# Perspective: Sloppiness and emergent theories in physics, biology, and beyond

Mark K. Transtrum, Benjamin B. Machta, Kevin S. Brown, Bryan C. Daniels, Christopher R. Myers, and James P. Sethna

Citation: The Journal of Chemical Physics **143**, 010901 (2015); doi: 10.1063/1.4923066
View online: http://dx.doi.org/10.1063/1.4923066
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/143/1?ver=pdfcov
Published by the AIP Publishing



# Perspective: Sloppiness and emergent theories in physics, biology, and beyond

Mark K. Transtrum,[1] Benjamin B. Machta,[2] Kevin S. Brown,[3,4] Bryan C. Daniels,[5] Christopher R. Myers,[6,7] and James P. Sethna[6]
[1]*Department of Physics and Astronomy, Brigham Young University, Provo, Utah 84602, USA*
[2]*Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544, USA*
[3]*Departments of Biomedical Engineering, Physics, Chemical and Biomolecular Engineering, and Marine Sciences, University of Connecticut, Storrs, Connecticut 06269, USA*
[4]*Institute for Systems Genomics, University of Connecticut, Storrs, Connecticut 06030-1912, USA*
[5]*Center for Complexity and Collective Computation, Wisconsin Institute for Discovery, University of Wisconsin, Madison, Wisconsin 53715, USA*
[6]*Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, New York 14853, USA*
[7]*Institute of Biotechnology, Cornell University, Ithaca, New York 14853, USA*

(Received 2 February 2015; accepted 4 June 2015; published online 1 July 2015)

Large scale models of physical phenomena demand the development of new statistical and computational tools in order to be effective. Many such models are "sloppy," i.e., exhibit behavior controlled by a relatively small number of parameter combinations. We review an information theoretic framework for analyzing sloppy models. This formalism is based on the Fisher information matrix, which is interpreted as a Riemannian metric on a parameterized space of models. Distance in this space is a measure of how distinguishable two models are based on their predictions. Sloppy model manifolds are bounded with a hierarchy of widths and extrinsic curvatures. The manifold boundary approximation can extract the simple, hidden theory from complicated sloppy models. We attribute the success of simple effective models in physics as likewise emerging from complicated processes exhibiting a low effective dimensionality. We discuss the ramifications and consequences of sloppy models for biochemistry and science more generally. We suggest that the reason our complex world is understandable is due to the same fundamental reason: simple theories of macroscopic behavior are hidden inside complicated microscopic processes.  [http://dx.doi.org/10.1063/1.4923066]

## I. PARAMETER INDETERMINACY AND SLOPPINESS

As a young physicist, Dyson paid a visit to Enrico Fermi[1] (recounted in Ditley, Mayer, and Loew[2]). Dyson wanted to tell Fermi about a set of calculations that he was quite excited about. Fermi asked Dyson how many parameters needed to be tuned in the theory to match experimental data. When Dyson replied there were four, Fermi shared with Dyson a favorite adage of his that he had learned from Von Neumann: "with four parameters I can fit an elephant, and with five I can make him wiggle his trunk." Dejected, Dyson took the next bus back to Ithaca.

As scientists, we are frequently in a similar position to Dyson. We are often confronted with a model—a heavily parameterized, possibly incomplete or inaccurate mathematical representation of nature—rather than a theory (e.g., the Navier-Stokes equations) with few to no free parameters to tune. In recent decades, fueled by advances in computing capabilities, the size and scope of mathematical models have exploded. Massive complex models describing everything from biochemical reaction networks to climate to economics are now a centerpiece of scientific inquiry. The complexity of these models raises a number of challenges and questions, both technical and profound, and demands development of new statistical and computational tools to effectively use such models.

Here, we review several developments that have occurred in the domain of sloppy model research. *Sloppy* is the term used to describe a class of complex models exhibiting large parameter uncertainty when fit to data. Sloppy models were initially characterized in complex biochemical reaction networks,[3,4] but were soon afterward found in a much larger class of phenomena including quantum Monte Carlo,[5] empirical atomic potentials,[6] particle accelerator design,[7] insect flight,[8] and critical phenomena.[9]

As a prototypical example, consider fitting decay data to a sum of exponentials with unknown decay rates,

$$y(t,\theta) = \sum_{\mu} e^{-\theta_\mu t}. \tag{1}$$

We denote the vector of unknown parameters by $\theta$. These parameters are to be inferred from data, for example, by nonlinear least squares. This inference problem is notoriously difficult.[10] Intuitively, we can understand why by noting that the effect of each *individual* parameter is obscured by our choice to observe only the sum. Parameters have compensatory effects relative to the system's *collective* behavior. A single

 

decay rate can be decreased, for example, provided other rates are appropriately increased to compensate.

This uncertainty can be quantified using statistical methods, as we detail in Section II. In particular, the Fisher Information Matrix (FIM) can be used to estimate the uncertainty in each parameter in our model. The result for the sum of exponentials is that each parameter is *almost completely undetermined*. Any parameter can be varied by an infinite amount and the model could still fit the data. This does not mean that all parameters can be varied independently of the others. Indeed, while the statistical uncertainty in each individual parameter might be infinite, the data place constraints on *combinations* of the parameters.

The eigenvalues of the FIM tell us which parameter combinations are well-constrained by the data and which are not. Most of the FIM eigenvalues are very small, corresponding to combinations of parameters that have little effect on model behavior. These unimportant parameter combinations are designated *sloppy*. A small number of eigenvalues are relatively large, revealing the few parameter combinations that are important to the model (known as *stiff*). It is generally observed that the FIM eigenvalues decay roughly log-linearly, with each parameter combination being less important than the previous by a fixed factor, as in Figure 1. Consequently, there is not a well-defined boundary between the stiff and sloppy combinations, and four parameters really can "fit the elephant."

The degree of parameter indeterminacy in the simple sum-of-exponentials model has been seen in many complex models of real life systems for many of the same reasons. The FIMs for 17 systems biology models have been shown to have the same characteristic eigenvalue structure,[12] and examples from other scientific domains abound.[5] In each case, observations measure a system's collective behavior, and this means that when parameters have compensatory effects, they cannot be individually identified.

The ubiquity of sloppiness would seem to limit the usefulness of complex parameterized models. If we cannot accurately know parameter values, how can a model be predictive?

