



Measuring the dimensionality of behavior

Surya Ganguli^{a,b,1}

The conjunction of high-speed video and machine learning advances now enables us to almost automatically extract detailed descriptions of highly complex animal behaviors across multiple species (1, 2) as they engage in complex tasks or naturalistic behaviors. This explosion of behavioral data, along with advances in our ability to record the simultaneous dynamics of many neurons or genes, raises profound questions about how we should efficiently organize our scientific efforts to conceptually understand the emergence of behavior from underlying neural or biochemical mechanisms. Bialek (3) addresses this question, arguing first for the importance of quantitative phenomenological characterizations of behavior and then providing a framework for studying one quantitative aspect of behavior: its dimensionality.

In Defense of Phenomenology in Biology

From the perspective of a biologist who can simultaneously measure neural or biochemical dynamics in conjunction with behavior, one might then be tempted to delve directly into the question of mechanism (i.e., how do neural circuits in the brain or biochemical circuits in cells generate such behavior?). In contrast, Bialek (3) urges us to also examine the question of what aspects of behavior we would like to explain. In convincingly arguing for this viewpoint, Bialek (3) marshals together a wealth of instructive examples across multiple fields of biology in which a purely phenomenological yet highly quantitative characterization of behaviors came long before the elucidation of their underlying mechanisms. Moreover, such detailed quantitative behavioral phenomenology provided essential clues as to what sorts of underlying mechanisms one ought to search for. For example, the quantitative observation that error rates in copying DNA are multiple orders of magnitude smaller than predicted by equilibrium thermodynamics suggests the existence of an important nonequilibrium or kinetic proofreading mechanism (4) underlying the transmission of genetic information. Also, a quantitative characterization of the dimmest flash of light a human can still perceive led to the remarkable conclusion that individual retinal photoreceptors can absorb single photons, implying that a fundamentally quantum mechanism underlies the first steps of vision (5). Hodgkin and Huxley's (6) detailed quantitative and purely phenomenological characterization of the dynamics of the neuronal action potential yielded fundamental clues as to its mechanistic basis in terms of ion channel conformational dynamics. More recently, limits of perceptual acuity for small changes in shape suggest that correlated noise in the nervous system may limit the information content in neural ensembles (7, 8).

Now, while history rarely exactly repeats itself, it certainly rhymes. Thus, in this modern day and age, given the seminal lessons of our past, it would certainly behoove us to engage in a careful and quantitative phenomenological

characterization of the much richer behaviors we can only now recently measure. Just as many times before, such a quantitative characterization will likely help us understand what aspects of behavior are most interesting to explain and provide essential clues as to what sorts of underlying mechanisms we should search for.

A General Definition of Dimensionality in Behavior

With this defense of quantitative behavioral phenomenology in hand, Bialek (3) goes on to provide a framework for quantifying, at a significant level of generality, what is perhaps one of the most fundamental aspects of behavior: its dimensionality. The dimensionality of behavior provides an important clue as to how many degrees of freedom an underlying neural or biochemical system may need in order to generate the behavior (9, 10). It can also suggest how many measurements might be required to accurately capture the underlying dynamics (11). Building on ref. 12, in ref. 3 a candidate general definition of the dimensionality of a single scalar behavioral time series (e.g., the joint angle of a single joint, the speed of a single bacterium, or a discrete dynamic behavioral state of an organism) is constructed in a sequence of steps starting from particular cases and culminating in a general information theoretic definition. Here, we start from the general definition.

