Statistical analysis and optimality of biological systems

Wiktor Młynarski*, Michal Hledík*, Thomas R. Sokolowski, Gašper Tkačik[1]

[1] Institute of Science and Technology Austria, Am Campus 1, A-3400 Klosterneuburg, Austria
*Equal contribution

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, utility, or fitness. 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 (characteristic of "bottom-up" statistical models), and a data-limited *ab inito* prediction regime (characteristic of "top-down" normative theory). We demonstrate the applicability of our framework using data from the visual cortex, and argue that the flexibility it affords is essential to address a number of fundamental challenges relating to inference and prediction in complex, high-dimensional biological problems.

Optimization | Statistical inference | Neural coding | Evolution

Correspondence: ***wiktor.mlynarski@ist.ac.at***

Ideas about optimization are at the core of how we approach biological complexity (1–3). Quantitative predictions about biological systems have been successfully derived from first principles in the context of efficient coding (4, 5), metabolic (6, 7), reaction (8, 9), and transport (10) networks, evolution (11), reinforcement learning (12), and decision making (13, 14), 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 (15, 16), or to define rigorously what that even means (17–19). Alternatively, data may be "close to" but not "at" optimality, and different instances of the system may show variation "around" optima (20, 21), 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.

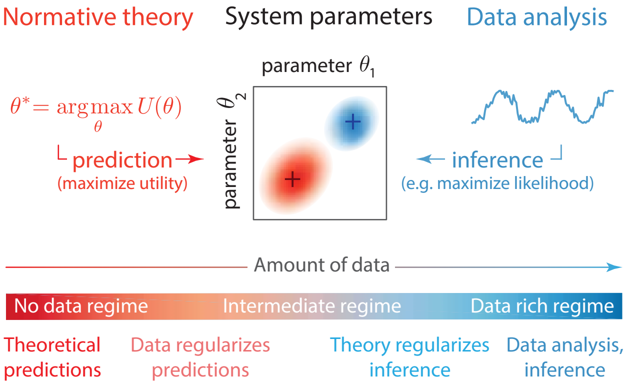

Fig. 1. Normative theories and statistical inference. Both approaches make statements about values of system parameters (middle row; center panel). Normative theories predict which parameters would be of highest utility to the system (middle row in red; left panel) without reference to experimental data. 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).

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 (22). 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. 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? We illustrate the application of these questions and the related concepts to simple model systems, and demonstrate their relevance to real-world data analysis by analyzing receptive fields of neurons in the visual cortex.

Results

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) posterior over parameters given the data:

$$\log P(\theta|\mathcal{D}) = \log \mathcal{L}(\theta) + \log P(\theta) + \text{const}, \qquad \textbf{(1)}$$

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 (23), priors that regularize the inference and thus help models generalize to unseen data (24, 25), or priors that can coarse-grain the model depending on the amount of data samples, T (26).

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, $U(\theta)$; optimality then amounts to the assumption that the real system operates at a point in parameter space, θ^*, that maximizes utility, $\theta^* = \text{argmax}_\theta\, U(\theta)$. Viewed in the Bayesian framework, the assertion that the system is optimal thus represents an infinitely strong prior that the parameters are concentrated at θ^*, i.e., $P(\theta) = \delta(\theta - \theta^*)$. In this extreme case, no data is needed: the prior fixes the values of parameters and typically no finite amount of data will suffice for the likelihood in Eq (1) to move the posterior away from θ^*. This concentrated prior can be however 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 (27), we therefore consider for our priors distributions that are as random and unstructured as possible while attaining a prescribed average utility:

$$P(\theta|\beta) = \frac{1}{Z(\beta)} \exp\left[\beta U(\theta)\right]. \qquad \textbf{(2)}$$

This is in fact a family of priors, 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 θ^* that maximizes the utility to $U_{\max}$ (if such a point is unique) irrespective of whether data supports this or not. At finite β, however, the prior is "smeared" around θ^* so that the average utility, $\bar{U}(\beta) = \int d\theta\, P(\theta|\beta)U(\theta) < U_{\max}$ increases monotonically with β. For this reason, we refer to β as the "optimization parameter," and to the family of priors in Eq (2) as "optimization priors."

The intermediate regime, $0 < \beta < \infty$, in the prior entering Eq (1) is interesting from an inference standpoint. It represents the belief that the system may be "close to" optimal with respect to the utility $U(\theta)$ 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 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. 1.

In the following, we apply this framework to a toy model system, a single linear-nonlinear neuron, which is closely related to a linear classifier. 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 real data from the visual cortex.

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.

Example: Efficient coding in a simple model neuron. Let us consider a simple probabilistic model of a spiking neuron (Fig. 2A), a broadly applied paradigm in sensory neuroscience (28–32). 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. 2B, left panel), by first summarizing the data with the likelihood, $\mathcal{L}(\theta)$ (Fig. 2B, right panel), followed either by maximization of the likelihood, $\hat{\theta} = \text{argmax}_\theta \mathcal{L}(\theta)$ in the maximum-likelihood (ML) paradigm, or by deriving $\hat{\theta}$ from the posterior, Eq (1), in the Bayesian paradigm.

To apply our reasoning, we must propose a normative theory for neural function, form the optimization prior, and combine it with the likelihood in Fig. 2B, as prescribed by the Bayes rule in Eq (1). An influential theory in neu-