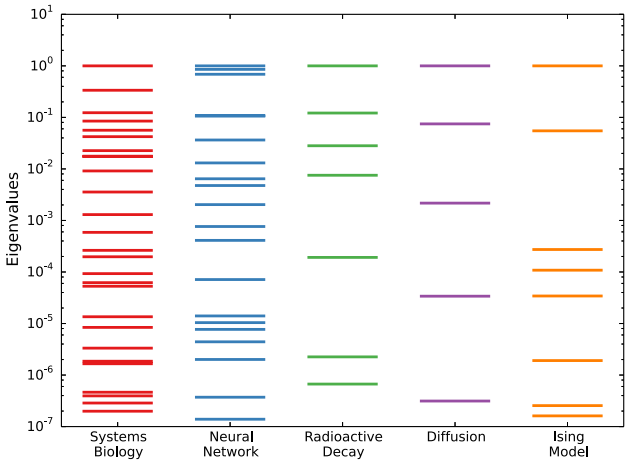


FIG. 1. Sloppy eigenvalue spectra of multiparameter models from various fields.[3–5,9,11] Eigenvalues of the FIM, indicating sensitivity to perturbations along orthogonal directions in parameter space, are roughly evenly spaced in log-space, extending over many orders of magnitude. Reprinted with permission from Machta *et al.,* Science **342**, 604-607 (2013). Copyright 2013 by AAAS.

Surprisingly, predictions are possible without precise parameter knowledge. As long as the model predictions depend on the same stiff parameter combinations as the data, the predictions of the model will be constrained in spite of large numbers of poorly determined parameters.

The existence of a few stiff parameter combinations can be understood as a type of *low effective dimensionality* of the model. In Section III, we make this idea quantitative by considering a geometric interpretation of statistics. This leads naturally to a new method of model reduction that constructs low-dimensional approximations to high-dimensional models (Section IV). These approximations are useful for extracting a simple emergent theory of the collective behavior from the larger, complex model.

Simple approximations to complex processes are common in physics (Section V). The ubiquity of sloppiness suggests that similarly simple models can be constructed for other complex systems. Indeed, sloppiness provides a number of new insights into the unreasonable effectiveness of mathematics[13] and the hierarchical structure of scientific theories.[14] We discuss some of these consequences specifically for modeling biochemical networks in Section VI. We discuss more generally the implications of sloppiness for mathematical modeling in Section VII. We argue that sloppiness is the underlying reason why the universe (a complete description of which would be indescribably complex) is comprehensible, i.e., well-approximated by relatively simple mathematical constructions.

## II. MATHEMATICAL FRAMEWORK

In this section, we use information theory to define key measures of sloppiness geometrically.[15] We first consider the special case of model selection for models fit to data by least squares. We then generalize to the case of arbitrary probabilistic models. The key insight is that the Fisher information defines a Riemannian geometry on the space of possible models.[15] The geometric picture allows us to show (in Section III) that this local sloppy structure in the metric is paralleled by a global hyper-ribbon structure of the entire space of possible models.

We begin with a simple case—a model $\mathbf{y}$ predicting data $\mathbf{d}$ at experimental conditions $\mathbf{u}$, with independent Gaussian errors; each of these are vectors whose length $M$ is given by the number of data points. Our model depends on $N$ parameters $\boldsymbol{\theta}$. In general, an arbitrary model is a mathematical mapping from a parameter space into predictions, so interpreting a model as a manifold of dimension $N$ embedded in a data space $\mathbb{R}^M$ is natural; the parameters $\boldsymbol{\theta}$ then become the coordinates for the manifold. If our error bars are independent and Gaussian all with the same width (say, $\sigma = 1$), finding the best fit of model to data is a least squares data fitting problem, as we illustrate in Figure 2. In this case, we assume that each experimental data point, $d_i$, is generated from a parameterized model, $y(u_i, \theta)$, plus random Gaussian noise, $\xi_i$,

$$d_i = y(u_i, \theta) + \xi_i. \tag{2}$$

Since the noise is Gaussian,

$$P(\xi) \propto e^{-\xi^2/2}, \tag{3}$$

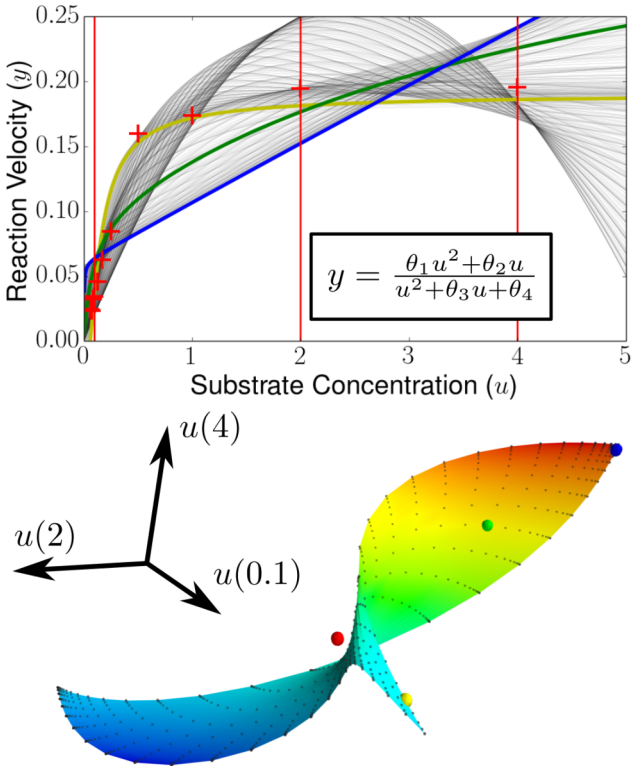


FIG. 2. The model manifold: a simple model[16,17] of an enzyme-catalyzed reaction can be expressed as a rational function in substrate concentration ($u$) with four parameters ($\theta$) predicting the reaction velocity ($y$) (inset, top). By varying $\theta$, the model can predict a variety of behaviors $y$ as a function of $u$ (top). The model manifold is constructed by collecting all possible predictions of the model at specific values of $u$ (red vertical lines at $u = 0.1, 2.0, 4.0$). To visualize the manifold, we take a two-dimensional cross section of the four dimensional manifold by choosing $\theta_1$ and $\theta_2$ to best fit the experimental data. Varying $\theta_3$ and $\theta_4$ then maps out a two-dimensional surface of possible values in three-dimensional data space (bottom). Each curve in the top figure corresponds to a point of the same color on the model manifold (bottom); the red crosses on top are data corresponding to the red dot below.

