information and mutational robustness. Regulatory parameters will evolve such that they associate high fitness with many, mutationally robust genotypes, and themselves become robust to mutation.

4.4 Discussion

Equilibrium population genetics suggests an optimization principle for regulatory parameters: they should be tuned such that across their target loci, high-fitness phenotypes are realized by as many random sequences as possible. An analogy with coding in information theory suggests what, approximately, the evolved regulatory architectures look like. Most of the sequence space will be devoted to the phenotypes required at many loci, and a lot of genetic information will be needed at loci that require unusual phenotypes. In ideal circumstances, the necessary information is given by the entropy associated with phenotype requirements across the genome, but this can be increased due to biophysical constraints or the need for robustness in high-mutation regimes.

The quantification of genetic information can also be used, at least under naive assumptions, to think about how much genetic information is needed for regulatory tasks. For example, evolving a specified set of TF-DNA interactions may require information that is proportional to the genome size and the number of TFs (Eq. (4.14)). However, this is dramatically reduced if an additional layer or regulation (e.g. closed chromatin) efficiently excludes all regions that should not bind any TFs (Eq. (4.15) and leaves open only a region scaling with the number of genes. In humans, this suggests order $2 \times 10^4 \times 1.5 \times 10^3 = 3 \times 10^7$ bits, plausible given that under 15% of the human genome (3×10^8 bp) is known to be selectively constrained (Ponting and Hardison, 2011; Rands et al., 2014).

We briefly comment on further limitations of the theory. First, the focus on stationary distributions may be problematic because approaching them takes too long to ever be relevant. This could be because selection changes on a faster time scale (e.g. due to ecological processes), or because populations might settle on local optima and never explore more of the sequence space. While this is indeed a serious limitation, our results might be nonetheless relevant.

TF binding sites are short and the difference between a functional binding site and random sequence is often only a few point mutations (Yona et al., 2018; Lagator et al., 2022), making exploration of the sequence space possible at least for individual regulatory elements. Also, even if is not reached, the information needed to achieve a phenotype can be an informative statistic. Phenotypes requiring less information can evolve faster ((Wagner, 2017), unpublished work by Reka Borbely, and Sec. 4.5.4) or be more likely to evolve when several alternatives with equal fitness are available. Regardless of the evolutionary scenario, it seems helpful for adaptation to have many genotypes that realize the high-fitness phenotypes, and future work can explore this in more detail.

Second, the existence of equilibrium distributions (associated with free fitness maximization) requires the assumption of detailed balance. In the low mutation regime, this restricts the form of mutation rates, excluding e.g. structural mutations such as duplications or deletions. In the high mutation regime, it further restricts us to scenarios with linkage equilibrium between small loci (with negligible standing variation at each), excluding also finite recombination rate. What happens outside these regimes is an open question. Structural mutation that produces repetitive sequences could, for example, lead to optimal architectures that employ such sequences for regulatory tasks, as seems to be the case with transposable elements (Trizzino et al., 2017). Methods from non-equilibrium statistical physics might be used to study these technically difficult situations.

A more conceptual problem is that optimal architectures can only be derived from assumptions about what selection is acting. If we knew what regulatory phenotypes are required along the genome, we could study the optimal maps and the necessary information – but the requirements are actually unknown. Instead, we typically make assumptions about selection based on what actually evolved.

4.5 Detailed calculations

4.5.1 Genetic information and entropy reduction

If the neutral distribution φ_i is uniform over all possible genotypes, $\varphi_i = 1/|G|$ where $|G|$ is the size of the genotype space, then the genetic information D can be rewritten as

$$D = \sum_i \psi_i \ln \frac{\psi_i}{1/|G|} = \ln|G| + \sum_i \psi_i \ln \psi_i, \tag{4.21}$$

where the two terms are equal to the Shannon entropy $H(\varphi)$ and (negative) Shannon entropy $H(\psi)$. If the distribution under selection is also uniform over a subset of genotypes G', i.e. $\psi_i = 1/|G'|$, then $H(\psi) = \ln|G'|$ and $D = \ln|G| - \ln|G'|$. This is the information formula used by Wagner (Wagner, 2017).

4.5.2 Free fitness is a non-decreasing function of time: low mutation regime

We start by taking a time derivative of F as defined in Eq. (4.7) and apply the CTMC dynamics from Eq. (4.2). After simplifying,

$$\frac{dF}{dt} = \sum_{ij;\, j\neq i} (\psi_j \alpha_{ij} - \psi_i \alpha_{ij}) \left(w_i - \frac{1}{2N} \ln \frac{\psi_i}{\varphi_i} \right) \tag{4.22}$$

$$= \sum_{ij;\, j\neq i} \psi_i \alpha_{ij} \left((w_j - w_i) - \frac{1}{2N} \ln \frac{\psi_j \varphi_i}{\psi_i \varphi_j} \right). \tag{4.23}$$

We can simplify further using the formula for fixation probabilities in Eq. (4.1) and detailed balance in Eq. (4.3),

$$\frac{dF}{dt} = \frac{1}{2N} \sum_{ij;\, j\neq i} \psi_i \alpha_{ij} \ln \frac{\psi_i \alpha_{ij}}{\psi_j \alpha_{ji}} \tag{4.24}$$

$$\geq 0, \tag{4.25}$$

where the inequality is proved using $\ln u \geq 1 - 1/u$.

4.5.3 Free fitness in genetically diverse populations

We model a population with n haploid loci, with locus l having m_l alleles. The state of the population is described by a list of frequencies x_{al} for allele a at locus l, normalized at each locus, $\sum_{a=1}^{m_l} x_{al} = 1$. Between loci, we assume linkage equilibrium, implying that the frequency of genotype $i = a_1, a_2, \ldots, a_n$ (consisting of an allele at each locus) is the product of frequencies of all alleles it contains, $x_i = x_{a_1 1} x_{a_2 2} \ldots x_{a_n n}$.

The population states have a distribution $\psi(x)$, and its dynamics under mutation, selection, free recombination and drift can be modeled with the diffusion equation,

$$\partial_t \psi(x) = -\sum_{l=1}^{l} \sum_{a=1}^{m_l - 1} \partial_{al} \left(m_{al}(x)\psi(x) + s_{al}(x)\psi(x) - \frac{1}{2N} \sum_{k=1}^{l} \sum_{b=1}^{m_k - 1} \partial_{bk} \left(b_{al,bk}(x)\psi(x) \right) \right). \tag{4.26}$$

We use short notation $\partial_t = \frac{\partial}{\partial t}$ and $\partial_{al} = \frac{\partial}{\partial x_{al}}$ for partial derivatives w.r.t. time and allele frequencies. The sums leave the last allele at each locus, because its frequency is determined by the frequencies of other alleles at the same locus. Genetic drift scales inversely with the effective population size N, and its covariance structure is given by

$$b_{al,bk}(x) = \begin{cases} x_{al}(1 - x_{al}); & a = b,\, l = k, \\ -x_{al} x_{bk}; & a \neq b,\, l = k, \\ 0; & l \neq k. \end{cases} \tag{4.27}$$

Mutation and selection enter via the expected changes in allele frequencies per generation, $m_{al}(x)$ and $s_{al}(x)$. These depend on the particular model, but for the diffusion process to have an equilibrium distribution, we require mutation and selection to have a gradient form,

$$s_{al}(x) = \sum_{k=1}^{l} \sum_{b=1}^{m_k - 1} b_{al,bk}(x)\, \partial_{bk} \ln \bar{\omega}(x) \tag{4.28}$$

$$m_{al}(x) = \sum_{k=1}^{l} \sum_{b=1}^{m_k - 1} b_{al,bk}(x)\, \partial_{bk} M(x). \tag{4.29}$$

For selection, the form in Eq. (4.28) comes naturally if we assign a fitness ω_i to each genotype i, allowing arbitrary epistasis. Selection is then controlled by the gradient of the logarithm of mean fitness, $\bar{\omega} = \sum_i x_i \omega_i$.

For mutation, the form in Eq. (4.29) emerges only under somewhat restrictive conditions. In general, we can parametrize it with mutation rates μ_{abl} from allele a to b at locus l, such that the $m_{al}(x) = \sum_{b=1}^{m_l}(x_{bl}\mu_{bal} - x_{al}\mu_{abl})$. Mustonen and Lassig (Mustonen and Lässig, 2010) indentified two situations when a suitable mutation potential M exists. One is when the mutation rates are independent of the source allele and depend only on the target allele, i.e. $\mu_{abl} = \tilde{\mu}_{bl}$, in which case $M = \sum_{l=1}^{n}\sum_{a=1}^{m_l}\tilde{\mu}_{al}\ln x_{al}$. The second situation requires a detailed balance condition at each locus, $\mu_{abl}/\mu_{bal} = \varphi_{bl}/\varphi_{al}$, but also that mutation rates are small enough at each locus ($N\ell\mu \ll 1$) so that at most two alleles per locus coexist in the population at the same time.

Under these conditions, the diffusion equation in Eq. (4.26) has an equilibrium solution (with zero probability flux),

$$\psi(x) = \frac{e^{2NM(x)}e^{2N\ln\bar{\omega}(x)}}{Z\prod_{l=1}^{n}\prod_{a=1}^{m_l}x_{al}} = \frac{\varphi(x)\,e^{2N\ln\bar{\omega}(x)}}{Z}, \tag{4.30}$$

analogous to Eq. (4.4), with $\varphi(x)$ being the equilibrium distribution under neutrality (U-shaped). This equilibrium distribution maximizes a population-level free fitness,

$$\tilde{F} = \int \psi(x)\ln\bar{\omega}(x)dx - \frac{1}{2N}\int \psi(x)\ln\frac{\psi(x)}{\varphi(x)}dx = \langle\ln\bar{\omega}\rangle - \frac{1}{2N}\tilde{D}, \tag{4.31}$$

similarly to Eq. (4.7). The population-level free fitness is also a non-decreasing function of time, as implied by the fitness flux theorem (Mustonen and Lässig, 2010).

4.5.4 Time to evolve a phenotype

While this paper focuses on stationary distributions, the concept of genetic information can also be helpful when considering the process of evolving a new genotype (Wagner, 2017).

This is easy to demonstrate in a very simple scenario. We assume only two phenotypes, such that the one favored by selection is realized by $|G'|$ out of $|G|$ possible genotypes; therefore $D = \ln|G| - \ln|G'|$ is required to satisfy selection. Assuming no other selection and sequential fixations regime, evolution will be a random walk in the genotype space, with a random step every $1/U$ generations on average, where U is the total mutation rate. Eventually, the favored phenotype will be proposed by a mutation. Then, if selection is strong ($Ns \gg 1$ where s is the log-fitness difference between the two phenotypes), it will be fixed. The time that this takes depends on where in the genotype space the $|G'|$ favored genotypes are located, and where the random walk starts. If, for simplicity, both are distributed uniformly, every proposed mutation has a probability of about $|G'|/|G|$ of finding the favored phenotype. Therefore, the expected time to evolve will be about $|G|/(U|G'|) = e^D/U$, exponential in the information D.

In more realistic settings, the $|G'|$ favored genotypes might be concentrated in a small part of the genotypes space, and if evolution starts far away, it may take longer for the random walk to reach it. On the other hand, in practice there may also be intermediate phenotypes with a smaller selective advantage leading up to the best phenotype. Then evolution can proceed faster by climbing up a fitness landscape rather than walking entirely blindly (Wilf and Ewens, 2010).

Acknowledgements. We thank Réka Borbély discussions. G.T. and M.H. acknowledge funding from the Human Frontier Science Program (HFSP RGP0032/2018). N.B. acknowledges funding from ERC grant 250152 “Information and Evolution”.

CHAPTER 5

A tight upper bound on mutual information

Abstract. We derive a tight lower bound on equivocation (conditional entropy), or equivalently a tight upper bound on mutual information between a signal variable and channel outputs. The bound is in terms of the joint distribution of the signals and maximum a posteriori decodes (most probable signals given channel output). As part of our derivation, we describe the key properties of the distribution of signals, channel outputs and decodes, that minimizes equivocation and maximizes mutual information. This work addresses a problem in data analysis, where mutual information between signals and decodes is sometimes used to lower bound the mutual information between signals and channel outputs. Our result provides a corresponding upper bound.

5.1 Introduction

The relationship between conditional entropy (equivocation) or mutual information, and best possible quality of decoding is an important concept in information theory. The best possible quality of a decoding scheme, when quantified by the minimal probability of error ϵ, does not uniquely determine the value of equivocation or mutual information, but various upper and lower bounds have been proved, see Sec. 5.1.1.

Here we discuss a scenario when not only ϵ, but the complete joint probability distribution $p(x, \hat{x})$ of signals x and maximum a posteriori decodes $\hat{x}$ is available. We refer to $p(x, \hat{x})$ as the confusion matrix. To our knowledge, such a scenario has not been extensively studied in the literature, despite having practical relevance for estimation of mutual information, as we point out in Sec. 5.1.2. In this article, we derive an upper bound on mutual information (and a corresponding lower bound on equivocation) that is based on the confusion matrix and is tighter than the known similar bound by Kovalevsky and others (Kovalevsky, 1968; Tebbe and Dwyer, 1968; Feder and Merhav, 1994) based on probability of error alone. The inequality in our bound can be proved quickly using the bound by Kovalevsky, as we show in Sec. 5.3.1. However, we also include a self-contained derivation in Sec. 5.4, where we construct the distribution of channel outputs that minimizes equivocation $H(X|Y)$ under our constraints.

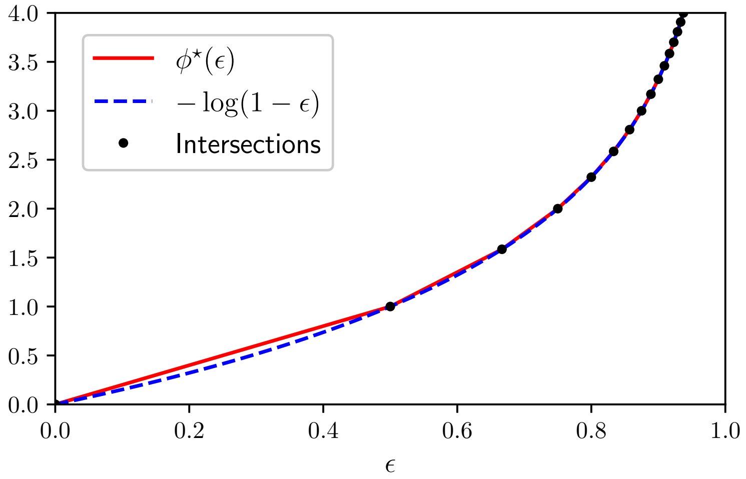

Figure 5.1: Plot of the functions $\phi^*(\epsilon)$ and $-\log(1-\epsilon)$. The two functions intersect at $\epsilon = 0$, $1/2$, $2/3$, $\dots$, $(|\mathcal{X}|-1)/|\mathcal{X}|$ (black dots), and in between $\phi^*(\epsilon)$ is piecewise linear.

5.1.1 Equivocation, mutual information and the minimal probability of error

We consider a signal variable (message) X that is communicated through a channel with output Y and then decoded, obtaining a "decode" $\hat{X}$ – forming a Markov chain $X \leftrightarrow Y \leftrightarrow \hat{X}$. The equivocation $H(X|Y)$ quantifies the uncertainty in X if the value of Y is given. Conversely, the mutual information $I(X;Y)$ measures how much information about X is contained in Y. It is not surprising that both $H(X|Y)$ and $I(X;Y)$ can be related to the minimal probability of error while decoding, $\epsilon = \Pr(X \neq \hat{X})$.

Accurate decoding, i.e., low ϵ, requires sufficiently low equivocation $H(X|Y)$. This is quantified by Fano's inequality (Cover and Thomas, 2006). The mutual information between the true signal and the channel output, $I(X;Y) = H(X) - H(X|Y)$, needs to be sufficiently high, and this is described by rate-distortion theory (Shannon, 1959).

Here we focus on the opposite bounds. If the minimal probability of error ϵ is specified, there is also a minimal possible equivocation. The following lower bound was derived for discrete X with finite support by Kovalevsky (Kovalevsky, 1968) and later Tebbe and Dwyer (Tebbe and Dwyer, 1968) and Feder and Merhav (Feder and Merhav, 1994) (see (Golic, 1999)). It reads

$$H(X|Y) \geq \phi^*(\epsilon), \tag{5.1}$$

where $\phi^*(\epsilon)$ is a piecewise linear function that coincides with $-\log(1-\epsilon)$ at points $\epsilon = 0$, $1/2$, $2/3$, $\dots$, $(|\mathcal{X}|-1)/|\mathcal{X}|$ (we use $\log = \log_2$ throughout the paper, and $\mathcal{X}$ is the support of X), and it can be written using the floor and ceiling functions,

$$\phi^*(\epsilon) = \alpha(\epsilon) \log \left\lfloor \frac{1}{1-\epsilon} \right\rfloor + (1-\alpha(\epsilon)) \log \left\lceil \frac{1}{1-\epsilon} \right\rceil, \tag{5.2}$$

$$\alpha(\epsilon) = \left\lfloor \frac{1}{1-\epsilon} \right\rfloor \left((1-\epsilon) \left\lceil \frac{1}{1-\epsilon} \right\rceil - 1 \right). \tag{5.3}$$

The function $\phi^*(\epsilon)$ is plotted in Fig. 5.1.

The bound Eq. (5.1) has been generalized to countably infinite support of X by Ho and Verdú (Ho and Verdú, 2010). Sason and Verdú (Sason and Verdú, 2018) proved a generalisation of Eq. (5.1) for Arimoto-Rényi conditional entropy of arbitrary order.

The bound Eq. (5.1) is tight when only the overall probability of error ϵ is available. However, when more constraints on the the joint distribution of X and Y are given, tighter bounds can be obtained. Prasad (Prasad, 2015) introduced two series of lower bounds on $H(X|Y)$ based on partial knowledge of the posterior distribution $p(x|y)$. The first is in terms of the k largest posterior probabilities $p(x|y)$ for each y, that we could label $p_1(y), p_2(y), \dots, p_k(y)$ in descending order (where $1 \leq k \leq |\mathcal{X}|$). The second series of bounds by Prasad is in terms of the averages of $p_1(y), p_2(y), \dots, p_k(y)$ across all y.

Hu and Xing (Hu and Xing, 2016) focused on a binary signal X and derived a bound tighter than Eq. (5.1) by taking into account the prior distribution of signals $p(x)$. Hu and Xing also discuss suboptimal (other than maximum a posteriori) decoding, which is otherwise rare in the related literature.

5.1.2 Motivation: estimation of mutual information

Here we extend the bound Eq. (5.1) to account for the situation when the complete confusion matrix – the joint distribution $p(x, \hat{x})$ is known. We are motivated by the following scenario: suppose that the goal is to estimate the mutual information $I(X;Y)$ from a finite set of (x, y) samples. Moreover, assume that the space of possible channel outputs $\mathcal{Y}$ is large (much larger than the space of signals, $|\mathcal{Y}| \gg |\mathcal{X}|$), making a direct calculation of $I(X;Y)$ by means of their joint distribution $p(x, y)$ infeasible due to insufficient sampling. In such a case, one approach (used e.g. in neuroscience (Borst and Theunissen, 1999)) is to construct a decoder, map each y into a decode $\hat{x}$ and estimate the confusion matrix $p(x, \hat{x})$. Then the post-decoding mutual information $I(X;\hat{X})$ can be calculated and used as a lower bound on $I(X;Y)$ due to the data processing inequality (Cover and Thomas, 2006). However, the gap between $I(X;\hat{X})$ and $I(X;Y)$ is not known (but see a discussion of this gap in (Samengo, 2002)), and an upper bound on $I(X;Y)$ based on $p(x, \hat{x})$ is desirable. Our result is such a bound, for the specific case of maximum a posteriori decoder.

While mutual information $I(X;Y)$ has this practical importance, we formulate our result as an equivalent lower bound on equivocation $H(X|Y) = H(X) - I(X;Y)$ first. This is simpler to state and prove.

5.2 Statement of the bound

Given the joint distribution $p(X, \hat{X})$ of signals X (discrete with finite support) and maximum a posteriori decodes $\hat{X}$ based on the channel output Y, the equivocation $H(X|Y)$ is bounded from below by

$$H(X|Y) \geq \sum_{\hat{x}} p(\hat{x})\, \phi^*(\epsilon_{\hat{x}}), \tag{5.4}$$

where $\epsilon_{\hat{x}} = p(X \neq \hat{X}|\hat{x}) = 1 - p(X = \hat{x}|\hat{X} = \hat{x})$ is the probability of error for the decode $\hat{x}$ and the function ϕ^* is defined in Eq. (5.2), Eq. (5.3).

Equivalently, we can bound the mutual information $I(X;Y)$ from above:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &\leq H(X) - \sum_{\hat{x}} p(\hat{x})\, \phi^*(\epsilon_{\hat{x}}). \end{aligned} \tag{5.5}$$

These bounds are tight, and we construct the distributions $p(y|\hat{x})$ and $p(x|y)$ that achieve equality in Sec. 5.4.

5.2.1 Comments on the bound

We note that since the function $\phi^*(\epsilon_{\hat{x}})$ is convex, we can apply Jensen's inequality to the right hand side of Eq. (5.4) and recover the bound Eq. (5.1) by Kovalevsky (Kovalevsky, 1968),

$$H(X|Y) \geq \phi^* \left(\sum_{\hat{x}} p(\hat{x})\, \epsilon_{\hat{x}} \right) = \phi^*(\epsilon). \tag{5.6}$$

Both bounds coincide in case of binary signal $|\mathcal{X}| = 2$, or any other case when the probability of error is less than $1/2$, $\epsilon_{\hat{x}} < 1/2$ for all $\hat{x}$. On this range, $\phi^*(\epsilon_{\hat{x}}) = 2\epsilon_{\hat{x}}$ and the bound simplifies to

$$H(X|Y) \geq 2 \sum_{\hat{x}} p(\hat{x})\, \epsilon_{\hat{x}} = 2\epsilon, \tag{5.7}$$

as has been noted in (Feder and Merhav, 1994) and before.

5.2.2 Example calculation

As an illustration, we apply our bound Eq. (5.4) to an example confusion matrix and compare it to the bound Eq. (5.1) that is in terms of error probability ϵ only.

The confusion matrix considered is depicted in Fig. 5.2 (A) for the case $|\mathcal{X}| = 5$. We vary the size $|\mathcal{X}|$ of the space of signals $\mathcal{X} = \{1, 2, \ldots, |\mathcal{X}|\}$, and the confusion matrix always takes the form

$$p(x, \hat{x}) = \begin{cases} \frac{1}{2|\mathcal{X}|}; & x = \hat{x} < |\mathcal{X}|, \\ \frac{1}{2|\mathcal{X}|}; & x < |\mathcal{X}|,\, \hat{x} = |\mathcal{X}|, \\ \frac{1}{|\mathcal{X}|}; & x = \hat{x} = |\mathcal{X}|, \\ 0; & x \neq \hat{x},\, \hat{x} < |\mathcal{X}|. \end{cases} \tag{5.8}$$

This distribution has the property that while most of the decodes have zero probability of being incorrect ($\epsilon_{\hat{x}} = 0$ for $\hat{x} < |\mathcal{X}|$), the last one has a high probability of being incorrect, $\epsilon_{\hat{x}} = (|\mathcal{X}| - 1)/(|\mathcal{X}| + 1)$ for $\hat{x} = |\mathcal{X}|$. Our bound Eq. (5.4) takes this into account – which makes it substantially tighter than the bound Eq. (5.1) based only on the overall probability of error ϵ. This can be seen in Fig. 5.2 (B), where both lower bounds are plotted. We also plot the post-decoding conditional entropy $H(X|\hat{X})$ which serves as the upper bound on the true value of $H(X|Y)$.

5.3 Proof of the bound

We offer two alternative proofs of the bound here. The first proves it as a simple consequence of the bound Eq. (5.1) by Kovalevsky. It is short, but it leaves open the question of tightness. We therefore focus on the second proof, which is self-contained, implies tightness and perhaps offers additional insights, since it includes a derivation of the distribution of channel outputs $p(y|\hat{x})$, $p(x|y)$ that minimizes $H(X|Y)$.

Throughout the proofs, the spaces of possible values of X and Y are written as $\mathcal{X}$ and $\mathcal{Y}$ respectively. The decoding function is denoted $g : \mathcal{Y} \to \mathcal{X}$ and is based on the maximum a posteriori rule, $g(y) \in \operatorname*{argmax}_x p(x|y)$. Finally, $\mathcal{Y}_{\hat{x}} = \{y \in \mathcal{Y} \,|\, g(y) = \hat{x}\}$ is the set of all y that decode into $\hat{x}$.

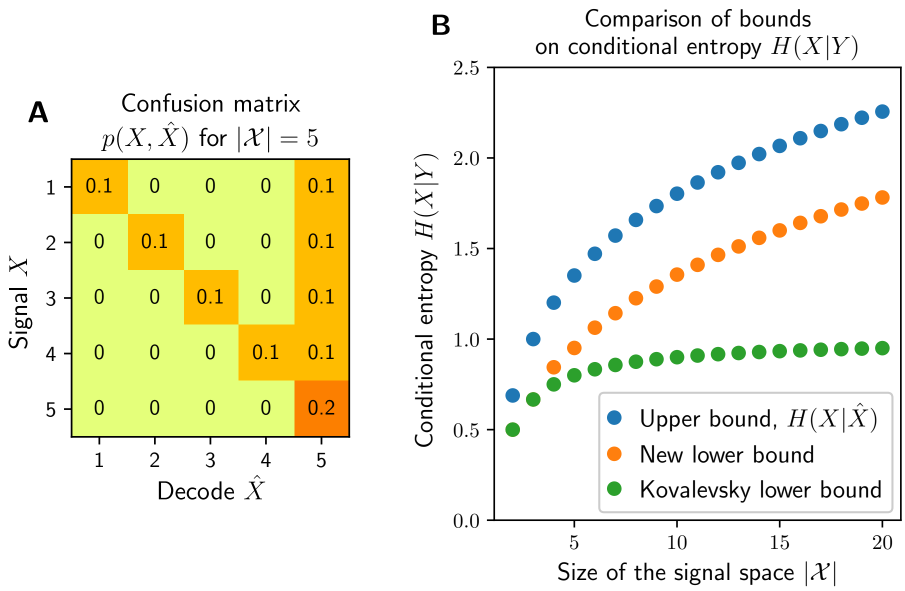

Figure 5.2: Example application of the bound. (A) The joint distribution of signals and decodes $p(x,\hat{x})$ for which we compute the bound, defined in Eq. (5.8). Here for the case $|\mathcal{X}|=5$. (B) Bounds on conditional entropy (equivocation) $H(X|Y)$ plotted for different sizes of signal space $|\mathcal{X}|$. $H(X|Y)$ is bounded from above by $H(X|\hat{X})$ (blue points). Our novel lower bound (Eq. (5.4)) is in orange and the bound by Kovalevsky (Eq. (5.1)) in green. Our bound Eq. (5.4) is the tightest possible given the confusion matrix.

5.3.1 A quick proof of inequality following Kovalevsky's bound

The left hand side of Eq. (5.4), the equivocation $H(X|Y)$ can be written as

$$H(X|Y) = \sum_{\hat{x}} p(\hat{x}) \int_{\mathcal{Y}_{\hat{x}}} H(X|Y=y)\, dp(y|\hat{x}), \tag{5.9}$$

where the term $\int_{\mathcal{Y}_{\hat{x}}} H(X|Y=y)\, dp(y|\hat{x})$ is the entropy of X conditional on Y, but with the values of Y only limited to $\mathcal{Y}_{\hat{x}}$. Since it has the form of conditional entropy, we can use the Kovalevsky bound Eq. (5.1) and obtain our result Eq. (5.4).

This establishes the inequality in our bound, but it does not tell us if equality can be achieved – and if it can, for what distribution of Y does it happen. We address this in the following section.