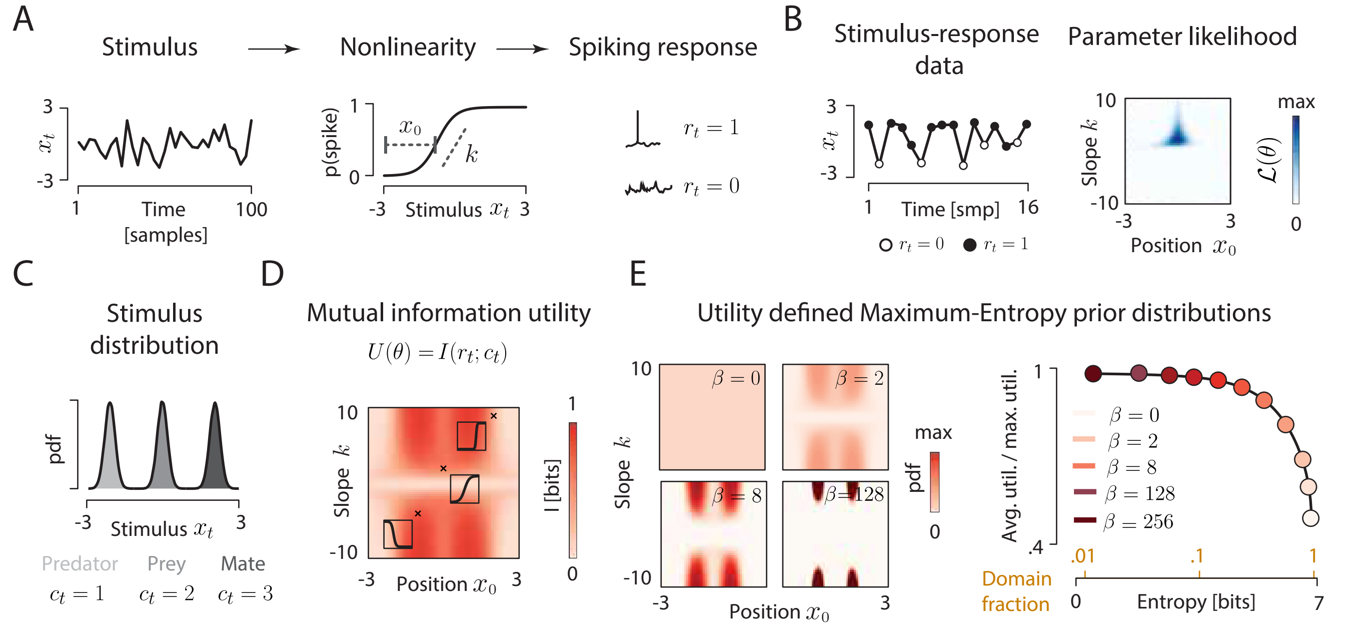

Fig. 2. Efficient coding in a toy model neuron and the corresponding optimization prior. (A) Model neuron uses a logistic nonlinearity (middle panel) to map continuous stimuli x_t (left panel) to a discrete spiking response r_t (right panel). The shape of the nonlinearity is described by two parameters: slope k and offset x_0. **(B)** An example dataset (left panel) consisting of stimulus values (gray 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). Orange numbers on the horizontal axis specify the fraction of the entire domain effectively occupied by parameters at given β.

roscience called "efficient coding" postulates that sensory neurons maximize the amount of information about natural stimuli they encode into spikes given biophysical constraints (5, 31, 33–36). This information-theoretic optimization principle (37) has correctly predicted neural parameters such as receptive field shapes (34, 38) and the distribution of tuning curves (17, 39), as well as other quantitative properties of sensory systems (4, 40–44), *ab initio*, from the distribution of ecologically relevant stimuli (2, 34). As such, efficient coding is also a suitable normative theory for our model neuron.

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. 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, $U(k, x_0) = I(r_t; c_t)$, relevant to our case. Figure 2D shows that U 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. 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 (18, 45, 46).

Given the utility function, the construction of the maximum-entropy optimization prior according to Eq (2) is straightforward. Explicit examples for different values of β are shown in Fig. 2E (left panel); more 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. 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. This completes our setup and allows us to address the four questions posed in the Introduction.

Question 1: Statistical test for the optimality hypothesis. Given a candidate normative theory and experimental

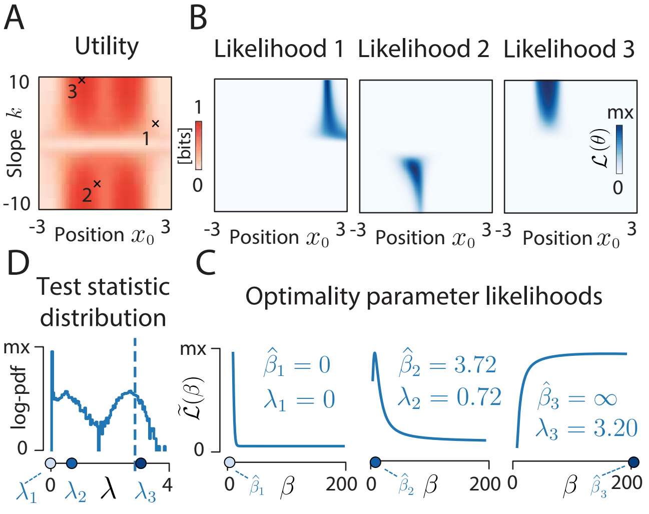

Fig. 3. Statistical test of optimality. **(A)** The utility function $U(k, x_0)$. The crosses and numbers show the locations of ground truth parameters. **(B)** Likelihood of the nonlinearity parameters obtained from 20 stimulus–response (x_i, r_i) pairs. The three examples correspond to three ground truth parameter values (black crosses in panel 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 the data in A). Maximum likelihood estimates of β, $\hat{\beta}_{1,2,3}$ (denoted by blue circles), indicate that the data would be most probable with no preference for high utility U (left panel, $\hat{\beta}_1 = 0$ – note that we do not allow negative $\hat{\beta}$), some preference for high U (middle panel, $\hat{\beta}_2 > 0$ finite) and strong preference for high U (right panel, $\hat{\beta}_3 \to \infty$; blue circle displayed at $\beta = 200$ for illustration purposes). The 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)** The null distribution of the test statistic λ. The point mass at $\lambda = 0$ corresponds to the cases when 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. The dashed vertical line shows the significance threshold at $\alpha = 0.05$, and blue circles are the values $\lambda_{1,2,3}$. Only λ_3 crosses the threshold, indicating that the corresponding data would be surprising without preference for high utility parameters.

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. 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. 3B): how should the test for optimality be formulated in the presence of such uncertainty?

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. 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).

The test statistic λ has a null distribution that can be estimated by sampling (Fig. 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 slightly different utility function or resource constraints than those postulated by our theory (21). Instead, the proposed test asks if the system has relatively high utility, compared to the utility distribution in the full parameter space.

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).

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 what value of the optimization parameter, β, of the prior 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. 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. 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.

