Weak universality in sensory tradeoffs

Sarah Marzen[1,2,*] and Simon DeDeo[3,4,5]

[1]*Department of Physics, Redwood Center for Theoretical Neuroscience, University of California at Berkeley, Berkeley, California 94720, USA*

[2]*Physics of Living Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

[3]*Center for Complex Networks and Systems Research, Department of Informatics, Indiana University, 919 East 10th Street, Bloomington, Indiana 47408, USA*

[4]*Department of Social and Decision Sciences, 5000 Forbes Avenue, BP 208, Pittsburgh, Pennsylvania 15213, USA*

[5]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA*

For many organisms, the number of sensory neurons is largely determined during development, before strong environmental cues are present. This is despite the fact that environments can fluctuate drastically both from generation to generation and within an organism's lifetime. How can organisms get by by hard coding the number of sensory neurons? We approach this question using rate-distortion theory. A combination of simulation and theory suggests that when environments are large, the rate-distortion function—a proxy for material costs, timing delays, and energy requirements—depends only on coarse-grained environmental statistics that are expected to change on evolutionary, rather than ontogenetic, time scales.

I. INTRODUCTION

The amount of sensory information potentially available to an organism is, for all practical purposes, infinite. This, taken together with the finite size of the brain, implies that we constantly operate in the lossy regime, transmitting only some of the information present in the environment. Optimal sensing, in other words, is optimal compression, and this means that core theorems of information theory constrain perception.

In the context of evolved sensors, the introduction of the need to compress and coarse-grain environmental signals extends the efficient coding hypothesis [1], which has guided experimental and theoretical neuroscience for the past five and a half decades (see Ref. [2] and references therein), to the lossy regime [3].

Considering the role of lossy compression in evolved sensory systems leads to interesting interpretations of existing experimental results concerning neurogenesis, or the dynamic creation of new neurons over an organism's lifetime [4–7]. While neurogenesis is widespread, neurogenesis in sensory regions is less commonly observed. Indeed, the number of neurons in sensory regions appears to be determined prior to the receipt of any environmental cues [8], though some famous counterexamples exist [9]. In other words, for many species, the number of neurons in a brain's sensory region is strongly determined by genetic effects [10].

Here, we provide an information-theoretic explanation for these facts by viewing early sensory regions as lossy perceptual feature extractors for which the number of sensory neurons limits the accuracy of the organism's internal representation of the environment. To show this, we use a model of the environment general enough to apply to a range of biological situations, but rich enough to capture the basic problem of perception and encoding, in which both the probability of observing a particular environmental symbol and the cost of misperceiving those symbols are randomly drawn [11].

In this minimal model, the tradeoff between neuron number and representational accuracy is essentially invariant to changes in the probability distribution over sensory inputs and the particular costs of misperceiving one sensory input for another; this is true even though the optimal internal coding of environmental inputs varies wildly from one environment to the next. These results lead to a different functional interpretation of phenotypic variability and neurogenesis in sensory brain regions: First, phenotypic variability in the sensory neuron number may be tied to phenotypic variability in the average heat dissipation rate of a sensory neuron; and second, neurogenesis may only be necessary when the organism-environment interactions change drastically, e.g., due to changes in action policy.

II. RELATING SENSORY COSTS TO THE RATE-DISTORTION FUNCTION

Confusing one environmental state for another can be costly due to a subsequent suboptimal choice of action. For example, mistaking a lion for a domesticated cat might lead to death, while mistaking a domesticated cat for a lion might lead to unnecessary energy spent running. However, correctly identifying a greater number of objects requires more mental effort, whether that be measured by a larger number of neurons devoted to object recognition or a correspondingly larger number of adenosine triphosphate (ATP) molecules consumed in their function.

Previous work suggests that resource constraints such as these are critical in shaping the neural code [12–19]. In this Rapid Communication, we also endeavor to describe the tradeoff between the neuron number and the quality of information transmission. However, our approach differs from these previous efforts in two major ways. First, we use rate-distortion theory (a branch of information theory) and recent advances in nonequilibrium statistical mechanics

————
*smarzen@berkeley.edu

to place lower bounds on material, timing, and energy costs without specifying a particular neural code. Second, we use a distortion measure (e.g., loss function) to quantify the quality of sensory information processing; confusing sensory input x for $\tilde{x}$ costs the organism $d(x,\tilde{x})$. This leads to a stark difference in objective function: Previous authors looked to maximize the shared information between stimulus and response $I[X;\tilde{X}]$ while minimizing a separately calculated energy expenditure; we look to *minimize* both shared information $I[X;\tilde{X}]$ and expected distortion $\langle d(x,\tilde{x}) \rangle_{p(x,\tilde{x})}$.

For concreteness, we consider m sensory neurons that form an information bottleneck between environmental information and downstream brain regions that decide organism actions based on the perceptual information. Though some work suggests that the neural code might be analog (based on spike timing), there is inherent noise in the neural circuitry that effectively imposes a minimal discretization time (a few milliseconds [20]) on which the neural code operates. We choose the time units to be that minimal discretization time, and think of the sensory neural code as a binary vector of length m in which 1 (0) in the ith position codes for a spike (no spike) from neuron i in that minimal discretization time. Here, sensory inputs x are independently and identically distributed (i.i.d.) with probability $p(x)$.