5.4 Proof by minimization of equivocation

For simplicity, we formulate the derivation for discrete Y. However, as we comment in Sec. 5.5, the derivation applies to continuous Y with only minor modifications. Two small steps of the proof are left out for brevity, but the reader will be referred to the preprint (Hledík et al., 2019) which includes these.

For clarity, let us state the minimization problem we are solving. We minimize

$$H(X|Y) = \sum_{\hat{x}} p(\hat{x}) \sum_{y\in\mathcal{Y}_{\hat{x}}} p(y|\hat{x}) H(X|Y=y) \tag{5.10}$$

with respect to $p(y|\hat{x})$ and $p(x|y)$, with the constraints given by the confusion matrix and maximum a posteriori decoding:

$$\forall x, \hat{x} : \qquad \sum_y p(x|y)p(y|\hat{x}) = p(x|\hat{x}), \tag{5.11}$$

$$\forall \hat{x}, \forall y \in \mathcal{Y}_{\hat{x}} : \qquad \hat{x} \in \operatorname*{argmax}_x p(x|y). \tag{5.12}$$

Note in Eq. (5.10) that the minimization can be done separately for each $\hat{x}$, since the corresponding $\mathcal{Y}_{\hat{x}}$ are disjoint. Hence we have $|\mathcal{X}|$ independent minimization problems with the objective function

$$\sum_{\mathcal{Y}_{\hat{x}}} p(y|\hat{x})H(X|Y = y). \tag{5.13}$$

Note also that we do not have any constraint on $|\mathcal{Y}|$, the number of elements of $\mathcal{Y}$. We actually exploit this flexibility in the proof. However, it turns out (see Propositions 1 and 2) that when the minimum is achieved, there can be only a limited number of y values with different distribution $p(x|y)$.

Our approach is based on update rules for $p(y|\hat{x})$ and $p(x|y)$ that decrease the objective function Eq. (5.13) while respecting the constraints Eq. (5.11), Eq. (5.12). In fact, the updates also change $|Y|$. The minimum of $H(X|Y)$ is achieved when the update rules can no longer be used to decrease it – and such situations can be characterized and the corresponding $H(X|Y)$ can be calculated.

It is instructive to have in mind the following visualization of our minimization problem, which we use to illustrate the update rules in Fig. 5.3. The distribution $p(x, y|\hat{x})$ for some $\hat{x}$, with y restricted to $y \in \mathcal{Y}_{\hat{x}}$ can be represented as a matrix, with a row for each x and a column for each y. Normalized columns correspond to $p(x|y)$ and the sum of each column is $p(y|\hat{x})$. The constraint Eq. (5.11) means that each row has a fixed sum, $p(x|\hat{x})$, and the constraint Eq. (5.12) means that one row (e.g. the first) contains the dominant elements of all columns. The objective function Eq. (5.13) is a weighted sum of entropies of all columns. Our minimization will consist of adding and removing columns, and moving probability mass within rows.

In the following, a probability distribution is called *flat* if all non-zero elements are equal, i.e. there are n non-zero elements and all have probabilities $1/n$. The number n is called its *length*.

Proposition 1: equivocation is minimized by flat $p(x|y)$

The minimum of the objective function Eq. (5.13), given constraints Eq. (5.11), Eq. (5.12) can only be achieved when the distributions $p(x|y)$ are flat for all y.

Proof. Suppose that there is a channel output y' with a non-flat distribution $p(x|y')$. Then, the following update rule, illustrated in Fig. 5.3 (A), will decrease the objective function Eq. (5.13).

We label the elements of $\mathcal{X}$ as $x_1, x_2, \ldots, x_{|\mathcal{X}|}$ such that

$$p(x_1|y') \geq p(x_2|y') \geq \cdots \geq p(x_{|\mathcal{X}|}|y') \geq 0, \tag{5.14}$$

where at least two of the inequalities are sharp (otherwise $p(x|y')$ would be flat). Note that x_1 must be the decode of y', i.e. $g(y') = \hat{x} = x_1$. The proposed update is to replace y' by $y'_1, y'_2, \dots, y'_{|\mathcal{X}|}$ with flat distributions $p(x|y'_i)$,

$$p(x_j|y'_i) = \begin{cases} 1/i; & j \leq i \\ 0; & j > i, \end{cases} \tag{5.15}$$

$$p(y'_i|\hat{x}) = \begin{cases} ip(y'|\hat{x})\left(p(x_i|y') - p(x_{i+1}|y')\right); & i < |\mathcal{X}| \\ ip(y'|\hat{x})\, p(x_i|y'); & i = |\mathcal{X}|. \end{cases} \tag{5.16}$$

Intuitively, this replaces y' by multiple elements y'_i with flat distributions $p(x|y'_i)$ covering the first $1, 2, \dots, |\mathcal{X}|$ elements of the ordered $x_1, x_2, \dots, x_{|\mathcal{X}|}$. It can be confirmed that this replacement respects the constraints Eq. (5.11). All y'_i still decode into $\hat{x} = x_1$, and the probability associated with y' is merely divided among the elements y'_i,

$$\sum_i p(y'_i|\hat{x}) = p(y'|\hat{x}), \tag{5.17}$$

$$\sum_i p(x_j|y'_i)p(y'_i|\hat{x}) = p(x_j|y')p(y'|\hat{x}). \tag{5.18}$$

See Fig. 5.3 for an example.

This replacement decreases the objective function Eq. (5.13). More detailed proof of this can be found in the preprint (Hledík et al., 2019), section IV.

The only case when the proposed replacement cannot be used to decrease the objective function is when $p(x|y)$ is flat for all y. Therefore flat $p(x|y)$ must be a characteristic of any solution to our minimization problem. □

Note that there are only $2^{|\mathcal{X}|-1}$ different possible flat distributions $p(x|y)$ with nonzero $p(X = \hat{x}|y)$, which means that we need at most $2^{|\mathcal{X}|-1}$ elements in $\mathcal{Y}_{\hat{x}}$ to achieve the minimum equivocation. However, as the following proposition will show, there are further restrictions on $p(x|y)$ at the minimum.

Reflecting that only flat $p(x|y)$ are of further interest in the minimization, we say that the channel output y has length l if $p(x|y)$ has length l.

Proposition 2: minimization restricts lengths of $p(x|y)$

Building on Proposition 1, we further claim that if equivocation is minimized, no two channel outputs $y_1, y_2 \in \mathcal{Y}_{\hat{x}}$ can have lengths differing by more than 1.

Proof. As before, we introduce an update rule. Recalling the visualization with a column for each y, this update rule will move a nonzero element from a longer column to a shorter column, as shown in Fig. 5.3 (B).

Take two elements $y_1, y_2 \in \mathcal{Y}_{\hat{x}}$ that have flat distributions $p(x|y_1)$ and $p(x|y_2)$ with lengths a and b respectively where $a > b$. Assume that a and b differ by more than one, $a - b > 1$. This means that we can choose an element $x' \in \mathcal{X}$ such that $p(x'|y_1) = 1/a$ and $p(x'|y_2) = 0$. Assume momentarily that $p(y_1|\hat{x})/a = p(y_2|\hat{x})/b$ (we will relax this assumption later). Then we can replace y_1, y_2 by y'_1 and y'_2, such that

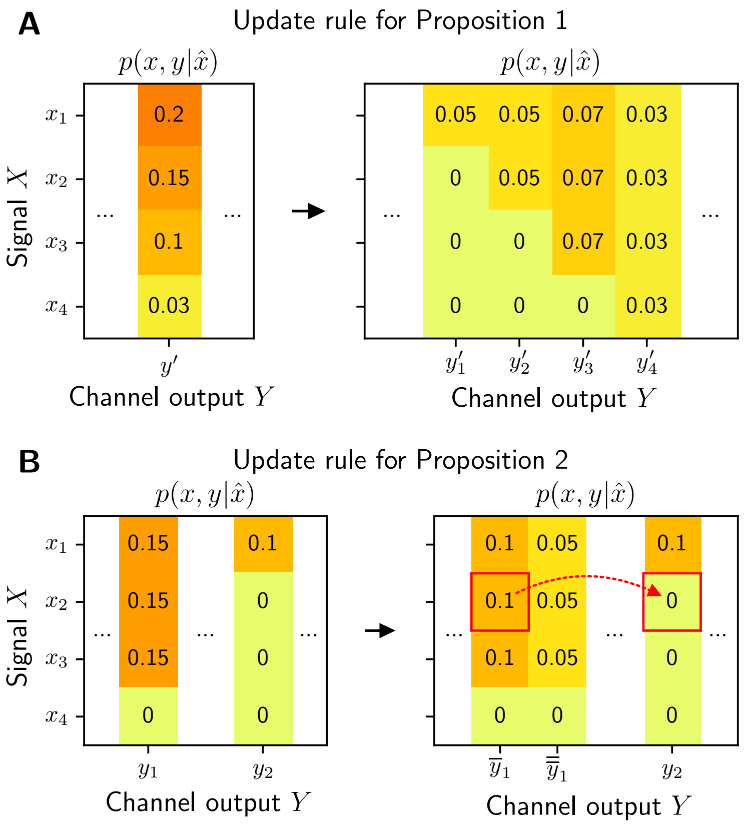

Figure 5.3: Illustrations of the update rules used to prove (A) Proposition 1 and (B) Proposition 2. Displayed is the joint distribution $p(x, y|\hat{x})$. (A) A channel output y' with a non-flat distribution $p(x|y')$ is replaced by $y'_1, y'_2, \ldots, y'_4$ with flat distributions $p(x|y'_i)$, such that $y'_1, y'_2, \ldots, y'_4$ still decode into x_1 and the confusion matrix is not affected. This replacement decreases $H(X|Y)$, our objective function. The elements of $\mathcal{X}$ are labeled in decreasing order of $p(x, y|\hat{x})$. (B) Two channel outputs, y_1 and y_2, have flat distributions $p(x|y_{1,2})$ with 3 and 1 nonzero elements respectively. We replace y_1 by $\overline{y}_1$ and $\overline{\overline{y}}_1$, and then transfer probability $p(x_2, \overline{y}_1|\hat{x})$ to $p(x_2, y_2|\hat{x})$ (dotted red arrow). The distributions $p(x|\overline{y}_1)$, $p(x|\overline{\overline{y}}_1)$ and $p(x|y_2)$ remain flat, and the objective function $H(X|Y)$ is decreased.

- $p(x|y'_1)$ is flat with length $a-1$. It is nonzero for the same x as $p(x|y_1)$, except for x' where it is zero.
- $p(x|y'_2)$ is flat with length $b+1$. It is nonzero for the same x as $p(x|y_2)$, and also for x'.

Given that $p(y_1|\hat{x})/a = p(y_2|\hat{x})/b$, we can also choose the probabilities $p(y'_1|\hat{x})$ and $p(y'_2|\hat{x})$ such that y'_1, y'_2 contribute the same amount to $p(x|\hat{x}) = \sum_y p(x|y)p(y|\hat{x})$ as y_1 and y_2 did, ensuring that constraints Eq. (5.11) are respected:

$$p(y'_1|\hat{x}) = \frac{a-1}{a} p(y_1|\hat{x}), \tag{5.19}$$

$$p(y'_2|\hat{x}) = \frac{b+1}{b} p(y_2|\hat{x}). \tag{5.20}$$

This update rule reduces the objective function Eq. (5.13), we show this in the extended version of the paper (Hledík et al., 2019) (section IV).

This update rule is applicable to any $y_1, y_2 \in \mathcal{Y}_{\hat{x}}$ with lengths a and b such that $a - b > 1$ respectively. We have, however, further required that $p(y_1|\hat{x})/a = p(y_2|\hat{x})/b$. This requirement can be avoided. If $p(y_1|\hat{x})/a > p(y_2|\hat{x})/b$, we first split y_1 into $\overline{y}_1$ and $\overline{\overline{y}}_1$ with

$$p(\overline{y}_1|\hat{x}) = a\, p(y_2|\hat{x})/b, \tag{5.21}$$

$$p(\overline{\overline{y}}_1|\hat{x}) = p(y_1|\hat{x}) - a\, p(y_2|\hat{x})/b, \tag{5.22}$$

$$p(x|\overline{y}_1) = p(x|\overline{\overline{y}}_1) = p(x|y_1), \tag{5.23}$$

such that the above mentioned update rule can be applied to $\overline{y}_1$ and y_2 while $\overline{\overline{y}}_1$ is left unchanged, see Fig. 5.3 (B). If $p(y_1|\hat{x})/a < p(y_2|\hat{x})/b$, we can proceed analogously by splitting y_2.

We can decrease the objective function by repeatedly applying this generalized update rule. Therefore, the minimum can only be achieved when the lengths of $p(x|y)$ for $y \in \mathcal{Y}_{\hat{x}}$ vary by no more than 1. □

Note that by repeated application of this update rule, in a finite number of steps we reach a state with only up to two lengths (per $\hat{x}$) that differ by at most 1. As shown in the next section, such a state implies a specific value of $H(X|Y)$. Together with the update rule in the proof of Proposition 1, this gives us an algorithm to find the distributions $p(y|\hat{x})$ and $p(x|y)$ that achieves the minimum $H(X|Y)$. The algorithm can start from an arbitrary initialization of $p(y|\hat{x})$ and $p(x|y)$ that follows the constraints Eq. (5.11), Eq. (5.12) and finishes in a finite number of steps.

It remains to be determined what are the (at most two) allowed lengths of $y \in \mathcal{Y}_{\hat{x}}$ and how the elements y with these lengths contribute to the equivocation $H(X|Y)$.

Admissible lengths of $p(x|y)$

Let us call the two admissible lengths $l_{\hat{x}}$ and $l_{\hat{x}} + 1$. Given $\hat{x}$, the total probability of all $y \in \mathcal{Y}_{\hat{x}}$ with length $l_{\hat{x}}$ is $\alpha_{\hat{x}}$, and those of length $l_{\hat{x}} + 1$ have probability $1 - \alpha_{\hat{x}}$. Then from the constraint Eq. (5.11), we can write the probability that $\hat{x}$ is the correct decode

$$1 - \epsilon_{\hat{x}} = \frac{\alpha_{\hat{x}}}{l_{\hat{x}}} + \frac{1 - \alpha_{\hat{x}}}{l_{\hat{x}} + 1}, \tag{5.24}$$

from which we can deduce that $\frac{1}{l_{\hat{x}}+1} \leq 1 - \epsilon_{\hat{x}} \leq \frac{1}{l_{\hat{x}}}$, and that the two admissible lengths must be

$$l_{\hat{x}} = \left\lfloor \frac{1}{1 - \epsilon_{\hat{x}}} \right\rfloor \text{ and } l_{\hat{x}} + 1 = \left\lceil \frac{1}{1 - \epsilon_{\hat{x}}} \right\rceil, \tag{5.25}$$

unless $\frac{1}{1-\epsilon_{\hat{x}}}$ is an integer – in that case the floor and ceiling coincide into a single admissible length.

Now, from equations Eq. (5.24) and Eq. (5.25) we can determine that

$$\alpha_{\hat{x}} = \left\lfloor \frac{1}{1 - \epsilon_{\hat{x}}} \right\rfloor \left((1 - \epsilon_{\hat{x}}) \left\lceil \frac{1}{1 - \epsilon_{\hat{x}}} \right\rceil - 1 \right) = \alpha(\epsilon_{\hat{x}}) \tag{5.26}$$

is the total probability (given $\hat{x}$) of $y \in \mathcal{Y}_{\hat{x}}$ with length $\lfloor \frac{1}{1-\epsilon_{\hat{x}}} \rfloor$.

Finally, the minimal value of equivocation is simply

$$H(X|Y) \geq \sum_{\hat{x}} p(\hat{x}) \left(\alpha_{\hat{x}} \log l_{\hat{x}} + (1 - \alpha_{\hat{x}}) \log \left(l_{\hat{x}} + 1 \right) \right), \tag{5.27}$$

which together with equations Eq. (5.25) and Eq. (5.26) constitutes our main bound, as stated in Eq. (5.4).

5.5 Discussion

We have introduced a tight lower bound on equivocation in terms of the maximum a posteriori confusion matrix, and proved it in two ways. The first is a proof of the inequality, starting from a similar bound by Kovalevsky (Kovalevsky, 1968), but it does not prove that the bound is tight. Therefore, we developed a second proof, in which we construct the distribution of channel outputs that minimizes the equivocation and achieves equality in our bound.

Central to the latter approach are two update rules for the distribution of the channel outputs. These update rules exploit the fact that equivocation can be, under our constraints, minimized by (1) making the posterior distributions $p(x|y)$ flat and (2) making sure that these flat distributions contain similar numbers of nonzero elements.

We formulated the proof for discrete random variables X and Y, but it can be extended. If X is discrete but Y continuous, application of a modified version of the first update rule would result in $2^{|\mathcal{X}|}$ regions in the $\mathcal{Y}_{\hat{x}}$ space corresponding to each of the $2^{|\mathcal{X}|}$ possible flat distributions $p(x|y')$. For example, the region associated with a flat distribution of length $|\mathcal{X}|$, that is $p(x|y') = 1/|\mathcal{X}|$, would have a total probability $\int_{\mathcal{Y}_{\hat{x}}} |\mathcal{X}| \min_x p(x|y) dp(y|\hat{x})$. These subsets of $\mathcal{Y}$ where $p(x|y)$ is constant can then be treated like discrete values, and the rest of our derivation applies.

Bounds on equivocation (or mutual information) in terms of the confusion matrix are, to our knowledge, not common – despite their relevance for estimation of mutual information. We hope that our result can be useful for these purposes, and that it sheds some light on the gap between mutual information before and after decoding. However, its applicability is restricted by the assumption of maximum a posteriori decoding, and relaxing this assumption remains an interesting challenge.

Acknowledgment

The authors would like to thank Sarah A. Cepeda-Humerez for helpful discussions.

Bibliography

R McNeill Alexander. *Principles of Animal Locomotion*. Princeton University Press, 2003.

John C. Baez and Blake S. Pollard. Relative entropy in biological systems. *Entropy*, 18(2), 2016. doi: 10.3390/e18020046.

H B Barlow. Possible Principles Underlying the Transformations of Sensory Messages. *Sensory communication*, 1961.

N H Barton. Linkage and the limits to natural selection. *Genetics*, 140(2), 1995. doi: 10.1093/genetics/140.2.821.

N. H. Barton. How does epistasis influence the response to selection? *Heredity*, 118(1), 2017. doi: 10.1038/hdy.2016.109.

N. H. Barton and J. B. Coe. On the application of statistical physics to evolutionary biology. *Journal of Theoretical Biology*, 259(2), 2009. doi: 10.1016/j.jtbi.2009.03.019.

N. H. Barton and H. P. de Vladar. Statistical mechanics and the evolution of polygenic quantitative traits. *Genetics*, 181(3), 2009. doi: 10.1534/genetics.108.099309.

N. H. Barton, A. M. Etheridge, and A. Véber. The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology*, 118, 2017. doi: 10.1016/j.tpb.2017.06.001.

Nick Barton and Tiago Paixão. Can quantitative and population genetics help us understand evolutionary computation? In *Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation*, GECCO '13, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 978-1-4503-1963-8. doi: 10.1145/2463372.2463568.

Nick Barton and Linda Partridge. Limits to natural selection. *BioEssays*, 22(12), 2000. doi: 10.1002/1521-1878(200012)22:12<1075::AID-BIES5>3.0.CO;2-M.

Richard Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics*, 6 (5), 1957.

Johannes Berg, Stana Willmann, and Michael Lässig. Adaptive evolution of transcription factor binding sites. *BMC Evolutionary Biology*, 4(1), 2004. doi: 10.1186/1471-2148-4-42.

Otto G. Berg and Peter H. von Hippel. Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters. *Journal of Molecular Biology*, 193(4), 1987. doi: 10.1016/0022-2836(87)90354-8.

Pietro Berkes and Laurenz Wiskott. Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of vision*, 5(6), 2005.

William Bialek. *Biophysics: Searching for Principles*. Princeton University Press, 2012.

William Bialek, Curtis G Callan, and Steven P Strong. Field theories for learning probability distributions. *Physical Review Letters*, 77(23), 1996.

Mark D. Biggin. Animal Transcription Networks as Highly Connected, Quantitative Continua. *Developmental Cell*, 21(4), 2011. doi: 10.1016/j.devcel.2011.09.008.

Adrian P. Bird and Alan P. Wolffe. Methylation-Induced Repression— Belts, Braces, and Chromatin. *Cell*, 99(5), 1999. doi: 10.1016/S0092-8674(00)81532-9.

Sean R Bittner, Agostina Palmigiano, Alex T Piet, Chunyu A Duan, Carlos D Brody, Kenneth D Miller, and John P Cunningham. Interrogating theoretical models of neural computation with deep inference. *bioRxiv*, 2019.

Katarína Bod'ová, Gašper Tkačik, and Nicholas H Barton. A general approximation for the dynamics of quantitative traits. *Genetics*, 202(4), 2016. doi: 10.1534/genetics.115.184127.

Bart G Borghuis, Charles P Ratliff, Robert G Smith, Peter Sterling, and Vijay Balasubramanian. Design of a neuronal array. *Journal of Neuroscience*, 28(12), 2008.

Alexander Borst and Frédéric E. Theunissen. Information theory and neural coding. *Nat Neurosci*, 2(11), 1999. doi: 10.1038/14731.

Evan A. Boyle, Yang I. Li, and Jonathan K. Pritchard. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169(7), 2017. doi: 10.1016/j.cell.2017.05.038.

Braden AW Brinkman, Alison I Weber, Fred Rieke, and Eric Shea-Brown. How do efficient coding strategies depend on origins of noise in neural circuits? *PLoS computational biology*, 12(10), 2016.

Sagie Brodsky, Tamar Jana, Karin Mittelman, Michal Chapal, Divya Krishna Kumar, Miri Carmi, and Naama Barkai. Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding Specificity. *Molecular Cell*, 79(3), 2020. doi: 10.1016/j.molcel.2020.05.032.

Tyler D. P. Brunet and W. Ford Doolittle. Getting "function" right. *PNAS*, 111(33), 2014. doi: 10.1073/pnas.1409762111.

Nicole L Carlson, Vivienne L Ming, and Michael Robert DeWeese. Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. *PLoS computational biology*, 8(7), 2012.

M Chalk, G Tkacik, and O Marre. Inferring the function performed by a recurrent neural network. *biorxiv*, 2019.

Matthew Chalk, Olivier Marre, and Gašper Tkačik. Toward a unified theory of efficient, predictive, and sparse coding. *Proceedings of the National Academy of Sciences*, 115(1), 2018.

Brian Charlesworth. Why we are not dead one hundred times over. *Evolution*, 67(11), 2013. doi: 10.1111/evo.12195.

Beth L Chen, David H Hall, and Dmitri B Chklovskii. Wiring optimization can relate neuronal structure and function. *Proceedings of the National Academy of Sciences*, 103(12), 2006.

Wei-Chia Chen, Ammar Tareen, and Justin B Kinney. Density estimation on small data sets. *Physical review letters*, 121(16), 2018.

Dmitri B Chklovskii. Exact solution for the optimal neuronal layout problem. *Neural computation*, 16(10), 2004.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2 edition, 2006. ISBN 978-0-471-24195-9.

Gavin E. Crooks. Path-ensemble averages in systems driven far from equilibrium. *Phys. Rev. E*, 61(3), 2000. doi: 10.1103/PhysRevE.61.2361.

James F. Crow. Shannon's brief foray into genetics. *Genetics*, 159(3), 2001.

Carl G. de Boer, Eeshit Dhaval Vaishnav, Ronen Sadeh, Esteban Luis Abeyta, Nir Friedman, and Aviv Regev. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nature Biotechnology*, 38(1), 2020. doi: 10.1038/s41587-019-0315-8.

Daniele De Martino, Anna MC Andersson, Tobias Bergmiller, Călin C. Guet, and Gašper Tkačik. Statistical mechanics for metabolic networks during steady state growth. *Nature Communications*, 9(1), 2018. doi: 10.1038/s41467-018-05417-9.

Harold P. de Vladar and Nicholas H. Barton. The contribution of statistical physics to evolutionary biology. *Trends in Ecology & Evolution*, 26(8), 2011. doi: 10.1016/j.tree.2011.04.002.

Aimée M. Deaton and Adrian Bird. CpG islands and the regulation of transcription. *Genes Dev.*, 25(10), 2011. doi: 10.1101/gad.2037511.

Stephane Deny, Ulisse Ferrari, Emilie Mace, Pierre Yger, Romain Caplette, Serge Picaud, Gašper Tkačik, and Olivier Marre. Multiplexed computations in retinal ganglion cells of a single type. *Nature communications*, 8(1), 2017.

Michael M Desai and Daniel S Fisher. Beneficial mutation–selection balance and the effect of linkage on positive selection. *Genetics*, 176(3), 2007. doi: 10.1534/genetics.106.067678.

Eizaburo Doi and Michael S Lewicki. A simple model of optimal population coding for sensory systems. *PLoS Comput Biol*, 10(8), 2014.

Eizaburo Doi, Jeffrey L Gauthier, Greg D Field, Jonathon Shlens, Alexander Sher, Martin Greschner, Timothy A Machado, Lauren H Jepson, Keith Mathieson, Deborah E Gunning, et al. Efficient coding of spatial information in the primate retina. *Journal of Neuroscience*, 32(46), 2012.

Alessandra Donato, Konstantinos Kagias, Yun Zhang, and Massimo A. Hilliard. Neuronal sub-compartmentalization: A strategy to optimize neuronal function. *Biological Reviews*, 94(3), 2019. doi: 10.1111/brv.12487.

Dawei W Dong and Joseph J Atick. Statistics of natural time-varying images. *Network: Computation in Neural Systems*, 6(3), 1995.

W. Ford Doolittle. Is junk DNA bunk? A critique of ENCODE. *PNAS*, 110(14), 2013. doi: 10.1073/pnas.1221376110.

J. H. Eaton and L. A. Zadeh. Optimal pursuit strategies in discrete-state probabilistic systems. *Journal of Basic Engineering*, 84(1), 1962. doi: 10.1115/1.3657260.

Jan Eichhorn, Fabian Sinz, and Matthias Bethge. Natural image coding in V1: How much use is orientation selectivity? *PLoS computational biology*, 5(4), 2009.

Manfred Eigen. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10), 1971. doi: 10.1007/BF00623322.

Anne Elk. My theory on brontosauruses. In *Monty Python's Flying Circus, the All-England Summarize Proust Competition*, volume 31, 1972.

W. J. Ewens. Remarks on the substitutional load. *Theoretical Population Biology*, 1(2), 1970. doi: 10.1016/0040-5809(70)90031-6.

M. Feder and N. Merhav. Relations between entropy and error probability. *IEEE Transactions on Information Theory*, 40(1), 1994. doi: 10.1109/18.272494.

Yair Field, Noam Kaplan, Yvonne Fondufe-Mittendorf, Irene K. Moore, Eilon Sharon, Yaniv Lubling, Jonathan Widom, and Eran Segal. Distinct Modes of Regulation by Chromatin Encoded through Nucleosome Positioning Signals. *PLOS Computational Biology*, 4(11), 2008. doi: 10.1371/journal.pcbi.1000216.

R. A. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(02), 1918. doi: 10.1017/S0080456800012163.

Ronald Aylmer Sir Fisher. *The Genetical Theory of Natural Selection.* Clarendon Press, Oxford, 1930. doi: 10.5962/bhl.title.27468.

Stephen J. Freeland and Laurence D. Hurst. The Genetic Code Is One in a Million. *J Mol Evol*, 47(3), 1998. doi: 10.1007/PL00006381.

Timothy Fuqua, Jeff Jordan, Maria Elize van Breugel, Aliaksandr Halavatyi, Christian Tischer, Peter Polidoro, Namiko Abe, Albert Tsai, Richard S. Mann, David L. Stern, and Justin Crocker. Dense and pleiotropic regulatory information in a developmental enhancer. *Nature*, 587(7833), 2020. doi: 10.1038/s41586-020-2816-5.