maximizing the log likelihood is equivalent to minimizing the sum of squared residuals, sometimes referred to as the cost or $\chi^2$ function,

$$\chi^2(\theta) = \sum_i r_i^2 = \sum_i (d_i - y(u_i, \theta))^2. \tag{4}$$

A sum of squares is reminiscent of a Euclidean distance. Fitting a model to data by least squares is therefore minimizing a distance in data space between the observed data and the model. Distance in data space measures the quality of a fit to experimental data (red point in Figure 2). Distance on the manifold is induced by, i.e., is the same as, the corresponding distance in data space and is measured using the metric tensor,[11,18]

$$g_{\mu\nu} = \sum_i \frac{\partial y(u_i,\theta)}{\partial \theta^\mu} \frac{\partial y(u_i,\theta)}{\partial \theta^\nu} = (J^T J)_{\mu\nu}, \tag{5}$$

where $J_{i\mu} = \partial y(u_i,\theta)/\partial\theta^\mu$ is the Jacobian matrix of partial derivatives. This metric tensor is precisely the well-known FIM defined below, specialized to our least-squares problem. The FIM plays a key role in optimal experimental design[19] and the selection of a particular kind of non-informative Bayesian prior.[20] The matrix in Eq. (5) is also equal to the least squares Hessian matrix,[3] measuring sensitivity of the fit to changes in parameters using second derivatives of $\frac{1}{2}\chi^2$ from Eq. (4), evaluated where the data point $\mathbf{d}$ is taken to be perfectly predicted by $\mathbf{y}(\boldsymbol{\theta})$. On the manifold, the least-squares distance between two alternate models is a measure of identifiability—how difficult it would be to distinguish nearby points on the manifold through their predictions.

This general approach to identifiability allows us to discuss multiparameter models for systems with non-Gaussian distributions or error estimates that vary with parameters. This can include the extreme case of models (like the Ising model[9]) that predict entire probability distributions. For the purpose of modeling, the output of our model is a probability distribution for $x$, the outcome of an experiment. A parameterized space of models is thus defined by $P(x|\theta)$. To define a geometry on this space, we must define a measure of how distinct two points $\theta_1$ and $\theta_2$ in parameter space are, based on their *predictions*.[21]

Imagine getting a sequence of assumed independent data $x_1, x_2, \ldots$ with the task of inferring the model which produced them. The likelihood that model $\theta$ would have produced these data is given by

$$P(x_1, x_2, \ldots|\theta) = \prod_i P(x_i|\theta) = \exp\left(\sum_i \log P(x_i|\theta)\right). \tag{6}$$

In maximum likelihood estimation, our goal is simply to find the parameter set $\theta$ which maximizes this likelihood. It is useful to talk about $\log P(x|\theta)$, the log-likelihood, as this is the unique measure which is additive for independent data points. The familiar Shannon entropy of a model's predictions $x$ is given by minus the expectation value of the log-likelihood,

$$S(\theta) = -\sum_x P(x|\theta) \log P(x|\theta). \tag{7}$$

The Shannon entropy is the average likelihood of the data being generated from the model. We can define an analogous quantity that measures the likelihood that model $\theta_2$ would produce typical data from $\theta_1$,

$$\sum_x P(x|\theta_1) \log P(x|\theta_2). \tag{8}$$

The Kullback-Leibler divergence between $\theta_1$ and $\theta_2$ measures how more likely $\theta_1$ is to produce typical data from $\theta_1$ than $\theta_2$ would be,

$$D_{\mathrm{KL}}(\theta_1||\theta_2) = \sum_x P(x|\theta_1)\big(\log P(x|\theta_1) - \log P(x|\theta_2)\big). \tag{9}$$

Thus, $D_{\mathrm{KL}}$ is an intrinsic measure of how difficult distinguishing these two models will be from their data.

The KL divergence does not satisfy the mathematical requirements of a distance measure. It is asymmetric and does not satisfy even a weak triangle inequality: in some cases, $D_{\mathrm{KL}}(\theta_1||\theta_3) > D_{\mathrm{KL}}(\theta_1||\theta_2) + D_{\mathrm{KL}}(\theta_2||\theta_3)$. However, for models whose parameters $\theta$ and $\theta + \delta\theta$ are quite close to one another, the leading terms are symmetric and can be written as

$$D_{\mathrm{KL}}(\theta||\theta + \delta\theta) = g_{\mu\nu}\delta\theta^\mu\delta\theta^\nu + O\delta\theta^3, \tag{10}$$

where $g_{\mu\nu}$ is the FIM, which can be written as

$$g_{\mu\nu}(P_\theta) = -\sum_x P_\theta(x) \frac{\partial}{\partial\theta^\mu} \frac{\partial}{\partial\theta^\nu} \log P_\theta(x), \tag{11}$$

providing the generalization of Eq. (5) to arbitrary $P_\theta(x)$. The FIM has all the properties of a metric tensor. It is symmetric and positive semi-definite (because no model can on average be better described by a different model) and it transforms properly under a coordinate reparameterization of $\theta$. Information geometry[11,18,22–27] is the study of the properties of the model manifold defined by this metric. In particular, it defines a space of models in a way that does not depend on the labels given to the parameters, which are presumably arbitrary; should one measure rate constants in seconds or hours, and more problematically, should one label these constants as rates or time constants? Information geometry makes clear that some aspects of a parameterized model can be defined in ways that are invariant to these arbitrary choices. Independence to reparameterization is often required of both computational and theoretical methods to guarantee consistency and robustness.[28–31]

## III. WHY SLOPPINESS? INFORMATION GEOMETRY

We noted previously that the characteristic eigenvalue spectrum of the FIM suggests a simpler, lower-dimensional "theory" embedded within larger, more complex "models," and in this section, we make this notion explicit. We will see that although this interpretation of sloppy models turns out to be correct, the eigenvalues of the FIM are not sufficient to make this conclusion. Instead, we use the geometric interpretation of modeling introduced in Section II that allows us to quantify important features of the model in a global and parameterization independent way. The effort to develop this formalism will pay further dividends when we consider model reduction in Section IV.