Rate-distortion theory places asymptotically achievable lower bounds on the *rate* of the sensor, the bits per input symbol required to communicate the sensor's state to a decoder, lower bounded by $I[X;\tilde{X}]$. The sensor's *distortion* is given by the expected value of a user-specified distortion measure $d(x,\tilde{x})$, which quantifies the cost of confusing x and $\tilde{x}$. Distortion can be connected to the reward function $\tilde{r}(x,a)$ and action policy $p(a|x)$ in a simplified reinforcement learning setup [21] where a are possible actions, via $d(x,\tilde{x}) = [\max_{\tilde{x}} \sum_a p(a|\tilde{x})\tilde{r}(x,a)] - \sum_a p(a|\tilde{x})\tilde{r}(x,a)$. This distortion measure is "normal", $d(x,x) = 0$, if the action policy uses all available information about the environmental state via sensory representation. The rate-distortion function $R(D)$ delineates the boundary between achievable and unachievable combinations of rate and distortion as shown in Fig. 1.

Material and timing costs both run into fundamental limits quantified by the rate-distortion function. If the number of sensory neurons is greater than the rate-distortion function $R(D) \leqslant m$, then one can instantaneously decode each estimated input $\tilde{x}_t$ from the tth binary vector of length m, but $R(D)$ places a lower bound on material cost m. If $R(D) \geqslant m$, then we can acquire additional expressiveness by coding each input x as a string of binary vectors of length m, resulting in timing delays. The expected length of the neuronal output string is no less than $R(D)/\log_2 2^m = R(D)/m$, which, when multiplied by the number of input symbols sensed thus far, is the timing delay between encoding and decoding.

Finally, a more generally applicable nonequilibrium thermodynamics viewpoint ties the rate-distortion function to power consumption. Memoryless channels implicitly have a measure-reset cycle: First, the channel senses the environment, and the channel communicates its measurement to some "homunculus"; and afterwards, the channel resets its internal state. The energy per reset required to maintain such a channel is lower bounded by $k_B T I[X;\tilde{X}]$ [22,23], which is lower bounded by $k_B T R(D)$, where $I[X;\tilde{X}]$ and $R(D)$ here are
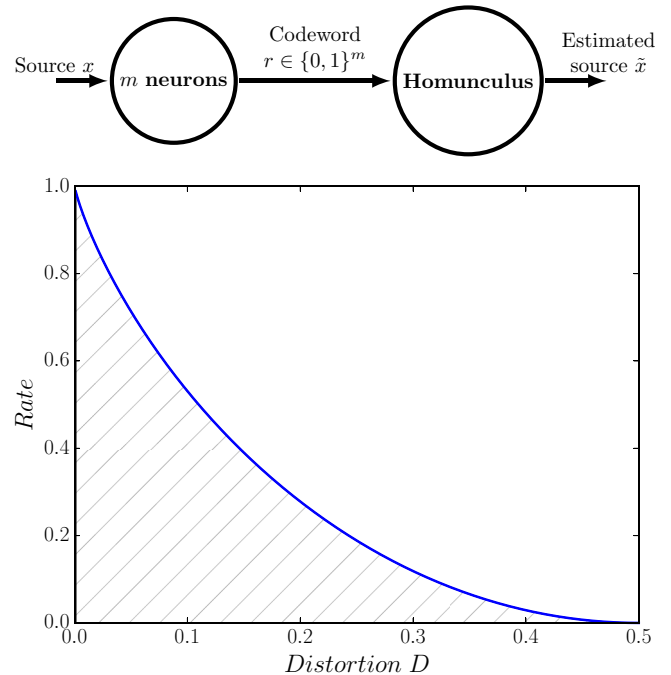

FIG. 1. Rate-distortion theory: Estimation with an information bottleneck. Top panel: m sensory neurons are asked to communicate an environmental input x to a "homunculus" using neuron spike code $p(r|x)$; the homunculus creates an estimate $\tilde{x}$ of the environmental input on the basis of the spiking pattern r. The quality of estimation is given by the expected distortion $D = \langle d(x,\tilde{x}) \rangle_{p(x,\tilde{x})}$, and the absolute rate of the information bottleneck R is m, the number of sensory neurons. Bottom panel: The rate-distortion function shown in blue delineates the boundary between achievable (white) and unachievable (hatched) combinations of rate and distortion. The rate-distortion function shown here is that for two equiprobable environmental inputs and Hamming distortion measure [25].

measured in nats. This is a different energetic consideration than that mentioned in Ref. [24].

In short, $R(D)$ places a lower bound on the size of the physical substrate, on timing delays between encoding and decoding environmental input, and on the power required to maintain the sensor. Alternatively, if some part of the brain is "transparent", the rate $R(D)$ places a lower bound on the channel capacity of downstream brain regions.

Researchers have used rate-distortion theory to study everything from chemotaxis [26] to genetic transcription [24] to prediction in the salamander retina [27] to human vision [28]. The appropriate choice of information source and distortion measure depends heavily on the particular biological system that one studies. Here, we use rate-distortion theory to model environments using the framework of Ref. [11]. Distortions $d(x,\tilde{x})$ and the probability distribution of inputs $p(x)$ are drawn from a probability distribution that represents the range of possible environments an organism might find itself born into. In this Rapid Communication, for simplicity, all off-diagonal distortions are drawn i.i.d. with probability density function $\rho(d)$, and the probability distribution of inputs is drawn from a Dirichlet distribution with concentration parameter α. The α parameter in Dirichlet distribution dictates how uncertain an environment is; as α increases, $p(x)$ is

more evenly distributed among the N possible states, and the uncertainty of the environment increases. Together, $\rho(d)$ and α specify a generative model for environments.