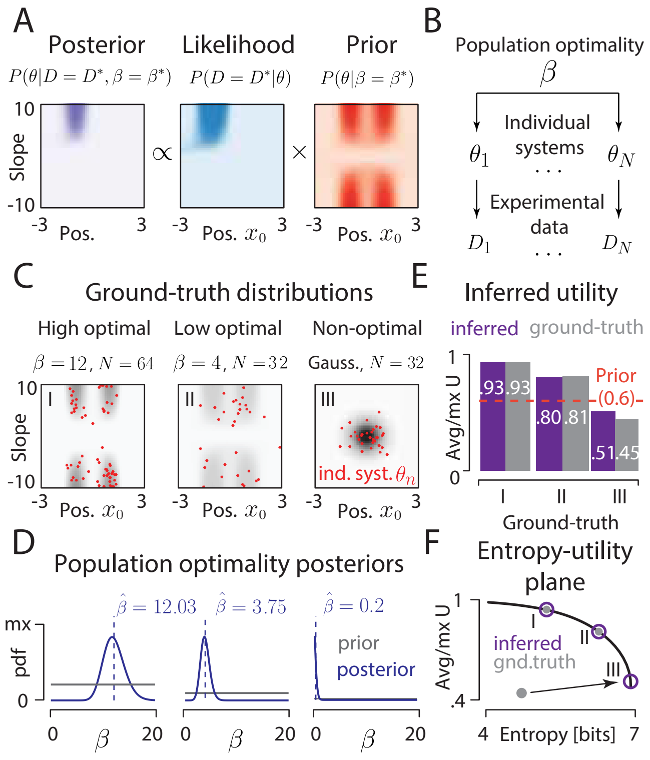

Fig. 4. Inferring the degree of population optimality. **(A)** Posterior over nonlinearity parameters, inferred for a single system with a utility-derived prior at fixed optimality parameter, $\beta=\beta^*$. **(B)** A hierarchical model of a population of optimized systems. Population optimality parameter 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 (red dots) sampled from different ground truth distributions. 64 samples from a strongly optimized population ($\beta=12$; left panel), 32 samples from a weakly optimized population ($\beta=4$; middle panel), 32 samples from a non-optimal distribution (Gaussian distribution; right panel). For each model neuron θ_n, data $\mathcal{D}_n$ consists of 100 stimulus-response pairs. **(D)** Results of hierarchical inference. Posteriors over population optimality β (purple lines) were obtained using simulated data from C). Posterior averages, $\hat{\beta}$, shown as dashed purple lines. Priors (gray lines) were uniform on the $[0,20]$ interval. **(E)** Average utility, $I(r_t;c_t)$, reported as a fraction of the maximum value. Estimated values (purple bars) closely match ground truth (gray bars). Roman numerals correspond to scenarios from panel C. **(F)** Entropy and relative average utility of ground truth distributions (gray, filled circles) and inferred distributions parametrized by $\hat{\beta}$ (purple, empty circles). Roman numerals correspond to scenarios from panel C.

To explore this possibility, we generated parameters θ_n of model neurons from three different distributions: strongly optimized ($\beta=12$; Fig. 4C, left panel), weakly optimized ($\beta=4$; Fig. 4C, middle panel) and non-optimal (Gaussian distribution of parameters; Fig. 4C, right panel). From each neuron we obtained an experimental dataset $\mathcal{D}_n$ of 100 stimulus-response pairs. Using standard hierarchical Bayesian inference we then computed the posterior distributions over the population optimality parameter, β (purple lines in Fig. 4D; see Methods). In each of the three cases, posterior averages, $\hat{\beta}$ (Fig. 4D; dashed purple lines) closely approximated ground truth values.

Following the hierarchical inference, we can interpret the inferred population parameter $\hat{\beta}$. We note that parameter estimate $\hat{\beta}$ can be mapped onto normalized average utility (as illustrated in Fig. 2E), which enables us to report the optimality on a $[0,1]$ scale. Normalized values of utility for three different ground truth are displayed in Fig. 4E (purple bars), side-by-side with relative utilities of corresponding ground-truth distributions (gray bars). These normalized utility values can be then compared to the average utility assuming no optimization, $\bar{U}(\beta=0)$.

The maximum-entropy, probabilistic inference framework enables us to draw unique inferences about system's optimality, which are not possible otherwise. For example, in addition to estimating average relative 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. 4F, horizontal axis). In example I from Fig. 4C-F, $\hat{\beta}=12.03$ 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 $\hat{\beta}=0.2$ is consistent with high entropy and indicates almost no such restriction. This is despite the fact that the parameters were sampled from a Gaussian highly concentrated in the parameter space — but not in a region with high utility. The average utility value equal to that from example III could be obtained if the parameters had been sampled uniformly. This implies, that such a system may be optimized for a different utility function or shaped by other processes. The system could also be anti-optimized, i.e. prefer negative values of U. Such cases could be easily identified if we allowed $\hat{\beta}$ to be negative — but we focus on positive β for simplicity. Taken together, the location of a system on the entropy-average utility plane Fig. 4F presents two different insights into optimality of the system's parameters which are not accessible by other means.

Finally, another clear benefit of the probabilistic framework is the possibility of computing uncertainty estimates of β and the associated average utility and entropy.

Question 3: Data resolves ambiguous theoretical predictions. When the predictions of a normative theory are degenerate, with multiple maxima of the utility function, the 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 our example, 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. 5A) and observe how their posteriors look like after seeing as few as $T=12$ stimulus-response pairs from each neuron

(Fig. 5B). 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. 5C). 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. 5D).

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. We recall that 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. 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. 5E). 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 can often be resolved in our Bayesian framework in the "data-regularized prediction" regime by a very small amount of data, which breaks the degeneracy of the theoretical predictions. 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 (9, 31, 42), except in those rare cases where strict guarantees exist (21). 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 (25), 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.

Question 4: Optimization priors improve inference for high-dimensional problems. To answer the last question we extend our toy model neuron with 2 parameters to a more realistic model with 16×16 parameters. The purpose of this exercise is two-fold: technically, we will show that an application of our framework to a realistically high-dimensional problem is feasible; formally, we will show that optimization priors can play a powerful role in regularizing such high-dimensional inference problems.

We simulated the responses of a Linear-Nonlinear-Poisson (LNP) neuron (30) to natural image stimuli (Fig. 6A). Natural image patches x_t (16×16 pixels each) are projected onto a linear filter ϕ, and the output of the filter s_t is transformed with a logistic nonlinearity into average neural firing rate λ_t. The number of spikes elicited by the neuron r_t is then drawn from a Poisson distribution, with mean λ_t. The goal of data analysis is to estimate the linear filter $\phi \in \mathbb{R}^{16\times16}$, which determines the sensory feature encoded by the neuron, from data consisting of stimulus-response pairs, $\mathcal{D} = \{x_t, r_t\}_{t=1}^{T}$. Experimentally observed filters ϕ have been suggested to maximize the sparsity of responses s_t to natural stimuli (34). 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 (38, 47). Filters optimized for sparse utility $U_{\mathrm{s}}(\phi)$ (see Methods) are oriented and localized in space and frequency (Fig. 6B, left panel). To simulate a realistic experiment, we constructed a model neural population of 64 neurons with 40 filters optimized under sparse utility, and 24 generated according to a related, but different criterion (Fig. 6C; see Methods for details). We generated neural responses by exposing each model neuron to a sequence of 2000 natural image patches. Using these simulated data we then inferred the filter estimates, $\hat{\phi}$, using Spike Triggered Average (STA) (18, 48), which under our assumptions is equivalent to the maximum likelihood (ML)