To understand the limitations of interpreting the eigenvalues of the FIM, we return to the question of model reparameterization. Something as trivial as changing the units of a rate constant from Hz to kHz changes the corresponding row and column of the FIM by a factor of 1000, in turn changing the eigenvalues. Of course, none of the model predictions are altered by such a change since a correcting factor of 1000 will be introduced throughout the model. More generally, the FIM can be transformed into any positive definite matrix by a simple linear transformation of parameters while model predictions are always invariant to such a reparameterization.

Although the FIM eigenvalues are not invariant to reparameterization, we can use information geometry to search for a parameterization independent measure of sloppiness. Specifically, the key geometric feature of the model manifolds of nonlinear sloppy systems is that they have *boundaries*. Many parameters and parameter combinations can be taken to extreme values (zero or infinity) without leading to infinite predictions.

These boundaries can be explored in a parameter independent way using geodesics. Geodesics are the analogs of straight lines on curved surfaces. They are one-dimensional curves through parameter-space that are constructed numerically as the solution of a differential equation using the methods of computational differential geometry. A review of these methods is beyond the scope of this paper, and we refer the interested reader to Refs. 18 and 11 or any standard text on differential geometry.[32,33] The geodesic curve in parameter space corresponds to a curve on the model manifold. The arc lengths of geodesics on the manifold are a measure of the manifold width in each direction. Measuring these arc lengths for a sloppy model shows that the widths of sloppy model manifolds are exponentially distributed, reminiscent of the exponential distribution of FIM eigenvalues. Indeed, when we use dimensionless model parameters (e.g., log-parameters), the square roots of the FIM eigenvalues are a reliable approximation to the widths of the manifold in the corresponding eigendirections.[11,18]

The exponential distribution of manifold widths has been described as a *hyperribbon*[11,18] (Fig. 3). A three-dimensional ribbon has a long dimension, a broad dimension, and a very thin dimension. The observed hierarchy of exponentially decreasing manifold widths is a high-dimensional generalization of this structure. We will explore the nature of these boundaries in more detail when we discuss model reduction in Section IV.

The observed hierarchy of widths can be demonstrated analytically for the case of a single independent variable (such as time or substrate concentration in Figure 2) by appealing to theorems for the convergence of interpolating functions (Fig. 3(a)). Consider removing a few degrees of freedom from a time series by fixing the output of the model at a few time points. The resulting model manifold corresponds to a cross section of the original. Next, consider how much the predictions at intermediate time points can be made to vary as the remaining parameters are scanned. As more and more predictions are fixed (i.e., considering higher-dimensional cross sections of the model manifold), we intuitively predict that the behavior of the model at intermediate time points will become more constrained. Interpolation theorems make this intuition formal; presuming smoothness or analyticity of the predictions as a function of time, one can demonstrate an exponential hierarchy of widths consistent with the hyperribbon structure observed empirically.[11,18] This argument is illustrated in Fig. 3(b). Briefly, an analytic function $f(t)$ with radius of convergence $R$ allows one to predict the behavior at a neighboring point $t+\epsilon$ using $n$ derivatives with error going as $(\epsilon/R)^n$. Interpolation theory tells us that one can use values $f(t_1),\ldots,f(t_n)$ separated by $\sim\Delta t$ to predict values with errors converging as $(\Delta t/R)^n$.[18,34] Hence, each successive data point (cross section) becomes thinner by a constant factor $\Delta t/R$, giving us our hyperribbon structure.

The exponential hierarchy of manifold widths reflects a low effective dimensionality in the model, which was hinted at by the eigenvalues of the FIM. It also helps illustrate how models can be predictive without parameters being tightly constrained. Only those parameter combinations that are required to fit the key features of the data need to be estimated accurately. The remaining parameter combinations (controlling, for example, the high-frequency behavior in our time series example) are unnecessary. In short, the model essentially functions as an interpolation scheme among observed data points. Models are predictive with unconstrained parameters when the predictions interpolate among observed data.

Understanding models as generalized interpolation schemes makes additional predictions about the generic structure of sloppy model manifolds. There is not only an exponential