To calculate $R(D)$ given a distortion measure and probability distribution over inputs, we find the $p_\beta(\tilde{x}|x)$ which minimizes the rate-distortion Lagrangian $\beta \langle d(x,\tilde{x}) \rangle_{p(x,\tilde{x})} + I[X;\tilde{X}]$ using the Blahut-Arimoto algorithm [25], and calculate the resultant rate R_β and expected distortion D_β. As β sweeps from 0 (high distortion) to ∞ (low distortion), R_β and D_β parametrically trace out the rate-distortion function $R(D)$.

The rate-distortion functional also has a physical interpretation as a total energetic cost. Distortion D quantifies the food energy that the organism failed to intake from the environment; rate r is correlated with the energy that the organism expended in doing so. We can loosely think of the rate r as a proxy for neuron number, so that the energy expenditure of this organism's brain is $\beta^{-1}r$, where β is the average rate of energy use for a single neuron [29]. The overall energetic cost to the organism is then $D + \beta^{-1}r$, and the fitness of an organism is some monotonically decreasing function of the organism's energetic cost.

III. WEAK UNIVERSALITY OF THE RATE-DISTORTION FUNCTION

Numerical experiments shown in Fig. 2 strongly suggest that, when there are many possible environmental inputs ($N \gg 1$), the rate-distortion function $R(D)$ does not depend
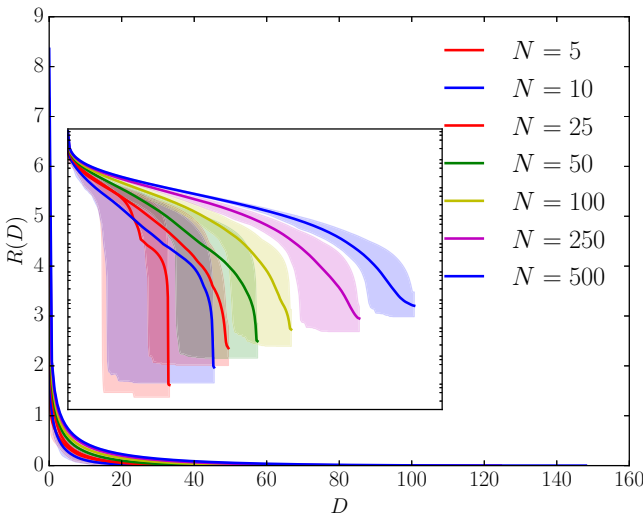

FIG. 2. Convergence in probability to a single rate-distortion function. Each line shows the average $R(D)$ for 100 rate-distortion functions calculated for worlds with equivalent generative models, with $\alpha = 1$ and $\rho(d) = \frac{1}{\sigma d \sqrt{2\pi}} e^{-(\log d - \mu)^2/2\sigma^2}$ with $\mu \approx 3$ and $\sigma \approx 2$. Lines denote estimates of $\bar{R}_N(D)$ and the surrounding transparent regions show 68% confidence intervals on $R(D)$ obtained by bootstrapping. Linear interpolation is used to find $R(D)$ at the desired D's from the distortions at which $R(D)$ was actually calculated. The inset is a log-log plot of the same. The size of the 68% confidence intervals appear to decrease, and the average rate-distortion function $\bar{R}_N(D)$ appears to increase, as N grows larger.

on the specific distortion measure or environmental input probabilities, but only on the distribution from which distortions were drawn, $\rho(d)$, and the distribution from which the input probabilities were drawn, characterized by concentration parameter α. Note that Ref. [11] considered the effects of a nonzero $d_{\min} = \inf\{x : \rho(x) > 0\}$. Here, we assume that $d_{\min} = 0$.

We refer to the insensitivity of the rate-distortion function to the particular distortion measure and probability distribution over inputs as "weak universality" [30]. In particular, we now argue that the rate-distortion function converges in probability to a curve which depends only on ρ and α. Let the subscripts of $R_{N,\mathbf{d}}(D)$ denote the number of sensory inputs N and the distortion measure $\mathbf{d}$. We wish to show that

$$\lim_{N \to \infty} \mathbb{P}\left(\left| R_{N,\mathbf{d}}(D) - \lim_{N \to \infty} \langle R_{N,\mathbf{d}}(D) \rangle_{p(\mathbf{d})} \right| \geqslant \epsilon \right) = 0 \; \forall \epsilon > 0. \tag{1}$$

If Eq. (1) holds, then (loosely speaking) the rate-distortion function $R_{N,\mathbf{d}}(D)$ depends only on ρ and α in the large N limit, *even though* optimal codebooks for distortions with the same ρ and α but different $\mathbf{d}$ tend to differ wildly.