At an intuitive level, the dimensionality d^* of a time series can be defined as how many functions of its past one needs to measure in order to accurately predict its future as well as possible. Consider a scalar behavioral time series $x(t)$ observed from $t = -T$ to T . The past of this time series is defined as $\mathbf{x}_{\text{past}} = x(-T \dots 0)$, and its future is defined as $\mathbf{x}_{\text{fut}} = x(0 \dots T)$ (Fig. 1A). The total amount of predictable information the past contains about the future can be quantified by the mutual information

$$I_{\text{pred}}(T) \equiv I(\mathbf{x}_{\text{past}}; \mathbf{x}_{\text{fut}}). \quad [1]$$

Now, to define dimensionality, consider predicting the future \mathbf{x}_{fut} not directly from an observation of the past \mathbf{x}_{past} but from d measurements or smooth functions of the past [i.e., from $\mathcal{F}_\mu(\mathbf{x}_{\text{past}})$ for $\mu = 1, \dots, d$] (Fig. 1A). Moreover, of the space \mathcal{M}_d of all such possible d measurements, consider the optimal set that maximizes the predictive information

$$I_{\text{pred}}(T, d) \equiv \max_{\{\mathcal{F}_\mu\} \in \mathcal{M}_d} I(\{\mathcal{F}_\mu\}; \mathbf{x}_{\text{fut}}). \quad [2]$$

Author affiliations: ^aMeta AI Research, Menlo Park, CA 94025; and ^bDepartment of Applied Physics, Stanford University, Stanford, CA 94305

Author contributions: S.G. wrote the paper.

The author declares no competing interest.

See companion article, "On the dimensionality of behavior," [10.1073/pnas.2021860119](https://doi.org/10.1073/pnas.2021860119).

Copyright © 2022 the Author(s). Published by PNAS. This article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

¹ Email: sganguli@stanford.edu.

Published October 20, 2022.

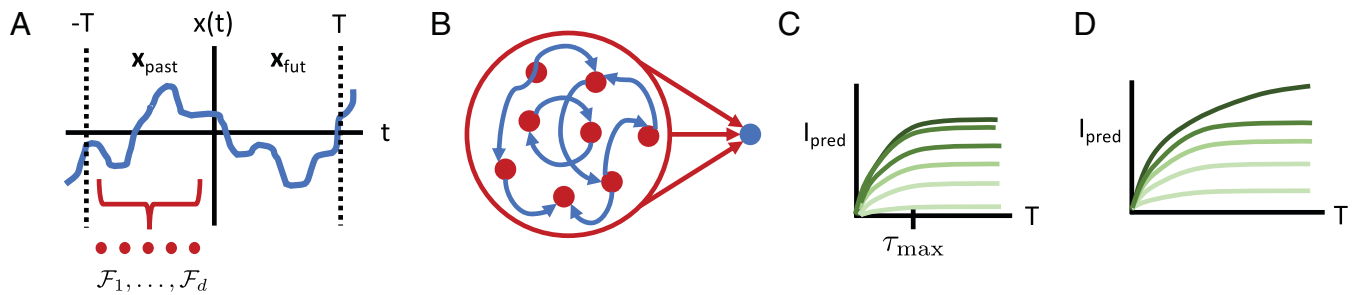


Fig. 1. (A) A single sample $x(t)$ of a behavioral time series (blue curve) can be divided into the past \mathbf{x}_{past} from $t = -T$ to $t = 0$ and the future \mathbf{x}_{fut} from $t = 0$ to $t = T$. One can then either predict \mathbf{x}_{fut} directly from \mathbf{x}_{past} , leading to the mutual information $I_{\text{pred}}(T)$ defined in Eq. 1. Alternatively, one could construct d optimal measurements of the past $\mathcal{F}_{\mu}(\mathbf{x}_{\text{past}})$ for $\mu = 1, \dots, d$ (red circles) and use these to predict the future \mathbf{x}_{fut} , leading to the mutual information $I_{\text{pred}}(T, d)$ defined in Eq. 2. (B) A schematic of a d^* -dimensional dynamical system consisting of d^* interacting scalar variables (red circles) that generates behavior (blue circle) through a set of outgoing connections (red arrows). (C) The top curve (darkest green) shows a schematic $I_{\text{pred}}(T)$ for behavior generated by a dynamical system as in B with a maximal autocorrelation time τ_{max} . $I_{\text{pred}}(T)$ saturates around $T \approx \tau_{\text{max}}$. The bottom curves show a schematic of $I_{\text{pred}}(T, d)$ for increasing d reflected by the deepening shade of green. In the situation where behavior is generated by a d^* -dimensional dynamical system with a maximum autocorrelation time, $I_{\text{pred}}(T, d^*)$ is close to the upper bound set by $I_{\text{pred}}(T)$. (D) The top curve shows a schematic of $I_{\text{pred}}(T)$ for a behavioral time series with temporal correlations that decay as a power law. Such long-range correlations lead to a logarithmic divergence of $I_{\text{pred}}(T)$ with T . The bottom curves show $I_{\text{pred}}(T, d)$ for increasing d reflected by the deepening shade of green. For any fixed d , $I_{\text{pred}}(T, d)$ often saturates with time T due to a finite number of measurements, with the saturation value necessarily increasing with d . Thus, given the divergence of $I_{\text{pred}}(T)$, an increasing number of measurements d will be required for $I_{\text{pred}}(T, d)$ to approach $I_{\text{pred}}(T)$ as T increases. This leads to an effective infinite dimensionality of long-range temporally correlated behavior.

