PNAS RESEARCH ARTICLE | NEUROSCIENCE
PSYCHOLOGICAL AND COGNITIVE SCIENCES OPEN ACCESS

Divisive normalization is an efficient code for multivariate Pareto-distributed environments

Stefan F. Bucher[a,b,c,1] and Adam M. Brandenburger[d,e,f]

Edited by Wilson Geisler, The University of Texas at Austin, Austin, Texas; received November 11, 2021; accepted July 11, 2022

Divisive normalization is a canonical computation in the brain, observed across neural systems, that is often considered to be an implementation of the efficient coding principle. We provide a theoretical result that makes the conditions under which divisive normalization is an efficient code analytically precise: We show that, in a low-noise regime, encoding an n-dimensional stimulus via divisive normalization is efficient if and only if its prevalence in the environment is described by a multivariate Pareto distribution. We generalize this multivariate analog of histogram equalization to allow for arbitrary metabolic costs of the representation, and show how different assumptions on costs are associated with different shapes of the distributions that divisive normalization efficiently encodes. Our result suggests that divisive normalization may have evolved to efficiently represent stimuli with Pareto distributions. We demonstrate that this efficiently encoded distribution is consistent with stylized features of naturalistic stimulus distributions such as their characteristic conditional variance dependence, and we provide empirical evidence suggesting that it may capture the statistics of filter responses to naturalistic images. Our theoretical finding also yields empirically testable predictions across sensory domains on how the divisive normalization parameters should be tuned to features of the input distribution.

divisive normalization | efficient coding | natural stimulus statistics | histogram equalization | Pareto distribution

Significance

Divisive normalization is a ubiquitous computation commonly thought to be an implementation of the efficient coding principle. Despite empirical evidence that it reduces statistical redundancy present in naturalistic stimuli, making the relationship between this neural code and the statistics of a stimulus precise has remained elusive. This paper closes this gap by providing a necessary and sufficient condition for divisive normalization to generate an efficient code. The multivariate Pareto distribution found to be efficiently encoded exhibits many stylized features of naturalistic stimulus statistics and provides testable predictions. In an empirical analysis, we find that the Pareto distribution captures the statistics of natural images well, suggesting that divisive normalization may have evolved to efficiently represent stimuli from such distributions.

The brain has to make efficient use of its limited resources to represent and respond to the wide range of stimuli in its environment. An important mechanism by which this can be achieved is divisive normalization (1, 2), which is thought to be a canonical computation in the brain (3). This gain control mechanism (according to which the response of a neuron to its preferred stimulus is suppressed by the intensity of nonpreferred stimuli) permits the representation of potentially unbounded stimuli by biophysically feasible bounded firing rates. Originally proposed for individual neurons in the primary visual cortex (1, 4, 5), this computation has since also been observed at the population level in the primary visual cortex (6–8) and throughout the visual hierarchy (9, 10), as well as in several other neural systems including olfactory pathways (11), the middle temporal area (12, 13), the inferotemporal cortex (14), the hippocampus (15), and in multisensory integration (16). In addition, divisive normalization has been shown to play an important role in value representations (17, 18) and for choice behavior, where it has been proposed to account for violations of the independence of irrelevant alternatives (IIA) axiom of rational choice (19–23; but see refs. 24 and 25). The nonlinear computation has also been suggested to play a role in attentional modulation (12, 26, 27), the modulation of response variability (28), the representation of visual uncertainty (29), and probabilistic inference (30, 31). It is further used in neural network models of the visual system (32, 33) as well as in computer vision and image compression (34).

This ubiquitous array of functions begs the question of what overarching objective the divisive normalization computation achieves. In this paper, we consider this computation's information-theoretic properties and provide testable conditions for its efficiency that are both simple and general, making them applicable across many of the aforementioned settings.

Since Schwartz and Simoncelli (35) showed empirically that divisive normalization reduces the statistical redundancy present in natural images, a common answer (36) has been that divisive normalization is an implementation of the efficient coding principle (37–41). This principle has been central to our understanding of the visual and other sensory systems (42–44), and it has also provided an account of biases in perception (45) and choice (46–49). Of course, divisive normalization has benefits beyond coding efficiency and redundancy reduction, such as permitting tuning curves that are invariant with respect to "nuisance" dimensions (e.g., maintaining discriminability of orientations regardless of contrast) or ensuring that population responses are easily decodable (e.g., by

Author contributions: S.F.B. designed research; S.F.B. and A.M.B. developed theoretical results; S.F.B. performed numerical and data analyses; and S.F.B. and A.M.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: stefan.bucher@nyu.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2120581119/-/DCSupplemental.

Published September 26, 2022.

a linear classifier or winner-take-all competition), among other features (3). Its widespread implementation in the nervous system may thus simultaneously achieve a number of purposes. Here, we focus on the question of whether divisive normalization is indeed an efficient computation, which arises naturally in both the sensory and choice domains.

Despite significant progress (50), an answer to this question in terms of testable conditions for efficiency has remained elusive, since formally relating neural computations to stimulus distributions has proved difficult: "The establishment of a precise quantitative relationship between environmental statistics and neural processing is important . . . [but] it has been surprisingly difficult to make the link quantitatively precise . . . [and] specification of a probability distribution over the space of input signals . . . is a difficult problem in its own right" (ref. 51, p. 1194). We close this gap with a theoretical result that makes precise the conditions on the input distribution under which divisive normalization encodes a stimulus efficiently.

Existing analytical work in the domain of vision has demonstrated that divisive normalization approximately (but not entirely) removes the statistical dependence in models of filter responses to natural images (52–54) such as the conditional normal (35) or lognormal (55) distributions. Moreover, divisive normalization can be viewed as an approximation of the nonlinear radial Gaussianization transformation that removes the statistical dependence of non-Gaussian elliptically symmetric distributions (56), yet divisive normalization itself can do so only imperfectly owing to its bounded range (57, 58). Lyu (50, 59) has quantified the extent to which divisive normalization reduces the statistical dependence of one such elliptical distribution: the multivariate Student's t distribution, which is in the class of Gaussian scale mixture models of natural images (60). He showed that even though divisive normalization approximates the transformation that eliminates this model's statistical dependencies, it can also increase them in low-dimensional settings.

This literature typically assumes a model of empirical stimulus statistics and derives the predictions of an (approximately) optimal code. It thus represents the first of two common approaches for testing the efficient coding hypothesis (51). The second approach is to examine the statistics of actual neural responses to naturalistic stimuli, in the spirit of Laughlin (40). Here, we pursue instead a third approach that consists of deriving analytically what stimulus distribution a given computation efficiently represents. This is in contrast to Malo and Laparra (54) or Lyu (50), for example, who use similar techniques but who start by assuming a given model of stimulus statistics. Instead, our approach is similar in spirit to that of Ballé et al. (61), who obtain a density model on images by inverting a generalized divisive normalization transform, except that we obtain the input density in analytical closed form. Without making any a priori assumptions about the stimulus statistics, the input distribution we find to be efficiently encoded captures many important features of naturalistic stimulus statistics, as we demonstrate in an analysis of image statistics. Our approach thus provides an additional perspective on the efficiency properties of divisive normalization.

We consider a setting in which an n-dimensional input is to be encoded by the divisively normalized firing rates of n neurons. The input can be either a stimulus or a representation of a stimulus coming from another neural system upstream. In the context of visual stimuli, the multivariate input could arise, for instance, as the responses of a population of linear filters convolved with the stimulus (35). At least two conditions have to be satisfied for the resulting multivariate representation to be efficient in a low-noise regime. First, it ought to adhere to histogram equalization (40) along each input dimension, which ensures that each output is used equally often. Second, maximizing the Shannon entropy of the output distribution requires—in the absence of constraints—that any statistical dependence across dimensions be removed. We use a formulation of the efficient coding principle that, in a low-noise regime, implies both of these desiderata and thus gives rise to a multivariate analog of the classic criterion of histogram equalization. Specifically, we consider a neural code to be efficient if and only if it maximizes the Shannon mutual information (62, 63) between the n-dimensional input and its representation. Since, for sufficiently small noise, this criterion can be approximated arbitrarily well by the requirement that the output distribution is entropy-maximizing, divisive normalization is then efficient whenever it transforms the input distribution into an output distribution that is uniform over the range of values divisive normalization can attain (39, 64).* This allows us to characterize the class of input distributions that are efficiently encoded in a low-noise regime.

We prove that divisive normalization maximizes the entropy of the output distribution if and only if the distribution of inputs in the environment is multivariate Pareto. This suggests that divisive normalization may have evolved as an efficient encoding strategy for heavy-tailed, scale-invariant power-law distributions of the kind that occur in many ecological contexts (66); see also *Discussion*.

The statistical dependence in the multivariate Pareto distribution is also consistent with the conditional variance dependence observed in natural image statistics (35), and it has a representation as a Gamma mixture of independent exponential random variables (providing a link to ref. 60). In an empirical analysis of naturalistic images, we demonstrate that the efficiently encoded Pareto distribution indeed captures the statistics of filter responses to natural images just as well as a common model of natural image statistics does. Divisive normalization may thus be an adaptation, in evolution or development, to various natural contexts with physical quantities whose distributions are characterized by heavy-tailed marginals and an empirically important form of statistical dependence.