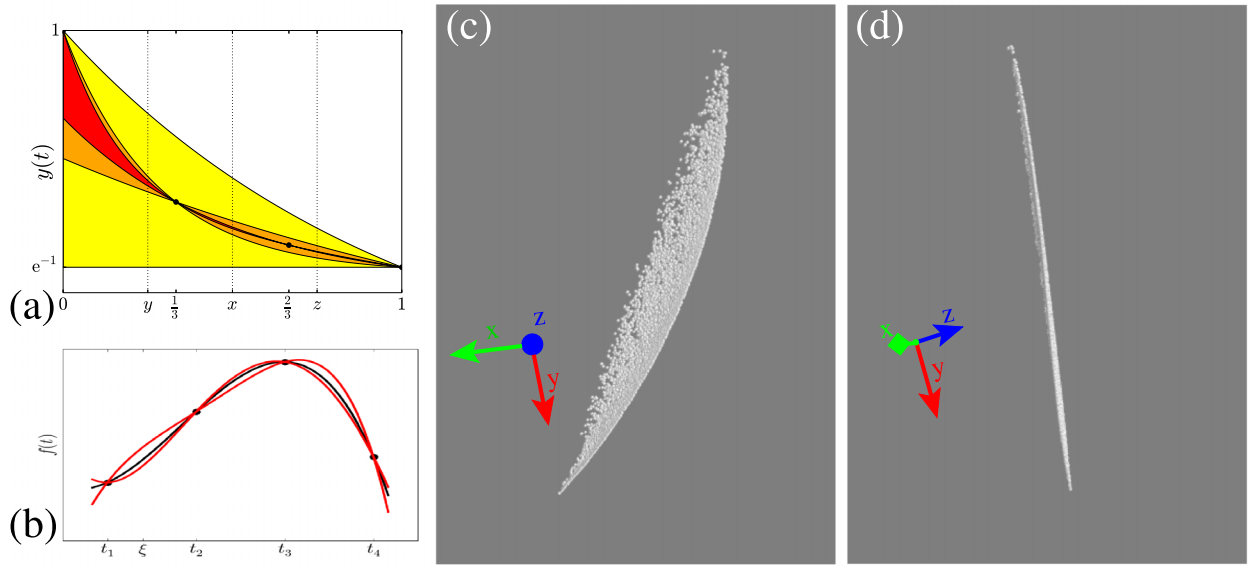


FIG. 3. Hyperribbon. (a), (c), and (d) are visualizations of a model for radioactive decay (a sum of exponentials), a famously ill-posed fitting problem.[5] (a) Given a multiparameter model for one-dimensional data $y(t)$ at different times $t$, the model manifold has a different dimension for every time $t$. Here, we successively specify the values at times $t = 0, 1$ (yellow), $\frac{1}{3}$ (orange), and $\frac{2}{3}$ (red); $f(0) = 1$, $f(1) = 1/e$, and successive points are constrained to the centers of their respective possible ranges. Specifying each data point $y(t_n)$ can also be viewed in data space as giving a cross section of the model manifold—each successive data point giving the later cross section narrower widths, implying a hyperribbon structure.[18] (b) Interpolation theory[11,18] can be used to quantify this qualitative argument, showing that each cross section of an analytic function $f(t)$ reduces the width by an approximately constant factor. (Reprinted with permission from Fig. 5 of M. K. Transtrum, B. B. Machta, and S. P. Sethna, Phys. Rev. E **83**, 036701 (2011). Copyright 2011 by American Physical Society.) (c) and (d) Two views of a hyperribbon cross section of a model manifold, with the values $y(t)$ at $t = 0, \frac{1}{3}, \frac{2}{3}$, and 1 set to measured values as in part (a). The $xyz$ coordinates are the values $y(\frac{1}{2})$, $y(\frac{1}{4})$, $y(\frac{3}{4})$; the width of the red band in (a) is the projection of the model manifold onto the corresponding axis. Notice the ribbon-like structure of these two projections: (c) long and narrow and (d) very thin. Notice that total range of $z$ in (a) is much larger than the range of $z$ given $x$ and $y$ depicted in (d); the addition of two more data points constrains $z$ more tightly.

distribution of widths but also an exponential distribution of extrinsic curvatures. Furthermore, these curvatures are relatively small in relation to the widths, making the model manifold surprisingly flat. Most of the nonlinearity of the model's parameters takes the form of "parameter effects curvature"[23,35–37] (equivalent to the connection coefficients[11]). The small extrinsic curvature of many models was a mystery first noted in the early 1980s,[23] which is explained by interpolation arguments.

## IV. MODEL REDUCTION

In this section, we leverage the power of the information geometry formalism to answer the question: how can a simple effective model be constructed from a (more-or-less) complete but sloppy representation of a physical system? Our goal is to construct a physically meaningful representation that reveals the simple "theory" that is hidden in the model.

The model reduction problem has a long history, and it would be impossible in this review to even approach a comprehensive survey of literature on the subject. Several standard methods have emerged that have proven effective in appropriate contexts. Examples include clustering components into modules,[38–40] mean field theory, various limiting approximations (e.g., continuum, thermodynamic, or singular limits), and the renormalization group (RG).[41,42] Considerable effort has been devoted by the control and engineering communities to approximate large-scale dynamical systems,[43–47] leading to the method of balanced truncation,[48–50] including several structure preserving variations[51–53] and generalizations to nonlinear cases.[54–56] Methods for inferring minimal dynamical models in cases for which the underlying structure is not known are also beginning to be developed.[57,58]

Unfortunately, many automatic methods produce "black box" approximations. For most scenarios of practical importance, a reduced representation alone has limited utility since attempts to engineer or control the system typically operate on the microscopic level. For example, mutations operate on individual genes and drugs target specific proteins. A method that explicitly reveals how microscopic components are "compressed" into a few effective degrees of freedom would be very useful. On the other hand, methods that do explicitly connect microscopic components to macroscopic behaviors have limited scope since they often exploit special properties of the model's functional form, such as symmetries. Consider, for example, the renormalization group, which operates on field theories with a scale invariance or conformal symmetry. Simplifying modular network systems, such as biochemical networks, is particularly challenging due to inhomogeneity and lack of symmetries.

The Manifold Boundary Approximation Method (MBAM)[59] is an approach to model approximation whose goal is to alleviate these challenges. As the name implies, the basic idea is to approximate a high-dimensional, but thin, model manifold by its boundary. The procedure can be summarized as a four-step algorithm. First, the least sensitive parameter combination is identified from an eigenvalue decomposition of the FIM. Second, a geodesic on the model manifold is constructed numerically to identify the boundary. Third, having found the edge of the model manifold, the corresponding

model is identified as an approximation to the original model. Fourth, the parameter values for this approximate model are calibrated by fitting the approximate model to the behavior of the original model.

The result of this procedure is an approximate model that has one less parameter and that is marginally less sloppy than the original. Iterating the MBAM algorithm therefore produces a series of models of decreasing complexity that explicitly connect the microscopic components to the macroscopic behavior. These models correspond to hyper-corners of the original model manifold. The MBAM requires that the model manifold is bounded with a hierarchy of boundaries (faces, edges, corners, hyper-corners, etc.). It makes no assumptions about the underlying physics or mathematical form of the model. As such, MBAM is a very general approximation scheme.

The key component that enables MBAM is the edges of the model manifold. The existence of these edges was first noted in the context of data fitting[18] and Markov Chain Monte Carlo (MCMC) sampling of Bayesian posterior distributions.[7] It was noted that algorithms would often "evaporate" parameters, i.e., allow them to drift to extreme, usually infinite, values. These extreme parameter values correspond to limiting behaviors in the model, i.e., manifold boundaries.

"Evaporated parameters" are especially problematic for numerical algorithms. Numerical methods often push parameters to the edge of the manifold and then become lost in parameter space. Consider the case of MCMC sampling of a Bayesian posterior. If a parameter drifts to infinity, there is an infinite amount of entropy associated with that region of parameter space and the sampling will never converge. Furthermore, the model behavior of such a region will always dominate the posterior distribution.[7] Because the entropic contributions from these evaporated parameters are essentially infinite, they can overwhelm a small, nonzero factor from the likelihood. The result is a posterior distribution that is dominated by unlikely parameter values, is inconsistent with the observed data, and makes poor predictions for new experiments.