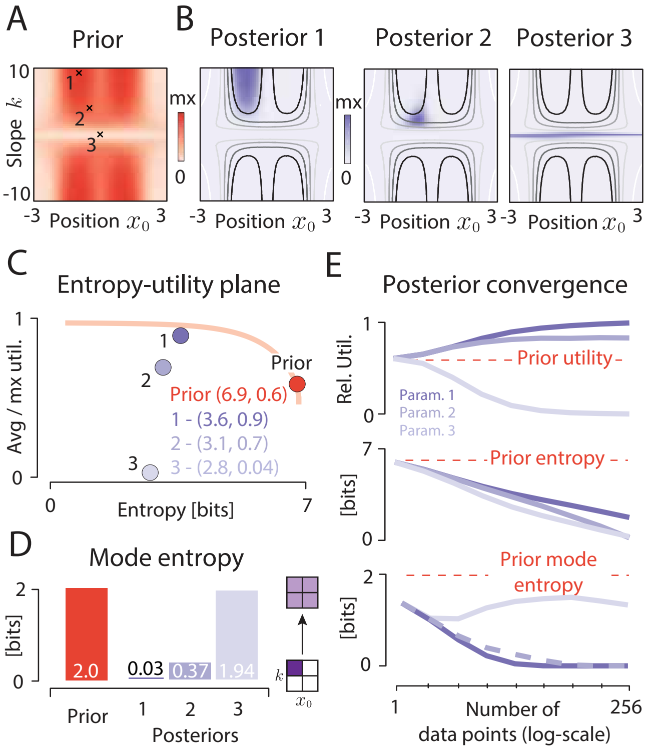

Fig. 5. Resolving ambiguities of theoretical predictions. (A) A maximum-entropy prior derived from the mutual information utility with $\beta = 1$. The prior has multiple maxima reflecting non-uniqueness of theoretical predictions. **(B)** Posteriors obtained by updating the prior with three example 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). **(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.

estimate (30) (see Methods). STAs computed from limited data recover noisy estimates of neural filters (Fig. 6C, right panel).

We then asked whether normative theories can provide powerful priors to aid inference in high-dimensional problems. Using our sparse utility, $U_s(\phi)$, we formulated optimization priors for various values of β and computed maximum-a-posteriori (MAP) filter estimates $\hat{\phi}(\beta)$ from simulated data (Fig. 6D; see Methods for details). Increasing values of β interpolate between pure data-driven ML estimation (Fig. 6D, second column from the left) that ignores the utility, and pure utility maximization (Fig. 6D, right column) at very high $\beta = 10^3$ where the predicted filters become almost completely decoupled from data; these two regimes seem to be separated by a sharp transition.

For intermediate $\beta = 10, 100$, MAP filter estimates show a large improvement in estimation performance relative to the ML estimate (as quantified by Pearson correlation) with the ground truth. Optimization priors achieve this boost in

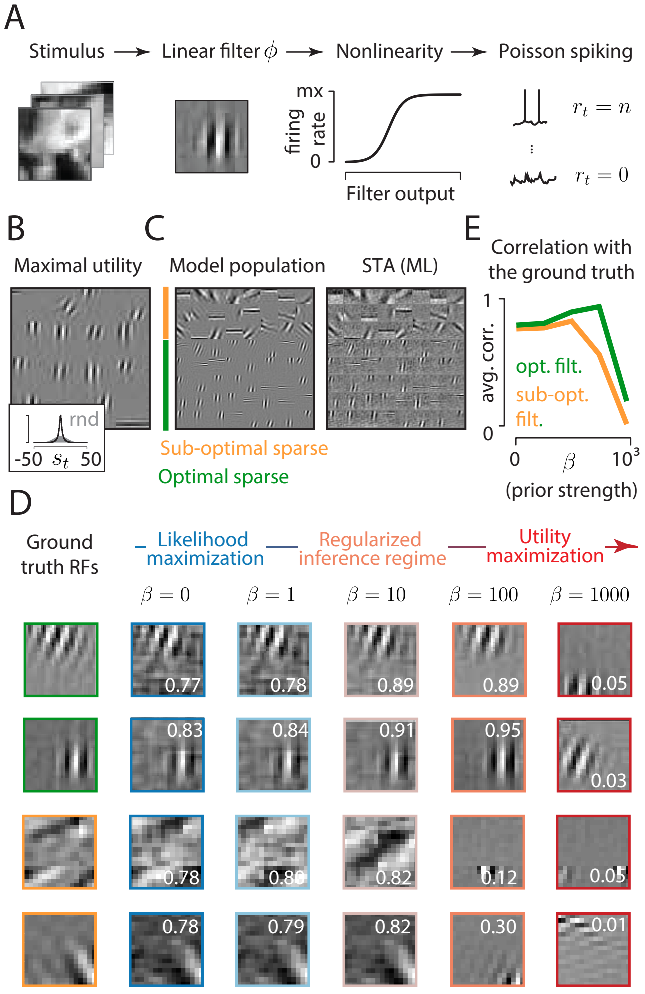

Fig. 6. Optimality analysis and inference of high-dimensional receptive fields. (A) A lineear-nonlinear-Poisson (LNP) neuron responding to natural scenes. High-dimensional stimuli (natural image patches, x_t) are projected onto a linear filter ϕ. Filter output is transformed with a logistic nonlinearity into average firing rate which drives Poisson spiking, r_t. **(B)** Filters optimized for maximally sparse response to natural stimuli. Inset depicts histograms of filter responses to natural images (black line) overlaid on top of histograms of random filter responses (gray-shaded histogram). **(C)** Model neural population consisting of 16 sub-optimally sparse (3 top rows, orange marker) and 40 optimally sparse filters (5 bottom rows, green marker; see Methods for details). Corresponding spike-triggered averages (maximum-likelihood filter estimates) computed from responses to 2000 natural stimuli (right panel). **(D)** MAP estimates of two optimally sparse filters (left column; green frames) and two sub-optimally sparse filters (left column; orange frames) obtained with optimality prior of increasing strength. White digits denote correlation with the corresponding ground truth. **(E)** Average correlations of filter estimates with the ground truth as a function of prior strength β. Green and orange lines correspond to optimally and sub-optimally sparse neuron sub-populations respectively.

performance because they implicitly encode many notions about how neural filters look like (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 specialized priors designed for receptive field estimation can capture some of these characteristics explicitly (18, 49), optimization priors grounded in the relevant normative theory represent the most succinct and complete way of summarizing our prior beliefs about receptive fields. Importantly, using an optimization prior does not imply that the neural data *must* have been generated by an optimal neuron: even if the real neuron is not optimal, the inference will benefit from the implied smoothness, localization, and orientation properties suggested by the prior (Fig. 6D, two bottom rows). This intuition is reflected in our analysis - on average optimality priors increased the quality of optimally-sparse filters by a larger amount and at higher β values (Fig. 6E, green line) than sub-optimal filters (Fig. 6E, orange line). Sub-optimal filter estimation, however also benefited from the presence of the prior. In practice, the value of β which determines how strongly the prior shapes the resulting inference can be set via the standard method of cross-validation to maximize the performance of the inferred model on withheld data.