We generalize our result by allowing for a representation to come at an arbitrary metabolic cost, which affects the shape of an efficient code (67, 68), and we show how this impacts the optimally encoded input distribution. For example, if costs are linear in the total number of spikes (which constrains the average firing rate), then the entropy-maximizing output distribution is exponential, and the associated input distribution changes accordingly. We provide necessary and sufficient conditions on the stimulus distribution for divisive normalization to be efficient under any member of a large family of cost functions.

Beyond providing a testable prediction on the shape of stimulus distributions that divisive normalization efficiently encodes, our theoretical result also yields empirically testable predictions across sensory domains on how the parameters of the divisive normalization transformation should be tuned to the parameters of the stimulus distribution (35). Specifically, the power index in the normalization function matches the shape parameter of the stimulus distribution, while the normalization weights are the inverses of the scale parameters of the stimulus distribution. Our theoretical predictions thus open the door to systematic experimental tests of the efficiency properties of empirically observed divisive normalization.

*Our method extends to other formalizations of the efficient coding principle that relax the small-noise assumption (65), since we provide an analytical formula relating any input distribution to the distribution of its representation.

Results

We consider a multivariate stimulus (or input from another neural system) modeled as an n-dimensional random vector $\boldsymbol{S} = (S_1, \ldots, S_n)$ taking values $\boldsymbol{s} \in \mathbb{R}^n$, where, for all i, $s_i > \mu_i$ for some $\mu_i \in \mathbb{R}$. The distribution of this stimulus is described by a continuous probability density function (pdf) $f_{\boldsymbol{S}}$. The support of this density is semi-infinite, bounded below by the constant $\boldsymbol{\mu} \in \mathbb{R}^n$, so that shifting the stimulus $\boldsymbol{s}$ by $\boldsymbol{\mu}$ gives rise to the positive input $\boldsymbol{x} = \boldsymbol{s} - \boldsymbol{\mu}$. This input is encoded by a population of n neurons whose (mean) firing rates are given by the divisive normalization function (3)

$$r_i(\boldsymbol{x}) = \gamma \frac{x_i^\alpha}{b^\alpha + \sum_{j=1}^n \lambda_j x_j^\alpha}, \qquad i = 1, \ldots, n, \qquad [1]$$

for $\boldsymbol{x} \geq \boldsymbol{0}$ and finite parameters $\gamma > 0$, $\alpha > 0$, $b > 0$, and $\lambda_j > 0$. The encoding $\boldsymbol{r} : \mathbb{R}_+^n \to \mathbb{R}_+^n$ gives rise to an n-dimensional representation. The parameter γ is often interpreted as the maximal firing rate, the parameter b as a semisaturation constant, and λ_j are normalization weights. One possible interpretation of α in the context of vision neuroscience is that x_i^α models the activation resulting from a linear filter response that is then normalized. While numerous functional forms have been proposed to describe divisive normalization (59, 69), our formulation generalizes the normalization equation in Carandini and Heeger (ref. 3, equation **10**) to include different weights λ_j in the normalization pool. For analytical tractability, we assume that these normalization weights can differ across input dimensions but not across output dimensions.

Constraints and Metabolic Costs. To assess the efficiency of the divisive normalization transform, we have to specify the class of permissible codes with which to compare it, as well as the efficiency criterion. We start by defining the former. Specifically, we compare divisive normalization with any encoding $\boldsymbol{g} : \mathbb{R}^n \to \Delta$ whose codomain is restricted to

$$\Delta \equiv \left\{ \boldsymbol{y} \in \mathbb{R}_+^n : \sum_{i=1}^n \lambda_i y_i < \gamma \right\} \qquad [2]$$

so that the output $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x})$ resulting from any feasible input $\boldsymbol{x}$ respects an upper bound γ on the $\boldsymbol{\lambda}$-weighted sum of firing rates $\boldsymbol{y}$. In addition to this constraint, we allow for a metabolic cost of a representation (70–72). We assume that a representation $\boldsymbol{y} \in \mathbb{R}_+^n$ comes at a cost $c(\boldsymbol{y})$, where c is an arbitrary cost function that is continuous and bounded on Δ.

The constraint is motivated by the fact that the range of values attained by divisive normalization is bounded by a linear constraint. The following proposition shows that Δ is exactly the range of values the divisive normalization function (Eq. **1**) attains:

Proposition 1. *The divisive normalization function $\boldsymbol{r}$ is invertible and its image is given by the simplex Δ.*

The proof of *Proposition 1* is based on an application of the matrix determinant lemma (e.g., ref. 73, theorem 18.1.1); it is given in *SI Appendix* along with all other proofs. Note that the bound γ also implies an upper bound γ/λ_i on the firing rate of each individual neuron, but not all neurons can simultaneously spike at their maximal firing rate. We take this constraint, which may arise from physiological limitations, as given and thus evaluate the efficiency of divisive normalization relative to the class of encodings that respect this same upper bound on the (weighted) sum of (nonnegative) firing rates.

Note that a bounded codomain is not only empirically plausible, but also mathematically necessary for considering "lossy compression," since an unbounded range of firing rates amounts to perfect coding capacity. Moreover, efficiency among all encodings respecting the constraint Δ is a necessary condition for efficiency among an even larger class of bounded representations with which divisive normalization could potentially be compared, because divisive normalization does adhere to the constraint Δ. Whether neural codes with image Δ are efficient even among those with more general representations is beyond the scope of this paper, but an interesting question for future research. Imposing suitable additional structure on the metabolic cost function, for example, would likely result in such a constraint.

Efficiency Criterion. To state the conditions under which a neural code is efficient, we employ a formalization of the efficient coding hypothesis whose criterion is based on the mutual information between the input $\boldsymbol{X}$ and its noisy representation

$$\tilde{\boldsymbol{Y}} = \boldsymbol{g}(\boldsymbol{X}) + \boldsymbol{\varepsilon},$$

where the additive noise $\boldsymbol{\varepsilon}$ is a random vector in $\mathbb{R}^n$ and is independent of $\boldsymbol{X}$. (The divisive normalization function $\boldsymbol{r}$ is one example of an encoding $\boldsymbol{g}$, so that the noise applies to the output of the transformation.)

We assume a low-noise regime in which the entropy $h(\boldsymbol{\varepsilon})$ of the noise is sufficiently small that the mutual information between the input and its noisy representation can be approximated by the entropy of the distribution of outputs $\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X})$.

Proposition 2. *Given independent random vectors $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}$ in $\mathbb{R}^n$ and a map $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^n$, the mutual information $I(\boldsymbol{X}; \boldsymbol{g}(\boldsymbol{X}) + \boldsymbol{\varepsilon})$ can be approximated arbitrarily closely by the entropy $h(\boldsymbol{g}(\boldsymbol{X}))$, given sufficiently small entropy $h(\boldsymbol{\varepsilon})$.*

According to this result, the efficient coding hypothesis amounts to maximization of the entropy of the output distribution (74, 75) net of metabolic costs.[†] We thus consider a stimulus encoding $\boldsymbol{g} : \mathbb{R}^n \to \Delta$ to be efficient if the resulting output distribution maximizes, among all distributions with support equal to Δ, the entropy of $\boldsymbol{Y}$ net of the expected cost $\mathbb{E}[c(\boldsymbol{Y})]$.

Characterizing the Optimal Output Distribution. Ignoring metabolic costs, our efficiency criterion amounts to a multivariate version of histogram equalization and requires that all feasible outputs occur equally often. Among all distributions attaining the same range of values as divisive normalization, the entropy-maximizing distribution of mean firing rates is uniform over its simplex support Δ. Note that, constrained to such a support, the uniform distribution is not statistically independent across dimensions (as would be the case in absence of constraints), yet this distribution does result from the optimal encoding.

The following proposition (cf. 76) characterizes the optimal output distribution taking into account any metabolic costs, allowing for potentially more empirically realistic firing-rate distributions.

Proposition 3. *Fix a random vector $\boldsymbol{Y}$ (the representation) with bounded support $\mathcal{C}$ in $\mathbb{R}_+^n$ and a continuous and bounded cost function $c : \mathcal{C} \to \mathbb{R}_+$. The pdf of the distribution that maximizes the entropy of $\boldsymbol{Y}$ net of the expected cost, i.e., that maximizes*

[†] Other, more general approximations are possible (65) but, for mathematical simplicity, are not considered here. However, *Theorem 1* is general and could be combined with variants of *Proposition 2*.