To do so, we must first argue that $\lim_{N \to \infty} \langle R_{N,\mathbf{d}}(D) \rangle_{p(\mathbf{d})}$ exists. Reference [11] showed that $\langle R_{N,\mathbf{d}}(D) \rangle_{p(\mathbf{d})} \leqslant \log \frac{1}{\int_D^\infty \rho(x)dx}$ in the large N limit, which implies that $\langle R_{N,\mathbf{d}}(D) \rangle_{p(\mathbf{d})}$ is bounded from above. Simulation results suggest that $\langle R_{N,\mathbf{d}}(D) \rangle_{p(\mathbf{d})}$ is strictly increasing with N; see Ref. [11] and Fig. 2 for examples. The monotone convergence theorem then implies that $\lim_{N \to \infty} \langle R_{N,\mathbf{d}}(D) \rangle_{p(\mathbf{d})}$ exists. For ease, we introduce the notation $\bar{R}_N(D) := \langle R_{N,\mathbf{d}}(D) \rangle_{p(\mathbf{d})}$ and

$$\bar{R}(D) := \lim_{N \to \infty} \bar{R}_N(D), \tag{2}$$

where $\bar{R}_N(D)$ and $\bar{R}(D)$ depend on both ρ and α.

An application of Markov's inequality with a non-negative random variable $X = |R_{N,\mathbf{d}}(D) - \bar{R}(D)|$ reveals that

$$\mathbb{P}(|R_{N,\mathbf{d}}(D) - \bar{R}(D)| \geqslant \epsilon)$$
$$\leqslant 2\bar{R}(D) \frac{\sqrt{\langle [R_{N,\mathbf{d}}(D) - \bar{R}_N(D)]^2 \rangle_{p(\mathbf{d})}}}{\epsilon}, \tag{3}$$

where we have used $\langle |R_{N,\mathbf{d}}(D) - \bar{R}(D)| \rangle_{p(\mathbf{d})} \leqslant \langle R_{N,\mathbf{d}}(D) \rangle_{p(\mathbf{d})} + \bar{R}(D) \leqslant 2\bar{R}(D)$. For Eq. (1) to hold, we must show that the right-hand side of Eq. (3) tends to 0 as $N \to \infty$. We argued above that $\bar{R}(D)$ was bounded. Figure 3 suggests that $\lim_{N \to \infty} \sqrt{\langle [R_{N,\mathbf{d}}(D) - \bar{R}_N(D)]^2 \rangle_{p(\mathbf{d})}} = 0$ for all D. Altogether, then, we have numerical evidence that Eq. (1) holds for any D, i.e., that $R_{N,\mathbf{d}}(D)$ converges in probability to $\bar{R}(D)$.

In some ways, the results presented above are unsurprising. The rate-distortion function is a one-dimensional projection of $N(N-1)$ i.i.d. distortions, so we might expect weak universality to emerge from some particular application of the weak law of large numbers. Indeed, we can show analytically that this happens in the two extreme limits (low and high distortion), thus providing some insight into the mechanism by which the rate-distortion function converges in probability
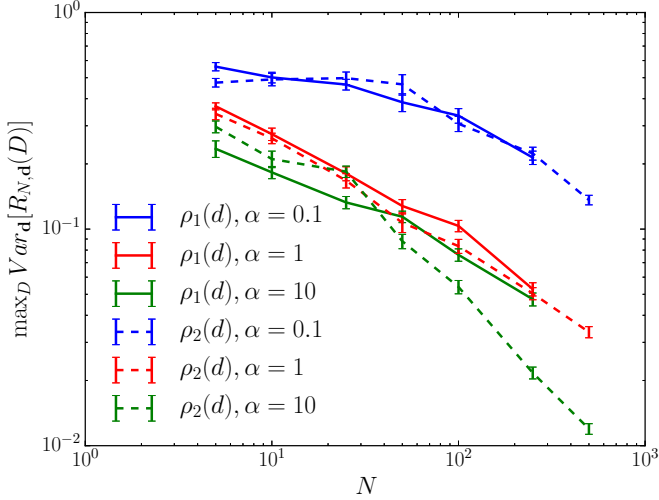
FIG. 3. Fluctuations in required resources diminish as the environmental complexity increases. Maximal variance in $R(D)$ calculated for three different world generative models at $N = 5, 10, 25, 50, 100, 250, 500$. The distortions are chosen from $\rho_1(d) = e^{-d}$ and $\rho_2(d) = \frac{1}{\sigma d \sqrt{2\pi}} e^{-(\log d - \mu)^2/2\sigma^2}$, with $\mu \approx 3$ and $\sigma \approx 2$; the input probability distributions are Dirichlet with concentration parameter α. The y-axis values, $\max_D \mathrm{Var}_{\mathbf{d}}[R_{N,\mathbf{d}}(D)]$, are estimated by bootstrapping given 100 samples of the rate-distortion function at 200 uniformly spaced distortions between 0 and the maximum $D_{\max}$. In all cases, deviations from the average case decline rapidly with N; this is true despite large differences in the moments of the different distributions considered.

to $\bar{R}(D)$. Proof of convergence at other distortions remains an open problem.