Benjamin Galeota-Sprung, Paul Sniegowski, and Warren Ewens. Mutational load and the functional fraction of the human genome. *Genome Biology and Evolution*, 12(4), 2020. doi: 10.1093/gbe/evaa040.

Rafael Galupa, Gilberto Alvarez-Canales, Noa Ottilie Borst, Timothy Fuqua, Lautaro Gandara, Natalia Misunou, Kerstin Richter, Mariana R. P. Alves, Esther Karumbi, Melinda Liu Perkins, Tin Kocijan, Christine A. Rushlow, and Justin Crocker. Enhancer architecture and chromatin accessibility constrain phenotypic space during Drosophila development. *Developmental Cell*, 58(1), 2023. doi: 10.1016/j.devcel.2022.12.003.

Deep Ganguli and Eero P Simoncelli. Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. *Neural computation*, 26(10), 2014.

Wilson S Geisler. Contributions of ideal observer theory to vision research. *Vision research*, 51 (7), 2011.

Andrew Gelman. Parameterization and bayesian modeling. *Journal of the American Statistical Association*, 99(466), 2004.

Julijana Gjorgjieva, Haim Sompolinsky, and Markus Meister. Benefits of Pathway Splitting in Sensory Coding. *J. Neurosci.*, 34(36), 2014. doi: 10.1523/JNEUROSCI.1032-14.2014.

Emily C Glassberg, Ziyue Gao, Arbel Harpak, Xun Lan, and Jonathan K Pritchard. Evidence for weak selective constraint on human gene expression. *Genetics*, 211(2), 2019. doi: 10.1534/genetics.118.301833.

Joshua I Gold and Michael N Shadlen. The neural basis of decision making. *Annual review of neuroscience*, 30, 2007.

J.D. Golic. Comment on "Relations between entropy and error probability". *IEEE Transactions on Information Theory*, 45(1), 1999. doi: 10.1109/18.746849.

S. J. Gould and R. C. Lewontin. The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 205(1161), 1979. doi: 10.1098/rspb.1979.0086.

Dan Graur. An upper limit on the functional fraction of the human genome. *Genome Biology and Evolution*, 9(7), 2017. doi: 10.1093/gbe/evx121.

Dan Graur, Yichen Zheng, Nicholas Price, Ricardo B.R. Azevedo, Rebecca A. Zufall, and Eran Elhaik. On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution*, 5(3), 2013. doi: 10.1093/gbe/evt028.

Paul E. Griffiths. Genetic Information: A Metaphor In Search of a Theory. *Philosophy of Science*, 68(3), 2001. doi: 10.1086/392891.

Ryan N Gutenkunst, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R Myers, and James P Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLoS computational biology*, 3(10), 2007.

David Haig and Laurence D. Hurst. A quantitative measure of error minimization in the genetic code. *J Mol Evol*, 33(5), 1991. doi: 10.1007/BF02103132.

J. B. S. Haldane. The effect of variation on fitness. *The American Naturalist*, 71(735), 1937. doi: 10.1086/280722.

J. B. S. Haldane. The cost of natural selection. *J Genet*, 55(3), 1957. doi: 10.1007/BF02984069.

Hiroshi Hasegawa. Thermodynamic properties of non-equilibrium states subject to Fokker-Planck equations. *Progress of Theoretical Physics*, 57(5), 1977. doi: 10.1143/PTP.57.1523.

Éléa Héberlé and Anaïs Flore Bardet. Sensitivity of transcription factors to DNA methylation. *Essays in Biochemistry*, 63(6), 2019. doi: 10.1042/EBC20190033.

Jody Hey. The neutralist, the fly and the selectionist. *Trends in Ecology & Evolution*, 14(1), 1999. doi: 10.1016/S0169-5347(98)01497-9.

W. G. Hill and Alan Robertson. The effect of linkage on limits to artificial selection. *Genetics Research*, 8, 1966.

Michal Hledík, Thomas R. Sokolowski, and Gašper Tkačik. A Tight Upper Bound on Mutual Information, 2019.

Michal Hledik, Nick H. Barton, and Gasper Tkacik. Accumulation and maintenance of information in evolution, 2022.

Siu-Wai Ho and Sergio Verdú. On the Interplay Between Conditional Entropy and Error Probability. *IEEE Transactions on Information Theory*, 56(12), 2010. doi: 10.1109/TIT.2010.2080891.

Bao-Gang Hu and Hong-Jie Xing. An Optimization Approach of Deriving Bounds between Entropy and Error from Joint Distribution: Case Study for Binary Classifications. *Entropy*, 18(2), 2016. doi: 10.3390/e18020059.

Aapo Hyvärinen, Jarmo Hurri, and Patrick O Hoyer. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.*, volume 39. Springer Science & Business Media, 2009.

RU Ibarra, JS Edwards, and BO Palsson. Escherichia coli k-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, 420, 2002.

Yoh Iwasa. Free fitness that always increases in evolution. *Journal of Theoretical Biology*, 135 (3), 1988. doi: 10.1016/S0022-5193(88)80243-1.

Edwin T Jaynes. *Probability Theory: The Logic of Science*. Cambridge university press, 2003.

E.T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9), 1982. doi: 10.1109/PROC.1982.12425.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007), 1946.

Jeffrey D. Jensen, Bret A. Payseur, Wolfgang Stephan, Charles F. Aquadro, Michael Lynch, Deborah Charlesworth, and Brian Charlesworth. The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. *Evolution*, 73(1), 2019. doi: 10.1111/evo.13650.

H Kacser and JA Burns. The control of flux. *Biochem Soc Trans*, 23, 1995.

Robert E Kass, Uri T Eden, and Emery N Brown. *Analysis of Neural Data*, volume 491. Springer, 2014.

David B Kastner, Stephen A Baccus, and Tatyana O Sharpee. Critical and maximally informative encoding between neural populations in the retina. *Proceedings of the National Academy of Sciences*, 112(8), 2015.

Manolis Kellis, Barbara Wold, Michael P. Snyder, Bradley E. Bernstein, Anshul Kundaje, Georgi K. Marinov, Lucas D. Ward, Ewan Birney, Gregory E. Crawford, Job Dekker, Ian Dunham, Laura L. Elnitski, Peggy J. Farnham, Elise A. Feingold, Mark Gerstein, Morgan C. Giddings, David M. Gilbert, Thomas R. Gingeras, Eric D. Green, Roderic Guigo, Tim Hubbard, Jim Kent, Jason D. Lieb, Richard M. Myers, Michael J. Pazin, Bing Ren, John Stamatoyannopoulos, Zhiping Weng, Kevin P. White, and Ross C. Hardison. Reply to Brunet and Doolittle: Both selected effect and causal role elements can influence human biology and disease. *PNAS*, 111(33), 2014b. doi: 10.1073/pnas.1410434111.

Manolis Kellis, Barbara Wold, Michael P. Snyder, Bradley E. Bernstein, Anshul Kundaje, Georgi K. Marinov, Lucas D. Ward, Ewan Birney, Gregory E. Crawford, Job Dekker, Ian Dunham, Laura L. Elnitski, Peggy J. Farnham, Elise A. Feingold, Mark Gerstein, Morgan C. Giddings, David M. Gilbert, Thomas R. Gingeras, Eric D. Green, Roderic Guigo, Tim Hubbard, Jim Kent, Jason D. Lieb, Richard M. Myers, Michael J. Pazin, Bing Ren, John A. Stamatoyannopoulos, Zhiping Weng, Kevin P. White, and Ross C. Hardison. Defining functional DNA elements in the human genome. *PNAS*, 111(17), 2014a. doi: 10.1073/pnas.1318948111.

Andrew D Kern and Matthew W Hahn. The Neutral Theory in light of natural selection. *Molecular Biology and Evolution*, 35(6), 2018. doi: 10.1093/molbev/msy092.

G. Kesidis and J. Walrand. Relative entropy between Markov transition rate matrices. *IEEE Transactions on Information Theory*, 39(3), 1993. doi: 10.1109/18.256516.

M. Kimura. Limitations of Darwinian selection in a finite population. *PNAS*, 92(6), 1995. doi: 10.1073/pnas.92.6.2343.

Motoo Kimura. Natural selection as the process of accumulating genetic information in adaptive evolution. *Genetics Research*, 2(1), 1961. doi: 10.1017/S0016672300000616.

Motoo Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47 (6), 1962. doi: 10.1093/genetics/47.6.713.

Motoo Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129), 1968. doi: 10.1038/217624a0.

Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, 1983. ISBN 978-0-521-31793-1.

Motoo Kimura and Takeo Maruyama. The mutational load with epistatic gene interactions in fitness. *Genetics*, 54(6), 1966.

Motoo Kimura, Takeo Maruyama, and James F. Crow. The mutation load in small populations. *Genetics*, 48(10), 1963.

Alexey S. Kondrashov. Deleterious mutations and the evolution of sexual reproduction. *Nature*, 336(6198), 1988. doi: 10.1038/336435a0.

Alexey S. Kondrashov. Contamination of the genome by very slightly deleterious mutations: Why have we not died 100 times over? *Journal of Theoretical Biology*, 175(4), 1995. doi: 10.1006/jtbi.1995.0167.

Fyodor A. Kondrashov and Alexey S. Kondrashov. Multidimensional epistasis and the disadvantage of sex. *PNAS*, 98(21), 2001. doi: 10.1073/pnas.211214298.

Vladimir A Kovalevsky. The problem of character recognition from the point of view of mathematical statistics. In *Character Readers and Pattern Recognition*. Spartan New York, 1968.

Mato Lagator, Srdjan Sarikas, Magdalena Steinrueck, David Toledo-Aparicio, Jonathan P Bollback, Calin C Guet, and Gašper Tkačik. Predicting bacterial promoter function and evolution from random sequences. *eLife*, 11, 2022. doi: 10.7554/eLife.64543.

Russell Lande. Genetic variation and phenotypic evolution during allopatric speciation. *The American Naturalist*, 116(4), 1980.

Áki J. Láruson, Sam Yeaman, and Katie E. Lotterhos. The Importance of Genetic Redundancy in Evolution. *Trends in Ecology & Evolution*, 35(9), 2020. doi: 10.1016/j.tree.2020.04.009.

Michael Lässig, Ville Mustonen, and Aleksandra M. Walczak. Predicting evolution. *Nature Ecology & Evolution*, 1(3), 2017. doi: 10.1038/s41559-017-0077.

Simon Laughlin. A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung c*, 36(9-10), 1981.

Stefan Linquist. Causal-role myopia and the functional investigation of junk DNA. *Biol Philos*, 37(4), 2022. doi: 10.1007/s10539-022-09853-2.

Alexander E. Lobkovsky and Eugene V. Koonin. Replaying the tape of life: Quantification of the predictability of evolution. *Front. Genet.*, 3, 2012. doi: 10.3389/fgene.2012.00246.

Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural dynamics. In *Advances in Neural Information Processing Systems*, 2017.

M. Lynch. *The Origins of Genome Architecture*. Sinauer Associates, 2007. ISBN 978-0-87893-484-3.

Michael Lynch and John S. Conery. The origins of genome complexity. *Science*, 302(5649), 2003. doi: 10.1126/science.1089370.

Michael Lynch and Kyle Hagner. Evolutionary meandering of intermolecular interactions along the drift barrier. *PNAS*, 112(1), 2015. doi: 10.1073/pnas.1421641112.

Benjamin B Machta, Ricky Chachra, Mark K Transtrum, and James P Sethna. Parameter space compression underlies emergent theories and predictive models. *Science*, 342(6158), 2013.

David John Cameron MacKay. 19. Why have sex? Information acquisition and evolution. In *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003a. ISBN 978-0-521-64298-9.

David John Cameron MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003b. ISBN 978-0-521-64298-9.

John S. Mattick and Marcel E. Dinger. The extent of functionality in the human genome. *The HUGO Journal*, 7(1), 2013. doi: 10.1186/1877-6566-7-2.

John S. Mattick, Ryan J. Taft, and Geoffrey J. Faulkner. A global view of genomic information – moving beyond the gene and the master regulator. *Trends in Genetics*, 26(1), 2010. doi: 10.1016/j.tig.2009.11.002.

John Maynard Smith. The Concept of Information in Biology. *Philosophy of Science*, 67(2), 2000. doi: 10.1086/392768.

Wiktor Młynarski. The opponent channel population code of sound location is an efficient representation of natural binaural sounds. *PLoS computational biology*, 11(5), 2015.

Wiktor Młynarski and Josh H McDermott. Learning midlevel auditory codes from natural sound statistics. *Neural computation*, 30(3), 2018.

Wiktor Młynarski, Michal Hledík, Thomas R. Sokolowski, and Gašper Tkačik. Statistical analysis and optimality of neural systems. *Neuron*, 109(7), 2021. doi: 10.1016/j.neuron.2021.01.020.

Wiktor F Młynarski and Ann M Hermundstad. Adaptive coding for dynamic sensory inference. *Elife*, 7, 2018.

Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. MIT press, 2012.

Ville Mustonen and Michael Lässig. Fitness flux and ubiquity of adaptive evolution. *PNAS*, 107(9), 2010. doi: 10.1073/pnas.0907953107.

Armita Nourmohammad and Ceyhun Eksin. Optimal evolutionary control for artificial selection on molecular phenotypes. *Phys. Rev. X*, 11(1), 2021. doi: 10.1103/PhysRevX.11.011044.

S. Ohno. So much "junk" DNA in our genome. *Brookhaven Symp Biol*, 23, 1972.

Timothy O'Leary, Alexander C Sutton, and Eve Marder. Computational models in the age of large datasets. *Current Opinion in Neurobiology*, 32, 2015. doi: 10.1016/j.conb.2015.01.006.

Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583), 1996.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research*, 37(23), 1997.

Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4), 2004.

Steven Hecht Orzack. *Adaptionism and Optimality*. Cambridge University Press, 2001.

Liam Paninski, Jonathan Pillow, and Jeremy Lewi. Statistical models for neural encoding, decoding, and optimal stimulus design. *Progress in brain research*, 165, 2007.

Il Memming Park and Jonathan W Pillow. Bayesian efficient coding. *bioRxiv*, 2017.

Mijung Park and Jonathan W Pillow. Receptive field inference with localized priors. *PLoS computational biology*, 7(10), 2011.

Dmitri S. Pavlichin, Yihui Quek, and Tsachy Weissman. Minimum power to maintain a nonequilibrium distribution of a Markov chain, 2019.

Joshua L. Payne and Andreas Wagner. The Robustness and Evolvability of Transcription Factor Binding Sites. *Science*, 343(6173), 2014. doi: 10.1126/science.1249046.

Joshua L. Payne and Andreas Wagner. Mechanisms of mutational robustness in transcriptional regulation. *Frontiers in Genetics*, 6, 2015.

Joel R. Peck and David Waxman. Is life impossible? Information, sex, and the origin of complex organisms. *Evolution*, 64(11), 2010. doi: 10.1111/j.1558-5646.2010.01074.x.

Alfonso Pérez-Escudero and Gonzalo G de Polavieja. Optimally wired subnetwork determines neuroanatomy of Caenorhabditis elegans. *Proceedings of the National Academy of Sciences*, 104(43), 2007.

Alfonso Pérez-Escudero, Marta Rivera-Alba, and Gonzalo G. de Polavieja. Structure of deviations from optimality in biological systems. *PNAS*, 106(48), 2009. doi: 10.1073/pnas.0905336106.

Michael Pheasant and John S. Mattick. Raising the estimate of functional human sequences. *Genome Res.*, 17(9), 2007. doi: 10.1101/gr.6406307.

Xaq Pitkow and Markus Meister. Decorrelation and efficient coding by retinal ganglion cells. *Nature neuroscience*, 15(4), 2012.

Chris P. Ponting and Ross C. Hardison. What fraction of the human genome is functional? *Genome Res.*, 21(11), 2011. doi: 10.1101/gr.116814.110.

Sudhakar Prasad. Bayesian Error-Based Sequences of Statistical Information Bounds. *IEEE Transactions on Information Theory*, 61(9), 2015. doi: 10.1109/TIT.2015.2457913.

Chris M. Rands, Stephen Meader, Chris P. Ponting, and Gerton Lunter. 8.2% of the human genome is constrained: Variation in rates of turnover across functional element classes in the human lineage. *PLOS Genetics*, 10(7), 2014. doi: 10.1371/journal.pgen.1004525.

Riccardo Rao and Stanislas Leibler. Evolutionary dynamics, evolutionary forces, and robustness: A nonequilibrium statistical mechanics perspective. *Proceedings of the National Academy of Sciences*, 119(13), 2022. doi: 10.1073/pnas.2112083119.

Charles P Ratliff, Bart G Borghuis, Yen-Hong Kao, Peter Sterling, and Vijay Balasubramanian. Retina is structured to process an excess of darkness in natural scenes. *Proceedings of the National Academy of Sciences*, 107(40), 2010.

Dario L Ringach. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of neurophysiology*, 88(1), 2002.

A. Robertson and Conrad Hal Waddington. A theory of limits in artificial selection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 153(951), 1960. doi: 10.1098/rspb.1960.0099.

Alan Robertson. Some optimum problems in individual selection. *Theoretical Population Biology*, 1(1), 1970. doi: 10.1016/0040-5809(70)90045-6.

Robert Rosen. *Optimality Principles in Biology*. Springer, 2013.

Rotem Ruach, Nir Ratner, Scott W. Emmons, and Alon Zaslaver. The synaptic organization in the Caenorhabditis elegans neural network suggests significant local compartmentalized computations. *Proceedings of the National Academy of Sciences*, 120(3), 2023. doi: 10.1073/pnas.2201699120.

Inés Samengo. Information Loss in an Optimal Maximum Likelihood Decoding. *Neural Computation*, 14(4), 2002. doi: 10.1162/089976602317318947.

Igal Sason and Sergio Verdú. Arimoto–Rényi Conditional Entropy and Bayesian M - Ary Hypothesis Testing. *IEEE Transactions on Information Theory*, 64(1), 2018. doi: 10.1109/TIT.2017.2757496.

Cristina Savin and Gasper Tkacik. Estimating nonlinear neural response functions using GP priors and Kronecker methods. In *Advances in Neural Information Processing Systems*, 2016.

Yonatan Savir, Elad Noor, Ron Milo, and Tsvi Tlusty. Cross-species analysis traces adaptation of Rubisco toward optimality in a low-dimensional landscape. *Proc Natl Acad Sci USA*, 107, 2010.

Thomas D. Schneider, Gary D. Stormo, Larry Gold, and Andrzej Ehrenfeucht. Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology*, 188(3), 1986. doi: 10.1016/0022-2836(86)90165-8.

Guy Sella and Aaron E. Hirsh. The application of statistical physics to evolutionary biology. *PNAS*, 102(27), 2005. doi: 10.1073/pnas.0501865102.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Claude E Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142-163), 1959.

Tatyana Sharpee and William Bialek. Neural decision boundaries for maximal information transmission. *PLoS One*, 2(7), 2007.

Tatyana O Sharpee. Computational identification of receptive fields. *Annual review of neuroscience*, 36, 2013.

Emmanuil E. Shnol, Elena A. Ermakova, and Alexey S. Kondrashov. On the relationship between the load and the variance of relative fitness. *Biology Direct*, 6(1), 2011. doi: 10.1186/1745-6150-6-20.

Evan C Smith and Michael S Lewicki. Efficient auditory coding. *Nature*, 439(7079), 2006.

Mandyam Veerambudi Srinivasan, Simon Barry Laughlin, and Andreas Dubs. Predictive coding: A fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 216(1205), 1982.

Yi Sun, Aljoscha Nern, Romain Franconville, Hod Dana, Eric R Schreiter, Loren L Looger, Karel Svoboda, Douglas S Kim, Ann M Hermundstad, and Vivek Jayaraman. Neural signatures of dynamic stimulus selection in Drosophila. *Nature neuroscience*, 20(8), 2017.

D. Tebbe and S. Dwyer. Uncertainty and the probability of error (Corresp.). *IEEE Transactions on Information Theory*, 14(3), 1968. doi: 10.1109/TIT.1968.1054135.

A Tero, S Takagi, T Saigusa, K Ito, DP Bebber, MD Fricker, K Yumiki, R Kobayashi, and T Nakagaki. Rules for biologically inspired adaptive network design. *Science*, 327, 2010.

The ENCODE Project Consortium, Ian Dunham, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R. Lajoie, Stephen G. Landt, Bum-Kyu Lee, Florencia Pauli, Kate R. Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shoresh, Jeremy M. Simon, Lingyun Song, Nathan D. Trinklein, Robert C. Altshuler, Ewan Birney, James B. Brown, Chao Cheng, Sarah Djebali, Xianjun Dong, Ian Dunham, Jason Ernst, Terrence S. Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C. Hardison, Robert S. Harris, Javier Herrero, Michael M. Hoffman, Sowmya Iyer, Manolis Kellis, Jainab Khatun, Pouya Kheradpour, Anshul Kundaje, Timo Lassmann, Qunhua Li, Xinying Lin, Georgi K. Marinov, Angelika Merkel, Ali Mortazavi, Stephen C. J. Parker, Timothy E.

Reddy, Joel Rozowsky, Felix Schlesinger, Robert E. Thurman, Jie Wang, Lucas D. Ward, Troy W. Whitfield, Steven P. Wilder, Weisheng Wu, Hualin S. Xi, Kevin Y. Yip, Jiali Zhuang, Bradley E. Bernstein, Ewan Birney, Ian Dunham, Eric D. Green, Chris Gunter, Michael Snyder, Michael J. Pazin, Rebecca F. Lowdon, Laura A. L. Dillon, Leslie B. Adams, Caroline J. Kelly, Julia Zhang, Judith R. Wexler, Eric D. Green, Peter J. Good, Elise A. Feingold, Bradley E. Bernstein, Ewan Birney, Gregory E. Crawford, Job Dekker, Laura Elnitski, Peggy J. Farnham, Mark Gerstein, Morgan C. Giddings, Thomas R. Gingeras, Eric D. Green, Roderic Guigó, Ross C. Hardison, Timothy J. Hubbard, Manolis Kellis, W. James Kent, Jason D. Lieb, Elliott H. Margulies, Richard M. Myers, Michael Snyder, John A. Stamatoyannopoulos, Scott A. Tenenbaum, Zhiping Weng, Kevin P. White, Barbara Wold, Jainab Khatun, Yanbao Yu, John Wrobel, Brian A. Risk, Harsha P. Gunawardena, Heather C. Kuiper, Christopher W. Maier, Ling Xie, Xian Chen, Morgan C. Giddings, Bradley E. Bernstein, Charles B. Epstein, Noam Shoresh, Jason Ernst, Pouya Kheradpour, Tarjei S. Mikkelsen, Shawn Gillespie, Alon Goren, Oren Ram, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael J. Coyne, Timothy Durham, Manching Ku, Thanh Truong, Lucas D. Ward, Robert C. Altshuler, Matthew L. Eaton, Manolis Kellis, Sarah Djebali, Carrie A. Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K. Marinov, Jainab Khatun, Brian A. Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Philippe Batut, Ian Bell, Kimberly Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jackie Dumais, Radha Duttagupta, Megan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha P. Gunawardena, Cédric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Guoliang Li, Oscar J. Luo, Eddie Park, Jonathan B. Preall, Kimberly Presaud, Paolo Ribeca, Brian A. Risk, Daniel Robyr, Xiaoan Ruan, Michael Sammeth, Kuljeet Singh Sandhu, Lorain Schaeffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Timothy J. Hubbard, Alexandre Reymond, Stylianos E. Antonarakis, Gregory J. Hannon, Morgan C. Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigó, Thomas R. Gingeras, Kate R. Rosenbloom, Cricket A. Sloan, Katrina Learned, Venkat S. Malladi, Matthew C. Wong, Galt P. Barber, Melissa S. Cline, Timothy R. Dreszer, Steven G. Heitner, Donna Karolchik, W. James Kent, Vanessa M. Kirkup, Laurence R. Meyer, Jeffrey C. Long, Morgan Maddren, Brian J. Raney, Terrence S. Furey, Lingyun Song, Linda L. Grasfeder, Paul G. Giresi, Bum-Kyu Lee, Anna Battenhouse, Nathan C. Sheffield, Jeremy M. Simon, Kimberly A. Showers, Alexias Safi, Darin London, Akshay A. Bhinge, Christopher Shestak, Matthew R. Schaner, Seul Ki Kim, Zhuzhu Z. Zhang, Piotr A. Mieczkowski, Joanna O. Mieczkowska, Zheng Liu, Ryan M. McDaniell, Yunyun Ni, Naim U. Rashid, Min Jae Kim, Sheera Adar, Zhancheng Zhang, Tianyuan Wang, Deborah Winter, Damian Keefe, Ewan Birney, Vishwanath R. Iyer, Jason D. Lieb, Gregory E. Crawford, Guoliang Li, Kuljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Oscar J. Luo, Atif Shahab, Melissa J. Fullwood, Xiaoan Ruan, Yijun Ruan, Richard M. Myers, Florencia Pauli, Brian A. Williams, Jason Gertz, Georgi K. Marinov, Timothy E. Reddy, Jost Vielmetter, E. Partridge, Diane Trout, Katherine E. Varley, and Clarke Gasper. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 2012. doi: 10.1038/nature11247.

Evangelos A. Theodorou. Nonlinear stochastic control and information theoretic dualities:

Connections, interdependencies and thermodynamic interpretations. *Entropy*, 17(5), 2015. doi: 10.3390/e17053352.

C. A. Thomas. The Genetic Organization of Chromosomes. *Annual Review of Genetics*, 5(1), 1971. doi: 10.1146/annurev.ge.05.120171.001321.

N. Tinbergen. On aims and methods of Ethology. *Zeitschrift für Tierpsychologie*, 20(4), 1963. doi: 10.1111/j.1439-0310.1963.tb01161.x.

Gašper Tkačik and William Bialek. Information processing in biological systems. *Annu Rev Cond Matt Phys*, 7, 2016.

Gašper Tkačik, Curtis G Callan, and William Bialek. Information flow and optimization in transcriptional regulation. *Proceedings of the National Academy of Sciences*, 105(34), 2008.

Gašper Tkačik, Jason S Prentice, Vijay Balasubramanian, and Elad Schneidman. Optimal population coding by noisy spiking neurons. *Proceedings of the National Academy of Sciences*, 107(32), 2010.

Emanuel Todorov. Linearly-solvable Markov decision problems. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

Marco Trizzino, YoSon Park, Marcia Holsbach-Beltrame, Katherine Aracena, Katelyn Mika, Minal Caliskan, George H. Perry, Vincent J. Lynch, and Christopher D. Brown. Transposable elements are the primary source of novelty in primate gene regulation. *Genome Res.*, 27 (10), 2017. doi: 10.1101/gr.218149.116.

Eeshit Dhaval Vaishnav, Carl G. de Boer, Jennifer Molinet, Moran Yassour, Lin Fan, Xian Adiconis, Dawn A. Thompson, Joshua Z. Levin, Francisco A. Cubillos, and Aviv Regev. The evolution, evolvability and engineering of gene regulatory DNA. *Nature*, 603(7901), 2022. doi: 10.1038/s41586-022-04506-6.

J Hans van Hateren and Dan L Ruderman. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1412), 1998.

J Hans Van Hateren and Arjen van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1394), 1998.

Johannes H van Hateren. Theoretical predictions of spatiotemporal receptive fields of fly LMCs, and experimental validation. *Journal of Comparative Physiology A*, 171(2), 1992.

Erik van Nimwegen, James P. Crutchfield, and Martijn Huynen. Neutral evolution of mutational robustness. *Proceedings of the National Academy of Sciences*, 96(17), 1999. doi: 10.1073/pnas.96.17.9716.

P H von Hippel and O G Berg. On the specificity of DNA-protein interactions. *Proceedings of the National Academy of Sciences*, 83(6), 1986. doi: 10.1073/pnas.83.6.1608.