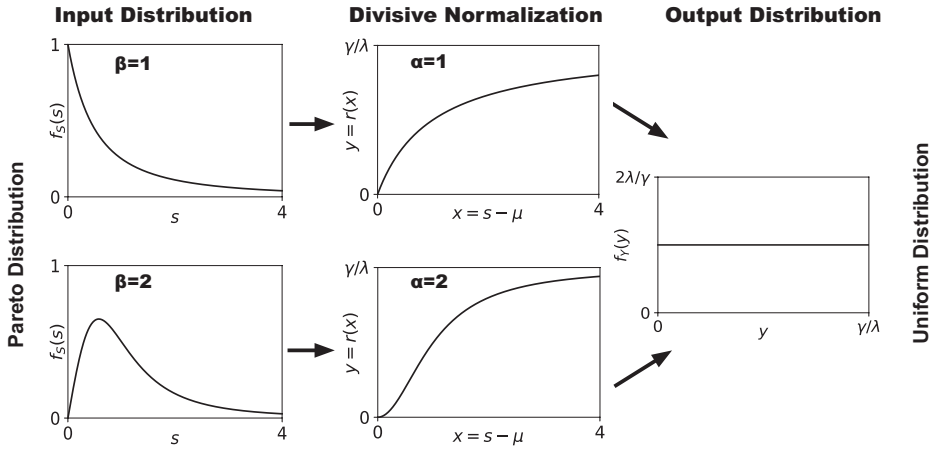

Fig. 1. Univariate example of how divisive normalization transforms a stimulus distribution into a distribution over outputs. If a stimulus s follows a Pareto distribution (*Left* column), with shape parameter β (and $\mu = 0$, $\sigma = 1$), the divisive normalization transform of $x = s$ (*Center* column), with appropriately chosen parameter α, ensures that the resulting representation $y = r(x)$ follows the entropy-maximizing uniform distribution (*Right* column).

$h(\boldsymbol{Y}) - \mathbb{E}\left[c\left(\boldsymbol{Y}\right)\right]$, *among all distributions with support* $\mathcal{C}$, *is given by*

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{e^{-c(\boldsymbol{y})}}{\int_{\mathcal{C}} e^{-c(\boldsymbol{z})}\, d\boldsymbol{z}} \qquad [3]$$

for $\boldsymbol{y}$ *in* $\mathcal{C}$.

For example, assuming that the cost of a representation is linear in the sum of all firing rates gives rise to a truncated exponential output distribution $f_{\boldsymbol{Y}}(\boldsymbol{y}) \propto \mathbf{1}_{\{y\in\mathcal{C}\}} \exp\left(-\kappa \sum_{i=1}^{n} y_i\right)$, which is in line with empirically observed firing rates in response to natural scenes (77, 78). Using high firing rates less frequently, it optimally balances the informational benefit of a wide range of firing rates with the cost of relying on metabolically costly high firing rates. Other special cases of interest include costs that take the quadratic form $c(\boldsymbol{y}) = \kappa \boldsymbol{y} \cdot \boldsymbol{y}$, in which case $f_{\boldsymbol{Y}}(\boldsymbol{y})$ is truncated normal on $\mathcal{C}$, and costs that are constant in $\boldsymbol{y}$, in which case $f_{\boldsymbol{Y}}(\boldsymbol{y})$ is uniform on $\mathcal{C}$.

Characterizing the Efficiently Encoded Input Distribution. By mapping inputs $\boldsymbol{x}$ to their representations $\boldsymbol{r}(\boldsymbol{x})$, divisive normalization transforms any stimulus distribution $f_{\boldsymbol{S}}$ into a corresponding distribution of outputs or firing rates. Our main mathematical result describes this transformation by relating any stimulus distribution to the output distribution that results under divisive normalization and vice versa.

Theorem 1. *Let* $\boldsymbol{X}$ *in* $\mathbb{R}^n_+$ *be a random vector with continuous pdf* $f_{\boldsymbol{X}}$ *and* $\boldsymbol{r} : \mathbb{R}^n_+ \to \Delta$ *be the divisive normalization mapping. Let* $\boldsymbol{Y} = \boldsymbol{r}(\boldsymbol{X})$ *and* $\boldsymbol{y} = (y_1, \ldots, y_n) = \boldsymbol{r}(\boldsymbol{x})$ *denote the output of the divisive normalization transformation, with pdf* $f_{\boldsymbol{Y}}$. *Then the two pdfs satisfy, for any* $\boldsymbol{x} \in \mathbb{R}^n_{++}$,

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \gamma^n \alpha^n \frac{b^\alpha \prod_{i=1}^{n} x_i^{\alpha-1}}{\left(b^\alpha + \sum_{i=1}^{n} \lambda_i x_i^\alpha\right)^{n+1}} \times f_{\boldsymbol{Y}}(\boldsymbol{y}). \qquad [4]$$

The proof uses the change-of-variables formula for random vectors (e.g., ref. 79, theorem 8.1.7) and again relies on the matrix determinant lemma to compute the determinant of the Jacobian of $\boldsymbol{r}$ in closed analytical form.

Given *Proposition 2* and the shape of the optimal output distribution provided by *Proposition 3*, *Theorem 1* lets us characterize, for any metabolic cost function, the stimulus distribution for which divisive normalization is an efficient encoding in a low-noise regime.

Fig. 1 illustrates this for the univariate special case ($n = 1$) absent metabolic costs. Histogram equalization on $[0, \gamma/\lambda]$ requires that the output distribution $f_Y(y)$ is uniform over this range. A univariate Pareto distribution with pdf $f_X(x) = \beta x^{\beta-1}/(1+x^\beta)^2$ is thus mapped into an entropy-maximizing uniform distribution by a transformation whose derivative is $\gamma\alpha\lambda x^{\alpha-1}/\left(\lambda(1+x^\alpha)\right)^2$ as in *Theorem 1* (with $\alpha = \beta$ and $\lambda = b^\alpha$), which indeed integrates to the univariate divisive normalization formula $\gamma x^\alpha/(b^\alpha + \lambda x^\alpha)$.

For the general case, the following result states the condition on the stimulus distribution for divisive normalization to maximize, among all distributions with support equal to Δ, the entropy of the resulting output distribution net of expected costs:

Theorem 2. *Divisive normalization of* $\boldsymbol{x} = \boldsymbol{s} - \boldsymbol{\mu}$ *is an efficient encoding of a stimulus* $\boldsymbol{S}$ *for a cost function* c *if and only if* 1) *the stimulus distribution has joint pdf, for* $\boldsymbol{s} > \boldsymbol{\mu}$,

$$\begin{aligned} &f_{\boldsymbol{S}}(s_1, \ldots, s_n; \boldsymbol{\mu}, \boldsymbol{\sigma}, \beta, \gamma, b, c) \\ &\quad = \gamma^n \beta^n \frac{\prod_{i=1}^{n} \left(s_i - \mu_i\right)^{\beta-1} / b^\beta}{\left(1 + \sum_{i=1}^{n} \left(\frac{s_i - \mu_i}{\sigma_i}\right)^\beta\right)^{n+1}} \\ &\qquad \times \frac{e^{-c(\boldsymbol{r}(\boldsymbol{s}-\boldsymbol{\mu}))}}{\int_\Delta e^{-c(\boldsymbol{z})}\, d\boldsymbol{z}}, \end{aligned} \qquad [5]$$

and 2) *the parameters satisfy* $\alpha = \beta$ *and* $\lambda_i = (b/\sigma_i)^\alpha$.

Note that the efficiently encoded distribution depends on the metabolic cost function c and also on the parameters of the divisive normalization transform. The efficiently encoded input distribution shares the parameters γ and b with the divisive normalization transform. Additionally, efficiency requires that the exponent α of the divisive normalization transform matches the shape parameter $\beta > 0$ of the input distribution and that the normalization weights λ_i are inversely proportional to the scale parameters $\sigma_i > 0$. This reflects the fact that scaling inputs requires adjusting normalization weights to maintain the efficiency condition of *Theorem 1*. Of course, other parameterizations of this distribution are possible, but the scale parameters σ_i reflect the extent to which λ_i and b are exchangeable, and they will be helpful below.

Fig. 2 illustrates *Theorem 2* for the bivariate case and metabolic costs that are constant (Fig. 2*A*) or linear (Fig. 2*B*) in the sum of the firing rates of all neurons. Note that the input distribution