Then, the ratio

$$f(d, T) = \frac{I_{\text{pred}}(T, d)}{I_{\text{pred}}(T)} \quad [3]$$

measures what fraction of the total information about the future one captures from the best possible d measurements of the past given a past-future time horizon of T . For any fixed T , $f(d, T)$ is necessarily a monotonically increasing function of d . One might expect that at very large T , $f(d, T)$ will first saturate to a fixed value for all d greater than d^* . The minimal value of d^* at which this saturation occurs is then a candidate measure for the dimensionality of the time series $x(t)$. Intuitively, this saturation means that at large time horizons T , taking more than d^* measurements does not help one predict the future any better.

To understand why d^* is a sensible measure of dimensionality, it is useful to consider the properties of $I_{\text{pred}}(T)$ and $I_{\text{pred}}(T, d)$ as a function of T and d for specific cases. Consider, for example, a dynamical system (i.e., a neural circuit) with d^* scalar degrees of freedom interacting via some generic connectivity subject to internal noise sources, and suppose further that the behavior $x(t)$ is some instantaneous function of these degrees of freedom (Fig. 1B). Thus, a d^* -dimensional stochastic dynamical system is generating behavior. Suppose, furthermore, that the interactions in this dynamical system are such that there is a maximum autocorrelation time τ_{max} in the dynamical system and therefore, also in the behavioral time series $x(t)$. This means that the autocorrelation between $x(t)$ and $x(t')$ is exponentially suppressed in $\frac{|t-t'|}{\tau_{\text{max}}}$. As a consequence, looking back into past behavior for any time T longer than τ_{max} cannot help predict future behavior any better since this distant past is uncorrelated with the future.* Thus, we expect $I_{\text{pred}}(T)$ to rise with T but saturate for T larger than τ_{max} (Fig. 1C). Similarly, we expect $I_{\text{pred}}(T, d)$ to saturate at least by $T \approx \tau_{\text{max}}$ but at a lower level than $I_{\text{pred}}(T)$ because of the limited number of d measurements. However, at a fixed large $T > \tau_{\text{max}}$, we also expect $I_{\text{pred}}(T, d)$ to increase with d but

saturate at $d = d^*$. The key intuition is since the behavior $x(t)$ is generated by a d^* -dimensional dynamical system, then with d^* measurements, the best one could do to predict future behavior is use the past behavioral time series \mathbf{x}_{past} to reconstruct the d^* -dimensional state of the generating dynamical system since the future behavior is independent of the past conditioned on this state. Taking more than this number of past measurements, assuming we chose them optimally, is unlikely to appreciably help our ability to better predict future behavior. In this fashion, $f(d, T)$ for $T > \tau_{\text{max}}$ will saturate with d when it increases past d^* . Thus, this particular statistical analysis of behavior $x(t)$ alone will recover the underlying dimensionality d^* of the dynamical system that generated it. We note that while we have provided a qualitative intuition for why this information theoretic characterization of dimensionality makes sense in the context of a given dynamical system that generates behavior, all of this intuition can be justified analytically when $x(t)$ obeys jointly Gaussian statistics across time (3).