In the low-distortion limit, there is an exact expression for R_β, D_β in Ref. [31]. Repeated applications of the weak law of large numbers in the large N limit yields

$$D_\beta = \frac{\langle d e^{-\beta d} \rangle_{\rho(d)}}{\langle e^{-\beta d} \rangle_{\rho(d)}}, \tag{4}$$

$$R_\beta = \psi(N\alpha) - \psi(\alpha) - \beta D_\beta$$
$$- \log(1 + N\langle e^{-\beta d} \rangle_{\rho(d)}), \tag{5}$$

to lowest order in $N\langle e^{-\beta d} \rangle_{\rho(d)}$; the first two terms in the expansion of R_β are equal to the expectation value of the entropy of the Dirichlet distribution with parameter α [32]. Unlike the histogram of eigenvalues of random matrices, these expressions show that there is *only* weak universality, as the moment-generating function $\langle e^{-\beta d} \rangle_{\rho(d)}$ generally uniquely specifies $\rho(d)$, and different α have different $\psi(N\alpha) - \psi(\alpha)$.

A similar result exists in the high-distortion limit. With zero rate, the minimal achievable distortion $[D_{\max} = \min_{I[X;\tilde{X}]=0} \sum_{x,\tilde{x}} p(x,\tilde{x})d(x,\tilde{x})]$ is equivalent to $\min_{\tilde{x}} \sum_x p(x)d(x,\tilde{x})$. The expected value of $\sum_x p(x)d(x,\tilde{x})$ over the ensemble of environments is $(1 - \frac{1}{N})\langle d \rangle_{\rho(d)}$ and the variance is $\frac{1}{N\alpha+1}(1 - \frac{1}{N})^2 \langle d^2 \rangle_{\rho(d)} - \frac{\langle d \rangle_{\rho(d)}^2}{N^2}$, as shown in the Appendix. As long as $\rho(d)$ has a finite second moment, this variance in $R(D)$ scales as $\sim \frac{\langle d^2 \rangle_{\rho(d)}}{\alpha} \frac{1}{N}$, which tends to 0 as N grows larger; therefore, $D_{\max}$ converges in probability to $\langle d \rangle_{\rho(d)}$ as N tends to infinity.

Although the rate-distortion function appears invariant to changes in the particular distortion measure and probability distribution over inputs, near-optimal codebooks vary wildly from one environment to the next. The statistics of near-optimal codebooks at an expected distortion D are dictated by the $p(\tilde{x}|x)$ for which $\langle d(x,\tilde{x}) \rangle_{p(x,\tilde{x})} \leqslant D$ and $I[X;\tilde{X}]$ is at a minimum [31], and numerical experiments show that the statistics of such $p(\tilde{x}|x)$ are heavily dependent on the environment.

IV. IMPLICATIONS FOR BIOLOGICAL ORGANISMS

Organisms can employ one of a few strategies to cope with wildly fluctuating environments. The first strategy derives from Kelly's classical analysis of gambling, applied to phenotypic bet hedging—that a population of organisms should develop into a range of phenotypes to maximize expected log growth rate [25,33]. A second strategy would involve delaying the development of key brain regions until the organism has received strong environmental cues.

A third strategy would be to essentially ignore environmental fluctuations. At first, this seems as a suboptimal strategy, in that a population of organisms that employ either of the two strategies listed above would have a higher log growth rate. However, the weak universality results presented here suggests that the necessary size of sensory brain regions, the minimum possible timing delays in sensory perception, and the minimal power required to maintain sensory brain regions all depend only on coarse environmental statistics, even though optimal neural wiring fluctuates wildly from environment to environment. In the examples discussed in the main text, these coarse environmental statistics are $\rho(d)$ and α. More generally, these coarse environmental statistics are the parameters specifying the distribution from which distortion measures are drawn and the distribution from which probability distributions over inputs are drawn.

In apparent agreement with these findings, environmental cues are scarce during development, and seem to have a limited effect on the neuron number [8], and there are few reports of neurogenesis in mammalian sensory brain regions [9].

The number may be fixed, but wiring is not, and there are many reports of synaptic plasticity in sensory brain regions; the particular wiring of neurons in sensory brain regions does depend on the details of environmental cues [34,35].

If weak-universality-type results mean that the sensory neuron number can be largely fixed ahead of time, two questions immediately suggest themselves. First, why do investigators find evidence of neurogenesis in nonvertebrate sensory brain regions [9]? And second, why is there high phenotypic variability in the sensory neuron number for many animals, including primates [8]?

First, Ref. [9] notes that animals with substantial neurogenesis in the sensory areas are also those that grow considerably postnatally, which—in our simple conception of organisms—corresponds to an increase in the possible actions a taken by the organism. Recall that one can connect the distortion measure directly to the reward function $r(x,a)$ and action policy $p(a|\tilde{x})$. Changes in the set of actions will thus change the distortion measure in a (possibly) more structured way than what was considered here. That, in turn, will likely lead to an increase

in the requisite sensor size, necessitating adult neurogenesis in sensory areas. We leave a delineation of the induced structure in the distortion measure to future research.

Meanwhile, phenotypic variability is explainable within our minimal model. Earlier, we identified the rate-distortion objective as a fitness function, implying that variability in the Lagrange multiplier β (representative of single neuron power usage) is tightly connected to variability in the observed number of sensory neurons. This explanation could be tested by correlating the sensory neuron number with the average heat dissipation rate of single neurons in sensory regions.