Andreas Wagner. Information theory, evolutionary innovations and evolvability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1735), 2017. doi: 10.1098/rstb.2016.0416.

Zhuo Wang, Alan A Stocker, and Daniel D Lee. Efficient neural codes that minimize lp reconstruction error. *Neural computation*, 28(12), 2016.

Wyeth W. Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4), 2004. doi: 10.1038/nrg1315.

C. Watkins. Selective breeding analysed as a communication channel: Channel capacity as a fundamental limit on adaptive complexity. In *2008 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, 2008. doi: 10.1109/SYNASC.2008.100.

Herbert S. Wilf and Warren J. Ewens. There's plenty of time for evolution. *Proceedings of the National Academy of Sciences*, 107(52), 2010. doi: 10.1073/pnas.1016207107.

Claus O. Wilke, Jia Lan Wang, Charles Ofria, Richard E. Lenski, and Christoph Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844), 2001. doi: 10.1038/35085569.

Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4), 2002.

Christopher S. Withers and Saralees Nadarajah. The spectral decomposition and inverse of multinomial and negative multinomial covariances. *Brazilian Journal of Probability and Statistics*, 28(3), 2014. doi: 10.1214/12-BJPS213.

R. P. Worden. A speed limit for evolution. *Journal of Theoretical Biology*, 176(1), 1995. doi: 10.1006/jtbi.1995.0183.

Sewall Wright. The distribution of gene frequencies in populations. *PNAS*, 23(6), 1937. doi: 10.1073/pnas.23.6.307.

Zeba Wunderlich and Leonid A. Mirny. Different gene regulation strategies revealed by analysis of binding motifs. *Trends in Genetics*, 25(10), 2009. doi: 10.1016/j.tig.2009.08.003.

Avihu H. Yona, Eric J. Alm, and Jeff Gore. Random sequences rapidly evolve into de novo promoters. *Nature Communications*, 9(1), 2018. doi: 10.1038/s41467-018-04026-w.

Joel Zylberberg, Jason Timothy Murphy, and Michael Robert DeWeese. A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of V1 simple cell receptive fields. *PLoS Comput Biol*, 7(10), 2011.

APPENDIX A

Methods and supplementary results: optimality and statistical analysis of biological systems

A.1 Methods

A.1.1 Model neuron and mutual information utility function

A model neuron elicits a spike at time t $(r_t = 1)$ with a probability:

$$P(r_t = 1|x_t) = \frac{1}{1 + \exp\left[-k(x_t - x_0)\right]}; \tag{A.1}$$

the stimuli x_t were distributed according to a Gaussian Mixture Model,

$$P(x_t) = \sum_{i=1}^{3} w_i \mathcal{N}(\mu_i, \sigma_i^2), \tag{A.2}$$

where $w_i = 1/3$ are weights of the mixture components, $\mu_{1,2,3} = -2, 0, 2$ are the means, and $\sigma_i = 0.2$ are standard deviations.

To estimate mutual information between class labels and neural responses, we generated $5 \cdot 10^4$ stimulus samples x_t from the stimulus distribution. Each sample was associated with a class label $c_t \in \{1, 2, 3\}$, corresponding to a mixture component. We created a discrete grid of logistic-nonlinearity parameters by uniformly discretizing ranges of slope $k \in [-10, 10]$ and position $x_0 \in [-3, 3]$ into 128 values each. For each pair of parameters on the grid, we simulated responses of the model neuron to the stimulus dataset and estimated the mutual information directly from a joint histogram of responses r_t and class labels c_t.

A.1.2 Likelihood ratio test of optimality

The proposed test uses the likelihood ratio statistic,

$$\lambda = 2 \log \frac{\max_{\beta > 0} P(\mathcal{D}|\beta)}{P(\mathcal{D}|\beta = 0)}. \tag{A.3}$$

The null hypothesis is rejected for high values of λ. The marginal likelihood of β, $\tilde{\mathcal{L}}(\beta) = P(\mathcal{D}|\beta)$, depends on the overlap of parameter likelihood and the optimization prior, $P(\mathcal{D}|\beta) = \int_\Theta P(\mathcal{D}|\theta)\, P(\theta|\beta)\, d\theta$, where Θ is the region of biophysically feasible parameter combinations.

The null distribution of λ is obtained by sampling in three steps: (i) sample a parameter combination θ from a uniform distribution on θ, i.e. $P(\theta|\beta = 0)$; (ii) sample a data set $\mathcal{D}$ according to the likelihood $P(\mathcal{D}|\theta)$; (iii) compute the test statistic λ according to Eq. (A.3). This computationally expensive process simplifies in two situations described below.

Data-rich-regime simplification. In the data-rich regime, when the parameter likelihood $P(\mathcal{D}|\theta)$ is concentrated at a sharp peak positioned at $\hat{\theta}_{ML}$, likelihood ratio depends only on the value of utility at $\hat{\theta}_{ML}$:

$$\lambda = 2 \log \frac{\max_{\beta>0} \int_\Theta P(\mathcal{D}|\theta)\, P(\theta|\beta)\, d\theta}{\int_\Theta P(\mathcal{D}|\theta)\, P(\theta|\beta = 0)\, d\theta} \tag{A.4}$$

$$= 2 \log \frac{\max_{\beta>0} P(\hat{\theta}_{ML}|\beta)}{P(\hat{\theta}_{ML}|\beta = 0)} \tag{A.5}$$

$$= 2 \log \left(Z(0) \max_{\beta>0} \frac{e^{\beta U(\hat{\theta}_{ML})}}{Z(\beta)} \right), \tag{A.6}$$

which is a non-decreasing function of the utility $U(\hat{\theta}_{ML})$. Thus, this test is equivalent to a test that uses the utility estimate itself, $U(\hat{\theta}_{ML})$, as the test statistic, making it possible to avoid the costly integration over Θ. The null distribution can then be obtained by computing $U(\theta)$ at uniformly sampled θ.

Multiple system instances simplification. If multiple instances of the system are available and we can assume that their parameters $\theta_1, \theta_2, \ldots, \theta_N$ are i.i.d. samples from the same distribution $P(\theta|\beta)$, then the datasets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N$ are also i.i.d., $P(\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N|\beta) = \prod_{n=1}^N P(\mathcal{D}_n|\beta)$. We test the hypotheses $\beta = 0$ vs. $\beta > 0$ with the likelihood ratio statistic

$$\lambda = 2 \log \frac{\max_{\beta>0} \prod_{n=1}^N P(\mathcal{D}_n|\beta)}{\prod_{n=1}^N P(\mathcal{D}_n|\beta = 0)}. \tag{A.7}$$

By Wilks' theorem, for large N the null distribution of λ approaches the χ^2_1 distribution (with a point mass of weight $1/2$ at $\lambda = 0$, because we only consider $\beta \geq 0$). This removes the need for sampling in order to obtain the null distribution.

A.1.3 Hierarchical inference of population optimality

Assuming that experimental datasets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N$ are i.i.d., the posterior over population optimality parameter β takes the form:

$$P(\beta|\mathcal{D}_1, \ldots, \mathcal{D}_N) \propto P(\beta) \prod_{n=1}^N \int_{\theta_n} P(\mathcal{D}_n|\theta_n) P(\theta_n|\beta) d\theta_n, \tag{A.8}$$

where $\theta = (k_n, x_{0,n})$ is a vector of neural parameters (slope and position), and $P(\beta)$ is a prior over β. We approximated integrals numerically via the method of squares. Neural parameter values were sampled from ground-truth distributions via rejection sampling.

A.1.4 Inference of receptive fields with optimality priors

We randomly sampled 16 × 16 pixel image patches from the van Hateren natural image database Van Hateren and van der Schaaf (1998) and standardized them to zero mean and unit standard deviation. Neural responses were simulated using a Linear-Nonlinear Poisson (LNP) model:

$$P(r_t|x_t, \phi, k, x_0) = \frac{\lambda_t^{r_t} e^{-\lambda_t}}{r_t!}, \tag{A.9}$$

where λ_t is the rate parameter equal to:

$$\lambda_t = \frac{L}{1 + \exp\Big[- \phi^T x_t \Big]}, \tag{A.10}$$

where $L = 20$ was the maximal firing rate.

Given a linear filter ϕ, we quantified sparsity of its responses to natural images using the following function:

$$U_{\mathrm{SC}}(\phi) = -\Big\langle |\phi^T x_t| \Big\rangle. \tag{A.11}$$

Filter sparsity was averaged across the natural image dataset consisting of $5 \cdot 10^4$ standardized image patches randomly drawn from the van Hateren image database. The mean and standard deviation of filters ϕ was set to be 0 and 1 respectively. We optimized filters which either maximize or minimize the sparse utility measure via gradient descent. Different random initializations led to different filter shapes.

The locality utility of neural filters was defined as follows:

$$U_{\mathrm{LO}}(\phi) = -\sum_{i,j}((i - i_{max})^2 + (j - j_{max})^2)\phi_{i,j}^2, \tag{A.12}$$

where i_{max}, j_{max} are positions of the RF pixel with the largest absolute value. This definition of locality was introduced in Doi et al. (2012).

Sparsity and locality utilities were combined into a single utility:

$$U(\phi; \xi) = U_{\mathrm{SC}}(\phi) + \xi U_{\mathrm{LO}}(\phi). \tag{A.13}$$

To estimate receptive fields (neural filters), we first simulated the responses of the model population to 2000 natural image patches. We estimated linear receptive fields from simulated data by computing the spike-triggered average (STA), a widely applied estimator of neural receptive fields Sharpee (2013). In the STA model, response of neuron n at time t is assumed to follow the normal distribution Park and Pillow (2017):

$$P(r_{t,n}|s_{t,n}, \phi_n) = \mathcal{N}(\phi_n^T s_t; \sigma^2) \tag{A.14}$$

where ϕ_n is the linear receptive field of the n-th neuron, and σ^2 is the noise variance.

To infer the receptive fields from simulated neural responses using our framework, we assumed the following optimization prior over receptive fields derived from the sparsity utility in Eq (A.11):

$$P(\phi_n|\beta) \propto \exp\Big[\beta(U_{\mathrm{SC}}(z(\phi_n)))\Big], \tag{A.15}$$

where $z(\phi_n)$ denotes normalization of the receptive field to zero mean and unit variance. The sparse utility was evaluated over 10^4 randomly sampled image patches. The resulting log-posterior took the following form:

$$E(\phi_n|D,S,\beta) \propto -\frac{1}{\sigma^2}\sum_{t=1}^{T}\left(\phi_n^T s_t - r_{t,n}\right)^2 + \beta U_{\mathrm{SC}}(z(\phi_n)). \tag{A.16}$$

MAP inference was performed via gradient ascent on the log-posterior. Receptive fields were inferred with different priors corresponding to following values of the β parameter: 0, 1, 10, 20, 100. Receptive fields were estimated after reducing the dimensionality of stimuli with Principal Component Analysis to 64 dimensions. Estimation via gradient ascent on the log-posterior was performed in the PCA domain. PCA preprocessing is equivalent to low-pass filtering the stimuli.

To estimate value of the locality constraint ξ as well as the prior strength β via cross-validation, we split the data into the training and testing datasets comprising of 80% and 20% of data respectively. We estimated receptive fields for a range of β and ξ values ($[0, 0.01, 0.1, 1, 10]$ and $[0, 0.05, 0.2, 1]$ respectively). For each MAP RF estimate, we predicted neural responses $\hat{r}_t$ using stimuli from the test dataset. We then computed the average error $\langle(\hat{r}_t - r_t)^2\rangle$ using neural responses in the test dataset. Combination of hyperparameters ξ, β which resulted in the smallest error value was taken to be the estimate of the correct one.

A.1.5 Analysis of V1 receptive fields

Receptive fields of 250 neurons in the Macaque V1 were published and analyzed in Ringach (2002). All receptive fields were downsampled to 32×32 pixels size and normalized to have zero mean and unit variance.

To evaluate sparseness of V1 receptive fields, we relied on the following sparse utility:

$$U_{\mathrm{SC}}(\phi) = \left\langle \log(1 + (z(\phi^T)x_t)^2), \right\rangle_t \tag{A.17}$$

where x_t are individual image patches and $z(\phi_n)$ denotes normalization of the receptive field to zero mean and unit variance. The sparse utility was evaluated over 5×10^4 randomly sampled image patches. This form of the sparse utility was proposed in Olshausen and Field (1997), and together with the measure specified in Eq (A.11) it belongs to a broad class of equivalent sparsity measures defined by convex functions Hyvärinen et al. (2009).

To test individual RFs for optimality, we generated the null distribution of utility values by bootstrapping 10^6 random filters as follows: (i) draw a random integer K between 1 and 128; (ii) superimpose K randomly selected principal components of natural image patches; each component is multiplied by a random coefficient $v \sim \mathcal{N}(0,1)$; (iii) generate a 2D Gaussian spatial mask centered at a random position on the image patch; lengths of horizontal and vertical axes of the Gaussian ellipse were drawn independently; (iv) multiply the random filter and the Gaussian mask. This procedure ensures that a range of filters of different sparsity and slowness will be randomly generated. Filters were standardized to zero mean and unit standard deviation.

To establish a measure of optimality at a population level, we needed to simplify the integration over all receptive field parameters, which was intractable due to their high-dimensionality. Computation of posteriors over β in Eq (A.8) was therefore approximated as follows:

$$P(\beta|\mathcal{D}_1, \dots, \mathcal{D}_N) \approx P(\beta) \prod_{n=1}^{N} \frac{1}{Z(\beta)} P(\hat{\theta}_n|\beta). \tag{A.18}$$

where $\hat{\theta}$ are receptive fields estimates computed in Ringach (2002).

We approximated $P(\hat{\theta}_n|\beta)$ via rejection sampling, noting that $P(\hat{\theta}_n|\beta) = P(U(\hat{\theta}_n)|\beta)$, i.e., the probability of a high dimensional receptive field is determined solely by a one-dimensional utility function.

For each β we randomly sampled 10^6 filters from the proposal distribution, as described above, and retained only those consistent with $P(U_{\text{SC}}(\theta)|\beta)$ via rejection sampling. Obtained utility values were fitted with a Gaussian distribution, used to evaluate posteriors over β, with point estimates being posterior maxima; the prior over β was uniform over the range displayed in the figures. For sparse utility, we discretized β values into 20 values equally spaced on the $[-5, 5]$ interval. Filters accepted for each β value were used to compute the average spatial autocorrelation.

For comparison we used optimally sparse receptive fields learned from natural image patches preprocessed with PCA. We note that this preprocessing step might not have a direct biological counterpart. To compare optimal solutions and neural data, we therefore evaluated sparsity of model and real V1 RFs in the domain of natural images without PCA preprocessing.

To cluster receptive fields according to optimality, we defined a mixture model:

$$P(\theta_n|\{w_1, \dots, w_K\}, \{\beta_1, \dots, \beta_K\}) = \sum_{k=1}^{K} w_k P(U_{\text{SC}}(\theta_n)|\beta_k) \tag{A.19}$$

where w_k is the weight of the kth mixture component and β_k is the optimality of that component. To approximate utility-defined distributions, we used the Gaussian approximation described above i.e.: $P(\theta_n|\beta) = P(U_{\text{SC}}(\theta_n)|\beta) = \mathcal{N}(U_{\text{SC}}(\theta_n); \mu_\beta, \sigma_\beta^2)$

Parameters of the model were learned via the standard expectation-maximization algorithm (EM).

A.1.6 Analysis of retinal receptive fields

Temporal receptive fields of retinal ganglion cells were published and analyzed in Deny et al. (2017). We analyzed RFs of 117 neurons selected by temporal smoothness. Each RF was normalized to unit norm and fitted with a parameteric biphasic filter model described in Sun et al. (2017).

We considered two different utility functions. First one was a generalization of the predictive coding objective introduced in Srinivasan et al. (1982). The predictive coding objective minimizes the squared difference between the stimulus value s_t at time t and the linear prediction of that stimulus value computed from N past values: $E(\phi) = \left[\sum_{\tau=0}^{N} \phi_\tau s_{t-\tau}\right]^2$, where ϕ_τ are the weights of the linear filter. In the classical approach it has been assumed that the linear weight of the current stimulus s_t is equal to 1 i.e. $\phi_0 = 1$. We note that such form makes it difficult to evaluate predictive coding filters adapted to stimuli of unknown temporal scale. In particular, we optimize and evaluate our filters on natural movies whose frame rate might be mismatched with the timescale of the retina. We therefore relax the assumption that the predictive coding filter reduces the dynamic range by subtracting only the current stimulus from its prediction, and assume that what is being predicted is itself a linear combination of stimulus values (e.g., integrating stimulus value over some recent period of time). In practice this means that we allow all values of the filter including ϕ_0 to vary freely. To avoid trivial solutions, where the residue $E(\phi)$ is minimized by setting all weights to 0, we impose a unit norm constraint on the filter ϕ. The utility function of a filter ϕ is then equal to:

$$U_{\mathrm{PC}}(\phi) = -\Big\langle \Big[\sum_{\tau=0}^{N} z(\phi)_\tau s_{n,t-\tau} \Big]^2 \Big\rangle_n \tag{A.20}$$

where z denotes the unit norm operator, and n indexes stimulus epochs s_n.

We evaluated the utility U_{PC} using 50000, 21-sample long excerpts of single-pixel luminance extracted from natural movies of scenes in the African savanna van Hateren and Ruderman (1998).

We used these natural stimulus data to learn the optimal predictive-coding filter, as described in Srinivasan et al. (1982) via gradient descent.

The second considered utility was measuring the amount of information between the stimulus and the instantaneous filter output in a low-noise regime. Under the Gaussian approximation of stimulus and output distribution this utility takes the form:

$$U_{\mathrm{II}}(\phi) = -\frac{1}{2}\log(1-\rho^2), \tag{A.21}$$

where ρ is the Pearson correlation coefficient between the stimulus s_t and the filter output r_t. This utility is high when the neural responses track the stimulus with high fidelity. Note that this is not the general solution to an efficient coding (infomax) problem, where the *full response trajectory*, not the instantaneous response, should encode high information about the stimulus, which leads to decorrelation / whitening in the low-noise regime. We evaluated U_{II} using a trajectory of 20000 samples of pixel intensity values extracted from the natural movie dataset.

To compute utility-defined distributions of the filter mode amplitude parameters c_1, c_2, we first discretized values of these parameters into 100 values uniformly spaced on the $[0.01, 13]$ interval, where 13 was the maximum amplitude parameter value among fits to normalized retinal RFs. For each filter we evaluated utility for each pair of discretized amplitude parameter values and a fixed value of the scale parameter a fitted to that filter. We used such utility surfaces to estimate the normalization constant of the utility-defined distribution parametrized by β and the scale parameter a.

We discretized the parameter β into 100 values uniformly spaced on the $[-10, 64]$ interval. We estimated the posterior over β by numerically integrating over filter parameters c_1, c_2, a. We assumed a uniform prior over β.

A.1.7 Analysis of connectivity in *C. elegans*

For our analysis we used the *C. elegans* neural wiring dataset available on Worm Atlas (www.wormatlas.org). This dataset has been published and analyzed before in Chen et al. (2006) as well as Pérez-Escudero et al. (2009); Pérez-Escudero and de Polavieja (2007) – for details about the dataset please refer to this prior work.

For the analyses depicted in Fig. 8 we selected two sets of neurons. The first set consisted of 126 neurons connected to at least one muscle, and the second set consisted of 86 neurons connected to at least one sensor. "i-th" neuron was therefore characterized by its position, x_i, number of landmark cells (muscles or sensors) it was connected to, N_i, vectors of positions of the landmark cells, m_i (muscles), and s_i (sensors), and vectors of the number of synapses in each neuron-to-landmark connection, n_i. For each neuron the utility of its position was defined as:

$$U_{\mathrm{WC}}(x_i; N_i, l_i, n_i) = -\sum_{j=1}^{N_i} n_{i,j}|x_i - l_{i,j}|^{\xi}. \tag{A.22}$$

where $l_i \in \{m_i, s_i\}$, denotes the vector of landmark cell positions. We evaluated the utility function on the $[0, 1]$ interval representing the linear extent of the worm body axis, discretized into 100 linearly spaced values. To compute the posterior distribution over parameters β and ξ we discretized them into 64 linearly spaced values. For neuron-muscle connections, β was defined over a [1.5, 4] interval and ξ over a [1.3, 1.9] interval. For neuron-sensor connections, β was defined over a [10, 25] interval and ξ over a [1.5, 2.2] interval. We assumed a uniform prior over parameters β, ξ.

A.1.8 Quantification and statistical analysis

Statistical test performed in Fig. 5D was a two-tailed t-test. Stars denote p-values lower than 0.001. Error bars in the figure denote standard errors of the mean.

A.2 Data disambiguates degenerate theoretical predictions

Related to Question 3: Data resolves ambiguous theoretical predictions and Fig. 4.

Ambiguity of the first kind Predictions of an optimization theory can be degenerate or ambiguous. Here we explore the *first kind of ambiguity*, where the utility function has multiple (possibly degenerate) maxima.

In this situation, biological context typically forces us to choose between two interpretations. On the one hand, we may observe multiple instances of the biological system and each instance could be an independent realization sampled from any of the maxima: statistical analyses of optimality thus need to consider and integrate over the whole parameter space, as in the approaches described above. On the other hand, we may observe a single (e.g., evolutionary) realization of the biological system which we hypothesize corresponds to a single optimum of the utility function. Our task is then first to identify that relevant maximum; if it exists, subsequent analyses can follow up on how well data agrees with that prediction and how surprising such an agreement might be in face of multiple alternative maxima.

In the Fig. 2 example of the main paper, multiple values of slope and offset yield optimal or close to optimal neural performance, resulting in ambiguous theoretical predictions. As a simple illustration of how data can break such ambiguities, we consider three example neurons with varying degree of optimality (Fig. A.1A) and observe how their posteriors look like after seeing as few as $T = 12$ stimulus-response pairs from each neuron (Fig. A.1B). All three simulated datasets reduced the uncertainty (entropy) about the neuron's parameters by a similar amount, as reflected by the entropy and utility of the posterior versus the entropy and utility of the prior (Fig. A.1C). Despite similar reductions in entropy, the resulting inferences were very different in terms of agreement with the theory. Only the posterior of the first neuron concentrated in a high-utility region of the parameter domain, thus clearly identifying one of the four peaks of the utility function as consistent with the operating regime of the simulated neuron. The two remaining posteriors are concentrated in regions of the parameter space which weakly overlap with the prior, or where prior probability is close to 0. To capture these qualitative differences mathematically, we define and compute the *mode entropy*, where each mode corresponds to the attraction basin of a local utility maximum. Optimality theories with degenerate maxima will allocate the prior probability relatively evenly among the modes, resulting in high mode entropy (here, 2 bits, i.e., 4 possible local maxima). A few observations of neuron 1 consistent with an optimal solution drastically collapsed this mode uncertainty and identified the single relevant utility maximum; this decrease was smaller for slightly suboptimal neuron 2 and vanished for neuron 3 (Fig. A.1D).

This is a very non-standard application of the Bayesian framework at small sample sizes, T: here, the structure of the prior (i.e., the normative theory) dominates the posterior, in what we refer to as the "data-regularized prediction" regime. In this regime, our goal is to derive *ab initio* theoretical predictions, not fit parameters to reproduce the data, and the data is only used to disambiguate the prediction – to identify which utility maximum, if any, is realized in nature. If we track the evolution of the average utility, full posterior entropy, and the mode entropy with the number of data points T, we clearly see the transition from such "data-regularized prediction" regime dominated by the prior normative theory, to the "theory-regularized inference" regime in the large sample limit (Fig. A.1E). In the first regime, data removes the theoretical ambiguity and collapses the mode entropy with $T < 10$ samples; in the second regime, the actual parameter values (k, x_0) are inferred with increasing precision, as evidenced by posterior entropy that continues to decrease linearly in the log sample size (corresponding to the standard asymptotic inverse scaling of the variance in parameter estimates with the sample size).

In the "data-regularized prediction" regime, β also serves a novel role: when the normative theory has multiple optima with a broader spectrum of utility values, β determines which of the peaks are considered as nearly degenerate candidate predictions. A peak with utility $U' < U_{\max}$ will be suppressed in the prior by $\sim \exp(-\beta(U_{\max} - U'))$, and, for sufficiently high β, the alternative theoretical prediction corresponding to U' will be disregarded irrespective of the data.

Here we showed that the ambiguities of normative theories resulting from degenerate utility maxima can often be resolved in our Bayesian framework in the "data-regularized prediction" regime by a very small amount of data. This power may appear trivial at first glance, because the parameter space of our example is two dimensional and so priors and posteriors can be evaluated explicitly and plotted across their whole domain. In more realistic cases involving tens of parameters, however, finding all (nearly) degenerate maxima of the utility function and deciding whether data is "close to" any one of them becomes a daunting task due to the curse of dimensionality. In the past, this has severely limited the application of optimality principles to complex systems with more than a few parameters Tkačik et al. (2010, 2008); Młynarski (2015), except in those rare cases where strict guarantees exist De Martino et al. (2018). In contrast, even in spaces of high dimensionality, posteriors resulting from our framework can be sampled with Monte-Carlo methods or optimized by well-developed methodology Murphy

(2012), with search concentrated around the unique peak of the normative theory that is simultaneously permitted by the chosen value of β and is consistent with the data, if such a peak exists. Intuitively, theory "proposes" possible optimal solutions *ab initio* while data "disposes" with those degenerate solutions for which there is no likelihood support.

Ambiguity of the second and third kind Here we further draw the distinction between ambiguities of the second and third kind.

In the second kind of ambiguity, parameters θ can be subdivided into optimized parameters θ^{opt} (e.g. the offset of the neural nonlinearity x_0 in our toy example) and constraints θ^{con} (e.g. the slope k, which determines the response reliability of the model neuron) i.e. $\theta = \{\theta^{opt}, \theta^{con}\}$. While the utility of the system depends on both sets of parameters, only the optimized parameters θ^{opt} depend on the optimality parameter β (Fig. A.2A) (i.e. $P(\theta^{opt}|\beta, \theta^{con}) \propto \exp\Big[\beta U(\theta)\Big]$.

In the case of ambiguity of the third kind, the utility function is parametrized by an additional parameter ξ, which is not set by the theory. The parameter probability distribution is therefore given by: $P(\theta|\beta, \xi) \propto \Big[\beta U(\theta; \xi)\Big]$. Resolving the ambiguity corresponds to inferring the value of that parameter directly from the data D. Thus here, in contrast to the ambiguity of the second kind, *all* parameters depend on the optimality parameter β and the additional parameter ξ (Fig. A.2B).

A.3 Sparse and slow utility in V1 receptive fields

Related to Application 1: Receptive fields in the visual cortex and Fig. 6.

To complement the analysis of V1 neurons in terms of the sparse utility U_{SC}, here we present additional analysis in terms of the slowness utility. Slowness utility U_{LC} assumes that neurons extract invariant properties of sensory data Wiskott and Sejnowski (2002). Given a linear filter ϕ, we quantified slowness of its responses to a set of natural image sequences using the following function:

$$U_{\mathrm{LC}}(\phi) = -\Big\langle \frac{1}{T-1} \sum_{t=2}^{T} (\phi^T x_{t,n} - \phi^T x_{t-1,n})^2 \Big\rangle_n . \tag{A.23}$$

where n is an index over image sequences, and t is a time index over images within a sequence. Filter slowness was averaged across a $5 \cdot 10^4$ artificially generated natural image sequences of length $T = 2$. Each sequence was generated by moving an image patch by a random distance $n_x \in [-8, 8]$ pixels in a horizontal direction and $n_y \in [-8, 8]$ pixels in vertical direction, and rotating it by a random angle $\alpha \in [-90^\circ, 90^\circ]$. The mean and standard deviation of filters ϕ and image patches $x_{t,n}$ was set to be 0 and 1 respectively.