A Data-Driven Approach to Estimating Dimensionality

We note that while it can be difficult to compute mutual information directly from data, one can obtain an approximate calculation of dimensionality d^* by reducing it to the computation of the eigenvalue spectrum of a certain kernel matrix connecting the past and the future of a behavior, as described in ref. 3. Moreover, this kernel itself can be derived from second-order temporal correlations in the behavioral time series. The key idea to achieve this reduction, applicable to both continuous and discrete behaviors in a unified fashion, is to first approximate the probability distribution over behavioral time series with a maximum entropy (MaxEnt) distribution constrained to have the same second-order temporal correlations as behavior. Then, one can carry through the information theoretic program for computing dimensionality, either analytically or numerically, on the MaxEnt distribution itself. Of course, in doing so, one must take care to correctly account for statistical significance of the results of estimating spectra of kernels

*More precisely, we make the stronger assumption that $\mathbf{x}(t > 0)$ is approximately conditionally independent of $\mathbf{x}(t < -\tau_{\text{max}})$ given $\mathbf{x}(t = -\tau_{\text{max}} \dots 0)$.

of MaxEnt distributions with finite amounts of data. The statistics of random matrices then become highly relevant in assessing statistical significance (13).

Interestingly, in the case of continuous behavior, linear dynamical systems driven by white noise generate exceedingly simple and instructive examples of behavioral time series with jointly Gaussian temporal statistics for which the MaxEnt approximation is exact (3). For example, when behavior $x(t)$ is simply a leaky linear integrator of white noise with time constant τ_{\max} , then $I_{\text{pred}}(T)$ saturates at $T > \tau_{\max}$, and for large T , $I_{\text{pred}}(T, d)$ saturates in d at $d = d^* = 1$. This is the correct dimension of behavior since a leaky integrator possesses a single degree of freedom. Moreover, the optimal measurement of the past to predict the future is itself a leaky integral of the past. Similarly, if behavior corresponds to the position of a damped harmonic oscillator driven by noise, the information theoretic characterization of dimensionality would again correctly predict $d^* = 2$ since two degrees of freedom are required to produce oscillations. Also, in this case, the two optimal measurements of the past would be smoothed linear filters that approximately estimate recent position and velocity, which are sufficient to predict the future of an oscillator.

Power-Law Correlations and Infinite Dimensionality

A key assumption underlying all examples so far is that temporal correlations in behavior are exponentially small in temporal separations greater than a maximum autocorrelation time τ_{\max} . This condition is sufficient for $I_{\text{pred}}(T)$ to saturate in T . However, a qualitatively distinct universality class of behavior can occur if $x(t)$ has temporal autocorrelations that decay as a power law with time. In such a setting, with long-range temporal correlations, $I_{\text{pred}}(T)$ need not saturate with T ; instead, it can grow logarithmically with time T . This means the further back in time one looks, the more one can predict about the future without bound. Behavioral time series for which this is the case are inevitably more interesting than the opposite universality class of leaky integrators and damped oscillators for which $I_{\text{pred}}(T)$ saturates. A classic example for how such long-range temporally correlated behavior might arise is, for example, written language. When looking back a few letters, the rules of phonology can help predict future letters. However, looking back a few words, qualitatively distinct rules of syntax emerge that can help predict future words. Finally, over longer timescales spanning sentences, rules of semantics, human intentions,

and story arcs might constrain the semantic content of future sentences.