Taken together, we suggest that whenever a normative theory for a high-dimensional system exists and the task at hand is to infer the system parameters from data – a task which is typically under-determined for biological complex systems and networks – optimization priors could lead to a crucial boost in the performance of our inferences. Beyond capturing the low-level statistical expectations for the parameters (such as their smoothness or sparsity), optimization priors impose a complex structure on the parameter space and, for example, *a priori* exclude swaths of parameter space that lead to non-functioning biological systems. 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.

Statistical analysis of optimality in the visual cortex

In this section we demonstrate the applicability of our framework to real biological data. We analyze receptive fields (RFs) of neurons in the primary visual cortex (V1) of the Macaque monkey (50) (Fig. 7A). This system is a particularly good test case, since multiple candidate optimality theories of V1 were developed and tested against data (34, 38, 51, 52).

For the purpose of our analysis, we consider two well known utility functions of neural RFs: sparseness and slowness. Sparse utility U_s, described in detail in the previous section, prioritizes receptive filters which rarely generate strong neural responses in natural conditions. Slowness utility U_l assumes that neurons extract invariant properties of sensory data (51) (see Methods for details). Optimally slow RFs minimize temporal variability of neural activity in natural sensory enviromnents (53). On the level of individual neurons, these two optimality criteria yield very different predictions. In contrast to optimally sparse RFs which are localized in space and frequency (Fig. 7B, left column), RFs optimized for slowness are broad and non-local (Fig. 7B, right column).

We first asked whether RFs of individual neurons support the optimality hypothesis, under both utilities, $U_\mathrm{s}(\phi)$ and $U_\mathrm{l}(\phi)$ as in Question 1. Given the high-quality of RFs estimates (Fig. 7A), we evaluate each utility function directly on inferred RFs and use it as a test statistic. In that way, we avoid costly marginalization of the high-dimensional likelihood. To construct null distributions for the test, we sampled 10^6 random filters consistent with optimization prior $P(\phi|\beta=0)$, and declared the 95th percentile of these distributions to be the optimality threshold (Fig. 7C left and right columns, dashed red lines). Orange and green dots in Fig. 7C denote utility values of non-significant and significant RFs, respectively. Examples of significant and non-significant RFs are displayed in green and orange frames, respectively. The large majority (204) of V1 neurons passes the sparse optimality thresholds, while only 3 pass the test of slowness optimality. This result is expected given the apparent visual similarity of neural RFs and optimally sparse filters.

We next asked whether all RFs can be used together to quantify the degree of population optimality, as in Question 2. We estimated approximate posteriors over parameter β via rejection sampling (see Methods), using all RFs in the population (Fig. 7D, purple lines). For comparison, we also computed posteriors using 250 utility-maximizing filters (Fig. 7D, red lines), and 250 utility-minimizing filters (Fig. 7D, gray lines). We next computed the maximum-a-posteriori (MAP) estimate of the population optimality parameter β (Fig. 7D, vertical dashed lines). MAP estimates obtained with simulated maximal and minimal utility RFs provide a reference for the interpretation of β estimated from real data ($\hat{\beta}_{V1}$). In case of sparse utility (Fig. 7D, left column), $\hat{\beta}_{V1}$ is very close to the parameter value of the optimally sparse filters, implying high degree of optimization. This is in contrast to slowness utility, where $\hat{\beta}_{V1}$ takes a small negative value, implying that individual V1 neurons are slightly "anti-optimized" for slowness. This means that a set of RFs drawn from a uniform distribution over the parameter domain would yield a higher average slow utility value than the one observed in the data.

Locating the distributions parametrized by infered values of β in the entropy-utility plane as in Fig. 4F was not feasible – estimating entropies is notoriously difficult in spaces of large dimension (though it may be possible using advanced Monte Carlo sampling methods (54)). However, the percentiles of the mean utility within the null distribution $p(U|\beta=0)$, Fig. 7C (blue dots) provide a similar insight into the restriction within the parameter (RF) domain. This percentile was 99.6% for the sparse utility U_s, meaning that only 0.4% of the parameter domain has higher utility than was observed on average. Taken together, this analysis illustrates how the “degree of population optimality” can be obtained from experimental data.

Population optimality β parametrizes the entire distribution of receptive fields with a particular average utility. Inference of β therefore enables us to draw conclusions and make predictions which would not be possible by other means (e.g. by averaging utilities of all RFs in the population). For example, we can predict the distribution of utility values which could be observed in future experiments and be consistent with the inferred degree of optimality $p(U(\phi)|\hat{\beta})$ (Fig. 7E, purple lines). This distribution can be different from the empirically observed one (Fig. 7E, blue lines) due to small sample size, but also if the system is optimized for a different utility or is subject to further constraints – violating the maximum entropy assumption. This is the case for slowness utility (Fig. 7E, right column). On the other hand, the predicted and empirical distributions are very similar for sparse utility (Fig. 7, left column), for which the system was found to be strongly optimized.

Another rich set of predictions can be made about correlations and other statistical features of system parameters (here–RF shapes), consistent with the given degree of optimality and average utility. As an example of such a complex statistical property, we consider spatial autocorrelation of RFs (Fig. 7F). Extreme values of β predict very different autocorrelation patterns (Fig. 7F, left and right column, left and right panels). Autocorrelation patterns consistent with inferred degrees of optimality resemble the average autocorrelation of V1 neurons more than the edge cases (Fig. 7F, left and right column, middle panels). These high-order statistics determined by the β estimate form a rigorous prediction about statistical properties of a system optimized to a given degree.

In this section, 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 a *population* of model neurons for maximal slowness yields filters which very closely resemble RFs of visual neurons (38, 53). 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.