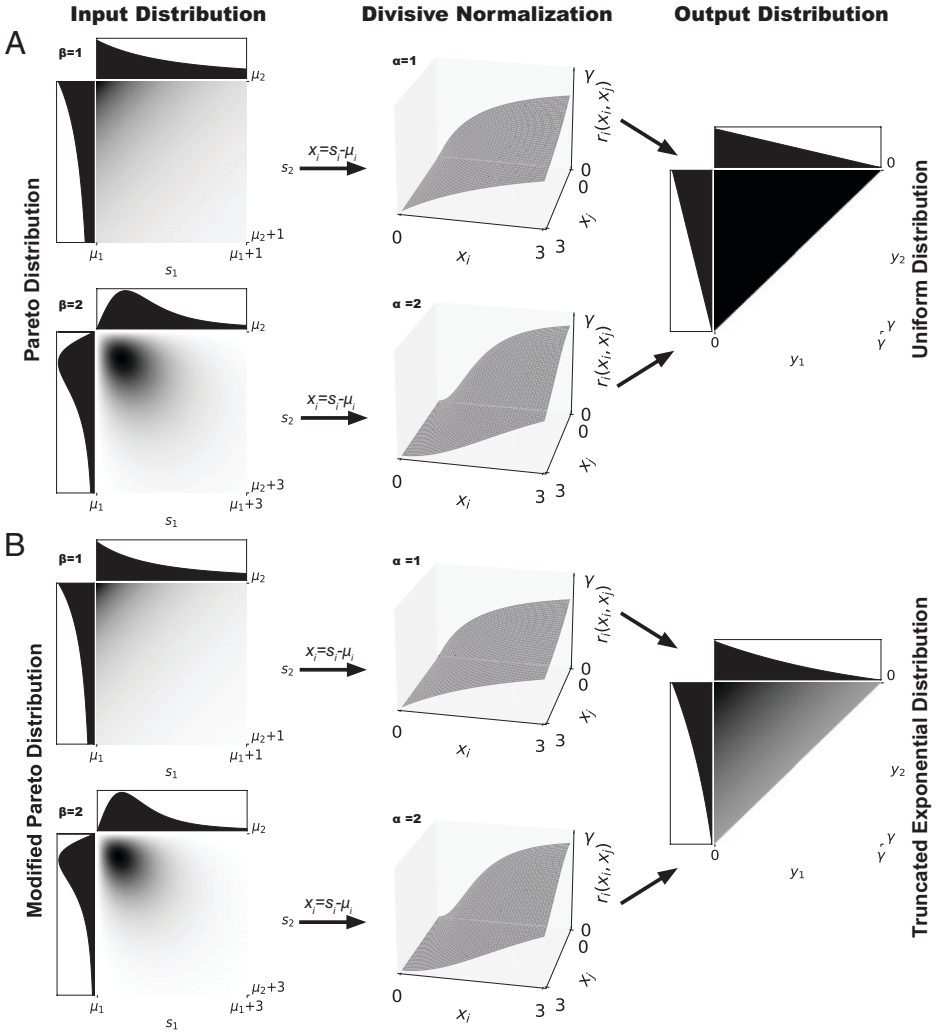

Fig. 2. Divisive normalization transforms the distribution of stimulus s into a distribution over outputs $\boldsymbol{y} = \boldsymbol{r}(\boldsymbol{x})$. In this bivariate example with $\boldsymbol{\sigma} = \boldsymbol{\lambda} = \mathbf{1}$, $b = 1$, and two values of α and β, joint probability densities are plotted with darker color representing higher density. Marginal densities are shown in the adjacent plots. The Pareto distributions in *A* are transformed into the uniform distribution over the simplex Δ (black triangle) that is efficient absent metabolic costs (*Theorem 3*). *B* shows the qualitatively similar stimulus distributions that are transformed into the truncated exponential distribution that is efficient under linear metabolic costs (*Theorem 2*).

that divisive normalization efficiently encodes has a higher density for small values for linear metabolic costs than it does absent metabolic costs. This reflects the fact that, in the presence of metabolic costs, higher firing rates should be used less frequently.

Constant Metabolic Costs. An important special case of interest is constant metabolic costs, since the problem then reduces to maximizing information transmission, which is a common formulation of the efficient coding hypothesis. *Theorem 3* states the result for this special case, which makes a particularly sharp prediction that can be viewed as a benchmark generalized by *Theorem 2*. Moreover, Fig. 2 demonstrates that the qualitative shape of the efficiently encoded stimulus distribution appears to be robust to the exact assumption on metabolic costs. Recall that the joint survival (or complementary cumulative distribution) function $\bar{F}_{\boldsymbol{S}}(\boldsymbol{s})$ of a random vector $\boldsymbol{S}$ is defined as

$$\bar{F}_{\boldsymbol{S}}(\boldsymbol{s}) = \mathbb{P}_{\boldsymbol{S}}(S_1 > s_1, \ldots, S_n > s_n).$$

Theorem 3. *Divisive normalization of $\boldsymbol{x} = \boldsymbol{s} - \boldsymbol{\mu}$ is an efficient encoding of a stimulus $\boldsymbol{S}$ under constant metabolic costs if and only if* 1) *the stimulus distribution is a multivariate Pareto type III distribution with joint survival function*

$$\bar{F}_{\boldsymbol{S}}(s_1, \ldots, s_n; \boldsymbol{\mu}, \boldsymbol{\sigma}, \beta) = \left[1 + \sum_{i=1}^{n} \left(\frac{s_i - \mu_i}{\sigma_i}\right)^{\beta}\right]^{-1} \quad [6]$$

and joint pdf

$$f_{\boldsymbol{S}}(s_1, \ldots, s_n; \boldsymbol{\mu}, \boldsymbol{\sigma}, \beta) = \beta^n \frac{n! \prod_{i=1}^{n} \frac{1}{\sigma_i} \left(\frac{s_i - \mu_i}{\sigma_i}\right)^{\beta-1}}{\left(1 + \sum_{i=1}^{n} \left(\frac{s_i - \mu_i}{\sigma_i}\right)^{\beta}\right)^{n+1}} \quad [7]$$

for $\boldsymbol{s} > \boldsymbol{\mu}$, and 2) *the parameters satisfy $\alpha = \beta$ and $\lambda_i = (b/\sigma_i)^{\alpha}$.*

Theorem 3 is a corollary of *Theorem 2* for the special case of constant (or zero) metabolic costs. According to *Theorem 3*, which is illustrated for the bivariate case in Fig. 2*A*, the efficiently encoded stimulus distribution is a particular multivariate Pareto type III distribution (80, 81).[‡] Note that this particular Pareto type III distribution is parameterized by a homogeneous shape parameter $\beta > 0$ along with location parameters μ_i and scale parameters

[‡]The minus sign in ref. 80, equation 6.1.17 appears to be a typo.

$\sigma_i > 0$. As above, efficiency requires that the parameters λ_i and α of the normalization formula should be tuned to the stimulus distribution as follows: The exponent α should be set to the shape parameter β of the distribution, and the normalization weights $\lambda_i = (b/\sigma_i)^\alpha$ should be inversely proportional to the (transformed) scales of the distribution. Note that Eq. **7** is obtained from Eq. **5** by imposing constant metabolic costs and evaluating the resulting integral in the denominator, which results in the cancellation of γ and b (see proof in *SI Appendix*). Under constant metabolic costs this leaves γ and b as free parameters. As mentioned above, the parameter γ can be interpreted as the constraint on the sum of firing rates, and b is a semisaturation constant.

The joint and conditional densities of this Pareto type III distribution are plotted in *SI Appendix* for a range of parameter values. Its marginal distribution is a univariate Pareto type III with cumulative distribution function (cdf)

$$F_{S_i}(s_i; \mu_i, \sigma_i, \beta) = \frac{1}{1 + \left(\frac{s_i - \mu_i}{\sigma_i}\right)^{-\beta}} \quad [8]$$

for $s_i > \mu_i$, whose mean is, for $\beta > 1$, given by

$$\mathbb{E}[S_i] = \mu_i + \sigma_i \frac{\pi/\beta}{\sin(\pi/\beta)}. \quad [9]$$

SI Appendix contains expressions for the variance and further moments. The marginal Pareto distribution is heavy tailed and for $\mu_i = \sigma_i = \beta = 1$ it is an exact power law

$$F_{S_i}(s_i; \mu_i = 1, \sigma_i = 1, \beta = 1) = 1 - 1/s_i \quad [10]$$

for $s_i \in [1, \infty)$. This observation is interesting given that many naturally occurring quantities exhibit approximate power-law characteristics (82–84). For $\mu_i = 0$, the marginal distribution is a univariate log-logistic distribution, sometimes also referred to as a Fisk (85) distribution.

Empirical Analysis of Natural Stimulus Statistics. If divisive normalization has evolved as an efficient computation, our result suggests that it may have adapted to environments whose stimulus statistics are well described by multivariate Pareto (type III) distributions. We examine this hypothesis in the visual domain where there is considerable empirical evidence on stimulus statistics to which we can relate our result. In particular, we ask whether the Pareto distribution we have found to be efficiently encoded exhibits the kind of statistical dependence that is commonly found in filter responses to natural images (35, 51, 56, 86). We first note that the pairwise covariance of the efficiently encoded Pareto distribution (Eq. **6**) is, for $\beta > 2$ and $i \neq j$, given by ref. 80, equation 6.1.29:

$$\mathrm{Cov}(S_i, S_j) = \sigma_i \sigma_j \left(\Gamma\left(\frac{\beta+1}{\beta}\right)\right)^2 \left(\Gamma\left(\frac{\beta-2}{\beta}\right) - \left(\Gamma\left(\frac{\beta-1}{\beta}\right)\right)^2\right), \quad [11]$$

where Γ is the gamma function.

To assess whether the Pareto distribution and its correlation structure are an empirically relevant model of natural image statistics, we performed a simple exploratory analysis (Fig. 3 *A–D*) using images from the van Hateren image dataset (87). We obtained the joint histogram of the responses of a pair of filters differing in orientation (shown in Fig. 3*C* for the example image of Fig. 3*B*), as well as the corresponding conditional histogram (Fig. 3*D*). The characteristic "bow-tie" shape of the conditional histogram is an empirical regularity observed in numerous naturalistic stimuli (35, 86).

Using the filter response data, we obtained maximum-likelihood estimates, for each image separately, of the parameters of a bivariate Pareto distribution (extended to $\mathbb{R}^2$; see *Materials and Methods*). For comparison, we also fitted a bivariate t-distribution (with the same number of parameters), as has been used to model natural image statistics (59, 88–90). Fig. 3*A* shows the distribution of the resulting log-likelihoods. The fact that the log-likelihoods of the Pareto model are greater than those of the multivariate t model (means: -5.19×10^6 for Pareto and -5.65×10^6 for t-distribution) suggests that the Pareto distribution is a strong contender as a model of natural image statistics. However, we do not consider our analysis to be definitive, and further research is required to determine how well the Pareto distribution describes image statistics quantitatively and qualitatively. But this analysis does demonstrate that the Pareto distribution, which we derived from first principles, is also likely to be an empirically relevant description of natural stimulus statistics.