For data fitting algorithms, methods such as Levenberg-Marquardt operate by fitting the key features of the data first (i.e., the stiffest directions), followed by successive refining approximations (i.e., progressively more sloppy components). While fitting the initial key features, algorithms often evaporate those parameters associated with less prominent features of the data. The algorithm is then unable to bring the parameters away from infinity in order to further refine the fit.[18]

Although problematic for numerical algorithms, manifold edges are useful for both approximating (as we have seen using the MBAM) and interpreting complex models. To illustrate, we consider an Epidermal Growth Factor Receptor (EGFR) signaling model.[3] Figure 4 illustrates components of one eigenparameter, corresponding in this case to the smallest eigenvalue of the FIM. Notice that the eigenparameters do not align with bare parameters of the model, but typically involve an unintuitive combination of bare parameters. However, by following a geodesic along the model manifold to the manifold edge (step 2 of the MBAM algorithm), these complex combinations slowly rotate to reveal relatively simple, interpretable combinations that correspond to a limiting approximation of

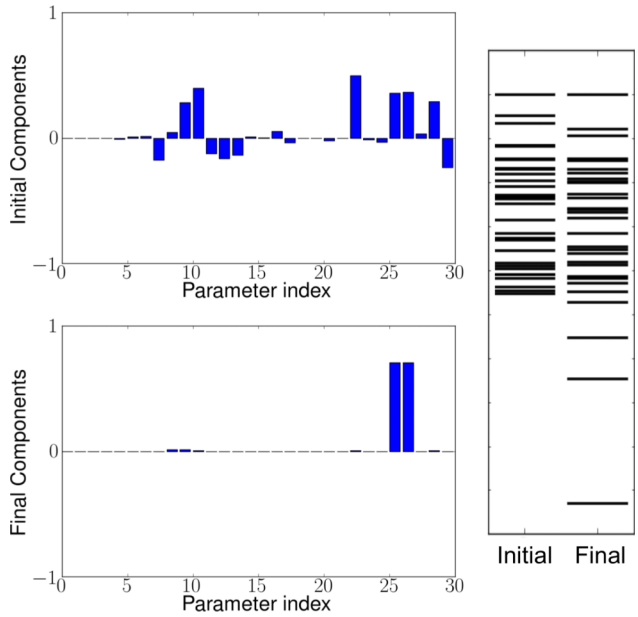


FIG. 4. Identifying the boundary limit.[59] The components of the smallest eigenvector of the FIM are often a complicated combination of bare parameters that is difficult to either interpret or remove from the model (top left). By following a geodesic to the manifold boundary, the combination rotates to reveal a limiting behavior (bottom left); here, two parameters (a reaction rate and a Michaelis-Menten constant) become infinite. The limiting behavior is revealed when the smallest eigenvalue has become separated from the other eigenvalues (right).

the model. For example, the EGFR model in Ref. 3 consists of a network of Michaelis-Menten reactions. The boundary revealed[59] in Figure 4 corresponds to the limit of a reaction rate and a Michaelis-Menten constant becoming infinite while their ratio is finite,

$$\frac{d}{dt}[A] = \frac{k[A][B]}{K_M + [A]} + \cdots \tag{12}$$

$$\rightarrow \left(\frac{k}{K_M}\right)[A][B] + \cdots, \tag{13}$$

where $[A]$ and $[B]$ are concentrations of two enzymes in the model and the ratio $k/K_M$ is the renormalized parameter in the approximate model.

Because the manifold edges correspond to models that are simple approximations of the original, the MBAM can be used to iteratively construct simple representations of otherwise complex processes. By combining several limiting approximations, simple insights into the system behavior emerge, which were obfuscated by the original model's complexity. Figure 5 compares network diagrams for the original and approximate EGFR models. The original consists of 29 differential equations and 48 parameters, while the approximate consists of 6 differential equations and 12 parameters and is notably *not sloppy*.

Because the MBAM process explicitly connects models through a series of limiting approximations, the parameters of the reduced model can be identified with (nonlinear) combinations of parameters in the original model. For example, one of the 12 variables in the reduced model of Fig. 5 is written as an explicit combination of 7 "bare" parameters of the original model,

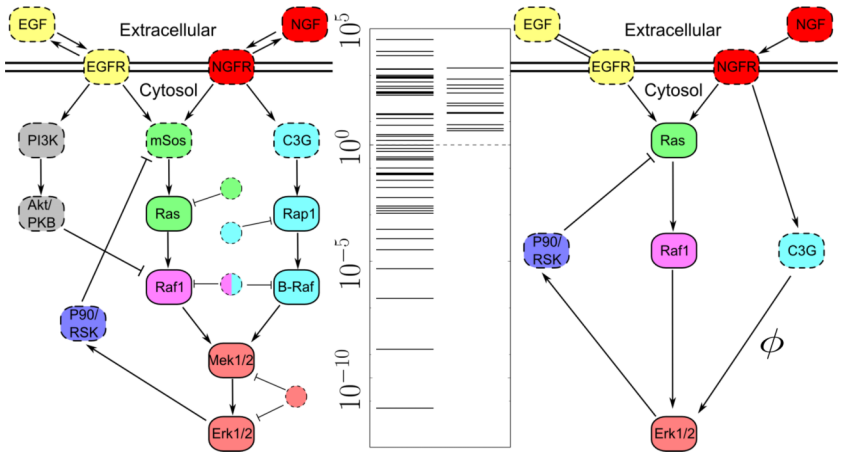


FIG. 5. Original[3] and reduced[59] EGFR models. Reprinted with permission from M. K. Transtrum and P. Qiu, Phys. Rev. Lett. **113**, 098701 (2014). Copyright 2014 by American Physical Society. The interactions of the EGFR signaling pathway[3,4] are summarized in the leftmost network. Solid circles are chemical species for which the experimental data were available to fit. Manifold boundaries reduce the model to a form (right) capable of fitting the same data and making the same predictions as in the original references.[3,4] The FIM eigenvalues (center) indicate that the simplified model has removed the irrelevant parameters identified as eigenvalues less than 1 (dotted line) while retaining the original model's predictive power.

$$\phi = \frac{(k_{Rap1ToBRaf})(K_{mdBRaf})(k_{pBRaf})(K_{mdMek})}{(k_{dBRaf})(K_{mRap1ToBRaf})(k_{dMek})}. \qquad (14)$$

Expressions such as this explicitly identify which combinations of microscopic parameters act as emergent control knobs for the system.