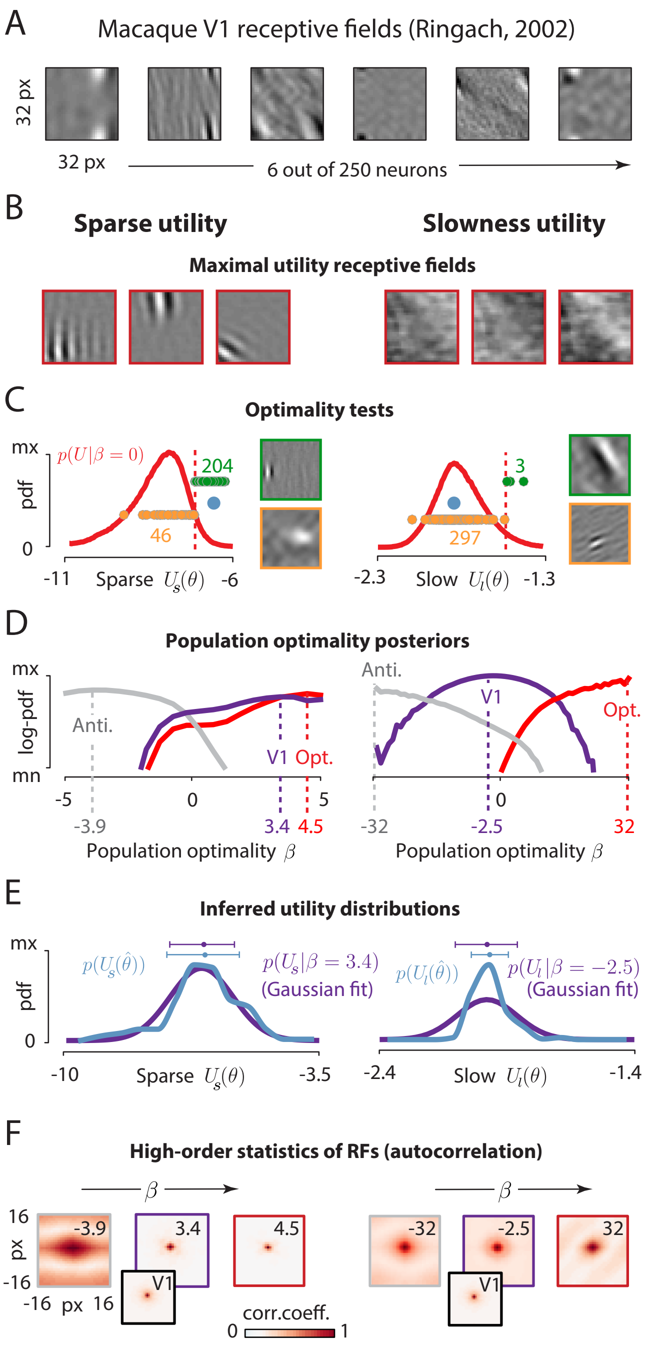

Fig. 7. Analysis of neural data. **(A)** Six example receptive fields (RFs) from Macaque visual cortex (courtesy of Dario Ringach; (50)). **(B)** Simulated RFs optimized for sparsity (left column) and slowness (right column). **(C)** Null distributions of utility values used to test for optimality under 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 99.6th percentile (sparse U_s) and 46th percentile (slow U_l) 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.

Discussion

In this paper, we presented a statistical framework that unifies normative, top-down models of complex systems which derive the system's parameters *ab initio* from an optimization principle, with bottom-up probabilistic models which fit the system's parameters to data. The union of these two approaches, often applied separately, becomes straightforward in the Bayesian framework, where the normative theory enters as the prior and data enters as the likelihood. The two traditional approaches are recovered as limiting cases; more importantly, interpolation between these two limits spans a mixed regime of optimization and inference that is highly relevant for understanding complex biological systems. We illustrated the relevance of our framework by describing how (i) measurements can be used to test a given system for consistency with an optimization theory; (ii) "closeness to optimality" can be defined and inferred; (iii) degeneracies of theoretical predictions can be broken by a small amount of data; (iv) optimization theories can provide powerful priors to aid inference in high-dimensional problems.

Our framework dovetails with other approaches which address the issues of ambiguity of theoretical predictions and model identifiability given limited data in biology. The framework of "sloppy-modelling" (55, 56), 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 o limited data to coarse-grain probabilistic models (26, 57, 58). Here, we demonstrate that breaking degeneracies of theoretical predictions with small data samples can be seen as a related coarse-graining approach.

Applications and extensions. In theoretical biology one is frequently confronted with a scenario where a biological system is hypothesized to be optimal (e.g., neuron maximizes information transmission) under some quantitative constraint (e.g., a limit to the maximal firing rate, or intrinsic noise; ([17], [28], [32])). When the value of the constraint is known, the prediction naturally emerges from the theory – but what if the constraint value is not known? One way of addressing the problem in our framework would be to consider the system parametrized by *all* parameters (including the constraints). In a pure optimization setup, the utility function reaches a nontrivial maximum within the allowed interval for some parameters, while the other parameters would be driven to extreme values (e.g. zero or infinity) by the optimization—even when that is physically impossible. In our classifier example, optimality sets the position of the nonlinearity, x_0, to a finite value, whereas it attempts to increase the slope, k, without bound—physically, this would imply reducing the noise in the classifier to zero. In contrast, in our framework, data will localize the otherwise-unbounded k value, reflecting the existence of a physical constraint in the real system. Thus, optimization prediction will correspond to finding the optimal x_0 given the value of k that is supported by data. In other words, our framework has the ability to jointly infer the parameters that correspond to constraints while simultaneously learning the remaining parameters from the normative theory. In more realistic settings, this ability could be greatly potentiated. For example, a standard neuron model could be parametrized by hundreds of parameters (corresponding to the receptive field) plus several parameters for the nonlinearity, with essentially all parameters determined by optimality except for the nonlinearity steepness (noise constraint) and/or maximum value (maximum firing rate constraint). Traditionally, these two values would be set manually and then optimization would be carried out for receptive field parameters for all values of the constraint(s) to test for match with data. A manual adjustment of the bounding intervals for those parameters that are unconstrained by optimality theory to yield consistency of optimality predictions with data is clearly problematic from the statistical viewpoint. Such manual "fine-tuning" of constraints would *de facto* amount to (over-)fitting that is not controlled for. Our framework solves such problems automatically in a single step, by reinterpreting constraints as the remaining model parameters to be *rigorously* inferred from data, formally reducing the dimensionality of the fitting problem from the number of all parameters to the number of those unconstrained by optimizing the utility function. Systematically assessing the interaction between fitting and optimization within our framework is an interesting topic for future research.