Optimally slow RFs minimize temporal variability of neural activity in natural sensory envirom-nents Berkes and Wiskott (2005). On the level of individual neurons, slowness and sparseness optimality criteria yield very different predictions. In contrast to optimally sparse RFs which are localized in space and frequency (Fig. A.3B, left column), RFs optimized for slowness are broad and non-local (Fig. A.3B, right column).

In the right column of Fig. A.3B-F, we present analysis of the optimality of V1 RFs in terms of the slow utility U_{LC}. This analysis is complementary to the analysis of the sparse utility U_{SC}

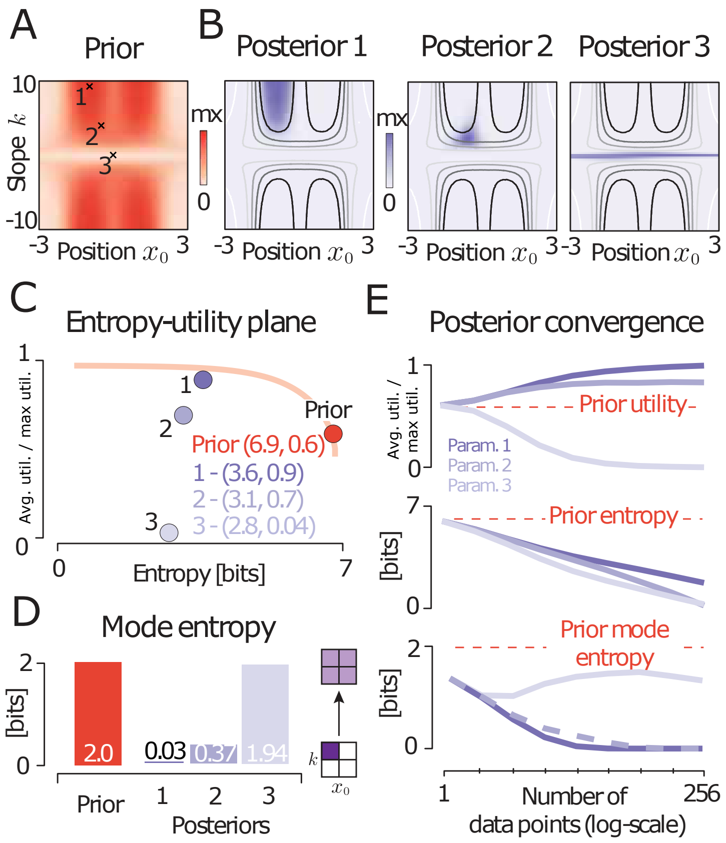

Figure A.1: **Disambiguating degenerate theoretical predictions,** related to Question 3: Data resolves ambiguous theoretical predictions and Fig. 4. **(A)** A maximum-entropy prior derived from the mutual information utility with $\beta = 1$. The prior has multiple maxima reflecting non-uniqueness of theoretical predictions. **(B)** Posteriors obtained by updating the prior with three example datasets $(\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3)$. Grayscale lines denote regions of different utility values (black – highest utility, white – lowest utility). Depending on the observed data, posteriors concentrate in regions of different utility value. **(C)** Distributions on the entropy-utility plane. Orange dot corresponds to the prior from A, purple dots to posteriors from B. Orange line is the entropy–average utility tradeoff in the maximum entropy optimization prior (analogous to Fig. 2E in the main text). **(D)** Mode entropy. In the prior (red bar), probability is equally distributed across 4 peaks of the distribution resulting in 2 bits of entropy. Mode entropy decreases significantly in posteriors 1 and 2. **(E)** Posterior convergence. Average utility (top row), posterior entropy (middle row) and posterior mode entropy (bottom row) are plotted against the number of data samples; shown are averages of 512 realizations for each data set size. Purple lines correspond to parameter settings 1-3 in panel A. Red dashed line denotes values of each statistic for the prior.

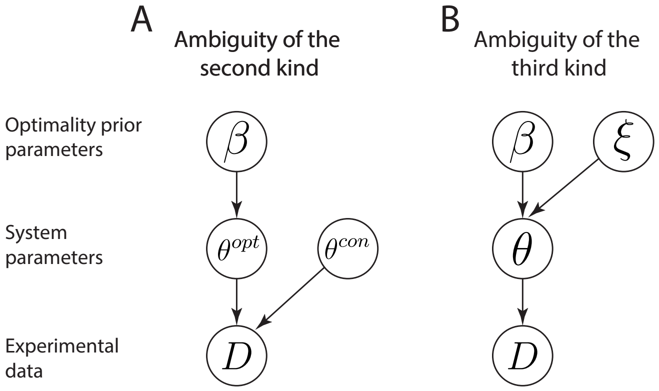

Figure A.2: **Graphical models depicting variable dependencies corresponding to the second and third kinds of ambiguity,** related to Question 3: Data resolves ambiguous theoretical predictions and Fig. 4. **(A)** Ambiguity of the second kind. **(B)** Ambiguity of the third kind.

presented in the main text. We present analysis of the sparse utility again in the left panels of Fig. A.3B-F for comparison. To compute posteriors over β (Fig. A.3D), for slow utility we used 64 β values equally spaced on the $[-32, 32]$ interval. The remaining details of the analysis are shared with the analysis of sparse coding utility and are described in the Methods.

Overall, RFs of individual neurons are much more consistent with the sparse, rather than with the slow utility. This is readily apparent in the outcome of the optimality test (Fig. A.3 C) and posteriors over the utility parameter β (Fig. A.3 D), where V1 RFs yield a slightly negative estimate $\hat{\beta}$. Similarly, the empirical distribution of utility values of V1 RFs as well as their spatial autocorrleation, is more consistent with the inferred distribution of sparse, rather than slow utilities (Fig. A.3E and F respectively).

Here we analyzed utility of *indiviudal* neurons, treating them as independent realizations from an underlying distribution of parameters. It is important to stress that simultaneous optimization of a *population* of model neurons for maximal slowness yields filters which very closely resemble RFs of visual neurons Berkes and Wiskott (2005); Hyvärinen et al. (2009). Moreover, slowness and sparseness are both affected by eye movements and natural stimulus dynamics, while the RFs used here were recorded in anesthesized and paralyzed animals. Our analysis is therefore not a proof of lack of optimization for slowness at the population level. It is rather a demonstration of applicability of the framework to real data. Analysis of optimality of neural populations is a subject of future work.

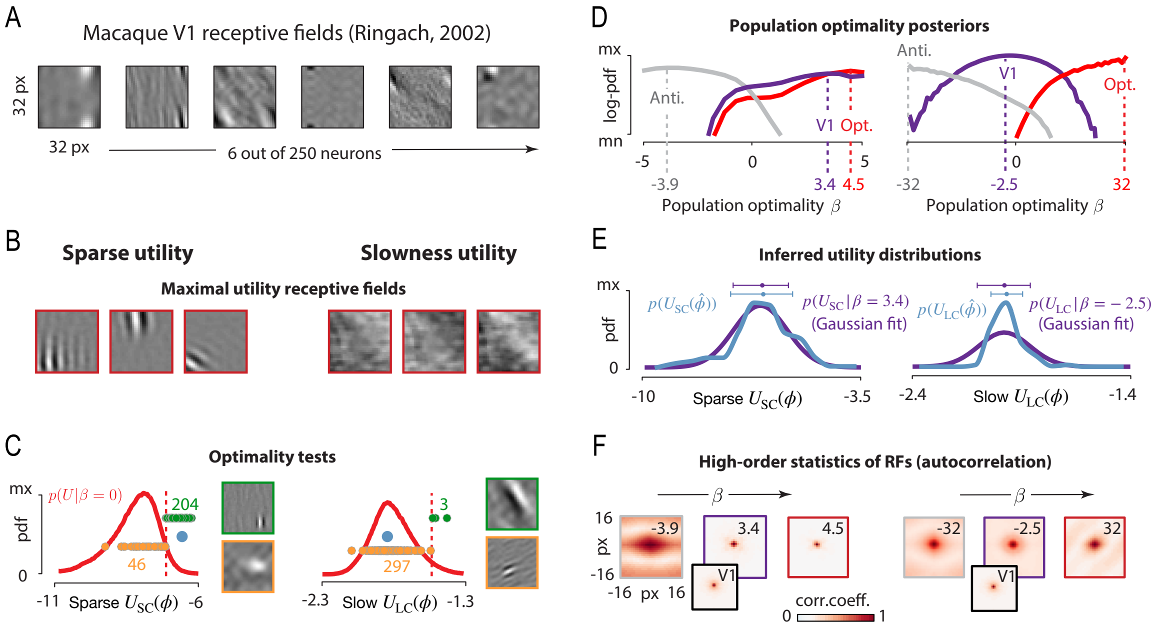

Figure A.3: **Sparse and slow utility analysis of V1 receptive fields,** related to Application 1: Receptive fields in the visual cortex and Fig. 6. **(A)** Six example receptive fields (RFs) from Macaque visual cortex (courtesy of Dario Ringach; Ringach (2002)). **(B)** Simulated RFs optimized for sparsity (left column) and slowness (right column). **(C)** Null distributions of utility values used to test for optimality under sparse (left column) and slow (right column) utilities. Red dashed lines denote the significance threshold (95th percentile). Green and orange circles correspond to significant and non-significant receptive fields (the axis was truncated for visualization purposes, and not all values are displayed). Example significant and non-significant receptive fields are displayed in green and orange frames respectively. Blue dots show the average utility of receptive fields, which are equal to the 99.6$^{\rm th}$ percentile (sparse $U_{\rm SC}$) and 46$^{\rm th}$ percentile (slow $U_{\rm LC}$) of $p(U|\beta=0)$. **(D)** Approximate log-posteriors over population optimality parameter β derived from 250 RFs estimates (purple line), 250 maximum-utility filters (red line) and 250 minimal-utility filters (gray line). Dashed lines mark MAP estimates of beta. **(E)** Empirical distributions of RF utilities (blue lines) compared with utility distributions consistent with the population optimality β inferred from V1 data (purple lines). **(F)** Spatial autocorrelation of RFs consistent with different average values of utility (determined by β parameter). Values of β are denoted in the top-right corner of each panel, and correspond to results of inference displayed in panel D. Middle plots (purple frame) in the left and the right column depict autocorrelation consistent with β inferred from V1 RFs. For comparison, autocorrelation of RFs is displayed as an inset.

APPENDIX B

Supplementary information: accumulation and maintenance of genetic information

B.1 Joint and conditional KL divergence, chain rule

For a single variable U, the KL divergence Cover and Thomas (2006) between its distributions with and without selection is

$$D(U) = \sum_u \psi^U(u) \log_2 \frac{\psi^U(u)}{\varphi^U(u)} \tag{B.1}$$

where U takes values u with probabilities $\psi^U(u)$ under selection and $\varphi^U(u)$ under neutrality.

To be well defined, the KL divergence requires that the support of ψ^U is a subset of the support of φ^U. In other words, if for some u we have $\varphi^U(u) = 0$, then also $\psi^U(u) = 0$ – outcomes impossible under neutrality are also impossible under selection. This condition needs to be respected when setting the initial conditions (ψ^U and ϕ^U at time zero). Over time, selection increases or decreases the probability of population states, genotypes or phenotypes that arise by reproduction with mutation, a by finite factor proportional to fitness. But selection cannot create entirely new states. On the other hand, some genotypes that arise by mutation can have zero fitness, and therefore be impossible under selection ($\psi^U(u) > 0$ but $\psi^U(u) = 0$). When this happens, the corresponding term $\psi^U(u) \log_2 \frac{\psi^U(u)}{\varphi^U(u)}$ is set to zero. Therefore this is a very natural assumption, and analogous arguments apply to joint/conditional distributions which we discuss next.

For a pair of variables U, V we can write their joint and conditional KL divergence Cover and Thomas (2006),

$$D(U,V) = \sum_{u,v} \psi^{U,V}(u,v) \log_2 \frac{\psi^{U,V}(u,v)}{\varphi^{U,V}(u,v)}, \tag{B.2}$$

$$D(U|V) = \sum_v \psi^V(v) \sum_u \psi^{U|V}(u|v) \log_2 \frac{\psi^{U|V}(u|v)}{\varphi^{U|V}(u|v)} \tag{B.3}$$

where $\psi^{U,V}(u,v)$ and $\psi^{U|V}(u|v)$ are the joint and conditional probabilities under selection, and $\varphi^{U,V}(u,v)$ and $\varphi^{U|V}(u|v)$ under neutrality. With these definitions, the chain rule states two possible decompositions of the joint KL divergence,

$$D(U,V) = D(U) + D(V|U) = D(V) + D(U|V). \tag{B.4}$$

In the case of the genotype frequencies X, genotype G and phenotype Z, the conditional KL divergences $D(G|X)$ and $D(Z|G)$ are both zero, implying the inequality Eq. (3.5). Along with some of the corresponding marginal distributions, this is also illustrated in Fig. B.1.

B.2 Allele frequencies, LD and information

When there is a fixed number of loci, instead of genotype frequencies, an alternative way to describe a population state is in terms of allele frequencies. Allele frequencies by themselves, however, do not capture correlations between loci and therefore can miss some of the information that selection can accumulate on the population level. This can be expressed using the chain rule,

$$D(X) = D(\text{Allele freq.}) + D(X|\text{Allele freq.}) \geq D(\text{Allele freq.}), \tag{B.5}$$

where the term $D(\text{Allele freq.}|X) = 0$ because regardless of selection, allele frequencies are fully determined by the genotype frequencies X. The term $D(X|\text{Allele freq.})$ quantifies how different from neutrality are the correlations between loci.

B.3 Violation of the bound by Worden 1995 by drift

The genotype-level information introduced by Worden Worden (1995) (see Eq. (11) there) is the KL divergence between the genotype frequencies and a uniform distribution,

$$I = \sum_g x_g \log_2(M x_g), \tag{B.6}$$

where M is the number of possible genotypes and x_g the frequency of genotype g (denoted q_j in Worden (1995)). Worden also introduced a similar genotype-level measure (Eq. (8) in Worden (1995), which was upper bounded by I. These measures of information can be seen as special cases of $D(G)$ and $D(Z)$ when there is no evolutionary stochasticity – ψ^X is concentrated at a single value x, and φ^X is concentrated at a single value of x that is uniform over all possible genotypes.

Worden proposes a bound on the rate of increase of I starting from a uniform x, and the maximal rate is proportional to a quantity similar to the genetic load, i.e. roughly a factor N (population size) times more stringent than the bound presented here.

The proof relies on the assumption that the population is large and x_g evolve deterministically, but later, validity in finite populations is claimed (Sec. 2.6 in Worden (1995)). This is mistaken: in a realistic population, I can hardly be zero to start with, as there will be more possible genotypes than individuals and x_g cannot be uniform. Starting from near uniform x, random drift will tend to remove variability from the population and concentrate all genotypes around some random ancestral genotype, and I will increase even without selection. This can also be seen as a consequence of the convexity of KL divergence: random fluctuations in x will, on

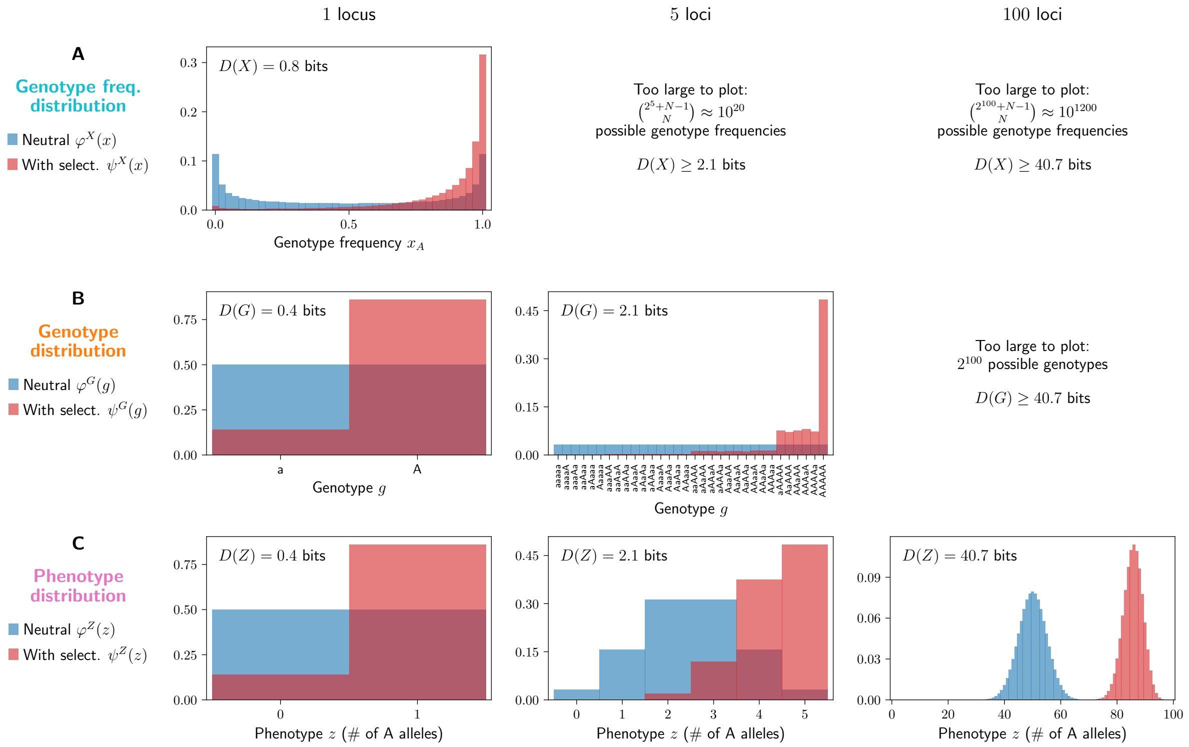

Figure B.1: An example of the distributions over genotype frequencies, genotypes and phenotypes, with and without selection and for a varying number of loci, along with the corresponding measures of information. Based on a Wright-Fisher model with parameters $N = 40$, $\mu = 0.005$, $s = 0.05$ (as in Main Text Fig. 3.1). Fitness is multiplicative across loci, $w = (1+s)^z$ where z is the number of beneficial A alleles carried – i.e. there is directional selection for an additive, fully heritable phenotype z (as in Main Text Fig. 3.6). All distributions are at the stationary state, computed using a transition matrix-based model (1 locus) or a long run (10^5 generations after 200 generations of burn-in) of an individual-based model (> 1 locus).
(A) Distributions over genotype frequencies ($\psi^X(x)$ with selection and $\varphi^X(x)$ without, red and blue) and the population-level information, $D(X)$. The distributions cannot be plotted and $D(X)$ cannot be directly estimated for 5 and 100 loci, due to the large number of possible genotype frequencies x, but we can still lower bound $D(X)$ by $D(G)$ or $D(Z)$.
(B) Distributions over genotypes ($\psi^G(g)$ with selection and $\varphi^G(g)$ without, red and blue) and the genotype-level information, $D(G)$. This information is less than $D(X)$, because selection not only gives preference to the fitter alleles, but also reduces the genetic variation within populations. The number of possible genotypes becomes too large for 100 loci, but we can still lower bound $D(G)$ by $D(Z)$.
(C) Distributions over the phenotype ($\psi^Z(z)$ with selection and $\varphi^Z(z)$ without, red and blue) and the phenotype-level information, $D(Z)$. The phenotype is simply the number of A alleles across all loci in an individual – an additive trait of varying polygenicity. In this example, selection favors the A allele at each locus, making fitness a function of z. In such cases, $D(G) \approx D(Z)$, since grouping genotypes into bins of equal z reduces the state space but preserves selective differences. In general, the trait might be unrelated to fitness or form only a component of it, leading to $D(G) > D(Z)$. Note that $D(Z)$ is approximately proportional to the number of loci, with each locus encoding about 0.4 bits. This is because loci evolve approximately independently, as there is zero epistasis, free recombination and little Hill-Robertson interference (see also Main Text Sec. 3.4.2-3.4.3).

average, increase I. This highlights the need to consider stochasticity as well as population variation when quantifying the intuitive notion of genetic information.

Worden's stringent bound does hold if the genotype frequencies evolve deterministically and there is no recombination. This is consistent with our observation that selection accumulates information less cost-efficiently when $Ns \gg 1$ and the fixation probability of a beneficial mutation is close to 1 (Main Text Fig. 3.3CD).

B.4 The single locus, two allele system used for figures

We use a haploid single locus, two allele system to produce Figures 3.1-3.5. The figures are produced with a Wright-Fisher model, Moran model (only the fitness flux in Fig. 3.4, B.2B and B.3), and some intuition can be gained by approximating it as diffusion under weak selection. Note that this is only an illustration, more general classes of models are discussed in sections B.5, B.6 and B.7.

The system has two alleles, a and A, where the latter is beneficial under selection. It is parametrized by the population size N, mutation rate μ and selection coefficient $s(x)$ (which is frequency dependent only in Fig. 3.3CD and B.4).

B.4.1 Wright-Fisher model

Under the Wright-Fisher model, the state space is a set of discrete frequencies of the A allele, $x_A = 0,\, 1/N,\, 2/N,\, \ldots,\, 1$, while the a allele always has the complementary frequency $x_a = 1 - x_A$. The two alleles have the following properties.

Allele g	frequency x_g	fitness $w_g(x)$	relative fitness $\hat{w}_g(x)$
a	$x_a = 1 - x_A$	$w_a(x) = 1$	$\hat{w}_a(x) = \frac{1}{1+s\,x_A}$
A	x_A	$w_a(x) = 1 + s$	$\hat{w}_A(x) = \frac{1+s}{1+s\,x_A}$

The probability of sampling allele A as a parent is $x_A \hat{w}_A(x)$, and the probability of sampling it as offspring is

$$q_A(x) = x_A(1-\mu) + x_a\mu \tag{B.7}$$

$$p_A(x) = x_A\hat{w}_A(x)(1-\mu) + x_a\hat{w}_a(x)\mu = \frac{x_A(1+s)(1-\mu) + (1-x_A)\mu}{1+sx_A}, \tag{B.8}$$

under neutrality and under selection respectively. The Wright-Fisher transition probabilities are given by the binomial distribution,

$$Q(x^{t+1}|x^t) = \binom{N}{Nx^{t+1}} q_A(x)^{Nx_A^{t+1}} (1 - q_A(x))^{N - Nx_A^{t+1}}. \tag{B.9}$$

$$P(x^{t+1}|x^t) = \binom{N}{Nx^{t+1}} p_A(x)^{Nx_A^{t+1}} (1 - p_A(x))^{N - Nx_A^{t+1}}. \tag{B.10}$$

This is a case of the Wright-Fisher model with selection among parents (see Eq. (B.27,B.28) and SI Sec. B.5.1). An analogous discrete-time Moran model can be written by plugging Eq. (B.7,B.8) into Eq. (B.43,B.44).

All calculations were done with $N \leq 200$. Given the small size of the system, we can compute the full matrix $P(x^{t+1}|x^t)$, and calculate the distribution over genotype frequencies over time by iterating $\psi^{X^{t+1}}(x^{t+1}) = \sum_{x^t} \psi^{X^t}(x^t) P(x^{t+1}|x^t)$.

B.4.2 The diffusion approximation

We write the diffusion approximation for the evolution of the frequency x_A. Following the notation in SI Sec. B.7, the first two moments of change of x_A are given by

$$a_A = \mu(1 - 2\,x_A) \qquad \text{expected change due to mutation,} \tag{B.11}$$

$$a_A^s = \frac{s\,x_A(1 - x_A)}{1 + s x_A} \qquad \text{expected change due to selection,} \tag{B.12}$$

$$b_{AA} = \frac{x_A(1 - x_A)}{N} \qquad \text{drift covariance.} \tag{B.13}$$

The diffusion equation for this system is a special case of Eq. (B.57,B.58).

Maintenance of information under weak selection. Since the diffusion process takes place along only one dimension, the stationary distribution can be determined by equating the probability flux to zero. For simplicity, we neglect the mean fitness $1 + s x_A \approx 1$ in the denominator in Eq. (B.12), by assuming that selection is weak, $s \ll 1$.

The stationary distributions under selection and under neutrality are Wright (1937)

$$\tilde{\psi} = \frac{(x_A(1 - x_A))^{2N\mu - 1} e^{2Nsx_A}}{Z(N, \mu, s)} \qquad \tilde{\varphi} = \frac{(x_A(1 - x_A))^{2N\mu - 1}}{Z(N, \mu, 0)} \tag{B.14}$$

with the normalization constant

$$Z(N, \mu, s) = \int_0^1 (x_A(1 - x_A))^{2N\mu - 1} e^{2Nsx_A} dx_A = \Gamma(2N\mu)^2\, {}_1\tilde{F}_1(2N\mu;\, 4N\mu;\, 2Ns), \tag{B.15}$$

where Γ is the Gamma function and ${}_1\tilde{F}_1$ is the regularized confluent hypergeometric function.

Similar integrals yield results for the maintained information $D(X)$, $D(G)$ and the associated expected cost at the stationary state. We calculate the expectation of the cost $C(x) = \frac{V(x)}{2 \ln 2} = \frac{s^2 x_A(1 - x_A)}{2 \ln 2}$ (see SI Sec. B.9 and B.7.1). For the genotype-level information $D(G)$, we also need the expected frequency of A which is equal to its marginal probability, $\langle x_A \rangle = \tilde{\psi}^G(A)$. For brevity, we only write the leading terms in s for each quantity.

$$\langle C \rangle = \int_0^1 \tilde{\psi}\, C(x)\, dx_A = \frac{N\mu\, s^2}{(4N\mu + 1)\, 2 \ln 2} + O\left(s^4\right) \tag{B.16}$$

$$\langle x_A \rangle = \tilde{\psi}^G(A) = \int_0^1 \tilde{\psi}\, x_A\, dx_A = \frac{1}{2} + \frac{Ns}{8N\mu + 2} + O\left(s^3\right) \tag{B.17}$$

$$D(G) = \langle x_A \rangle \log_2 \frac{\langle x_A \rangle}{1/2} + (1 - \langle x_A \rangle) \log_2 \frac{1 - \langle x_A \rangle}{1/2} = \frac{(Ns)^2}{(4N\mu + 1)^2\, 2 \ln 2} + O\left(s^4\right) \tag{B.18}$$

$$D(X) = \int_0^1 \tilde{\psi} \log_2 \frac{\tilde{\psi}}{\tilde{\varphi}}\, dx_A = \frac{(Ns)^2}{(4N\mu + 1)\, 2 \ln 2} + O\left(s^4\right) \tag{B.19}$$

Notably, at weak selection, both the cost $\langle C \rangle$ and the information $D(X)$, $D(G)$ scale with s^2. Their ratio is therefore given by the population size and the mutation rate,

$$\frac{D(G)}{\langle C \rangle} = \frac{N}{\mu\, (4N\mu + 1)} + O(s^2), \tag{B.20}$$

$$\frac{D(X)}{\langle C \rangle} = \frac{N}{\mu} + O(s^2). \tag{B.21}$$

The ratio $\frac{N}{\mu\,(4N\mu + 1)}$ is shown in Fig. 3.5C.

B.5 The bound on information accumulation rate – Markov chains

The bound on information accumulation rate, as stated in Eq. (3.10,3.11), holds across several different model classes. Here we derive it for models that are Markov chain, in particular the Wright-Fisher model and the discrete Moran model. The two following sections contain similar derivations for continuous time Markov chains and the diffusion approximation. Note that all of the model parameters, such as those that describe selection, mutation or population size, can be time dependent, but we do not write it explicitly as we only need to focus on a single time step.