Our minimal model of sensory tradeoffs in biological organisms lead to different questions at the intersection of random matrix theory, information theory, and sensory processing. Extensions of this approach—to distortion measures that change as the animal grows, or to distortion measures and probability distributions over inputs with more structure—may predict and mathematically explain other observed similarities and differences between species.

APPENDIX: WEAK UNIVERSALITY IN THE LOW- AND HIGH-DISTORTION LIMITS

First we tackle the low-distortion limit. When $p_\beta(\tilde{x})$ has full support, an exact expression exists for R_β and D_β from Ref. [31]. Let $\vec{p}(x)$ be a vector of input probabilities $p(x)$, let **d** be the distortion matrix, and let $Q_{x,\tilde{x}} = e^{-\beta d(x,\tilde{x})}$. Then

$$R_\beta = -\beta D_\beta + H[X] + \vec{p}(x)^\top \log(Q^{-1}\vec{1}), \quad \text{(A1)}$$

$$D_\beta = \left[\frac{\vec{p}(x)}{Q^{-1}\vec{1}}\right]^\top Q^{-1}(\mathbf{d} \odot Q)(Q^{-1}\vec{1}). \quad \text{(A2)}$$

When β is sufficiently large, then the entries of $Q - I$ are much smaller than 1 with high probability, suggesting the expansion

$$Q = I + (Q - I),$$
$$Q^{-1} = [I + (Q - I)]^{-1}$$
$$= \sum_{m=0}^{\infty}(-1)^m(Q - I)^m.$$

By the weak law of large numbers, $(Q - I)\vec{1}$ is highly concentrated around $(N - 1)\langle e^{-\beta d}\rangle_{\rho(d)}\vec{1}$ as long as the probability

density function for $e^{-\beta d}$ has finite variance, so that

$$(Q - I)^m\vec{1} \approx [(N - 1)\langle e^{-\beta d}\rangle_{\rho(d)}]^m\vec{1},$$

showing that

$$Q^{-1}\vec{1} = \sum_{m=0}^{\infty}[-(N - 1)\langle e^{-\beta d}\rangle_{\rho(d)}]^m\vec{1}$$
$$\approx [1 + (N - 1)\langle e^{-\beta d}\rangle_{\rho(d)}]^{-1}\vec{1}.$$

Then we find that

$$\vec{p}(x)^\top \log(Q^{-1}\vec{1}) \approx \sum_x p(x)\log[1 + (N - 1)\langle e^{-\beta d}\rangle_{\rho(d)}]^{-1}$$
$$= -\log[1 + (N - 1)\langle e^{-\beta d}\rangle_{\rho(d)}],$$

so that

$$R_\beta = -\beta D_\beta + H[X] - \log[1 + (N - 1)\langle e^{-\beta d}\rangle_{\rho(d)}]. \quad \text{(A3)}$$

Similar manipulations, again based on the weak law of large numbers, reveal that

$$D_\beta \approx \frac{(N - 1)\langle de^{-\beta d}\rangle_{\rho(d)}}{1 + (N - 1)\langle e^{-\beta d}\rangle_{\rho(d)}}. \quad \text{(A4)}$$

When the probability distribution over inputs is drawn from a Dirichlet distribution with concentration parameter α, then $H[X]$ is very peaked around $\psi(N\alpha) - \psi(\alpha)$ with corrections of $O(1/N)$ [36], yielding

$$R_\beta \approx -\beta D_\beta + \psi(N\alpha) - \psi(\alpha) - \log[1 + (N - 1)\langle e^{-\beta d}\rangle_{\rho(d)}]. \quad \text{(A5)}$$

Thus D_β and R_β are independent of the particular distortion matrix and dependent only on $\rho(d),\alpha$. Unlike the histogram of eigenvalues of random matrices, there is *only* weak universality, as the moment-generating function $\langle e^{-\beta d}\rangle_{\rho(d)}$ specifies $\rho(d)$ and different α have different $\psi(N\alpha) - \psi(\alpha)$. Note that this formula holds only when $p_\beta(\tilde{x})$ has full support, or (roughly speaking) when $N\langle e^{-\beta d}\rangle_{\rho(d)} \ll 1$.

Next, we tackle the high-distortion limit, i.e., find weak universality in $D_{\max}$. We look to show that $\min_{\tilde{x}} \sum_x p(x)d(x,\tilde{x})$ converges in probability $\langle d\rangle_{\rho(d)}$, and so we look to show that the expected value of $\min_{\tilde{x}} \sum_x p(x)d(x,\tilde{x})$ over worlds with the same $\rho(d),\alpha$ converges in probability to $\langle d\rangle_{\rho(d)}$. For any $\tilde{x}$ we have

$$\left\langle \sum_x p(x)d(x,\tilde{x})\right\rangle_{p(\vec{p}(x),\mathbf{d})} = (N - 1)\left(\frac{\alpha}{N\alpha}\right)\langle d\rangle_{\rho(d)}$$
$$= \left(1 - \frac{1}{N}\right)\langle d\rangle_{\rho(d)}$$