MBAM naturally includes many other approximation methods as special cases.[59] By an appropriate choice of parameterization, it is both a natural language for model reduction and a systematic method that in practice can be mostly automated.

The MBAM is a semi-global approximation method. Manifold boundaries are a non-local feature of the model. However, MBAM only explores the region of the manifold in the vicinity of a single hyper-corner. More generally, it is possible to identify *all* of the edges of a particular model (and by extension, all possible simplified models). This analysis is known as information topology.[60]

## V. SLOPPINESS IN PHYSICS: EMERGENCE AND EFFECTIVE THEORIES

Unlike in systems biology, physics is dominated by effective models and theories whose forms are often deduced long before a microscopic theory is available. This is in large part due to the great success of continuum limit arguments and RG procedures in justifying the expectation and deriving the form of simple emergent theories. These methods show that many different multi-parameter microscopic theories typically collapse onto one coarse-grained model, with the complex microscopic details being summarized into just a few "relevant" coarse-grained parameters. This explains why an effective theory, or an oversimplified "cartoon" microscopic theory, can often make quantitatively correct predictions. Thus, while three dimensional liquids have enormous microscopic diversity, in a certain regime (lengths and times large compared to molecules and their vibration periods), their behavior is determined entirely by their viscosity and density. Although two different liquids can be microscopically completely different, their effective behavior is determined only by the projection of their microscopic details onto these two control parameters. This parameter space compression underlies the success of renormalizable and continuum limit models.

This connection has been made explicit, by examining the FIM for typical microscopic models in physics.[9] A microscopic hopping model for the continuum diffusion equation quickly develops "stiff" directions corresponding to the parameters of the continuum theory—the total number of particles, net mean velocity, and diffusion constant. As time proceeds, all other microscopic parameter combinations become increasingly irrelevant for the prediction of long-time behavior. Similarly, a microscopic long-range Ising model for ferromagnetism, when observed on long length scales, develops stiff directions along precisely those parameter combinations deemed "relevant" under the renormalization group.

Consider a model of stochastic motion as a stand-in for a molecular level description of particles moving through a possibly complicated fluid. Such a fluid's properties depend on many parameters such as the bond angle of the molecules which make it up, all of which enter into the probability distribution for motion within the fluid. However, the law of large numbers says that as many of these random steps are added together, the long-time movement of particles will lead to them being distributed in space according to a Gaussian. As this happens, diverse microscopic details must become compressed into the two parameters of a Gaussian distribution—its mean and width. As a concrete example, in the top of Figure 6, two very different microscopic motions are considered. In each time step, red particles take a random step from a triangular distribution, while blue particles step according to a square distribution. While these motions lead to very different distributions after a single time step, as time proceeds they become indistinguishable precisely because their first and second moments are matched.

This indistinguishability can be quantified by considering the model manifold of possible microscopic models of stochastic motion. When probed at the intrinsic time and length scales of these fluids, we should make few assumptions about the type of motion we expect; in particular, we should allow for behaviors more complicated than diffusion, by analogy with the square and triangle described in two dimensions above. Following Ref. 9, we consider a one dimensional "molecular level"

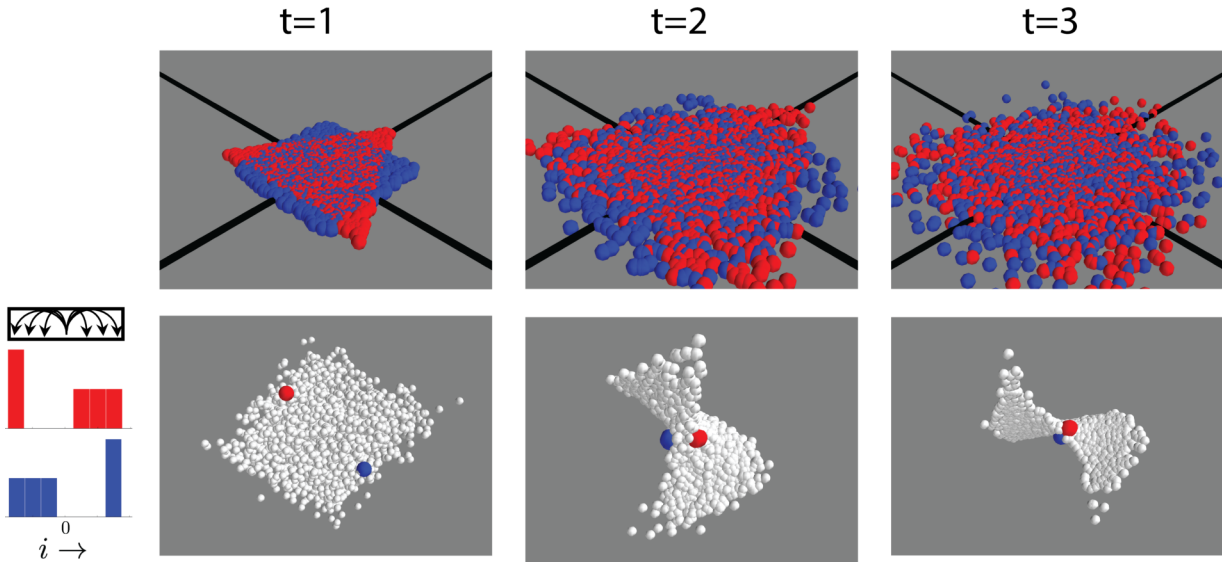


FIG. 6. Microscopic motion becomes diffusive. Top: Simulated particles undergo stochastic motion in discrete time. Red particles hop according to a triangular kernel, while blue particles hop according to a square kernel. After a single time step, the particles have very different distributions in space, and neither resemble the distribution predicted by the diffusion equation. However, as time evolves, most of the information about this kernel is lost, and only the diffusion tensor and average drift enter into a continuum description. Here, the drift is 0, and their respective diffusion tensors are matched, so that the resulting distributions become quantifiably indistinguishable as time proceeds. The compression of microscopic details mirrors the compression of molecular level detail in the emergence of diffusion as a continuum limit of motion in real fluids. Bottom: A three dimensional projection of the model manifold of a one dimensional lattice version of this diffusion example as in Ref. 9. Six parameters describe the probability of hopping to one of 6 nearby neighbor sites (shown pictorially in arrows at left). The red and blue kernels shown on the left have drift 0 and identical second moments, though higher moments are very different. Other (white) points are taken from a uniform distribution in the 6 dimensional parameter space which is bounded only by the constraint that no hopping probabilities are negative and that the total probability of a hop sums to a number less than 1. After a single time step, the model manifold resembles a "hyper-blob," with a large diversity of behaviors. In particular, the red and blue points are not initially close to each other. However, after several stochastic steps, the model manifold takes on a hyperribbon structure. Models for which all effective parameters are matched, like the red and blue kernels, rapidly move close to each other. At late times, any model sufficiently flexible to capture the two remaining extended directions is adequate to describe effective behavior, explaining the ubiquity and success of the continuum diffusion equation.