Our framework provides a new approach to handle scenarios where the optimization theory formulates degenerate, nonunique predictions. A frequent solution is to postulate further constraints within the theory itself, which disambiguate the predictions ([15]). Our proposed mechanism for breaking the degeneracy of normative theories is different, yet complementary: using a small amount of data to localize the theoretical predictions to the relevant optimum, against which further statistical tests can be carried out. A possible extension would be to formally incorporate into the prior the knowledge that the data is, for example, drawn from at most one local optimum (whose identity is, however, unknown) of the normative theory.

Another advantage provided by the maximum-entropy framework is an explicit, parametric form of the distribution of system parameters with a specific average utility. Having access to this distribution enables rigorous predictions about high-order statistical properties of optimal parameters. Traditional approaches to biologial optimality focused on properties of individual optima in the parameter space are not able to make such predictions.

We foresee additional applications of the proposed framework which are beyond the scope of this paper. For example, it is often difficult to determine which optimality criterion is plausibly implemented by the biological system of interest ([17], [59], [60]). Because we leverage the well-understood machinery of Bayesian inference, our framework could be used to 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 slowness. Such analysis would be analogous to the one we performed at a single-neuron level.

Outlook. Theories of biological function are currently less structured than physical theories of nonliving matter. This is partially due to the inherent properties of biological systems such as intrinsic complexity and lack of clear symmetries. It is also partially due to the lack of theoretical approaches to systematically coarse-grain across scales and identify relevant parameters. We hope that our approach which synthesizes statistical physics, inference, and optimality theories, can provide novel ways in tackling these fundamental issues.

Materials and Methods

Model neuron and mutual information utility function. A model neuron elicits a spike at time t ($r_t = 1$) with a probability:

$$P(r_t = 1|x_t) = \frac{1}{1 + \exp\left[-k(x_t - x_0)\right]}; \tag{3}$$

the stimuli x_t were distributed according to a Gaussian Mixture Model, $P(x_t) = \sum_{i=1}^{3} w_i \mathcal{N}(\mu_i, \sigma_i^2)$, 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.

Likelihood ratio test of optimality. The proposed test uses the likelihood ratio statistic,

$$\lambda = 2\log\frac{\max_{\beta>0} P(\mathcal{D}|\beta)}{P(\mathcal{D}|\beta=0)}. \quad \textbf{(4)}$$

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 normative 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. (4). 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 $\hat{\theta}_{ML}$:

$$\lambda = 2\log\frac{\max_{\beta>0}\int_\Theta P(\mathcal{D}|\theta)\,P(\theta|\beta)\,d\theta}{\int_\Theta P(\mathcal{D}|\theta)\,P(\theta|\beta=0)\,d\theta} \quad \textbf{(5)}$$

$$= 2\log\frac{\max_{\beta>0} P(\hat{\theta}_{ML}|\beta)}{P(\hat{\theta}_{ML}|\beta=0)} \quad \textbf{(6)}$$

$$= 2\log\left(Z(0)\max_{\beta>0}\frac{e^{\beta U(\hat{\theta}_{ML})}}{Z(\beta)}\right), \quad \textbf{(7)}$$

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

$$\lambda = 2\log\frac{\max_{\beta>0}\prod_{n=1}^N P(\mathcal{D}_n|\beta)}{\prod_{n=1}^N P(\mathcal{D}_n|\beta=0)}. \quad \textbf{(8)}$$

By Wilks' theorem, for large N the null distribution of λ approaches the χ_1^2 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.

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:

$$P(\beta|\mathcal{D}_1, \ldots, \mathcal{D}_N) \propto P(\beta)\prod_{n=1}^N \int_{\theta_n} P(\mathcal{D}_n|\theta_n)P(\theta_n|\beta)d\theta_n, \quad \textbf{(9)}$$

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.

Inference of receptive fields with optimality priors. We randomly sampled 16×16 pixel image patches from the van Hateren natural image database (52) and standardized them to zero mean and unit standard deviation. Neural responses were simulated using a Linear-Nonlinear Poisson (LNP) model:

$$P(r_t|x_t, \phi, k, x_0) = \frac{\lambda_t^{r_t} e^{-\lambda_t}}{r_t!}, \quad \textbf{(10)}$$

where λ_t is the rate parameter equal to:

$$\lambda_t = \frac{L}{1 + \exp\left[-\phi^T x_t\right]}, \quad \textbf{(11)}$$

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:

$$U_s(\phi) = -\left\langle 1 + \log(\phi^T x_t)^2)\right\rangle. \quad \textbf{(12)}$$

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 generated a model population of 64 neurons. We learned 64 linear filters using the logistic Independent Component Analysis (38). We sorted learned filters according to their sparse utility. We then retained 24 least sparse filters. The remaining 40 filters in the population were obtained by maximizing the sparse utility U_s, starting from different random conditions. Prior to learning, we reduced the dimensionality of stimuli to 64 dimensions using Principal Component Analysis.

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 (16):

$$P(\phi_n|\beta) \propto \exp\Big[\beta(U_s(z(\phi_n)))\Big], \quad \textbf{(13)}$$

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:

$$E(\phi_n|D,S,\beta) \propto -\frac{1}{\sigma^2}\sum_{t=1}^{T}\left(\phi_n^T s_t - r_{t,n}\right)^2 - \beta U_{\mathrm{s}}(z(\phi_n)). \quad \textbf{(14)}$$

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, 100, 1000. 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.

Analysis of V1 receptive fields. Receptive fields of 250 neurons in the Macaque V1 were published and analyzed in (50). All receptive fields were downsampled to 32×32 pixels size and normalized to have zero mean and unit variance.

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 standarized 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 (9) was therefore approximated as follows:

$$P(\beta|\mathcal{D}_1,\ldots,\mathcal{D}_N) \approx P(\beta)\prod_{n=1}^{N}\frac{1}{Z(\beta)}P(\hat{\theta}_n|\beta). \quad \textbf{(15)}$$

where $\hat{\theta}$ are receptive fields estimates computed in (50).

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 $P(U_{\mathrm{s}}(\theta)|\beta)$ or $P(U_{\mathrm{l}}(\theta)|\beta)$ 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. For slow utility we used 64 β values equally spaced on the $[-32,32]$ interval.

Filters accepted for each β value were used to compute the average spatial autocorrelation.

Given a linear filter ϕ, we quantified slowness of its responses to a set of natural image sequences using the following function:

$$U_{\mathrm{l}}(\phi) = -\left\langle \frac{1}{T-1}\sum_{t=2}^{T}(\phi^T x_{t,n} - \phi^T x_{t-1,n})^2 \right\rangle_n. \quad \textbf{(16)}$$

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 $\alpha \in [-90°, 90°]$. The mean and standard deviation of filters ϕ and image patches $x_{t,n}$ was set to be 0 and 1 respectively.