while the variance of $\sum_x p(x)d(x,\tilde{x})$ is

$$\left\langle \left(\sum_x p(x)d(x,\tilde{x}) - \left\langle\sum_x p(x)d(x,\tilde{x})\right\rangle_{p(\vec{p}(x),\mathbf{d})}\right)^2\right\rangle_{p(\vec{p}(x),\mathbf{d})}$$
$$= \left\langle\left(\sum_x p(x)d(x,\tilde{x})\right)^2\right\rangle_{p(\vec{p}(x),\mathbf{d})} - \left\langle\sum_x p(x)d(x,\tilde{x})\right\rangle_{p(\vec{p}(x),\mathbf{d})}^2$$

$$= \left\langle \sum_{x,x'} p(x)p(x')d(x,\tilde{x})d(x',\tilde{x}) \right\rangle_{p(\vec{p}(x),\mathbf{d})} - \left(1 - \frac{1}{N}\right)^2 \langle d\rangle^2_{\rho(d)}$$

$$= \left\langle \sum_{x} p(x)^2 d(x,\tilde{x})^2 \right\rangle_{p(\vec{p}(x),\mathbf{d})} + \sum_{x \neq x'} \langle p(x)p(x')d(x,\tilde{x})d(x',\tilde{x})\rangle_{p(\mathbf{d})} - \left(1 - \frac{1}{N}\right)^2 \langle d\rangle^2_{\rho(d)}$$

$$= (N-1)\langle p(x)^2\rangle_{p(\vec{p}(x))}\langle d^2\rangle_{\rho(d)} + (N-1)(N-2)\langle p(x)p(x')\rangle_{p(\vec{p}(x))}\langle d\rangle^2_{\rho(d)} - \left(1 - \frac{1}{N}\right)^2 \langle d\rangle^2_{\rho(d)},$$

where $x \neq x'$ in the second term. The first of these terms is relatively easy to evaluate using the fact that if $x_1, \ldots, x_N$ are drawn from a Dirichlet distribution with concentration parameter α, then $\rho(x_1) = B(x_1; \alpha, (N-1)\alpha)$ and $\rho(x_2|x_1) = B(\frac{x_2}{1-x_1}; \alpha, (N-2)\alpha)$ [36]. As such, we find that

$$\langle p(x)^2\rangle_{p(\vec{p}(x))} = \int_0^1 p(x_1)^2 B(p(x_1); \alpha, N\alpha)dp(x_1)$$

$$= \left(1 - \frac{1}{N}\right)^2 \frac{1}{N\alpha + 1}.$$

The second term is evaluated by noticing

$$\langle p(x)p(x')\rangle_{p(\vec{p}(x))} = \int \Pr(p(x_1), \ldots, p(x_N))p(x_1)p(x_2)dp(x_1)\cdots dp(x_N)$$

$$= \int\int \Pr(p(x_1), p(x_2))p(x_1)p(x_2)dp(x_1)dp(x_2)$$

$$= \int_0^1\int_0^1 p(x_1)p(x_2)B(p(x_1); \alpha, (N-1)\alpha)B\left(\frac{p(x_2)}{1-p(x_1)}; \alpha, (N-2)\alpha\right)dp(x_1)dp(x_2).$$

After some algebra, we find that $\langle p(x)p(x')\rangle = \frac{1}{N(N-1)}$ for $x \neq x'$, and so

$$\left\langle \left(\sum_x p(x)d(x,\tilde{x}) - \left\langle \sum_x p(x)d(x,\tilde{x})\right\rangle_{p(\vec{p}(x),\mathbf{d})}\right)^2 \right\rangle_{p(\vec{p}(x),\mathbf{d})}$$

$$= \left(1 - \frac{2}{N}\right)\langle d\rangle^2_{\rho(d)} + \left(1 - \frac{1}{N}\right)^2 \frac{\langle d^2\rangle_{\rho(d)}}{N\alpha + 1} - \left(1 - \frac{2}{N} + \frac{1}{N^2}\right)\langle d\rangle^2_{\rho(d)}$$

$$= \frac{1}{N\alpha + 1}\left(1 - \frac{1}{N}\right)^2 \langle d^2\rangle_{\rho(d)} - \frac{\langle d\rangle^2_{\rho(d)}}{N^2}.$$

In the large N limit, we have $\langle(\sum_x p(x)d(x,\tilde{x}) - \langle\sum_x p(x)d(x,\tilde{x})\rangle_{p(\vec{p}(x),\mathbf{d})})^2\rangle_{p(\vec{p}(x),\mathbf{d})} \sim \frac{\langle d^2\rangle_{\rho(d)}}{N\alpha}$. Chebyshev's inequality implies that $\sum_x p(x)d(x,\tilde{x})$ tends to $\langle d\rangle_{\rho(d)}$ in probability as $N \to \infty$ for all $\tilde{x}$, and so $D_{\max} = \min_{\tilde{x}} \sum_x p(x)d(x,\tilde{x})$ converges in probability to $\langle d\rangle_{\rho(d)}$ in probability as $N \to \infty$.

[1] H. B. Barlow, Possible principles underlying the transformations of sensory messages, in *Sensory Communications*, edited by W. A. Rosenblith (MIT Press, Cambridge, MA, 2012), Chap. 13.

[2] E. P. Simoncelli and B. A. Olshausen, Natural image statistics and neural representation, Annu. Rev. Neurosci. **24**, 1193 (2001).

[3] J. M. Salisbury and S. E. Palmer, Optimal prediction in the retina and natural motion statistics, J. Stat. Phys. **162**, 1309 (2016).