An intriguing consequence of the divergence of $I_{\text{pred}}(T)$ with T is that the dimensionality d^* of such behavior may also become effectively infinite (Fig. 1D). For example, $I_{\text{pred}}(T, d)$ may require progressively larger and larger d to approach the progressively larger and larger value of $I_{\text{pred}}(T)$ as T increases (Fig. 1D). Therefore, as $T \rightarrow \infty$, d may also have to grow without bound in order for the ratio in Eq. 3 to saturate with d . In essence, for such long-range temporally correlated behaviors, more and more measurements are required to predict the future as accurately as possible over longer and longer time horizons, leading to effectively infinite dimensionality.

Outlook

As Bialek notes in ref. 3, “the possibility that behavioral correlations decay as a power of time has a long and sometimes contentious history.” Recent work, however, suggests that discrete behavioral states visited by *Drosophila* over time exhibit such long-range correlations (14). However, further studies that search for such intriguingly complex behaviors and accurately characterize the nature of their temporal correlations and growing dimensionality with time are certainly a direction for future research.

Interestingly, beyond biology, machine learning attempts to model complex time series, such as language itself, also exhibit greater and greater accuracy in predicting future language given past language and more complex measurements (15), although the details of the analysis are quite different from those of ref. 3. Nevertheless it is intriguing to hypothesize that scaling laws exhibited by human generated behaviors may be intricately related to scaling laws associated with machine learning attempts to model them.

More generally, zooming out from the analysis of temporal correlations and the dimensionality of behavior (3) reminds us in a compelling way about the primacy of seeking to quantitatively understand complex behavior before diving into measurements that seek to relate neural or biochemical activity to behavior. This reminder is timely in an age where remarkable advances in technology almost taunt us to take detailed large-scale measurements underlying neural or biochemical activity patterns or even connectomes. However, before we do, an important guiding question should always remain at the fore. What is it about behavior that we would like to explain?

1. M. W. Mathis, A. Mathis, Deep learning tools for the measurement of animal behavior in neuroscience. *Curr. Opin. Neurobiol.* **60**, 1–11 (2020).
2. S. R. Datta, D. J. Anderson, K. Branson, P. Perona, A. Leifer, Computational neuroethology: A call to action. *Neuron* **104**, 11–24 (2019).
3. W. Bialek, On the dimensionality of behavior. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e20211860119 (2022).
4. J. J. Hopfield, Kinetic proofreading: A new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Natl. Acad. Sci. U.S.A.* **71**, 4135–4139 (1974).
5. M. A. Bouman, History and present status of quantum theory in vision. *Sensory Communication*, W. A. Rosenblith, Ed. (MIT Press, Cambridge, MA, 2012), chap. 21.
6. A. L. Hodgkin, A. F. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* **117**, 500–544 (1952).
7. R. Moreno-Bote *et al.*, Information-limiting correlations. *Nat. Neurosci.* **17**, 1410–1417 (2014).
8. O. I. Rumyantsev *et al.*, Fundamental bounds on the fidelity of sensory cortical coding. *Nature* **580**, 100–105 (2020).
9. P. Gao, S. Ganguli, On simplicity and complexity in the brave new world of large-scale neuroscience. *Curr. Opin. Neurobiol.* **32**, 148–155 (2015).
10. P. Gao *et al.*, A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv* [Preprint] (2017). <https://www.biorxiv.org/content/10.1101/214262v2> (Accessed 14 April 2022).
11. E. M. Trautmann *et al.*, Accurate estimation of neural population dynamics without spike sorting. *Neuron* **103**, 292–308.e4 (2019).
12. W. Bialek, I. Nemenman, N. Tishby, Predictability, complexity, and learning. *Neural Comput.* **13**, 2409–2463 (2001).
13. M. Potters, J. P. Bouchaud, *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists* (Cambridge University Press, 2020).
14. V. Alba, G. J. Berman, W. Bialek, J. W. Shaevitz, Exploring a strongly non-Markovian animal behavior. *arXiv* [Preprint] (2020). <https://arxiv.org/abs/2012.15681> (Accessed 14 April 2022).
15. J. Kaplan *et al.*, Scaling laws for neural language models. *arXiv* [Preprint] (2020). <https://arxiv.org/abs/2001.08361> (Accessed 14 April 2022).