In the Markov chains class of models, the population state X^t takes discrete values x^t at discrete time steps t. The distribution over states is governed by

$$\varphi^{X^{t+1}}(x^{t+1}) = \sum_{x^t} Q(x^{t+1}|x^t)\,\varphi^{X^t}(x^t) \qquad \text{under neutrality,} \tag{B.22}$$

$$\psi^{X^{t+1}}(x^{t+1}) = \sum_{x^t} P(x^{t+1}|x^t)\,\psi^{X^t}(x^t) \qquad \text{under selection,} \tag{B.23}$$

where $\varphi^{X^t}(x^t)$ and $\psi^{X^t}(x^t)$ are the marginal distributions over population states at time t, and $Q(x^{t+1}|x^t) = \varphi^{X^{t+1}|X^t}(x^{t+1}|x^t)$ and $P(x^{t+1}|x^t) = \psi^{X^{t+1}|X^t}(x^{t+1}|x^t)$ are the transition probabilities. $Q(x^{t+1}|x^t)$ and $P(x^{t+1}|x^t)$, as well as all the parameters that we later introduce to specify them, can be time-dependent, but we do not write it explicitly. The population-level information at time t is

$$D(X^t) = \sum_{x^t} \psi^{X^t}(x^t) \log_2 \frac{\psi^{X^t}(x^t)}{\varphi^{X^t}(x^t)}. \tag{B.24}$$

In general, the chain rule Eq. (3.2) yields a bound

$$\Delta D(X^t) = D(X^{t+1}) - D(X^t) = D(X^{t+1}|X^t) - D(X^t|X^{t+1}) \tag{B.25}$$

$$\leq D(X^{t+1}|X^t) = \sum_{x^t} \psi^{X^t}(x^t) \sum_{x^{t+1}} P(x^{t+1}|x^t) \log_2 \frac{P(x^{t+1}|x^t)}{Q(x^{t+1}|x^t)}. \tag{B.26}$$

The expression $\leq D(X^{t+1}|X^t)$ corresponds to the expected KL cost of control Todorov (2006); Theodorou (2015). In the special case when $Q(x^{t+1}|x^t)$ and $P(x^{t+1}|x^t)$ are independent of time $\psi^{X^t}(x^t)$ is the stationary distribution of $P(x^{t+1}|x^t)$, $D(X^{t+1}|X^t)$ is also the KL divergence rate between $P(x^{t+1}|x^t)$ and $Q(x^{t+1}|x^t)$. We now examine specific forms of the transition probabilities given by the Wright-Fisher model and the Moran model.

B.5.1 Wright-Fisher model

In this general model, each time step t represents a generation, and consists of sampling a new population of N offspring genotypes that constitute the population at $t+1$. The basic assumption is that the offspring genotypes are sampled independently with with probabilities $q_g(x^t)$ without selection or $p_g(x^t)$ with selection, leading to multinomial probability distributions over the frequencies x^{t+1} in the next generation,

$$Q(x^{t+1}|x^t) = \binom{N}{Nx^{t+1}} \prod_g q_g(x^t)^{Nx_g^{t+1}}, \tag{B.27}$$

$$P(x^{t+1}|x^t) = \binom{N}{Nx^{t+1}} \prod_g p_g(x^t)^{Nx_g^{t+1}}, \tag{B.28}$$

where $\binom{N}{Nx^{t+1}} = \frac{N!}{\prod_g (Nx_g^{t+1})!}$ is the multinomial coefficient. We can use these expressions to write down the general bound Eq. (B.26) as

$$\Delta D(X^t) \leq D(X^{t+1|X^t}) = N \sum_{x^t} \psi^{X^t}(x^t) \sum_g p_g(x^t) \log_2 \frac{p_g(x^t)}{q_g(x^t)} \tag{B.29}$$

The probabilities $q_g(x^t)$ capture arbitrary mutation and recombination, and we show examples of the form $q_g(x^t)$ can take below. Selection is modelled by the relationship between $q_g(x^t)$ and $p_g(x^t)$, and this can be done in two ways, which we discuss in the following subsections.

Asexual reproduction. In an asexual population with mutation, $q_g(x)$ will have the form

$$q_g(x) = \sum_{g'} x_{g'} \, \rho^{\text{asex}}_{g'g}, \tag{B.30}$$

where $\rho^{\text{asex}}_{g'g}$ is the probability that a parent with genotype g' produces offspring with genotype g, and includes arbitrary mutation.

Sexual reproduction, random mating. Provided that recombination happens always between two parental genotypes, we sum over possible pairs of parental genotypes,

$$q_g(x) = \sum_{g_1 g_2} x_{g_1} x_{g_2} \, \rho^{\text{rec}}_{g_1 g_2 g} \tag{B.31}$$

with $x_{g_1} x_{g_2}$ being the probability of a parental pair g_1, g_2 and $\rho^{\text{rec}}_{g_1 g_2 g}$ the probability that this pair produces offspring with genotype g. This includes arbitrary mutation and recombination.

Alternatively, we can distinguish between two sexes, classifying each genotype as male ($g \in \mathcal{G}^M$) of female ($g \in \mathcal{G}^F$). We sum over all male-female pairs,

$$q_g(x) = \sum_{g_m \in \mathcal{G}^M} \sum_{g_f \in \mathcal{G}^F} \frac{x_{g_m} x_{g_f}}{Z(x)} \, \rho^{\text{sex}}_{g_m g_f g} \tag{B.32}$$

with $\rho^{\text{sex}}_{g_m g_f g}$ being the probability that parents g_m, g_f give rise to offspring g and

$$Z(x) = \sum_{g_m \in \mathcal{G}^M} \sum_{g_f \in \mathcal{G}^F} x_{g_m} x_{g_f} \tag{B.33}$$

is a normalization factor such that $x_{g_m} x_{g_f} / Z(x)$ is the probability of a parental pair g_m, g_f.

Nonrandom mating. In the case of sexual reproduction, we can replace the expression $x_{g_m} x_{g_f} / Z(x)$ by a different expression for the probability of sampling a mating pair g_m, g_f. For example, we can include a factor $0 \leq \sigma_{g_m g_f} \leq 1$ corresponding to the probability that individuals with this pair of genotypes will mate. Then $q_g(x)$ will have the form

$$q_g(x) = \sum_{g_m \in \mathcal{G}^M} \sum_{g_f \in \mathcal{G}^F} \frac{x_{g_m} x_{g_f} \sigma_{g_m g_f}}{\tilde{Z}(x)} \, \rho^{\text{sex}}_{g_m g_f g}, \tag{B.34}$$

with normalization

$$\tilde{Z}(x) = \sum_{g_m \in \mathcal{G}^M} \sum_{g_f \in \mathcal{G}^F} x_{g_m} x_{g_f} \sigma_{g_m g_f}. \tag{B.35}$$

Selection in an infinite offspring pool

Here we assume that all individuals (under asexual reproduction) or all pairs of individuals (under sexual reproduction) in the population at time t contribute a large number of offspring to a common pool. The next generation then consists of N individuals sampled from this pool, with or without selection.

The genotype frequencies in the pool will be equal to $q_g(x^t)$, since this is the probability that a random invididual (or pair of individuals) from the population x^t has offspring with genotype g. Under neutrality, the genotypes that survive and constitute the next generation are sampled at random. Under selection, genotypes from the pool are sampled with probabilities proportional to fitness $w_g(x^t)$, leading to

$$p_g(x^t) = \frac{q_g(x^t) w_g(x^t)}{\sum_{g'} q_{g'}(x^t)\, w_{g'}(x^t)} = q_g(x^t)\, \tilde{w}_g(x^t). \tag{B.36}$$

where $\tilde{w}_g(x^t) = \frac{w_g(x^t)}{\sum_{g'} q_{g'}(x^t)\, w_{g'}(x)}$ is the relative fitness of g calculated within the offspring pool where selection takes place. Combining it with Eq. (B.29), we obtain the bound

$$\Delta D(X^t) \leq D(X^{t+1}|X^t) = N \sum_{x^t} \psi^{X^t}(x^t) \sum_g q_g(x^t)\, \tilde{w}_g(x^t) \log_2 \tilde{w}_g(x^t) = N \langle C^t_{\text{pool}} \rangle, \tag{B.37}$$

where the last expression coincides with the definition of the cost of selection in Main Text Eq. (B.90), calculated at time t within the offspring pool where selection takes place, and averaged over possible population states x^t.

Selection among parents

Here we assume that selection takes place before reproduction. From the population at time t, we first sample N genotypes (under asexual reproduction) or N pairs of genotypes (under sexual reproduction) as parents. These are sampled independently, with probabilities proportional to fitness. Then we sample one offspring genotype for each parent/pair of parents, with mutation and recombination, and these constitute the population at time $t+1$.

When sampling genotypes as parents, g gets picked with probability given by its frequency x^t_g under neutrality, and $x^t_g \hat{w}_g(x^t)$ under selection where $\hat{w}_g(x^t) = \frac{w_g(x^t)}{\sum_{g'} x_{g'}\, w_{g'}(x)}$ is the relative fitness of genotype g, now computed within the adult population at time t.

We can rewrite Eq. (B.29) with

$$p_g(x^t) = q_g\left(\hat{w}(x^t) \circ x^t\right), \tag{B.38}$$

where $(\hat{w}(x^t) \circ x^t)_g = \hat{w}_g(x^t)\, x^t_g$ is the vector of genotype frequencies that are weighted by their relative fitness. However, Eq. (B.38) is difficult to analyse. An easier approach is to introduce an intermediate variable Y^t, which signifies the genotype frequencies among parents (or pairs of parents) at generation t. The Wright-Fisher model is a Markov chain of the form

$$\cdots \to X^t \to Y^t \to X^{t+1} \to Y^{t+1} \to \cdots . \tag{B.39}$$

Selection operates in the step $X^t \to Y^t$ when parents are sampled from the existing population, but not in the step $Y^t \to X^{t+1}$ where reproduction takes place with mutation and recombination. This can be expressed by writing the chain rule

$$D(X^{t+1}|X^t) = D(Y^t|X^t) + D(X^{t+1}|Y^t, X^t) - D(Y^t|X^{t+1}, X^t) \leq D(Y^t|X^t) \tag{B.40}$$

where the term $D(X^{t+1}|Y^t, X^t) = D(X^{t+1}|Y^t) = 0$ because $Y^t \to X^{t+1}$ is the reproduction step with no selection. The term $D(Y^t|X^{t+1}, X^t)$ is nonnegative and can be dropped since an upper bound on $D(X^{t+1}|X^t)$ is sufficient for our purposes, but deserves some attention as it could make the bound loose. $D(Y^t|X^{t+1}, X^t)$ can only be large when, given the genotype frequencies X^t, X^{t+1} at two subsequent generations, there is uncertainty about the genotype frequencies among the parents Y^t sampled from X^t that gave rise to X^{t+1}. In diverse populations with low mutation rates or large genomes, parents tend to be easy to identify, and $D(Y^t|X^{t+1}, X^t)$ will be small.

Finally, to calculate $D(Y^t|X^t)$ we note that conditionally on X^t, Y^t is a multinomial variable with kN trials ($k = 1$ under asexual reproduction or $k = 2$ under sexual reproduction) and probabilities x^t_g under neutrality or $x^t_g \hat{w}_g(x^t)$ under selection. Then the bound can be written as

$$\Delta D(X^t) \leq D(Y^t|X^t) = kN \sum_{x^t} \psi^{X^t}(x^t) \sum_g x^t_g \, \hat{w}_g(x^t) \log_2 \hat{w}_g(x^t) = kN\langle C^t \rangle. \tag{B.41}$$

Here we have assumed no distinction between sexes, but we can extend to that case by sampling N parents of each sex separately and find

$$\Delta D(X^t) \leq D(Y^t|X^t) = N\langle C^t_{\text{male}} \rangle + N\langle C^t_{\text{female}} \rangle. \tag{B.42}$$

B.5.2 Discrete-time Moran model

This model can be defined similarly to the Wright-Fisher model, but under the Moran model each time step consists of only one birth and one death, and there are N such time steps per generation.

The genotype that dies is chosen at random from the population x^t, and the probability that it will be g is equal to its frequency x^t_g. The probability that the genotype born is g' is $q_{g'}(x^t)$ under neutrality and $p_{g'}(x^t)$ under selection. These can take the same form as under the Wright-Fisher model, with selection within an infinite offspring pool or among parents. This gives rise to the transition probabilities

$$Q(x^{t+1}|x^t) = \begin{cases} x^t_g q_{g'}(x^t); & x^{t+1} = x^t + \frac{e^{g'}}{N} - \frac{e^g}{N},\ g' \neq g \quad (g \text{ dies, } g' \text{ born}), \\ \sum_g x^t_g q_g(x^t); & x^{t+1} = x^t, \\ 0; & \text{otherwise}, \end{cases} \tag{B.43}$$

$$P(x^{t+1}|x^t) = \begin{cases} x^t_g p_{g'}(x^t); & x^{t+1} = x^t + \frac{e^{g'}}{N} - \frac{e^g}{N},\ g' \neq g \quad (g \text{ dies, } g' \text{ born}), \\ \sum_g x^t_g p_g(x^t); & x^{t+1} = x^t, \\ 0; & \text{otherwise}, \end{cases} \tag{B.44}$$

where e^g is a vector of genotype frequencies with one at element g and zeros elsewhere. Now the bound on information accumulation rate, Eq. (B.26), simplifies to

$$\Delta D(X^t) \leq D(X^{t+1}|X^t) = \sum_{x^t} \psi^{X^t}(x^t) \left(\sum_g p_g(x^t) \log_2 \frac{p_g(x^t)}{q_g(x^t)} - \sum_g \beta_g(x^t) \log_2 \frac{\beta_g(x^t)}{\alpha_g(x^t)} \right) \tag{B.45}$$

$$\leq \sum_{x^t} \psi^{X^t}(x^t) \sum_g p_g(x^t) \log_2 \frac{p_g(x^t)}{q_g(x^t)}, \tag{B.46}$$

$$\leq kN\langle C^t \rangle \tag{B.47}$$

where we have used $\alpha_g(x^t) = \frac{x_g^t q_g(x^t)}{\sum_{g'} x_{g'}^t q_{g'}(x^t)}$ and $\beta_g(x^t) = \frac{x_g^t p_g(x^t)}{\sum_{g'} x_{g'}^t p_{g'}(x^t)}$ to denote the probability that if the genotype that is born and dies is the same, it is the genotype g. Even though selection makes $\beta_g(x^t)$ different from $\alpha_g(x^t)$, such replacements leave the population unchanged regardless of g, and this is associated with the nonpositive second term inside the brackets in Eq. (B.45), which reduces the amount of information that can be accumulated.

Eq. (B.46) is almost identical to the bound on information accumulation rate under the Wright-Fisher model, Eq. (B.29). This means that the discussion of the two models of selection in Sec. B.5.1 and B.5.1 also applies to the Moran model, which allows us to write Eq. (B.47). The only difference is that in each time step, we sample only one genotype from the offspring pool, or one parent (or pair of parents). This is reflected in the missing factor N. As there are N time steps per generation, the bound on information accumulated per generation again scales with $kN\langle C^t\rangle$.

Note, however, that the effective population size of the Moran model (i.e. the population size of a Wright-Fisher model with the same covariance in allele frequency change per generation) is $N_e = N/2$. This is because in the Moran model, both the births and deaths are random events, whereas the Wright-Fisher model only has random births. We can therefore expect the tighter bound $\Delta D(X^t) \lesssim kN_e\langle C^t\rangle$ to hold for large enough populations, as both models approach the same diffusion limit. We also prove the bound under the diffusion approximation separately in Sec. B.7.

B.6 The bound on information accumulation rate – continuous-time Markov chains

In this class of models, the genotype frequencies X^t are discrete, but the time t is continuous. The distributions $\varphi^{X^t}(x)$, $\psi^{X^t}(x)$ over X^t are governed by the master equations

$$\frac{d}{dt}\varphi^{X^t}(x) = \sum_{x'} \bar{Q}(x,x')\,\varphi^{X^t}(x') \qquad \text{under neutrality,} \tag{B.48}$$

$$\frac{d}{dt}\psi^{X^t}(x) = \sum_{x'} \bar{P}(x,x')\,\psi^{X^t}(x') \qquad \text{under selection,} \tag{B.49}$$

where $\bar{Q}(x,x')$ and $\bar{P}(x,x')$ are transition rates from x' to x under neutrality and selection respectively. These can be time dependent. The population-level information $D(X^t)$ is defined as in Eq. (B.24), but now changes continuously in time. The rate of this change is upper bounded as

$$\frac{d}{dt}D(X^t) \leq \sum_x \psi^{X^t}(x) \sum_{x,\,x'\neq x}\left(\bar{P}(x',x)\log\frac{\bar{P}(x',x)}{\bar{Q}(x',x)} + \bar{Q}(x',x) - \bar{P}(x',x)\right). \tag{B.50}$$

When $\bar{P}$ and $\bar{Q}$ are independent of time and $\psi^{X^t}(x)$ is the stationary distribution associated with $\bar{P}$, the right hand side corresponds to the KL divergence rate between $\bar{P}$ and $\bar{Q}$, as derived in Kesidis and Walrand (1993). The bound Eq. (B.50) can be verified algebraically, or derived from the discrete bound Eq. (B.26) by taking $Q = e^{\epsilon\bar{Q}}$, $P = e^{\epsilon\bar{P}}$ and the limit $\epsilon \to 0$. We now show an example of the form that $\bar{P}$ and $\bar{Q}$ can take.

B.6.1 Continuous-time Moran model

This model is based on its discrete-time counterpart in Sec. B.5.2. Only transitions consisting of replacing one genotype (g, death) by another (g', birth) are allowed, and the transition rates have the form

$$\bar{Q}(x', x) = \begin{cases} N x_g q_{g'}(x); & x' = x - \frac{e^g}{N} + \frac{e^{g'}}{N},\ g' \neq g, \\ -\sum_{g, g' \neq g} N x_g q_{g'}(x); & x' = x, \\ 0; & \text{otherwise}, \end{cases} \tag{B.51}$$

$$\bar{P}(x', x) = \begin{cases} N x_g p_{g'}(x); & x' = x - \frac{e^g}{N} + \frac{e^{g'}}{N},\ g' \neq g, \\ -\sum_{g, g' \neq g} N x_g p_{g'}(x); & x' = x, \\ 0; & \text{otherwise}, \end{cases} \tag{B.52}$$

where time is measured in generations, i.e. there are on average N replacement events per unit time. With this form of the transition rates, we can rewrite the general bound Eq. (B.50) as

$$\frac{d}{dt} D(X^t) \leq N \sum_x \psi^{X^t}(x) \sum_{g, g' \neq g} x_g \left(p_{g'}(x) \log \frac{p_{g'}(x)}{q_{g'}(x)} + q_{g'}(x) - p_{g'}(x) \right) \tag{B.53}$$

$$= N \sum_x \psi^{X^t}(x) \sum_g \left(p_g(x) \log \frac{p_g(x)}{q_g(x)} - x_g q_g(x) \left(\frac{p_g(x)}{q_g(x)} \log \frac{p_g(x)}{q_g(x)} + 1 - \frac{p_g(x)}{q_g(x)} \right) \right) \tag{B.54}$$

$$\leq N \sum_x \psi^{X^t}(x) \sum_g p_g(x) \log \frac{p_g(x)}{q_g(x)}. \tag{B.55}$$

This is the same bound as for the discrete Moran model in Eq. (B.47), up to the factor N, which is due to the different unit of time.

B.7 The bound on information accumulation rate – diffusion approximation

Here we show an upper bound on the rate of accumulation of information under the diffusion approximation. The approach is similar to calculations by Iwasa Iwasa (1988) (who assumed detailed balance) and Hasegawa Hasegawa (1977) (who did not), but here we distinguish between the processes with and without selection. We start by deriving a general bound for a pair of diffusion processes, and then apply it to the population genetics context in Sec. B.7.1.

For brevity, we write the probability density over population states x at time t as $\varphi = \varphi(x, t)$ under neutrality and $\psi = \psi(x, t)$ under selection. The population-level information is now determined by integration,

$$D(X) = \int \psi \log_2 \frac{\psi}{\varphi}\, dx = \frac{1}{\ln 2} \int \psi \ln \frac{\psi}{\varphi}\, dx. \tag{B.56}$$

Note that while we stick to measuring information in bits, it is more convenient to use the natural logarithm during the derivation.

The diffusion equation is parametrized by the first and second moment of change in x_g. Selection is assumed to only exert control through the first moment, which we label as a_g

under neutrality and $a_g + a^s_g$ under selection. The second moment is $b_{gg'}$, both under neutrality and under selection. All these are functions of x and t, e.g. $a_g = a_g(x,t)$, but we do not write this for brevity. We sum over any index that appears twice in a term, e.g. $\partial_g b_{g'g} = \sum_{g=1}^{\mathcal{G}-1} \partial_g b_{g'g}$. Note that the diffusion is described in the subspace of $\mathcal{G}-1$ genotype frequencies, where $\mathcal{G}$ is the number of genotypes – the last frequency is determined by normalization, $x_{\mathcal{G}} = 1 - \sum_{g=1}^{\mathcal{G}-1} x_g$. The diffusion equation is

$$\partial_t \varphi = -\partial_g \left(u_g \varphi\right), \qquad u_g = a_g \qquad - \frac{1}{2}\partial_{g'} b_{gg'} - \frac{1}{2} b_{gg'} \partial_{g'} \ln \varphi \qquad \text{under neutrality,} \tag{B.57}$$

$$\partial_t \psi = -\partial_g \left(v_g \psi\right), \qquad v_g = a_g + a^s_g - \frac{1}{2}\partial_{g'} b_{gg'} - \frac{1}{2} b_{gg'} \partial_{g'} \ln \psi \qquad \text{under selection,} \tag{B.58}$$

where we introduced the velocity fields $u_g = u_g(x,t)$ and $v_g = v_g(x,t)$ such that $u_g\varphi$ and $v_g\psi$ are the probability fluxes under neutrality and under selection respectively. From their definition, it follows that

$$\partial_g \ln \frac{\psi}{\varphi} = -2 b^{-1}_{gg'} \left(v_{g'} - u_{g'} - a^s_{g'}\right). \tag{B.59}$$

The rate of change of $D(X)$ can be written as

$$\frac{d}{dt} D(X) = \frac{1}{\ln 2} \int \partial_t \left(\psi \ln \frac{\psi}{\varphi}\right) dx, \tag{B.60}$$

and the integrand can be written as

$$\begin{aligned}
\partial_t \left(\psi \ln \frac{\psi}{\varphi}\right) &= \ln \frac{\psi}{\varphi}\, \partial_t \psi + \partial_t \psi - \frac{\psi}{\varphi}\, \partial_t \varphi && \text{(B.61)}\\
&= -\ln \frac{\psi}{\varphi}\, \partial_g \left(v_g \psi\right) - \partial_g \left(v_g \psi\right) + \frac{\psi}{\varphi}\, \partial_g \left(u_g \varphi\right) && \text{(B.62)}\\
&= -\partial_g \left(v_g \psi \ln \frac{\psi}{\varphi} + \left(v_g - u_g\right)\psi\right) + \psi \left(v_g - u_g\right) \partial_g \ln \frac{\psi}{\varphi}. && \text{(B.63)}
\end{aligned}$$

The first term is a divergence and vanishes after integration in Eq. (B.60) because $u_g\varphi$ and $v_g\psi$ cannot cross the domain boundary (assuming $\psi/\varphi < \infty$). Therefore

$$\begin{aligned}
\frac{d}{dt} D(X) &= \frac{1}{\ln 2} \int \psi \left(v_g - u_g\right) \partial_g \ln \frac{\psi}{\varphi}\, dx && \text{(B.64)}\\
&= -\frac{2}{\ln 2} \int \psi \left(v_g - u_g\right) b^{-1}_{gg'} \left(v_{g'} - u_{g'} - a^s_{g'}\right) dx && \text{(B.65)}\\
&= -\frac{2}{\ln 2} \int \psi \left(v_g - u_g - \frac{1}{2} a^s_g\right) b^{-1}_{gg'} \left(v_{g'} - u_{g'} - \frac{1}{2} a^s_{g'}\right) dx + && \text{(B.66)}\\
&\quad + \frac{1}{2 \ln 2} \int \psi\, a^s_g\, b^{-1}_{gg'}\, a^s_{g'}\, dx. && \text{(B.67)}
\end{aligned}$$

The last expression has two terms, both of which are quadratic forms. The first one makes a nonpositive contribution, leading to the upper bound in information accumulation rate,

$$\frac{d}{dt} D(X) \leq \frac{1}{2 \ln 2} \int \psi\, a^s_g\, b^{-1}_{gg'}\, a^s_{g'}\, dx. \tag{B.68}$$

On the right hand side we can identify the KL cost of control Theodorou (2015) in bits. This bound holds for any pair of diffusion processes with the same fluctuations covariance $b_{gg'}$. We will now discuss it in the context of population genetics.

B.7.1 Application to population genetics

The bound Eq. (B.68) does not depend on the form of a_g, and therefore it can be used to model arbitrary mutation and recombination, for example by taking

$$a_g = q_g - x_g \tag{B.69}$$

with $q_g = q_g(x)$ as introduced in the discrete models above. Notably, a_g does not need to be the gradient of a scalar potential (an assumption in free fitness Iwasa (1988) and fitness flux Mustonen and Lässig (2010) theories).

The fluctuations in genotype frequencies can modeled based on multinomial sampling in the Wright-Fisher model (Eq. (B.27,B.28)),

$$b_{gg'} = \frac{\delta_{gg'} x_g - x_g x_{g'}}{N}, \tag{B.70}$$

with no summation over g, and where the (effective) population size N can be time-dependent. This has the inverse Withers and Nadarajah (2014)

$$b_{gg'}^{-1} = N\left(\frac{\delta_{gg'}}{x_g} + \frac{1}{x_{\mathcal{G}}}\right). \tag{B.71}$$

The control term a_g^s imposed by selection can be written as

$$a_g^s = (\hat{w}_g - 1)\, x_g \qquad \text{where} \qquad \hat{w}_g = \frac{w_g}{\sum_g w_g x_g} \tag{B.72}$$

where w_g is the fitness of genotype g, possibly time and frequency dependent, and $\hat{w}_g$ is the relative fitness. With these definitions, we find that

$$\sum_{g,g'=1}^{\mathcal{G}-1} a_g^s\, b_{gg'}^{-1}\, a_{g'}^s = N \sum_{g=1}^{\mathcal{G}} \frac{(a_g^s)^2}{x_g} = N \sum_{g=1}^{\mathcal{G}} (\hat{w}_g - 1)^2\, x_g = NV(x) \tag{B.73}$$

and the bound on information accumulation rate is

$$\frac{d}{dt} D(X) \le \frac{N\langle V\rangle}{2\ln 2} = N\langle C\rangle. \tag{B.74}$$

In the last equation, we identify $\frac{\langle V\rangle}{2\ln 2} = \langle C\rangle$ – this can be derived under weak selection in discrete time, see Sec. B.9.

We can get intuition about the tightness of this bound by analyzing the first, nonpositive term in Eq. (B.67). The bound is only tight when $v_g - u_g - \frac{1}{2}a_g^s = 0$ for all x with nonzero ψ. An interesting specific case is when the neutral process is at an equilibrium with detailed balance, such that $u_g = 0$. We note that v_g can be decomposed as $v_g = v_g' + a_g^s$ into the contribution from selection a_g^s and all other evolutionary forces, v_g'. Our bound is then tight when $a_g^s = -2v_g'$, i.e. when selection induces a probability flux in exactly the opposite direction and exactly twice the magnitude as the all the other evolutionary forces combined. While this might occasionally and approximately be the case, it will only be a transient phenomenon.

Suppose that the population starts at the neutral equilibrium with $v_g' = 0$ and then selection starts to act, e.g. after a change in the environment. For any nonzero a_g^s, the bound cannot be tight as $a_g^s \neq -2v_g' = 0$. After some time a new equilibrium might be reached, with $v_g = v_g' + a_g^s = 0$, where again, the bound is not tight $a_g^s \neq -2v_g' = 2a_g^s$. In this case, maintenance costs are incurred but no further adaptation takes place. The bound can only be tight for a moment when adaptation is taking place, selection pulls the population in the opposite direction as the other evolutionary forces combined, but selection is twice as strong, $a_g^s = -2v_g'$.