Bibliography

1. Robert Rosen. *Optimality principles in biology.* Springer, 2013.
2. William Bialek. *Biophysics: searching for principles.* Princeton University Press, 2012.
3. Gašper Tkačik and William Bialek. Information processing in biological systems. *Annu Rev Cond Matt Phys*, 7:89–117, 2016.
4. Simon Laughlin. A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung c*, 36(9-10):910–912, 1981.
5. Johannes H van Hateren. Theoretical predictions of spatiotemporal receptive fields of fly lmcs, and experimental validation. *Journal of Comparative Physiology A*, 171(2):157–170, 1992.
6. H Kacser and JA Burns. The control of flux. *Biochem Soc Trans*, 23:341–366, 1995.
7. RU Ibarra, JS Edwards, and BO Palsson. Escherichia coli k-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, 420:186–189, 2002.
8. 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*, 107:3475–3480, 2010.
9. Gašper Tkačik, Curtis G Callan, and William Bialek. Information flow and optimization in transcriptional regulation. *Proceedings of the National Academy of Sciences*, 105(34): 12265–12270, 2008.
10. 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*, 327:439–442, 2010.
11. Steven Hecht Orzack. *Adaptionism and Optimality.* Cambridge University Press, 2001.
12. R McNeill Alexander. *Principles of animal locomotion.* Princeton University Press, 2003.
13. Wilson S Geisler. Contributions of ideal observer theory to vision research. *Vision research*, 51(7):771–781, 2011.
14. Joshua I Gold and Michael N Shadlen. The neural basis of decision making. *Annual review of neuroscience*, 30, 2007.
15. 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):16256–16264, 2012.
16. 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*, page 837567, 2019.
17. Zhuo Wang, Alan A Stocker, and Daniel D Lee. Efficient neural codes that minimize lp reconstruction error. *Neural computation*, 28(12):2656–2686, 2016.
18. Il Memming Park and Jonathan W Pillow. Bayesian efficient coding. *bioRxiv*, page 178418, 2017.
19. Jan Eichhorn, Fabian Sinz, and Matthias Bethge. Natural image coding in v1: how much use is orientation selectivity? *PLoS computational biology*, 5(4):e1000336, 2009.
20. Alfonso Pérez-Escudero, Marta Rivera-Alba, and Gonzalo G de Polavieja. Structure of deviations from optimality in biological systems. *Proceedings of the National Academy of Sciences*, 106(48):20544–20549, 2009.
21. 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):2988, 2018.
22. Mijung Park and Jonathan W Pillow. Receptive field inference with localized priors. *PLoS computational biology*, 7(10):e1002219, 2011.
23. 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):453–461, 1946.
24. David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms.* Cambridge university press, 2003.
25. Kevin P Murphy. *Machine learning: a probabilistic perspective.* MIT press, 2012.

26. Benjamin B Machta, Ricky Chachra, Mark K Transtrum, and James P Sethna. Parameter space compression underlies emergent theories and predictive models. *Science*, 342 (6158):604–607, 2013.
27. Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.
28. Tatyana Sharpee and William Bialek. Neural decision boundaries for maximal information transmission. *PLoS One*, 2(7):e646, 2007.
29. 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):2533–2538, 2015.
30. Liam Paninski, Jonathan Pillow, and Jeremy Lewi. Statistical models for neural encoding, decoding, and optimal stimulus design. *Progress in brain research*, 165:493–507, 2007.
31. 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):14419–14424, 2010.
32. Julijana Gjorgjieva, Haim Sompolinsky, and Markus Meister. Benefits of pathway splitting in sensory coding. *Journal of Neuroscience*, 34(36):12127–12144, 2014.
33. Horace B Barlow et al. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1:217–234, 1961.
34. 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):607, 1996.
35. Evan C Smith and Michael S Lewicki. Efficient auditory coding. *Nature*, 439(7079):978, 2006.
36. 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): 186–191, 2018.
37. Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.
38. 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.
39. Deep Ganguli and Eero P Simoncelli. Efficient sensory encoding and bayesian inference with heterogeneous neural populations. *Neural computation*, 26(10):2103–2134, 2014.
40. 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):17368–17373, 2010.
41. Bart G Borghuis, Charles P Ratliff, Robert G Smith, Peter Sterling, and Vijay Balasubramanian. Design of a neuronal array. *Journal of Neuroscience*, 28(12):3178–3189, 2008.
42. Wiktor Młynarski. The opponent channel population code of sound location is an efficient representation of natural binaural sounds. *PLoS computational biology*, 11(5):e1004294, 2015.
43. Wiktor Młynarski and Josh H McDermott. Learning midlevel auditory codes from natural sound statistics. *Neural computation*, 30(3):631–669, 2018.
44. 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):e1002594, 2012.
45. 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):e1005150, 2016.
46. Xaq Pitkow and Markus Meister. Decorrelation and efficient coding by retinal ganglion cells. *Nature neuroscience*, 15(4):628, 2012.
47. Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4):481–487, 2004.
48. Tatyana O Sharpee. Computational identification of receptive fields. *Annual review of neuroscience*, 36:103–120, 2013.
49. Cristina Savin and Gasper Tkacik. Estimating nonlinear neural response functions using gp priors and kronecker methods. In *Advances in Neural Information Processing Systems*, pages 3603–3611, 2016.
50. Dario L Ringach. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of neurophysiology*, 88(1):455–463, 2002.
51. Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002.
52. 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):359–366, 1998.
53. Pietro Berkes and Laurenz Wiskott. Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of vision*, 5(6):9–9, 2005.
54. Fugao Wang and D. P. Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.*, 86:2050–2053, Mar 2001. doi: 10.1103/PhysRevLett.86.2050.
55. Timothy O'Leary, Alexander C Sutton, and Eve Marder. Computational models in the age of large datasets. *Current opinion in neurobiology*, 32:87–94, 2015.
56. 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):e189, 2007.
57. William Bialek, Curtis G Callan, and Steven P Strong. Field theories for learning probability distributions. *Physical review letters*, 77(23):4693, 1996.
58. Wei-Chia Chen, Ammar Tareen, and Justin B Kinney. Density estimation on small data sets. *Physical review letters*, 121(16):160605, 2018.
59. Wiktor F Młynarski and Ann M Hermundstad. Adaptive coding for dynamic sensory inference. *Elife*, 7:e32055, 2018.
60. M Chalk, G Tkacik, and O Marre. Inferring the function performed by a recurrent neural network. *biorxiv*, page 598086, 2019.