[4] C. G. Gross, Neurogenesis in the adult brain: Death of a dogma, Nat. Rev. Neurosci. **1**, 67 (2000).

[5] J. T. Paridaen and W. B. Huttner, Neurogenesis during development of the vertebrate central nervous system, EMBO Rep. **15**, 351 (2014).

[6] H. F. Green and Y. M. Nolan, Inflammation and the developing brain: Consequences for hippocampal neurogenesis and behavior, Neurosci. Biobehav. Rev. **40**, 20 (2014).

[7] I. Kirste, Z. Nicola, G. Kronenberg, T. L. Walker, R. C. Liu, and G. Kempermann, Is silence golden? Effects of auditory stimuli and their absence on adult hippocampal neurogenesis, Brain Struct. Funct. **220**, 1221 (2015).

[8] R. W. Williams and K. Herrup, The control of neuron number, Annu. Rev. Neurosci. **11**, 423 (1988).

[9] J. Kaslin, J. Ganz, and M. Brand, Proliferation, neurogenesis and regeneration in the non-mammalian vertebrate brain, Philos. Trans. R. Soc., B **363**, 101 (2008).

[10] F. Lallemend and P. Ernfors, Molecular interactions underlying the specification of sensory neurons, Trends Neurosci. **35**, 373 (2012).

[11] S. E. Marzen and S. DeDeo, The evolution of lossy compression, arXiv:1506.06138.

[12] S. B. Laughlin, Energy as a constraint on the coding and processing of sensory information, Curr. Opin. Neurobiol. **11**, 475 (2001).

[13] L. R. Varshney, P. J. Sjöström, and D. B. Chklovskii, Optimal information storage in noisy synapses under resource constraints, Neuron **52**, 409 (2006).

[14] V. Balasubramanian, D. Kimber, and M. J. Berry II, Metabolically efficient information processing, Neural Comput. **13**, 799 (2001).

[15] S. B. Laughlin, R. R. de Ruyter van Steveninck, and J. C. Anderson, The metabolic cost of neural information, Nat. Neurosci. **1**, 36 (1998).

[16] W. B. Levy and R. A. Baxter, Energy efficient neural codes, Neural Comput. **8**, 531 (1996).

[17] S. Schreiber, C. K. Machens, A. V. M. Herz, and S. B. Laughlin, Energy-efficient coding with discrete stochastic events, Neural Comput. **14**, 1323 (2002).

[18] D. B. Chklovskii and A. A. Koulakov, Maps in the brain: What can we learn from them? Annu. Rev. Neurosci. **27**, 369 (2004).

[19] L. Yu, C. Zhang, L. Liu, and Y. Yu, Energy-efficient population coding constrains network size of a neuronal array system, Sci. Rep. **6**, 19369 (2016).

[20] I. Nemenman, G. D. Lewen, W. Bialek, and R. R. de Ruyter van Steveninck, Neural coding of natural stimuli: Information at sub-millisecond resolution, PLoS Comput. Biol. **4**, e1000025 (2008).

[21] A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 1998).

[22] T. Sagawa and M. Ueda, Minimal Energy Cost for Thermodynamic Information Processing: Measurement and Information Erasure, Phys. Rev. Lett. **102**, 250602 (2009).

[23] D. H. Wolpert, The free energy requirements of biological organisms; implications for evolution, Entropy **18**, 138 (2016).

[24] T. Tlusty, Rate-Distortion Scenario for the Emergence and Evolution of Noisy Molecular Codes, Phys. Rev. Lett. **100**, 048101 (2008).

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley-Interscience, New York, 2006).

[26] B. A. Andrews and P. A. Iglesias, An information-theoretic characterization of the optimal gradient sensing response of cells, PLoS Comput. Biol. **3**, 1489 (2007).

[27] S. E. Palmer, O. Marre, M. J. Berry, and W. Bialek, Predictive information in a sensory population, Proc. Natl. Acad. Sci. U.S.A. **112**, 6908 (2015).

[28] C. A. Sims, The cost of misremembering: Inferring the loss function of visual working memory, J. Vision **15**, 1 (2015).

[29] S. Herculano-Houzel, Scaling of brain metabolism with a fixed energy budget per neuron: Implications for neuronal activity, plasticity and evolution, PLoS One **6**, e17514 (2011).

[30] "Strong universality" would imply that the rate-distortion function was equivalent for different distortion measures and probability distribution over inputs.

[31] T. Berger, *Rate Distortion Theory* (Prentice-Hall, New York, 1971).

[32] D. H. Wolpert and D. R. Wolf, Estimating functions of probability distributions from a finite set of samples, Phys. Rev. E **52**, 6841 (1995).

[33] M. C. Donaldson-Matasci, C. T. Bergstrom, and M. Lachmann, The fitness value of information, Oikos **119**, 219 (2010).

[34] J. H. Kaas, Plasticity of sensory and motor maps in adult mammals, Annu. Rev. Neurosci. **14**, 137 (1991).

[35] D. V. Buonomano and M. M. Merzenich, Cortical plasticity: From synapses to maps, Annu. Rev. Neurosci. **21**, 149 (1998).

[36] I. Nemenman, F. Shafee, and W. Bialek, Entropy and inference, revisited, *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA, 2001), pp. 471–478.