B.8 Relationship with free fitness and statistical physics

Stochastic models in population genetics show some mathematical properties analogous to statistical physics. In particular, a quantity called free fitness, analogous to free energy in physics, can be defined and shown to monotonically increase over time. In this section we provide some background about free fitness and discuss two connections with our work. First, the increasing property of free fitness can be proved by a method similar to the proof of our bound on information accumulation rate. Second, free fitness can be written as the difference between mean log fitness and genetic information as defined in this paper (e.g. on the genotype or population level), implying that evolution tends to maximize mean log fitness at a given amount of information.

B.8.1 Boltzmann form of stationary distributions

Under suitable conditions, the stationary distributions in population genetics models take a form similar to the Boltzmann distribution. Models on both the population level and the genotype level display this property.

- If mutation is weak ($NU \ll 1$ where U is the total mutation rate across the studied genomic region), populations are mostly monomorphic, with only occasional fixations of a different genotype. The system can then be described with the most recently fixed genotype g and the distribution $\psi^G(g)$. The stationary distribution $\tilde{\psi}^G(g)$ takes the form Berg et al. (2004); Sella and Hirsh (2005)

$$\tilde{\psi}^G(g) = \frac{1}{Z^G} e^{2N \ln w_g}, \tag{B.75}$$

  where $2N$ is again analogous to inverse temperature, log fitness $\ln w_g$ is analogous to negative energy, and Z^G is normalization constant.

- Assuming many biallelic loci under linkage equilibrium, we can describe the system with the vector of allele frequencies p and the joint distribution $\psi^P(p)$ over them. The stationary distribution $\tilde{\psi}^P(p)$ can be derived from the diffusion approximation and takes the form Wright (1937); de Vladar and Barton (2011)

$$\tilde{\psi}^P(p) = \frac{1}{Z^P} \prod_i (p_i q_i)^{2N\mu - 1} e^{2N \ln \bar{w}(p)}, \tag{B.76}$$

  where p_i and $q_i = 1 - p_i$ are the allele frequencies of the two alleles at locus i and Z^P is a normalization constant. Twice the population size $2N$ takes the role of inverse temperature and log mean fitness $\ln \bar{w}(p)$ takes the role of negative energy. The factors $(p_i(1-p_i))^{2N\mu-1}$ correspond to mutation and drift potential (similar to e.g. chemical potential), which will be made clearer in the next subsection.

The formulas apply to haploids, but similar formulas apply to diploids or when mutation coefficients vary across loci or alleles (see e.g. the SI of Sella and Hirsh (2005)). Importantly, they depend on the assumption of detailed balance. At stationarity, net probability flux between any two allele frequency vectors or genotypes must be zero, making $\tilde{\psi}^G(g)$ and $\tilde{\psi}^P(p)$ equilibrium distributions. This can be violated under certain forms of mutation, recombination and strong selection, leading to additional terms related to robustness Rao and Leibler (2022).

B.8.2 Free fitness

When a system starts from an arbitrary initial distribution and approaches the stationary distribution in Eq. (B.76) or Eq. (B.75), we can track this progress using free fitness Iwasa (1988); Sella and Hirsh (2005), which increases monotonically in time as was shown previously and as we can also prove in more generality here.

- On the genotype level, we can define free fitness F^G at any distribution $\psi^G(g)$ away from equilibrium as a sum of expected log fitness and entropy terms,

$$F^G = \underbrace{\langle \ln w_g \rangle_{\psi^G}}_{\text{Selection}} + \underbrace{\frac{1}{2N}\left\langle \ln \frac{1}{\psi^G(g)} \right\rangle_{\psi^G}}_{\text{Entropy}} \tag{B.77}$$

$$= \underbrace{\frac{\ln Z^G}{2N}}_{\substack{\text{Equilibrium}\\ \text{free fitness}}} - \underbrace{\frac{1}{2N}\left\langle \ln \frac{\psi^G(g)}{\tilde{\psi}^G(g)} \right\rangle_{\psi^G}}_{\substack{\text{KL divergence}\\ \text{from equilibrium}}} = \tilde{F}^G - \frac{1}{2N} D_{KL}(\psi^G || \tilde{\psi}^G) \tag{B.78}$$

  where $\langle \cdot \rangle_{\psi^G}$ denotes an expectation over $g \sim \psi^G(g)$. In the special case of equilibrium $\psi^G = \tilde{\psi}^G$, we obtain $F^G = \tilde{F}^G = \frac{\ln Z^G}{2N}$, and away from equilibrium, free fitness is reduced by an amount proportional to the KL divergence between the actual distribution ψ^G and the equilibrium $\tilde{\psi}^G$.

- On the population level with linkage equilibrium, free fitness at a distribution ψ^P can be defined as a sum of three terms – a negative potential for selection, mutation and drift, and entropy:

$$F^P = \underbrace{\langle \ln \bar{w}(p) \rangle_{\psi^P}}_{\text{Selection}} + \underbrace{\frac{2N\mu - 1}{2N}\left\langle \sum_i \ln(p_i q_i) \right\rangle_{\psi^P}}_{\text{Mutation and drift}} + \underbrace{\frac{1}{2N}\left\langle \ln \frac{1}{\psi^P(p)} \right\rangle_{\psi^P}}_{\text{Entropy}} \tag{B.79}$$

$$= \underbrace{\frac{\ln Z^P}{2N}}_{\substack{\text{Equilibrium}\\ \text{free fitness}}} - \underbrace{\frac{1}{2N}\left\langle \ln \frac{\psi^P(p)}{\tilde{\psi}^P(p)} \right\rangle_{\psi^P}}_{\substack{\text{KL divergence}\\ \text{from equilibrium}}} = \tilde{F}^P - \frac{1}{2N} D_{KL}(\psi^P || \tilde{\psi}^P) \tag{B.80}$$

  where the expectations $\langle \cdot \rangle$ are taken over $p \sim \psi^P(p)$. While mutation and drift now appear as additional terms in free fitness, free fitness can again be decomposed into its value at equilibrium and a difference proportional to the KL divergence away from it.

The key property of the free fitness is that it is a non-decreasing function of time, until it is maximized at equilibrium. This was proved by Iwasa Iwasa (1988) for the case of F^P and Sella and Hirsh Sella and Hirsh (2005) for the case of F^G in low mutation regime. In the next section we show that both results can also be derived by the same method as our bound on information accumulation rate.

B.8.3 Convergence to stationary distributions

The general bounds on information accumulation rate (Eq. (B.26) for Markov chains, Eq. (B.50) for continuous time Markov chains and Eq. (B.68) for the diffusion approximation) apply for any pair of stochastic processes, provided that they have compatible support such that the KL divergence is well defined. To make this explicit, we focus on the case of discrete Markov chains, consider some general process ξ instead of the neutral process φ, and introduce notation that generalizes $D(X^t)$,

$$D_{\psi||\xi}(X^t) = \sum_{x^t} \psi^{X^t}(x^t) \, \log_2 \frac{\psi^{X^t}(x^t)}{\xi^{X^t}(x^t)} \tag{B.81}$$

$$D_{\psi||\xi}(X^{t+1}|X^t) = \sum_{x^t} \psi^{X^t}(x^t) \sum_{x^{t+1}} \psi^{X^{t+1}|X^t}(x^{t+1}|x^t) \, \log_2 \frac{\psi^{X^{t+1}|X^t}(x^{t+1}|x^t)}{\xi^{X^{t+1}|X^t}(x^{t+1}|x^t)}. \tag{B.82}$$

The KL divergence chain rule now yields the inequality

$$\Delta D_{\psi||\xi}(X^t) = D_{\psi||\xi}(X^{t+1}) - D_{\psi||\xi}(X^t) \leq D_{\psi||\xi}(X^{t+1}|X^t). \tag{B.83}$$

If $\xi = \varphi$ is the neutral process, this is the KL cost of selection bound on the information accumulation rate (Eq. (3.9) and Eq. (B.26)). But we can also choose ξ such that it does contain selection and has the same transition probabilities as ψ, but starts from a different initial condition, i.e. $\xi^{X^{t+1}|X^t} = \psi^{X^{t+1}|X^t}$ and $\xi^{X^0} \neq \psi^{X^0}$. Then we find that $D_{\psi||\xi}(X^{t+1}|X^t) = 0$ and $\Delta D_{\psi||\xi}(X^t) \leq 0$, i.e. the divergence between ψ^{X^t} and ξ^{X^t} is non-increasing over time, because relative to ξ, there is no control exerted on ψ.

If, in addition, the system has a unique stationary distribution $\tilde{\psi}^X$ and ξ^{X^0} is initialized there (and therefore stays there indefinitely, $\xi^{X^0} = \xi^{X^t} = \tilde{\psi}^X$ for any t), we find that ψ^{X^t} converges to this stationary distribution $\tilde{\psi}^X$ monotonically in terms of the KL divergence $D_{\psi||\xi}(X^t)$. Similar proofs apply to continuous time Markov chains and diffusion, since we only need to replace φ by ξ and repeat the derivation leading to Eq. (B.50) and Eq. (B.68).

In the two regimes discussed above in Sec. B.8.1, the population state X corresponds to some fixed genotype G or a vector of allele frequencies P. Therefore $D_{\psi||\xi}(X^t) = D_{KL}(\psi^{G^t}||\tilde{\psi}^G)$ or $D_{\psi||\xi}(X^t) = D_{KL}(\psi^{P^t}||\tilde{\psi}^P)$ are non-increasing functions of time. Together with Eq. (B.78,B.80), this implies that F^G or F^P are non-decreasing functions of time.

Iwasa Iwasa (1988) and Sella and Hirsh Sella and Hirsh (2005) proved the same result by different methods. In our framework it emerges as a special case of the information accumulation bound with zero control. The key part of our proof, stating that $D_{\psi||\xi}(X^t)$ is non-increasing, is also more general (regarding the state space, the form of the stationary distribution, and detailed balance – although free fitness is not defined so generally). A similarly general proof for continuous time Markov chains, as well as several related results for replicator dynamics and reaction networks, is reviewed in reference Baez and Pollard (2016).

B.8.4 Free fitness as a trade-off between fitness and information

We can rewrite the expressions for free fitness using the genotype and population-level information respectively. We first write down the neutral stationary distributions. On the genotype level, we assume that it is uniform over 4^l possible sequences of length l,

$$\tilde{\varphi}^G(g) = \frac{1}{4^l}. \tag{B.84}$$

On the population level,

$$\tilde{\varphi}^P(p) = \frac{1}{Z^{\varphi,P}} \prod_i (p_i q_i)^{2N\mu - 1}, \tag{B.85}$$

where $Z^{\varphi,P}$ is the normalization constant, to be distinguished from Z^P in Eq. (B.76) which includes selection.

Using $\tilde{\varphi}^G(g)$ and $\tilde{\varphi}^P(p)$, we can rewrite free fitness, Eq. (B.78,B.80), as

$$F^G = \underbrace{\langle \ln w_g \rangle_{\psi^G}}_{\text{Selection}} - \frac{1}{2N} \underbrace{\left\langle \ln \frac{\psi^G(g)}{\tilde{\varphi}^G(g)} \right\rangle_{\psi^G}}_{\text{Genotype-level information}} - \underbrace{\frac{1}{2N} \ln(4^l)}_{\text{Independent of } \psi^G} = \langle \ln w_g \rangle_{\psi^G} - \frac{1}{2N} D(G) + \text{const.} \quad \text{(B.86)}$$

on the genotype level and

$$F^P = \underbrace{\langle \ln \bar{w}(p) \rangle_{\psi^P}}_{\text{Selection}} - \frac{1}{2N} \underbrace{\left\langle \ln \frac{\psi^P(p)}{\tilde{\varphi}^P(p)} \right\rangle_{\psi^P}}_{\text{Population-level information}} + \underbrace{\frac{\ln Z^{\varphi,P}}{2N}}_{\text{Independent of } \psi^P} = \langle \ln \bar{w}(p) \rangle_{\psi^P} - \frac{1}{2N} D(P) + \text{const.} \quad \text{(B.87)}$$

on the population level. In both cases we have emphasized that terms independent of ψ^G or ψ^P are constant in time and therefore not important for the dynamics of free fitness. Up to the constant, this formula for free fitness has also been used in the paper on fitness flux Mustonen and Lässig (2010). The fitness flux theorem (ref. Mustonen and Lässig (2010) and Sec. B.10) then provides perhaps the most elegant proof that free fitness is a non decreasing function of time, as it relates changes in $D(P)$ to changes in expected fitness.

Free fitness tends to increase over time until it is maximized at the Boltzmann-like equilibrium distribution $\tilde{\psi}^G$ or $\tilde{\psi}^P$. In other words, evolution maximizes the expected log fitness while constraining the amount of genetic information, with $1/(2N)$ serving as a Lagrange multiplier that controls the trade-off.

B.9 Properties of measures of cost of selection

Here we prove general inequalities between the genetic load $L(x)$, relative fitness variance $V(x)$, and the information theoretic cost $C(x)$. We also derive the form of $C(x)$ for the special cases of weak selection and truncation selection. The three measures are defined as

$$L(x) = 1 - \frac{1}{\hat{w}_{\max}(x)} \quad \text{(B.88)}$$

$$V(x) = \sum_g x_g \left(\hat{w}_g(x) - 1\right)^2, \quad \text{(B.89)}$$

$$C(x) = \sum_g x_g \hat{w}_g(x) \log_2 \hat{w}_g(x), \quad \text{(B.90)}$$

where x_g is the frequency of genotype g in the population and $\hat{w}_{\max}(x) = \max_{g;\, x_g > 0} \hat{w}_g(x)$ is the relative fitness of the fittest individual that is present in the population ($x_g > 0$).

We note that some previous work has defined $\hat{w}_{\max}(x)$ to be the maximum fitness possible, i.e. the fitness of an ideal genotype with no deleterious mutations regardless of whether such an individual exists. Load computed with such a definition is higher, and this has led to claims of severe restrictions on the rate of adaptive substitutions Kimura (1968) and the functional fraction of the human genome Graur (2017). However, load under this definition has been criticized as irrelevant, since the ideal genotype has a vanishing probability of existing in the population, and if only the fitness values likely to be present in the population are considered, load-based restrictions are more permissive Ewens (1970); Galeota-Sprung et al. (2020). Our definitions of $L(x)$, $V(x)$ and $C(x)$ all focus on the existing variation of fitness in the population x. $L(x)$ is also related to the concept of lead, which was defined as the difference between the maximum and the mean log fitness in a traveling wave Desai and Fisher (2007).

Truncation selection and limitations by reproductive capacity. Under truncation selection, a fraction α of individuals in the population has constant relative fitness, equal to the maximum $\hat{w}_g(x) = \hat{w}_{\max}(x)$ and the remaining fraction $1-\alpha$ has relative fitness zero $\hat{w}_g(x) = 0$. By definition, the mean relative fitness must be $\sum_g x_g \hat{w}_g(x) = 1$, which requires $\hat{w}_{\max}(x) = 1/\alpha$. The three measures of cost of selection then are

$$
\begin{aligned}
L^{\textsf{trunc}}(x) &= 1-\alpha, && \text{(B.91)}\\
V^{\textsf{trunc}}(x) &= \alpha\left(\frac{1}{\alpha}-1\right)^2 + (1-\alpha)\,(0-1)^2 = \frac{1}{\alpha}-1, && \text{(B.92)}\\
C^{\textsf{trunc}}(x) &= -\log_2 \alpha. && \text{(B.93)}
\end{aligned}
$$

At a constant population size, the expected number of offspring of an individual is equal to their relative fitness $\hat{w}_g(x)$ (or $2\hat{w}_g(x)$ under sexual reproduction, with two parents per offspring). In a species with a reproductive capacity R, we have $\hat{w}_{\max}(x) \le R$ and the load is limited as $L \le 1 - 1/R$. $V(x)$ and $C(x)$ at given R are maximized under truncation selection, when only the most extreme relative fitness values available are occupied ($N\alpha$ individuals have relative fitness $\hat{w}_g(x) = 1/\alpha = \hat{w}_{\max}(x) = R$ and $N - N\alpha$ individuals have fitness 0). This implies upper bounds $V(x) \le R-1$ and $C(x) \le \log_2 R$.

General inequality between $L(x)$ and $V(x)$. From Eq. (B.88), the genetic load $L(x)$ determines the maximum relative fitness in the population, $\hat{w}_{\max}(x) = 1/(1-L(x))$. Given that relative fitness of all individuals in the population must lie between 0 and $\hat{w}_{\max}(x)$, its variance $V(x)$ is maximized when only these extreme values are occupied, i.e. under truncation selection. In that case we have $V^{\textsf{trunc}}(x) = 1/\alpha - 1$ with $\alpha = 1/\hat{w}_{\max}(x)$. This implies a general bound,

$$
V(x) \le V^{\textsf{trunc}}(x) = \frac{L(x)}{1-L(x)}, \tag{B.94}
$$

with equality under truncation selection. The same inequality was derived in ref. Shnol et al. (2011) by other means.

General inequality between $L(x)$ and $C(x)$. Since logarithm is an increasing function and $x_g \hat{w}_g(x) \le 0$, we can upper bound each term of the form $x_g \hat{w}_g(x) \log_2 \hat{w}_g(x)$ in Eq. (B.90) by $x_g \hat{w}_g(x) \log_2 \hat{w}_{\max}(x)$. Summing over g, we obtain

$$
C(x) \le \sum_g x_g \hat{w}_g(x) \log_2 \hat{w}_{\max}(x) = \log_2 \hat{w}_{\max}(x) = \log_2 \frac{1}{1-L(x)}. \tag{B.95}
$$

Equality is again achieved under truncation selection.

General inequality between $V(x)$ and $C(x)$. We use the inequality $\log_2 u \le \frac{u-1}{\ln 2}$ in Eq. (B.90) to obtain

$$
C(x) \le \sum_g x_g \hat{w}_g(x) \frac{\hat{w}_g(x)-1}{\ln 2} = \frac{1}{\ln 2}\left(\sum_g x_g \hat{w}_g(x)^2 - 1\right) = \frac{V(x)}{\ln 2}. \tag{B.96}
$$

Equality is approached under truncation selection when $\alpha \to 1$.

$C(x)$ **under weak selection.** Here we assume that for all genotypes g present in the population ($x_g > 0$), the relative fitness $\hat{w}_g(x)$ is close to 1. We can then use the Taylor expansion

$$\hat{w}_g(x)\log_2 \hat{w}_g(x) = \frac{1}{\ln 2}\left(\hat{w}_g(x) - 1 + \frac{1}{2}\left(\hat{w}_g(x)-1\right)^2 + O\left(\left(\hat{w}_g(x)-1\right)^3\right)\right). \tag{B.97}$$

Combining this with Main Text Eq. (B.90), we find

$$C(x) = \frac{1}{\ln 2}\sum_g x_g \left(\hat{w}_g(x) - 1 + \frac{1}{2}\left(\hat{w}_g(x)-1\right)^2 + O\left(\left(\hat{w}_g(x)-1\right)^3\right)\right) \tag{B.98}$$

$$= \frac{V(x)}{2\ln 2} + \frac{1}{\ln 2}\sum_g x_g \, O\left(\left(\hat{w}_g(x)-1\right)^3\right), \tag{B.99}$$

or in short, $C(x) \approx V(x)/(2\ln 2)$ under weak selection. This is particularly relevant for the diffusion limit, when the population size is sent to infinity, and the selection strength is rescaled inversely to the population size.

B.10 Fitness flux theorem

In this section we compare the newly introduced bound on information accumulation rate (the *cost of selection bound*) and a similar bound implied by the fitness flux theorem Mustonen and Lässig (2010) (the *fitness flux bound*). The fitness flux theorem was originally derived under the diffusion approximation. For better comparison, we also derive an analogous result for discrete-time Markov chains. We then discuss the distinct interpretation of the two bounds, and illustrate them (using both the discrete and diffusion expressions) in Fig. B.2.

B.10.1 Discrete-time Markov chains

The fitness flux theorem, like its counterparts in statistical physics (e.g. Crooks (2000)), is based on the comparison of forward and reverse path probabilities. For simplicity, we will not derive the fitness flux theorem in its general form, but rather the form that allows a direct comparison with the cost of selection bound.

We focus on short paths consisting of only one step, (X^t, X^{t+1}). The probability of the forward path (x^t, x^{t+1}) is $\psi^{X^t}(x^t)P(x^{t+1}|x^t)$. We consider a probability distribution over reverse paths, $\psi^{X^{t+1}}(x^{t+1})P(x^t|x^{t+1})$, which is normalized to 1 – we can write this as

$$1 = \sum_{x^t\, x^{t+1}} \psi^{X^t}(x^t)\, P(x^{t+1}|x^t) \exp \ln \frac{\psi^{X^{t+1}}(x^{t+1})\, P(x^t|x^{t+1})}{\psi^{X^t}(x^t)\, P(x^{t+1}|x^t)} \tag{B.100}$$

By Jensen's inequality,

$$0 \geq \sum_{x^t\, x^{t+1}} \psi^{X^t}(x^t)\, P(x^{t+1}|x^t) \ln \frac{\psi^{X^{t+1}}(x^{t+1})\, P(x^t|x^{t+1})}{\psi^{X^t}(x^t)\, P(x^{t+1}|x^t)}, \tag{B.101}$$

Next, inside the logarithm, we divide and multiply by the neutral probabilities,

$$0 \geq \sum_{x^t\, x^{t+1}} \psi^{X^t}(x^t)\, P(x^{t+1}|x^t) \ln \frac{\psi^{X^{t+1}}(x^{t+1})\, P(x^t|x^{t+1})\, \frac{\varphi^{X^{t+1}}(x^{t+1})\, Q(x^t|x^{t+1})}{\varphi^{X^{t+1}}(x^{t+1})\, Q(x^t|x^{t+1})}}{\psi^{X^t}(x^t)\, P(x^{t+1}|x^t)\, \frac{\varphi^{X^t}(x^t)\, Q(x^{t+1}|x^t)}{\varphi^{X^t}(x^t)\, Q(x^{t+1}|x^t)}} \tag{B.102}$$

$$= \sum_{x^t\, x^{t+1}} \psi^{X^t}(x^t)\, P(x^{t+1}|x^t) \ln \frac{\frac{\psi^{X^{t+1}}(x^{t+1})\, P(x^t|x^{t+1})}{\varphi^{X^{t+1}}(x^{t+1})\, Q(x^t|x^{t+1})}}{\frac{\psi^{X^t}(x^t)\, P(x^{t+1}|x^t)}{\varphi^{X^t}(x^t)\, Q(x^{t+1}|x^t)}}, \tag{B.103}$$

where we assumed that the neutral process is at a stationary distribution with detailed balance, i.e. $\varphi^{X^t}(x^t)\, Q(x^{t+1}|x^t) = \varphi^{X^{t+1}}(x^{t+1})\, Q(x^t|x^{t+1})$. Finally, we rearrange terms and divide by $\ln 2$ to get an expression in bits,

Fitness flux bound: (B.104)

$$\Delta D(X^t) \leq \sum_{x^t} \psi^{X^t}(x^t) \sum_{x^{t+1}} P(x^{t+1}|x^t) \log_2 \frac{P(x^{t+1}|x^t)\, Q(x^t|x^{t+1})}{Q(x^{t+1}|x^t)\, P(x^t|x^{t+1})} \quad = 2N\langle\phi\rangle_t, \tag{B.105}$$

Cost of selection bound: (B.106)

$$\Delta D(X^t) \leq \sum_{x^t} \psi^{X^t}(x^t) \sum_{x^{t+1}} P(x^{t+1}|x^t) \log_2 \frac{P(x^{t+1}|x^t)}{Q(x^{t+1}|x^t)} \quad \leq kN\langle C\rangle_t, \tag{B.107}$$

where $\langle\phi\rangle_t$ is the discrete analog of the fitness flux, averaged over the possible transitions (x^t, x^{t+1}),

$$\phi(x^t, x^{t+1}) = \frac{1}{2N} \log_2 \frac{P(x^{t+1}|x^t)\, Q(x^t|x^{t+1})}{Q(x^{t+1}|x^t)\, P(x^t|x^{t+1})}. \tag{B.108}$$

The interpretation of this expression and the relationship to fitness accumulation is most clear in the diffusion approximation, see below. To help compare the fitness flux bound with the cost of selection bound, we take the expectation over the final state x^{t+1} and compute the expected fitness flux from any initial state x^t, $\phi(x^t) = \sum_{x^{t+1}} P(x^{t+1}|x^t)\phi(x^t, x^{t+1})$. An example plot of $\phi(x^t)$ is in Fig. B.2A. For comparison, we also included in Eq. (B.107) the cost of selection bound.

While the cost of control is a non-negative conditional KL divergence, the fitness flux bound contains an additional term related to the reverse transition probabilities, and can be negative (this is more easily interpretable as the mutation term in the diffusion approximation). The fitness flux bound relies on the additional assumption that the neutral process is at a stationary distribution with detailed balance. This can be satisfied in the single locus, two allele system when using the Moran model, but it is violated by the Wright-Fisher model which we use throughout most of the paper.

B.10.2 Diffusion approximation

Mustonen and Lässig Mustonen and Lässig (2010) derive the fitness flux theorem using a similar method but under the diffusion approximation, where the fitness flux is related to the rate accumulation of fitness. We include here an informal account of how that relates to the formula in Eq. (B.107).

In continuous time, we can generalize the definition of fitness flux in Eq. (B.108) to an arbitrary time interval Δt and the transition $(x^t, x^{t+\Delta t})$.

$$\phi(x^t, x^{t+\Delta t}) = \frac{1}{\Delta t}\frac{1}{2N} \log_2 \frac{P(x^{t+\Delta t}|x^t)\, Q(x^t|x^{t+\Delta t})}{Q(x^{t+\Delta t}|x^t)\, P(x^t|x^{t+\Delta t})}, \tag{B.109}$$

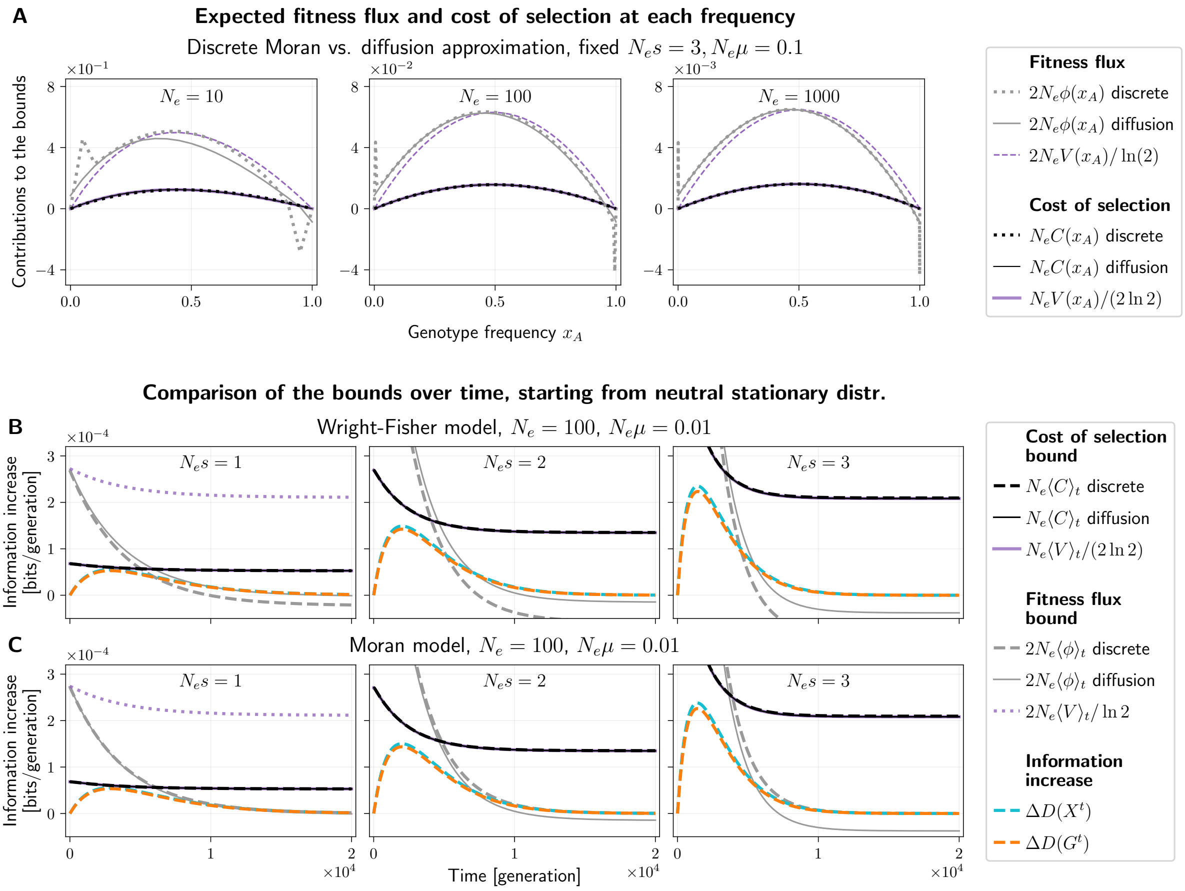

Figure B.2: Comparisons between the fitness flux, the information-theoretic cost of selection and fitness variance, using the single locus, two allele systems under the Wright-Fisher model, the equivalent discrete-time Moran model and the diffusion approximation.