Fig. 4 illustrates why the Pareto distribution may describe natural stimulus statistics well. The depicted conditional histogram of a bivariate Pareto type III distribution with $\beta = 1$ (extended to $\mathbb{R}^2$) closely matches the empirically observed statistical dependence in filter responses to natural images. We conclude that the efficiently encoded Pareto distribution exhibits key features of naturalistic stimulus distributions, including this kind of statistical dependence as well as heavy-tailed marginal distributions. We further note that the average estimate for β was 1.2 (*SI Appendix*, Fig. S1), which is comparable to estimates of the divisive normalization exponent (α) from neural data (6). This may suggest that the divisive normalization parameter α is indeed tuned to the shape parameter β of the Pareto distribution.

Relation to Existing Models of Natural Stimulus Statistics. This section explores the connections to existing models of natural stimulus statistics, with the goal of providing a foundation for future research. We first show how the statistical dependence of the Pareto distribution can capture the conditional variance dependence that is commonly observed in bow-tie plots (86). To see this, assume that $S_1, \ldots, S_n$ are filter responses that are distributed according to a multivariate Pareto type III distribution. The variance of a filter response S_i conditional on all other filter responses is then, for $\beta = 1$ (assumed for tractability) and $n > 2$, given by

$$\begin{aligned}\mathrm{Var}(S_i | \{S_j = s_j\}_{j \neq i}) &= \frac{\sigma_i^2 n}{(n-1)^2(n-2)} \left[1 + \sum_{j \neq i} \left(\frac{s_j - \mu_j}{\sigma_j}\right)\right]^2 \\ &= \frac{\sigma_i^2 n}{(n-1)^2(n-2)} \left[1 + 2\sum_{j \neq i} \frac{s_j - \mu_j}{\sigma_j} + \sum_{j \neq i} \sum_{k \neq i, k \neq j} \frac{s_j - \mu_j}{\sigma_j} \frac{s_k - \mu_k}{\sigma_k} + \sum_{j \neq i} \left(\frac{s_j - \mu_j}{\sigma_j}\right)^2\right]. \end{aligned} \quad [12]$$

The conditional variance dependence is thus quadratic, in accordance with Schwartz and Simoncelli (35) who use a similar quadratic model of the variance dependence observed in conditional histograms. Unlike in their model, the conditional

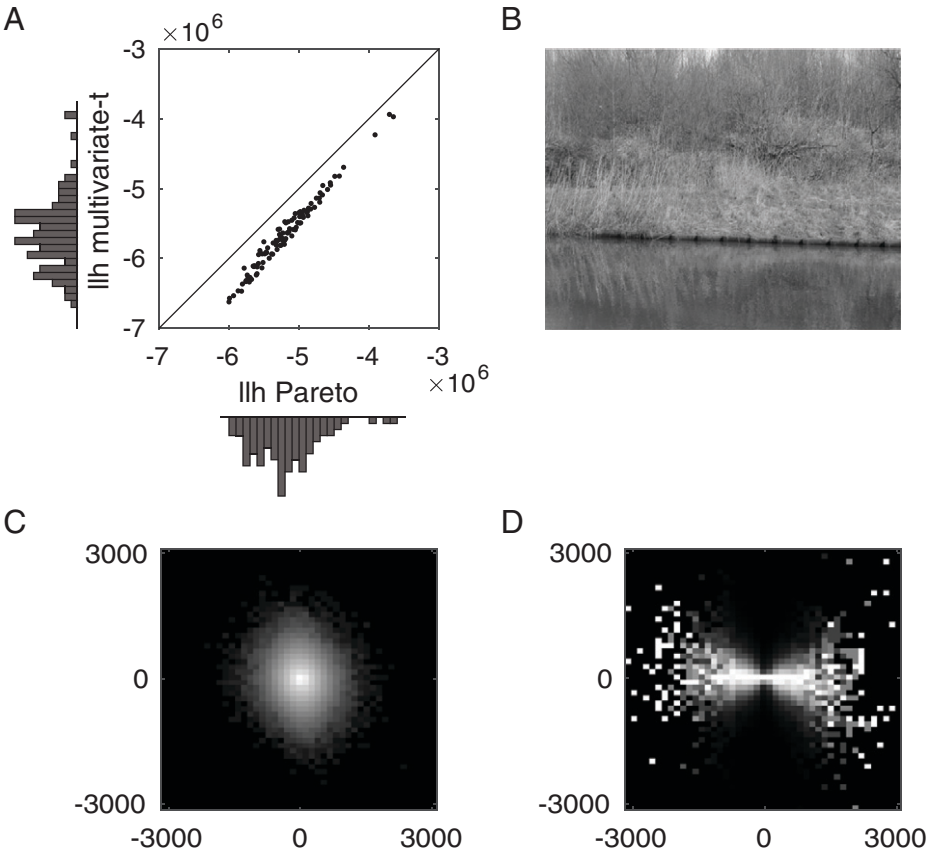

Fig. 3. Fitting natural stimulus statistics (see *Materials and Methods* for details). (*A*) Scatterplot showing the log-likelihood (llh) of the best-fitting Pareto (*x* axis) and multivariate-*t* (*y* axis) models to the statistics of naturalistic images from the van Hateren dataset (87). Each dot corresponds to one image and the histograms show the corresponding marginal distributions. (*B*) An (additional) example image from the van Hateren dataset, with log-transformed brightness for better visibility. (*C*) Joint histogram for the example image of the responses of two filters with orientations of 45° and 135° counterclockwise, respectively (brightness on log scale). (*D*) Corresponding conditional histogram of the same filter responses, showing the distribution of responses of one filter (*y* axis) conditional on the response of the other filter (*x* axis). The brightness is rescaled in each column to use the full range of intensities, in line with the literature (35).

distribution giving rise to this variance dependence is not Gaussian. The conditional distribution of the Pareto distribution in Eq. **6** has cdf

$$F_{S_i|\{S_j=s_j\}_{j\neq i}}(s_i;\{s_j\}_{j\neq i},\boldsymbol{\mu},\boldsymbol{\sigma},\beta) = 1-\left[1+\frac{\left(\frac{s_i-\mu_i}{\sigma_i}\right)^{\beta}}{1+\sum_{j\neq i}\left(\frac{s_j-\mu_j}{\sigma_j}\right)^{\beta}}\right]^{-n} \quad [13]$$

for $s_i > \mu_i$, as we show in *SI Appendix*.[§] For suitable parameter values this reduces to a log-logistic distribution, closely related to the conditionally lognormal distribution used by Wainwright et al. (55).

Next, we show that the Pareto distribution also has a connection with models of image statistics based on Gamma-weighted scale mixtures of Gaussian random variables (60). It turns out that the Pareto type III distribution can be expressed as a Gamma mixture of transformed exponential (or Weibull) random variables (ref. 80, chap. 6.2). The following proposition shows that the efficiently coded distribution of Eq. **6** is a particularly simple mixture of transformed standard exponential random variables:

Proposition 4. *Let* $U_i \overset{iid}{\sim} Exp(\lambda=1)$ *for* $i=1,\ldots,n$, *and let* $Z \sim Exp(\lambda=1)$ *independently of all* U_i. *Then the distribution of* $\boldsymbol{S}=(S_1,\ldots,S_n)$ *with*

$$S_i = \mu_i + \sigma_i\,(U_i/Z)^{1/\beta} \qquad \text{for } i=1,\ldots,n, \quad [14]$$

is a Pareto type III distribution with the joint survival function $\bar{F}_{\boldsymbol{S}}(s_1,\ldots,s_n;\boldsymbol{\mu},\boldsymbol{\sigma},\beta)$ *of Eq.* **6**.

This result not only facilitates comparisons with existing models of naturalistic stimuli, it is also of practical importance, since it provides a simple way to draw samples from the multivariate Pareto type III distribution.

Discussion

We have characterized the family of stimulus distributions for which divisive normalization maximizes mutual information in the low-noise limit. Absent metabolic costs, this family consists of particular multivariate Pareto type III distributions, which divisive normalization transforms into an entropy-maximizing output distribution. Taking into account metabolic costs of an arbitrary shape, the efficiently encoded input distributions take a generalized form in which inputs resulting in costlier outputs occur less frequently. We note that our result does not imply that divisive normalization is the only neural encoding to satisfy the necessary and sufficient condition for efficiency, even with a Pareto stimulus distribution. Rather, any encoding whose Jacobian

[§]Note that the conditional distribution is a univariate Pareto distribution of type IV (rather than III). Pareto type III distributions are not closed under conditioning.