(A) We compute the expected fitness flux per generation ($\phi(x_A)$, gray dotted) and the cost of selection ($C(x_A)$, black dotted) for each frequency x_A under the discrete Moran-model for three different effective population sizes N_e (left to right; census population size $N_{\text{Moran}} = 2N_e$ to account for the additional stochasticity in the Moran model compared to the Wright-Fisher model). We fixed $N_e s = 3$ and $N_e\mu = 0.1$ to obtain models with a similar behavior but different time scales and granularity. Both $\phi(x_A)$ and $C(x_A)$ are multiplied by N_e and the same numerical factor as in the bounds on information accumulation rate (the bounds are obtained by averaging these values with respect to $\psi^X(x_A)$). The discrete formulas are compared with their diffusion approximation (full gray and black lines). As expected, the diffusion approximation is closer to the discrete model at higher N_e. The fitness flux $\phi(x_A)$ converges to the diffusion approximation non-uniformly across x_A, due to the spikes next to the domain boundaries. These spikes do not disappear but get "squeezed out" at high N_e. We also plot multiples of the relative fitness variance $V(x_A)$ (purple dashed and full lines), which approximate the fitness flux and the cost of selection. The cost $C(x_A)$ can be approximated very closely by $V(x_A)/(2\ln 2)$ as long as $s \ll 1$ (here, the largest value is $s = 0.3$ for $N_e = 10$). Fitness flux is the sum of a selection term proportional to $V(x_A)$ which is largest at intermediate x_A, and a mutation term which dominates near $x_A = 0$ or 1 and causes the discrepancy between $\phi(x_A)$ and $2V(x_A)/\ln 2$. Even when mutation rate is small compared to selection, $\mu \ll s$, mutation is important as the system approaches the stationary distribution concentrated near $x_A = 0$ and 1. (Caption continues on the next page.)

Figure B.2: (Continued from previous page:)
(B) The Wright-Fisher model uses the same parameters as Main Text Fig. 3.4, namely $N = N_e = 100; N_e\mu = 0.01$ and $N_e s$, varied across columns. We plot the increase of information per generation (population level $\Delta D(X^t)$, blue dashed; genotype level $\Delta D(G^t)$, orange dashed), the upper bound in terms of cost of selection $N_e\langle C\rangle_t$, computed using the discrete formula (Eq. (B.107), black dashed) and the diffusion formula (Eq. (B.115), black full), and the upper bound in terms of the fitness flux computed using the discrete formula (Eq. (B.105), gray dashed) and the diffusion formula (Eq. (B.114), gray full). We also show the fitness variance approximations of the two bounds (Eq. (B.116), purple solid and the fitness variance term in Eq. (B.117), purple dotted, outside the plot range for $N_e s = 2$ and 3). The discrete fitness flux bound is violated, since the Wright-Fisher model does not satisfy detailed balance under neutrality. The continuous fitness flux bound holds under weak selection, but fails when selection gets stronger, as differences grow between the discrete system and the diffusion approximation.
(C) The Moran model has the same effective parameters but double the census population size, $N_{\text{Moran}} = 2N_e$. The curve descriptions are the same as for the Wright-Fisher model, but note that each generation consists of $2N_e$ replacements, and the information increase as well as the upper bounds are rescaled accordingly. Also note that we plot the cost of selection bound $N_e\langle C\rangle_t$ using the effective population size N_e rather than the census population size $N_{\text{Moran}} = 2N_e$, which would lead to a twice as large (and less tight) bound. The Moran morel satisfies detailed balance under neutrality, and the discrete fitness flux bound holds under arbitrary selection. The continuous fitness flux bound again fails under strong selection.
Note that the diffusion formula for the fitness flux bound would correctly upper bound the accumulation of information in a system modeled fully using the diffusion approximation. These figures show that it does not always upper bound the accumulation of information in discrete models, especially not when selection is strong. The Main Text Fig. 3.4 uses the information accumulation curves and cost of selection bounds based on the Wright-Fisher model, and the fitness flux bound based on the Moran model.

where we also divided by Δt to get the fitness flux per generation. Under the diffusion approximation, if Δt is small, the transition probabilities P, Q will be approximately normal with parameters given by $a(x^t)$, $a^s(x^t)$ and $b(x^t)$ (see Sec. B.7; we drop the dependence on x^t for brevity),

$$P(x^{t+\Delta t}|x^t) \approx \mathcal{N}(x^{t+\Delta t};\, x^t + a + a^s,\, b), \tag{B.110}$$

$$Q(x^{t+\Delta t}|x^t) \approx \mathcal{N}(x^{t+\Delta t};\, x^t + a,\, b). \tag{B.111}$$

Then in Eq. (B.109) we recover the definition of fitness flux from Mustonen and Lässig (2010),

$$\phi(x^t, x^{t+\Delta t}) \approx \frac{1}{N\Delta t \ln 2}\,(x^{t+\Delta t} - x^t)_g\, b^{-1}_{gg'}\, a^s_{g'} \tag{B.112}$$

where we sum over repeated indices as in Sec. B.7. Note that the factor $1/\ln 2$ appears because we use base 2 logarithms throughout. If the process takes place in a fitness landscape/seascape F, the vector $b^{-1}_{gg'}\, a^s_{g'}/N = \partial_g F$ is its gradient, and $\phi(x^t, x^{t+\Delta t})$ is the rate at which the system climbs it up and therefore accumulates fitness (Fig. 1 and Eq. S8-S10 in Mustonen and Lässig (2010)). This interpretation of fitness flux is exact in the diffusion approximation as $\Delta t \to 0$, but only approximate for the discrete formulas in Eq. (B.108) and Eq. (B.105).

We can take the expectation over $x^{t+\Delta t}$ to obtain the expected fitness flux per generation from any starting position x^t,

$$\phi(x^t) \approx \frac{1}{N \ln 2}(a + a^s)_g b^{-1}_{gg'}\, a^s_{g'}, \tag{B.113}$$

which is also plotted in Fig. B.2A. Finally, we can take the expectation over x^t to obtain the fitness flux bound in the diffusion approximation. It compares with the cost of selection bound as follows,

$$\begin{aligned}\text{Fitness flux bound:}\qquad & \frac{d}{dt}D(X^t) \leq \frac{2}{\ln 2}\int \psi \left(a_g + a^s_g\right) b^{-1}_{gg'}\, a^s_{g'}\, dx = 2N\langle\phi\rangle_t, && \text{(B.114)}\\ \text{Cost of selection bound:}\qquad & \frac{d}{dt}D(X^t) \leq \frac{1}{2\ln 2}\int \psi\, a^s_g\, b^{-1}_{gg'}\, a^s_{g'}\, dx \qquad\quad \leq N\langle C\rangle_t. && \text{(B.115)}\end{aligned}$$

Note that in Eq. (B.114) we added the factor 2 which was missing in Mustonen and Lässig (2010), as pointed out in Barton (2017).

Again, the fitness flux bound requires the neutral process to be at a stationary distribution with detailed balance, which means zero neutral flux $u_g = 0$ (see Eq. (B.57)). (There is a more general flux theorem, which does not require detailed balance Mustonen and Lässig (2010), but it does not provide a bound on $\Delta D(X^t)$.) The single locus, two allele system satisfies the detailed balance, since diffusion only takes place along a single dimension. However, detailed balance is rare in systems with multiple loci with recombination and general forms of mutation.

B.10.3 Comparisons of the discrete and the diffusion formulas

The two bounds, computed using both the Markov chain and the diffusion formulas, are compared in Fig. B.2BC for the single locus, two allele system. Note that the bounds are obtained by averaging the functions $2N_e\phi(x_A)$ and $N_eC(x_A)$, such as those plotted in Fig. B.2A, with respect to the distribution $\psi^X(x_A)$.

In Fig. B.2B, we use the Wright-Fisher model to compute the distribution over allele frequencies and the information increments. The population size is $N_{\mathrm{WF}} = N_e = 100$, mutation strength $N\mu = 0.01$ and selection strength varies across columns. The cost of selection bound is in black, and the diffusion formula (full line, based on Eq. (B.115)) is in an agreement with the discrete formula (dashed line, based on Eq. (B.107)). The discrete version of the fitness flux bound is violated, as the Wright-Fisher model does not satisfy detailed balance (grey dashed, based on Eq. (B.105)). This is because cycles such as $0 \to 1 \to 2 \to 0$ copies of the A allele take place more often than the reverse cycle, since two alleles can get lost by drift in a single generation with a high probability, but are unlikely to arise by mutation. However, detailed balance holds in the diffusion approximation. If fitness flux is computed according to the diffusion formula, the bound holds under weak selection when the diffusion approximation is close to the Wright-Fisher model, but fails when selection is stronger (full grey line, based on Eq. (B.114)).

In Fig. B.2C, we use the Moran model of the same system to compute the distributions and information increment over time. Note that in order to have the same magnitude of genetic drift as the Wright-Fisher model, the Moran model needs twice as large a census population size, $N_{\mathrm{Moran}} = 2N_e$. Time is measured in generations ($N_{\mathrm{Moran}} = 2N_e$ replacements each), and the increments in information are also computed per generation.

The discrete and diffusion formulas are again in agreement for the cost of selection bound. Note that we computed it as $N_e\langle C\rangle_t$ rather than $N_{\mathrm{Moran}}\langle C\rangle_t$, to account for the additional stochasticity due to random deaths, even though $N_{\mathrm{Moran}} = 2N_e$ parents are sampled with selection in each generation. The fitness flux bound now holds in its discrete version – the Moran model only allows allele frequency changes by $\pm 1/N_{\mathrm{Moran}}$ and therefore satisfies detailed balance. If fitness flux is computed using the diffusion formula in Eq. (B.114), it upper bounds the Moran model accumulation of information when selection is weak, but again fails when selection is strong and diffusion departs from the discrete model.

In conclusion, care is needed when applying the fitness flux bound to discrete models. In Main Text Fig. 3.4, we plot the information accumulation and the cost of selection bound based on the Wright-Fisher model, and the fitness flux bound based on the Moran model.

B.10.4 Interpretation of the bounds under diffusion

The cost of control is non-negative and determined solely by the magnitude of the selection term a^s_g, with the inverse drift covariance $b^{-1}_{gg'}$ acting as the metric. As shown in Sec. B.7 and Fig. B.2A, this is proportional to the variance in fitness,

$$\frac{d}{dt}D(X^t) \leq N\langle C\rangle_t \approx \frac{N}{2\ln 2}\langle V\rangle_t. \tag{B.116}$$

In Fig. B.2BC, this bound starts off large and reduces slightly as selection removes variation from the population. It remains positive at the stationary state, where it represents the cost of maintenance.

In contrast, the fitness flux bound in Eq. (B.114) contains the sum $a^s_g + a_g$, corresponding to selection and mutation contributions to the fitness flux. As a result, the fitness flux bound can be written as a sum of a selection term and a mutation term,

$$\frac{d}{dt}D(X^t) \leq 2N\langle \phi\rangle_t = \underbrace{\frac{2N}{\ln 2}\langle V\rangle_t}_{\text{Selection term}} + \underbrace{\frac{2}{\ln 2}\int \psi\, a_g\, b^{-1}_{gg'}\, a^s_{g'}\, dx.}_{\text{Mutation term}} \tag{B.117}$$

The selection term is proportional to the fitness variance – like the cost of selection bound, but with a 4 times higher numerical coefficient. When the selection term in Eq. (B.117) dominates (a regime also discussed in Mustonen and Lässig (2010)), the two bounds are proportional to each other, but the cost of selection bound is tighter. In general, the mutation term in Eq. (B.117) causes the two bounds to behave in qualitatively different ways. Notably, the mutation term can be substantial and comparable to the selection term even when the parameters N, s, μ suggest that selection is strong and mutation is weak (i.e. under any or all of the conditions $Ns \gg 1$, $N\mu \ll 1$ and $s \gg \mu$). We illustrate and explain this in the following paragraphs.

Fig. B.2A shows that the selection term in Eq. (B.117) dominates especially at intermediate allele frequencies (around $x_A = 0.5$), where the fitness variance is high and mutation in opposing directions cancels out. Near $x_A = 0$, fitness variance is low and mutation towards the fitter allele A leads to a positive expected fitness flux. Near $x_A = 1$, mutation towards the deleterious allele a dominates and leads to a negative expected fitness flux (Fig. B.2A).

This enables a more detailed understanding of the fitness flux bound in the scenario in Main Text Fig. 3.4 and Fig. B.2C. The system is initialized at the neutral stationary distribution, which is symmetric. Therefore the mutation term in fitness flux vanishes, because the positive and negative contributions (at $x_A < 0.5$ and $x_A > 0.5$ respectively) cancel out. The bound is therefore proportional to the average fitness variance. Over time, as selection shifts the distribution towards higher x_A, mutation contributes more and more negatively, until the mutation and selection terms exactly cancel at stationarity. This happens regardless of N, μ and s. At stationarity, populations fluctuate around frequencies where mutation and selection are balanced, typically close to $x_A = 0$ or $x_A = 1$.

The only regime when the selection term dominates fitness flux for an extended period of time is when not only mutation is weak overall, but also the population is initialized at an intermediate frequency. This is shown in Fig. B.3, which uses the same parameters as Fig. B.2C, but the population is initialized at the frequency $x_A = 0.5$.

Note that only the process with selection can be initialized at $x_A = 0.5$. The neutral process must always be at the neutral stationary distribution to satisfy the assumptions of the fitness flux theorem. As a result, the information $D(X^t)$ is very high initially and decreases until drift spreads out the distribution ψ^{X^t} towards $x_A = 0$ and 1. Later, $D(X^t)$ slightly increases over a much longer (mutation-limited) time scale, see Fig. B.3A. The early phase is useful for illustrating the behavior of the fitness flux and the cost of selection, but neither bound is very informative there, as information is being lost.

In Fig. B.3B we plot the increments in information, the fitness flux bound and the cost of selection bound per generation. Both bounds are proportional to the relative fitness variance in the early phase at intermediate frequencies, albeit with different proportionality constants (full and dotted purple lines in Fig. B.3B). In the later phase, when the population is mostly fixed for one of the alleles, the mutation term in Eq. (B.117) becomes important and negative, and the fitness flux bound departs from the fitness variance approximation.

B.11 Frequency dependent selection that maximizes fixation probability

In Main Text Fig. 3.3CD, we compare the efficiency of selection (accumulated information per unit cost of selection) under constant selection, and under a specific form of frequency dependent selection. This frequency dependence is optimal in the sense that it maximizes the fixation probability of A at a given cumulative cost of selection. Below we describe the optimization procedure.

The calculation is done using a single locus, two allele Wright-Fisher model as described in Sec. B.4, but instead of a single selection coefficient, we have a vector

$$\boldsymbol{s} = \left(s\left(0\right), s\left(\frac{1}{N}\right), s\left(\frac{2}{N}\right), \ldots, s\left(1\right)\right)^{\intercal} \tag{B.118}$$

of selection coefficients $s(x_A)$ for each possible allele frequency x_A. Given $\boldsymbol{s}$, we can compute

- The (right stochastic) transition matrix $P(\boldsymbol{s})$ according to Eq. (B.10), using the respective selection coefficient for each starting frequency (rows of $P(\boldsymbol{s})$).
- The vector $\boldsymbol{\psi}^{\text{fix}}(\boldsymbol{s})$ of fixation probabilities for each possible starting frequency. It is given by the last column of the matrix power $P(\boldsymbol{s})^t$ at infinite time $t \to \infty$. Numerically, we keep doubling t, until the total probability that neither allele is fixed is less than 10^{-6} for all starting frequencies.
- The vector $\boldsymbol{c}(\boldsymbol{s})$ of cost of selection at each frequency,

$$\boldsymbol{c}(\boldsymbol{s}) = \left(C\left(0\right), C\left(\frac{1}{N}\right), C\left(\frac{2}{N}\right), \ldots, C\left(1\right)\right)^{\intercal}. \tag{B.119}$$

  Note that $C(0) = C(1) = 0$ since there is no fitness variation when one of the alleles is fixed.

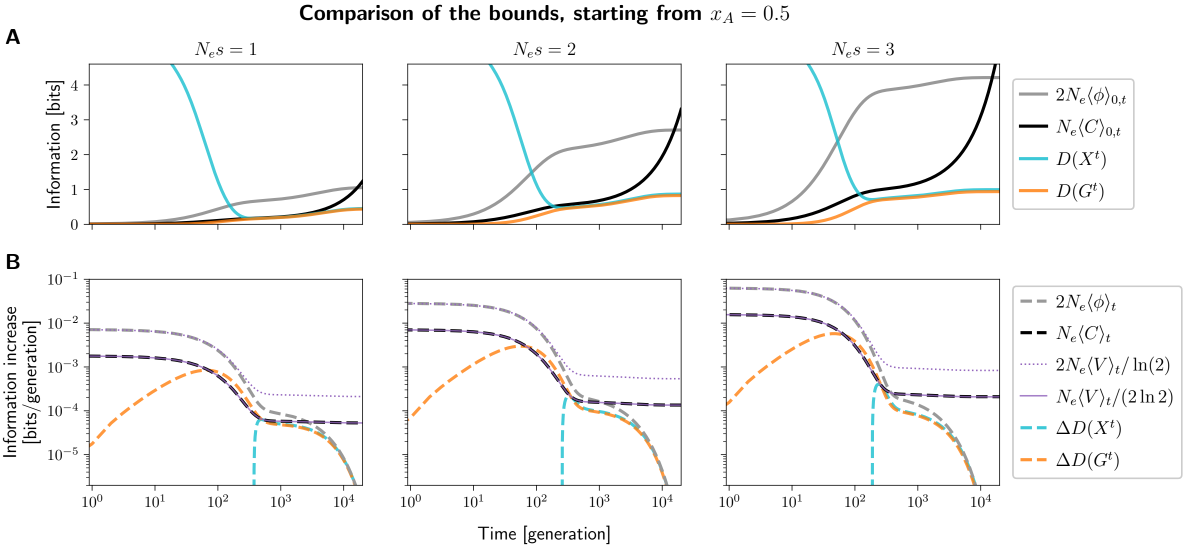

Figure B.3: Demonstration of the fitness flux and the cost of selection bounds in the single locus, two allele system initialized at an intermediate frequency $x_A = 0.5$. Note that only the process with selection is initialized at $x_A = 0.5$; the neutral process must always be at the neutral stationary distribution to satisfy the assumptions behind the fitness flux theorem. This leads to a high initial value of $D(X^t)$ at $t = 0$. Calculated using the discrete Moran model with the same parameters as in Fig. B.2C.

(A) The cumulative information at the population level ($D(X^t)$, blue) and genotype level ($D(G^t)$, orange), as well the cumulative fitness flux ($2N_e\langle\phi\rangle_{0,t}$, gray) and cumulative cost of selection ($N_e\langle C\rangle_{0,t}$, black). $D(X^t)$ starts out high due to the initial distribution being different under selection than under neutrality ($D(X^0) \approx 12.3$ bits, outside the plot range, in all three cases). At first, drift causes the distribution under selection to spread out towards extreme frequencies, $D(X^t)$ decreases on the time scale of $\sim N_e = 100$ generations. Meanwhile, $D(G^t)$ accumulates as the mean frequency x_A increases from 0.5 due to selection. The two measures of information eventually reach similar values, since mutation is low and the population is mostly fixed for one of the alleles. The cumulative fitness flux and cost of selection upper bound the cumulative information increase $D(X^t) - D(X^0)$, but are not very informative in this case, because despite selection acting, $D(X^t)$ started at a value that is higher than can be maintained and is lost rather than accumulated.

(B) The increase of information per generation on the population and genotype levels ($\Delta D(X^t)$, blue dashed; $\Delta D(G^t)$, orange dashed), and the upper bounds in terms of fitness flux ($2N_e\langle\phi\rangle_t$, gray dashed) and the cost of selection ($N_e\langle C\rangle_t$, black dashed). Note the log scales on both axes. Initially, $\Delta D(X^t)$ is negative and falls outside of the plot as drift spreads out the distribution over allele frequencies. Meanwhile, $\Delta D(G^t)$ is positive as the mean frequency x_A increases due to selection – this is associated with a positive fitness flux and cost of selection. After about $2N_e = 200$ generations, one of the alleles is likely to be fixed by drift, but the mean frequency continues to slowly increase at a mutation-limited rate (time scale $1/\mu = 10^4$ generations), which also causes modestly positive $\Delta D(X^t)$ and $\Delta D(G^t)$. In this phase, the cost of selection remains nearly constant, while fitness flux slowly decays, providing a fairly tight bound on $\Delta D(X^t)$. The cost of selection bound can be very well approximated with the relative fitness variance ($N_e\langle V\rangle_t/(2\ln 2)$, purple solid line). The fitness flux is proportional to the fitness variance in the first phase, when x_A is near 0.5 ($2N_e\langle V\rangle_t/\ln 2$, purple dotted line), but departs from it later as x_A tends to take values near 0 or 1 where mutation is important (see Fig. B.2A). We note that similar behavior can also observed for different values of $N_e\mu, N_e s$.

- The vector $\boldsymbol{\gamma}(\boldsymbol{s})$ of expected total cost of selection until either allele is fixed, for each starting frequency. The first and last elements are again equal to zero, since one of the alleles is fixed and no more cost is incurred. The remaining elements can be computed using the recurrence relation

$$\boldsymbol{\gamma}(\boldsymbol{s}) = \boldsymbol{c}(\boldsymbol{s}) + P(\boldsymbol{s})\boldsymbol{\gamma}(\boldsymbol{s}), \tag{B.120}$$

  where $\boldsymbol{c}(\boldsymbol{s})$ is the immediate cost and $P(\boldsymbol{s})\boldsymbol{\gamma}(\boldsymbol{s})$ is the expected future cost. Eq. (B.120) can be solved for $\boldsymbol{\gamma}(\boldsymbol{s})$ as a system of linear equations.

- The value vector $\boldsymbol{v}(\boldsymbol{s}) = \boldsymbol{\psi}^{\text{fix}}(\boldsymbol{s}) - \lambda\boldsymbol{\gamma}(\boldsymbol{s})$, where λ is the Lagrange multiplier which quantifies the constraint on the total cost.

We look for $\boldsymbol{s}$ which maximizes value $\boldsymbol{v}(\boldsymbol{s})$. This is an instance of a Markov decision process Bellman (1957) similar to the pursuit/first passage problem Eaton and Zadeh (1962) with a small modification to include an unwanted absorbing state (loss of the A allele). The frequency-dependent selection $\boldsymbol{s}$ corresponds to the decision policy. The optimal policy does not depend on time or the initial state, and maximizes all elements of $\boldsymbol{v}(\boldsymbol{s})$ simultaneously Eaton and Zadeh (1962).

We optimize $\boldsymbol{s}$ iteratively. It is initialized at all zeros, and we alternate between a value update and a policy (selection) update,

$$\text{Value update:} \qquad \boldsymbol{v} := \boldsymbol{v}(\boldsymbol{s}) = \boldsymbol{\psi}^{\text{fix}}(\boldsymbol{s}) - \lambda\boldsymbol{\gamma}(\boldsymbol{s}) \tag{B.121}$$

$$\text{Policy update:} \qquad \boldsymbol{s} := \operatorname{argmax}_{\boldsymbol{s}} \left(P(\boldsymbol{s})\boldsymbol{v} - \lambda\boldsymbol{c}(\boldsymbol{s})\right) \tag{B.122}$$

The value update uses the current estimate of $\boldsymbol{s}$. The policy update uses the current estimate of $\boldsymbol{v}$ to compute the selection coefficient at each frequency, which maximizes the expected value in the next step, minus the immediate cost. The maximization is independent for each element of $\boldsymbol{s}$ and is done by binary searching for zero gradient ($\log_{10} s$ in range from 10^{-5} to 10^{2}, binary search depth 10).

To produce Fig. 3.3CD, we vary the cost constraint λ and compute $\boldsymbol{s}$ by 60 value-policy update iterations. Examples of the frequency dependent $\boldsymbol{s}$ are shown in Fig. B.4A. Notably, selection is strongest at low frequencies of x_A when A is at the greatest risk of being lost, but weak at higher frequencies to reduce costs. Fig. B.4B,C show the fixation probability $\boldsymbol{\psi}^{\text{fix}}(\boldsymbol{s})$ and the expected total cost $\boldsymbol{\gamma}(\boldsymbol{s})$ for each starting frequency, from which the Main Text Fig. 3.3CD uses only the values for the initial frequency $1/N$.

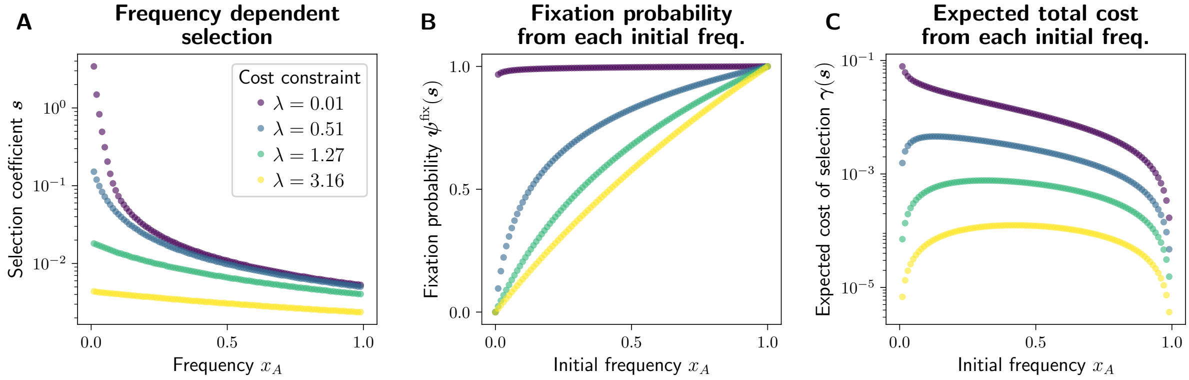

Figure B.4: Frequency dependent selection that optimizes fixation probability of an allele at a constrained total cost.

(A) The frequency dependent selection coefficient $\boldsymbol{s}$, computed as described in Sec. B.11, under various cost constraints λ. When this constraint is greater, selection is weaker overall.

(B) Fixation probability $\psi^{\text{fix}}(\boldsymbol{s})$ of the allele A for each starting frequency x_A. The fixation probability is close to the frequency itself under weak selection (large λ), and higher when selection is overall stronger.

(C) The expected total cost of selection, $\gamma(\boldsymbol{s})$ associated with trajectories starting from each initial frequency x_A. It is low for high frequencies, where the allele A is expected to be fixed soon, and for low frequencies under weak selection, where it is expected to be lost soon.