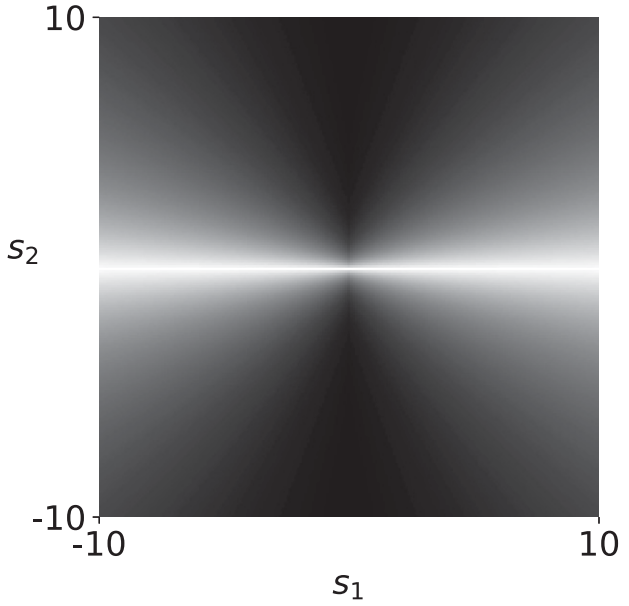

Fig. 4. Conditional histogram of a bivariate Pareto distribution extended to $\mathbb{R}^2$, with $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\sigma} = \mathbf{1}$, and $\beta = 1$. Brightness is proportional to the probability of s_2 conditional on s_1, rescaled in each column to use the full range of intensities, as is customary in the literature. The statistical dependence closely resembles the bow-tie shape observed empirically for many naturalistic stimuli (35).

satisfies, given some input distribution, the equivalent of Eq. **4** in *Theorem 1* attains the same output distribution and will thus be equally efficient. Yet for divisive normalization to be efficient, the stimulus distribution must necessarily be of the prescribed type.

We further demonstrated that the Pareto distribution is consistent with empirical findings on naturalistic stimulus statistics. In the context of vision, the Pareto distribution captures key features observed in the statistics of natural images that are commonly modeled with Gaussian scale mixtures (60, 91). Our empirical analysis demonstrated that the Pareto distribution may be a similarly good model of natural-image statistics. *Proposition 4* identifies the Pareto distribution as a mixture of transformed exponential random variables and may thus prove helpful in further exploring relations to existing Gaussian mixture models of natural stimuli. While sharing some features with such models, the Pareto distribution differs in other respects. For example, unlike Gaussian scale mixtures, it is not elliptically symmetric. This means that it can be skewed but is not invariant to rotations of the coordinate system and may thus capture joint histograms resembling a diamond or rhombus rather than an ellipse. Furthermore, the Pareto distribution also captures features of typical luminance and contrast distributions (92). We leave exploring these issues for future research.

Proposition 4 also opens the door to future work exploring more general coding mechanisms that are efficient for a wider class of input distributions. The distribution we find to be efficiently encoded is a special case of a Pareto type III distribution that is restrictive in the sense that a single shape parameter β governs the distribution's dependence structure as well as the shape of its marginals. *Proposition 4* suggests that another fruitful direction, beyond considering more general Pareto distributions, may be to consider mixtures of an exponential random vector whose components are not independent, allowing for a more general covariance structure.

It would also be desirable to generalize our result to allow normalization weights to differ across pairs of neurons. Ruling out the case where different neurons in a population interact with different weights eliminates the possibility that neurons representing more proximal inputs normalize each other more strongly than those representing more distant inputs. Such distance-dependent normalization would be an efficient code of stimuli whose statistics reflect a notion of proximity. For example, this is the case for contrast in images, where correlation is decreasing as a function of distance. Our current framework does not incorporate such a distance notion for reasons of analytical tractability, but an extension would be of interest.

Our results provide a sharp testable prediction that can potentially be used to test empirically the efficient coding hypothesis. In a neural system whose response function is well described by divisive normalization, testing whether the inputs are Pareto distributed amounts to testing a necessary condition for the efficient coding hypothesis to hold. A more rigorous and demanding test would also estimate the normalization parameters as well as the parameters of the distribution and test whether they satisfy the conditions imposed by *Theorems 2* and *3*. Formally, if the efficient coding hypothesis holds, then the hypothesis that such a system's inputs are Pareto distributed with $\alpha = \beta$ and $\lambda_i = (b/\sigma_i)^\alpha$ must be true (under the maintained hypothesis of constant metabolic costs). Rejecting this null hypothesis would thus result in the rejection of the efficient coding hypothesis. Experiments that systematically manipulate the distribution of sensory stimuli (or choice sets) in a subject's environment could even examine whether the divisive normalization parameters (such as α) optimally adapt to different stimulus distribution contexts (e.g., Pareto distributions with different values for β). Of course, less demanding tests are possible and informative.

It is important to stress that according to *Theorem 2*, testing the input distribution is a test of the joint hypothesis that coding is efficient and metabolic costs are of a particular form. For instance, Pareto-distributed inputs are only necessary for the efficient coding hypothesis under the maintained assumption of constant metabolic costs. This is a feature, not a bug. *Theorem 2* can be viewed as a representation theorem, analogous to the economics tradition of testing whether choice behavior is consistent with maximizing some utility function. This provides additional degrees of freedom: If the input distribution adheres to Eq. **5** for some metabolic cost function, then the efficient coding hypothesis would not be rejected.

Our work raises the possibility that divisive normalization is ubiquitous because it is an adaptation to Pareto distributions that are themselves widespread—in stimulus statistics but also, perhaps, in the statistics of other quantities. One example of such a quantity is firing rates of neurons, which in many neural systems appear to follow lognormal distributions (93, 94, 95), whose heavy-tailed shape closely resembles that of Pareto distributions. If firing rates of an upstream system are approximately Pareto distributed, it is conceivable that divisive normalization may be an adaptation to the firing-rate distribution of neurons from which it receives input. But, of course, the resulting outputs would then not be Pareto distributed, raising the question of whether the observed cascade of normalization-like modules in the brain is to some extent redundant. If each stage in a series of divisive normalization levels in the brain receives the preceding stage's output as its input, then each stage faces a different distributional structure. Understanding how several stages of divisive normalization may work together to produce an efficient code is an open issue that is beyond the scope of this paper. Perhaps, applying our result to distributions of firing rates will help shed light on such issues.

The fact that divisive normalization has been implicated in settings beyond sensory processing and perception, particularly in value representations (17, 18) and choice (19, 21), raises the

question of whether the Pareto distribution is also an ecologically relevant description of value-based choice environments. The univariate Pareto distribution of Eq. **8** (with $\mu = 0$) is often used to model wealth and income distributions in economics, which is where it originated (96, 97). Moreover, the Pareto distribution is intimately related to Zipf's law and is observed in many complex systems (82). Interestingly, it is also remarkably consistent (98) with Benford's law (99, 100), according to which the leading digits in many naturally arising sets of numbers are likely to be small.

The heavy-tailed power-law characteristic of the Pareto distribution is relevant not only in sensory contexts, but also in many economic contexts (101, 102) ranging from city-size distributions to stock returns and trading volumes (103). While empirical evidence from choice environments is lacking, the widespread occurrence of Pareto distributions in economic contexts hints that divisive normalization might be observed in value representations and choice behavior because it is an efficient code of Pareto-distributed values within choice sets. An important caveat to this hypothesis is that mutual information is less relevant as an objective function for a decision maker than it is in the sensory domain. The efficient code for a utility-maximizing decision maker differs from the information-maximizing code (104–106). Conditions under which divisive normalization is efficient for choice remain to be determined.

In conclusion, our theoretical results make a simple and sharp prediction about why divisive normalization may be observed across a wide range of settings. This prediction can be tested in different domains, ranging from electrophysiological studies to empirical studies of value distributions and experiments on choice behavior. Our findings provide a framework for future research to test empirically and experimentally whether divisive normalization occurs in environments in which it is an efficient computation.

Materials and Methods

All proofs are given in *SI Appendix*. The empirical analysis examined the statistics of filter responses to naturalistic images. For a set of 100 images from the (linear) van Hateren image dataset (87), we computed filter responses using a steerable pyramid (107) with four levels and four bands differing in orientation, using the matlabPyrTools package. (See https://github.com/LabForComputationalVision/matlabPyrTools and www.cns.nyu.edu/∼eero/STEERPYR/ for more information.) We examined the statistical dependency between a pair of filters of the pyramid's second level with orientations of 45° and 135° counterclockwise, respectively (108, figure 1.9). For each image separately, we fitted the joint distribution of these filter responses by obtaining maximum-likelihood estimates of the parameters of two bivariate distributions. The first distribution is a Pareto type III distribution extended to $\mathbb{R}^2$, in the sense that the density at any $\boldsymbol{s} \in \mathbb{R}^2$ is given by $\frac{1}{4} f_{\boldsymbol{S}}(|\boldsymbol{s}|; \boldsymbol{\mu} = \boldsymbol{0}, \boldsymbol{\sigma}, \beta)$, where $f_{\boldsymbol{S}}$ is as in Eq. **7**. We imposed the restriction $\sigma_1 = \sigma_2$ to ensure that the number of free parameters is not larger than for the second distribution, which was a bivariate *t*-distribution with correlation parameter *c* and degrees of freedom *df*. The estimation was performed using the mle function of Matlab R2022a.

Data, Materials, and Software Availability. Computer code can be accessed on GitHub (https://github.com/stefan-f-bucher/divisive-normalization-efficiency) (109). The van Hateren image dataset (87) is available at https://pirsquared.org/research/vhatdb/.

ACKNOWLEDGMENTS. We are grateful to Paul Glimcher for very important conceptual input; to Corey Ziemba for generous help with the empirical analysis; to Shervin Safavi for code review; to Barry Arnold, Peter Dayan, Ambuj Dewan, Pierre-Étienne Fiquet, Ann Hermundstad, Vered Kurtz, Kenway Louie, Barry Nalebuff, Eero Simoncelli, Fabian Sinz, and Michael Woodford; to members of the Glimcher Laboratory; to audience members at Cosyne 2020, the 2020 Annual Meeting of the Society for Neuroeconomics, the 2020 International Conference on Mathematical Neuroscience, and the 2021 Annual Meeting of the Society for Neuroscience for helpful discussions; and to NYU Graduate School of Arts & Science, NYU School of Medicine, NYU Stern School of Business, NYU Shanghai, and J. P. Valles for financial support. This article is based on a chapter of S.F.B.'s dissertation at New York University (110).

Author affiliations: [a]Department of Economics, New York University, New York, NY 10012; [b]Department of Computer Science, University of Tübingen, 72076 Tübingen, Germany; [c]Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany; [d]Stern School of Business, New York University, New York, NY 10012; [e]Tandon School of Engineering, New York University, Brooklyn, NY 11201; and [f]New York University Shanghai, Shanghai, China 200122

1. D. J. Heeger, Normalization of cell responses in cat striate cortex. *Vis. Neurosci.* **9**, 181–197 (1992).
2. W. S. Geisler, D. G. Albrecht, Cortical neurons: Isolation of contrast gain control. *Vision Res.* **32**, 1409–1410 (1992).
3. M. Carandini, D. J. Heeger, Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* **13**, 51–62 (2011).
4. M. Carandini, D. J. Heeger, J. A. Movshon, Linearity and normalization in simple cells of the macaque primary visual cortex. *J. Neurosci.* **17**, 8621–8644 (1997).
5. D. J. Tolhurst, D. J. Heeger, Comparison of contrast-normalization and threshold models of the responses of simple cells in cat striate cortex. *Vis. Neurosci.* **14**, 293–309 (1997).
6. L. Busse, A. R. Wade, M. Carandini, Representation of concurrent stimuli by population activity in visual cortex. *Neuron* **64**, 931–942 (2009).
7. D. L. Ringach, Population coding under normalization. *Vision Res.* **50**, 2223–2232 (2010).
8. T. K. Sato, B. Haider, M. Häusser, M. Carandini, An excitatory basis for divisive normalization in visual cortex. *Nat. Neurosci.* **19**, 568–570 (2016).
9. M. Aqil, T. Knapen, S. O. Dumoulin, Divisive normalization unifies disparate response signatures throughout the human visual hierarchy. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2108713118 (2021).
10. J. J. Foster, S. Ling, Normalizing population receptive fields. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2118367118 (2021).
11. S. R. Olsen, V. Bhandawat, R. I. Wilson, Divisive normalization in olfactory population codes. *Neuron* **66**, 287–299 (2010).
12. E. P. Simoncelli, D. J. Heeger, A model of neuronal responses in visual area MT. *Vision Res.* **38**, 743–761 (1998).
13. K. H. Britten, H. W. Heuer, Spatial summation in the receptive fields of MT neurons. *J. Neurosci.* **19**, 5074–5084 (1999).
14. D. Zoccolan, D. D. Cox, J. J. DiCarlo, Multiple object response normalization in monkey inferotemporal cortex. *J. Neurosci.* **25**, 8150–8164 (2005).
15. A. Bhatia, S. Moza, U. S. Bhalla, Precise excitation-inhibition balance controls gain and timing in the hippocampus. *eLife* **8**, e43415 (2019).
16. T. Ohshiro, D. E. Angelaki, G. C. DeAngelis, A neural signature of divisive normalization at the level of multisensory integration in primate cortex. *Neuron* **95**, 399–411.e8 (2017).
17. K. Louie, L. E. Grattan, P. W. Glimcher, Reward value-based gain control: Divisive normalization in parietal cortex. *J. Neurosci.* **31**, 10627–10639 (2011).
18. H. Yamada, K. Louie, A. Tymula, P. W. Glimcher, Free choice shapes normalized value signals in medial orbitofrontal cortex. *Nat. Commun.* **9**, 162 (2018).
19. K. Louie, M. W. Khaw, P. W. Glimcher, Normalization is a general neural mechanism for context-dependent decision making. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 6139–6144 (2013).
20. K. Steverson, A. Brandenburger, P. Glimcher, Choice-theoretic foundations of the divisive normalization model. *J. Econ. Behav. Organ.* **164**, 148–165 (2019).
21. R. Webb, P. W. Glimcher, K. Louie, Divisive normalization does influence decisions with multiple alternatives. *Nat. Hum. Behav.* **4**, 1118–1120 (2020).
22. R. Webb, P. W. Glimcher, K. Louie, The normalization of consumer valuations: Context-dependent preferences from neurobiological constraints. *Manage. Sci.* **67**, 93–125 (2020).
23. B. K. Chau, C. K. Law, A. Lopez-Persem, M. C. Klein-Flügge, M. F. S. Rushworth, Consistent patterns of distractor effects during decision making. *eLife* **9**, e53850 (2020).
24. S. Gluth, N. Kern, M. Kortmann, C. L. Vitali, Value-based attention but not divisive normalization influences decisions with multiple alternatives. *Nat. Hum. Behav.* **4**, 634–645 (2020).
25. S. Gluth, N. Kern, C. L. Vitali, Reply to: Divisive normalization does influence decisions with multiple alternatives. *Nat. Hum. Behav.* **4**, 1121–1123 (2020).
26. J. H. Reynolds, D. J. Heeger, The normalization model of attention. *Neuron* **61**, 168–185 (2009).
27. I. M. Bloem, S. Ling, Normalization governs attentional modulation within human visual cortex. *Nat. Commun.* **10**, 5660 (2019).
28. R. Coen-Cagli, S. S. Solomon, Relating divisive normalization to neuronal response variability. *J. Neurosci.* **39**, 7344–7356 (2019).
29. O. J. Hénaff, Z. M. Boundy-Singer, K. Meding, C. M. Ziemba, R. L. T. Goris, Representation of visual uncertainty through neural gain variability. *Nat. Commun.* **11**, 2513 (2020).
30. S. Deneve, P. E. Latham, A. Pouget, Reading population codes: A neural implementation of ideal observers. *Nat. Neurosci.* **2**, 740–745 (1999).
31. J. M. Beck, P. E. Latham, A. Pouget, Marginalization in neural circuits with divisive normalization. *J. Neurosci.* **31**, 15310–15319 (2011).
32. B. Vintch, J. A. Movshon, E. P. Simoncelli, A convolutional subunit model for neuronal responses in macaque V1. *J. Neurosci.* **35**, 14829–14841 (2015).
33. M. F. Burg *et al.*, Learning divisive normalization in primary visual cortex. *PLOS Comput. Biol.* **17**, e1009028 (2021).
34. J. Ballé, V. Laparra, E. P. Simoncelli, "End-to-end optimized image compression" in *International Conference on Learning Representations* (Toulon, France, 2017).
35. O. Schwartz, E. P. Simoncelli, Natural signal statistics and sensory gain control. *Nat. Neurosci.* **4**, 819–825 (2001).

36. K. Louie, P. W. Glimcher, Efficient coding and the neural representation of value. *Ann. N. Y. Acad. Sci.* **1251**, 13–32 (2012).
37. F. Attneave, Some informational aspects of visual perception. *Psychol. Rev.* **61**, 183–193 (1954).
38. H. B. Barlow, "Possible principles underlying the transformations of sensory messages" in *Sensory Communication*, W. A. Rosenblith, Ed. (MIT Press, 1961), pp. 217–234.
39. S. B. Laughlin, R. C. Hardie, Common strategies for light adaptation in the peripheral visual systems of fly and dragonfly. *J. Comp. Physiol.* **128**, 319–340 (1978).
40. S. Laughlin, A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch. C* **36**, 910–912 (1981).
41. H. Barlow, Redundancy reduction revisited. *Network* **12**, 241–253 (2001).
42. E. P. Simoncelli, Vision and the statistics of the visual environment. *Curr. Opin. Neurobiol.* **13**, 144–149 (2003).
43. M. Chalk, O. Marre, G. Tkačik, Toward a unified theory of efficient, predictive, and sparse coding. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 186–191 (2018).
44. W. F. Młynarski, A. M. Hermundstad, Efficient and adaptive sensory codes. *Nat. Neurosci.* **24**, 998–1009 (2021).
45. X. X. Wei, A. A. Stocker, A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nat. Neurosci.* **18**, 1509–1517 (2015).
46. R. Bhui, S. J. Gershman, Decision by sampling implements efficient coding of psychoeconomic functions. *Psychol. Rev.* **125**, 985–1001 (2018).
47. R. Polanía, M. Woodford, C. C. Ruff, Efficient coding of subjective value. *Nat. Neurosci.* **22**, 134–142 (2019).
48. M. Woodford, Modeling imprecision in perception, valuation, and choice. *Annu. Rev. Econ.* **12**, 579–601 (2020).
49. C. Frydman, L. J. Jin, Efficient coding and risky choice. *Q. J. Econ.* **137**, 161–213 (2022).
50. S. Lyu, Divisive normalization: Justification and effectiveness as efficient coding transform. *Adv. Neural Inf. Process. Syst.* **23**, 1522–1530 (2010).
51. E. P. Simoncelli, B. A. Olshausen, Natural image statistics and neural representation. *Annu. Rev. Neurosci.* **24**, 1193–1216 (2001).
52. R. Valerio, R. Navarro, Optimal coding through divisive normalization models of V1 neurons. *Network* **14**, 579–593 (2003).
53. R. Valerio, R. Navarro, Input-output statistical independence in divisive normalization models of V1 neurons. *Network* **14**, 733–745 (2003).
54. J. Malo, V. Laparra, Psychophysically tuned divisive normalization approximately factorizes the PDF of natural images. *Neural Comput.* **22**, 3179–3206 (2010).
55. M. J. Wainwright, O. Schwartz, E. P. Simoncelli, "Natural image statistics and divisive normalization: Modeling nonlinearities and adaptation in cortical neurons" in *Probabilistic Models of the Brain*, R. P. Rao, B. A. Olshausen, M. S. Lewicki, Eds. (MIT Press, Cambridge, MA, 2002), pp. 203–222.
56. S. Lyu, E. P. Simoncelli, Nonlinear extraction of independent components of natural images using radial Gaussianization. *Neural Comput.* **21**, 1485–1519 (2009).
57. F. Sinz, M. Bethge, The conjoint effect of divisive normalization and orientation selectivity on redundancy reduction in natural images. *Adv. Neural Inf. Process. Syst.* **21**, 1521–1528 (2009).
58. F. H. Sinz, M. Bethge, What is the limit of redundancy reduction with divisive normalization? *Neural Comput.* **25**, 2809–2814 (2013).
59. S. Lyu, Dependency reduction with divisive normalization: Justification and effectiveness. *Neural Comput.* **23**, 2942–2973 (2011).
60. M. J. Wainwright, E. P. Simoncelli, Scale mixtures of Gaussians and the statistics of natural images. *Adv. Neural Inf. Process. Syst.* **12**, 855–861 (2000).
61. J. Ballé, V. Laparra, E. P. Simoncelli, "Density modeling of images using a generalized normalization transformation" in *International Conference on Learning Representations* (San Juan, Puerto Rico, 2016).
62. C. E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
63. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (Wiley-Interscience, Hoboken, NJ, 2006).
64. N. Brady, D. J. Field, Local contrast in natural images: Normalisation and coding efficiency. *Perception* **29**, 1041–1055 (2000).
65. X. X. Wei, A. A. Stocker, Mutual information, Fisher information, and efficient coding. *Neural Comput.* **28**, 305–326 (2016).
66. H. Attias, C. Schreiner, Temporal low-order statistics of natural sounds. *Adv. Neural Inf. Process. Syst.* **9**, 27–33 (1997).
67. W. B. Levy, R. A. Baxter, Energy efficient neural codes. *Neural Comput.* **8**, 531–543 (1996).
68. Z. Wang, W. X. Wei, A. A. Stocker, D. D. Lee, Efficient neural codes under metabolic constraints. *Adv. Neural Inf. Process. Syst.*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett, Eds. (2016) vol. 29, pp. 4619–4627.
69. T. Sawada, A. A. Petrov, The divisive normalization model of V1 neurons: A comprehensive comparison of physiological data and model predictions. *J. Neurophysiol.* **118**, 3051–3091 (2017).
70. P. Lennie, The cost of cortical computation. *Curr. Biol.* **13**, 493–497 (2003).
71. A. Hasenstaub, S. Otte, E. Callaway, T. J. Sejnowski, Metabolic cost as a unifying principle governing neuronal biophysics. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 12329–12334 (2010).
72. G. Yi, W. M. Grill, Average firing rate rather than temporal pattern determines metabolic cost of activity in thalamocortical relay neurons. *Sci. Rep.* **9**, 6940 (2019).
73. D. A. Harville, *Matrix Algrebra From a Statistician's Perspective* (Springer-Verlag, New York, NY, 2008).
74. J. P. Nadal, N. Parga, Nonlinear neurons in the low-noise limit: A factorial code maximizes information transfer. *Network Comput. Neural Syst.* **5**, 565–581 (1994).
75. A. J. Bell, T. J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* **7**, 1129–1159 (1995).
76. E. T. Jaynes, Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630 (1957).
77. R. Baddeley, Visual perception. An efficient code in V1? *Nature* **381**, 560–561 (1996).
78. R. Baddeley *et al.*, Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proc. Biol. Sci.* **264**, 1775–1783 (1997).
79. J. K. Blitzstein, J. Hwang, *Introduction to Probability* (Chapman and Hall/CRC, ed. 2, 2019).
80. B. C. Arnold, *Pareto Distributions* (Chapman and Hall/CRC, Boca Raton, FL, ed. 2, 2015).
81. S. Kotz, N. Balakrishnan, N. L. Johnson, *Continuous Multivariate Distributions* (Wiley Series in Probability and Statistics, Wiley, New York, ed. 2, 2000), vol. 1.
82. M. Newman, Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **46**, 323–351 (2005).
83. A. Clauset, C. R. Shalizi, M. E. J. Newman, Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703 (2009).
84. A. Klaus, S. Yu, D. Plenz, Statistical analyses support power law distributions found in neuronal avalanches. *PLoS One* **6**, e19779 (2011).
85. P. R. Fisk, The graduation of income distributions. *Econometrica* **29**, 171 (1961).
86. R. W. Buccigrossi, E. P. Simoncelli, Image compression via joint statistical characterization in the wavelet domain. *IEEE Trans. Image Process.* **8**, 1688–1701 (1999).
87. J. H. van Hateren, A. van der Schaaf, Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. Lond. B Biol. Sci.* **265**, 359–366 (1998).
88. M. Welling, S. Osindero, G. Hinton, "Learning sparse topographic representations with products of student-t distributions" in *Advances in Neural Processing Information Systems*, S. Becker, S. Thrun, K. Obermayer, Eds. (2002), vol. 15, pp. 1359–1366.
89. S. Roth, M. J. Black, "Fields of experts: A framework for learning image priors" in *IEEE Computer Society Conference on Computer Vision Pattern Recognition* (2005), vol. 2, pp. 860–867.
90. G. Chantas, N. Galatsanos, A. Likas, M. Saunders, Variational Bayesian image restoration based on a product of t-distributions image prior. *IEEE Trans. Image Process.* **17**, 1795–1805 (2008).
91. R. Coen-Cagli, P. Dayan, O. Schwartz, Cortical surround interactions and perceptual salience via natural scene statistics. *PLOS Comput. Biol.* **8**, e1002405 (2012).
92. W. S. Geisler, Visual perception and the statistical properties of natural scenes. *Annu. Rev. Psychol.* **59**, 167–192 (2008).
93. K. Mizuseki, G. Buzsáki, Preconfigured, skewed distribution of firing rates in the hippocampus and entorhinal cortex. *Cell Rep.* **4**, 1010–1021 (2013).
94. G. Buzsáki, K. Mizuseki, The log-dynamic brain: How skewed distributions affect network operations. *Nat. Rev. Neurosci.* **15**, 264–278 (2014).
95. V. Kurtz-David, D. Persitz, R. Webb, D. J. Levy, The neural computation of inconsistent choice behavior. *Nat. Commun.* **10**, 1583 (2019).
96. V. Pareto, La legge della domanda. *Giornale degli Econ.* **10**, 59–68 (1895).
97. V. Pareto, *Cours d'Economie Politique* (F. Rouge, Lausanne, Switzerland, 1896).
98. L. M. Leemis, B. W. Schmeiser, D. L. Evans, Survival distributions satisfying Benford's law. *Am. Stat.* **54**, 236–241 (2012).
99. S. Newcomb, Note on the frequency of use of the different digits in natural numbers. *Am. J. Math.* **4**, 39–40 (1881).
100. F. Benford, The law of anomalous numbers. *Proc. Am. Philos. Soc.* **78**, 551–572 (1938).
101. X. Gabaix, Power laws in economics and finance. *Annu. Rev. Econ.* **1**, 255–294 (2009).
102. X. Gabaix, Power laws in economics: An introduction. *J. Econ. Perspect.* **30**, 185–206 (2016).
103. X. Gabaix, P. Gopikrishnan, V. Plerou, H. E. Stanley, A theory of power-law distributions in financial market fluctuations. *Nature* **423**, 267–270 (2003).
104. A. J. Robson, The biological basis of economic behavior. *J. Econ. Lit.* **39**, 11–33 (2001).
105. N. Netzer, Evolution of time preferences and attitudes toward risk. *Am. Econ. Rev.* **99**, 937–955 (2009).
106. J. Schaffner, P. Tobler, T. Hare, R. Polania, Neural codes in early sensory areas maximize fitness. *bioRxiv* [Preprint] (2021). https://doi.org/10.1101/2021.05.10.443388. Accessed 2 June 2022.
107. J. Portilla, E. P. Simoncelli, A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis.* **40**, 49–70 (2000).
108. C. Ziemba, Neural representation and perception of naturalistic image structure, Ph.D. thesis (New York University, New York, NY) (2016).
109. S. Bucher, Code for Bucher & Brandenburger: "Divisive normalization is an efficient code for multivariate Pareto-distributed environments." GitHub. https://github.com/stefan-f-bucher/divisive-normalization-efficiency. Accessed 6 September 2022.
110. S. Bucher, Information Constraints in Decision-Making, Ph.D. thesis, New York University, New York, NY (2021).