model for stochastic motion, in which parameters describe the rates at which a particle hops to one of its close-by neighbors. After a single time step, the corresponding model manifold is a "hyper-blob" (Fig. 6, bottom) and two particular models, marked in red and blue, are distinguishable; they are not nearby on the model manifold. The prediction space of a model is truly multidimensional in this regime—it cannot be described by the two parameter diffusion equation. In this "ballistic" regime, motion is not described by the diffusion equation and is presumably not just different, but genuinely more complicated. However, as time proceeds, the model manifold contracts onto a hyper-ribbon, in which just two parameter combinations distinguish behavior. In this regime, all points lie close to the two dimensional manifold predicted by the diffusion equation, and the red and blue points have become indistinguishable; they are now in close proximity on the manifold.

Using information geometry, approximations analogous to continuum limits or the renormalization group can be found and used to construct similarly simple theories in fields for which effective theories have historically been difficult to construct or justify.

## VI. SLOPPINESS IN SYSTEMS BIOLOGY: PARAMETER ESTIMATION AND "WRONG" MODELS

Unlike in physics, where the value of effective theories has long been recognized, the field of systems biology has focused on the development of detailed microscopic models and has wrestled with the associated challenges of parameter estimation. In Secs. II–IV, we have highlighted the structure of the model manifold in data space, as in Figure 2: thin, sloppy dimensions of the hyperribbon correspond to behavior that is minimally dependent on parameters. The dual picture in parameter space, sketched in Figure 7, is one in which the set of parameters that fits some given data sufficiently well is stretched to extend far along sloppy dimensions. This picture is important to understanding implications for biochemical modeling with regard to parameter uncertainty.

For instance, using the full EGFR signal transduction network (left side of Figure 5), we may wish to make a prediction about an unmeasured experimental condition, e.g., the time-course of Extracellular signal-Regulated Kinases (ERK) activity upon EGF stimulation. In general, if there are large uncertainties about the model's parameters, we expect our uncertainty about this time-course to also be large. If we view the problem of uncertainties in model predictions as coming from a lack of precision measurements of individual parameters, we may try to carefully and independently measure each parameter. This can work if such measurements are feasible, but can fail if even one relevant parameter remains unknown: as in the right plot of Figure 7, a large uncertainty along the direction of a single parameter corresponds to large changes in system-level behavior, leading to large predictive uncertainties.[12]

In contrast, we can instead constrain the model parameters with system-level measurements that are similar to the types of measurements we wish to predict. Due to sloppiness, we expect

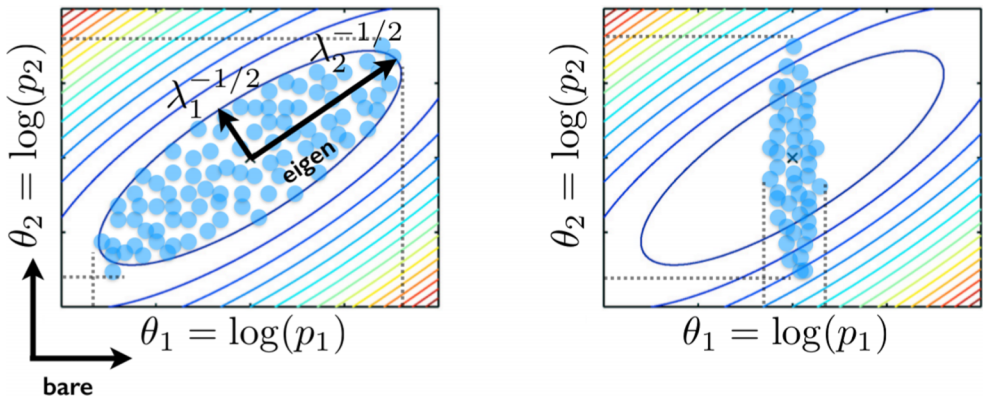


FIG. 7. Sloppiness in parameter space. Left: A schematic of a typical sloppy model ensemble, pictured in two dimensions for clarity. The underlying cost surface (with constant cost contours illustrated as ellipses) is generated by the fit to data. Eigenvectors of the FIM correspond to principal axes of the ellipse, with widths of the ellipse inversely proportional to the square roots of the corresponding eigenvalues $\lambda_i$. Points inside the ellipse each represent a set of parameters that fits the data within a given tolerance (in practice often created using a Monte Carlo approach), forming an ensemble representing uncertainty about the true values of parameters. Sloppiness can result in good fits to data despite enormous uncertainties in "bare" $\theta$ parameters (dotted lines intersecting the axes). Right: Careful measurements of individual parameters (like $\theta_1$) can shrink uncertainty, but if even a single parameter remains unknown (like $\theta_2$), large predictive uncertainty can still result.

that this will produce a subspace of acceptable parameters that will include large uncertainties in individual parameter values (left plot of Figure 7). And somewhat counterintuitively, because of correlations among parameter sensitivities in sloppy models, useful predictions of interest can be made without precisely knowing any single parameter.

Thus, from the perspective of sloppy models, working to estimate precise parameter values in a model is useless. This does not mean that anything goes; indeed, the region of acceptable parameters may be small compared to prior knowledge about their values. Yet it does validate a common approach to modeling such systems, in which educated guesses are made for most parameters, and a remaining handful is fit to the data. In the common situation in which there are a small number $m$ of important "stiff" directions, with remaining sloppy directions extending to cover the full range of feasible parameters, fitting $m$ parameters will be enough to locate the sloppy subspace. (And if using a maximum likelihood approach, this is in fact statistically preferred to fitting all parameters, in order to avoid overfitting.) Unfortunately, it is hard to know $m$ ahead of time, in general requiring a sampling scheme like MCMC or a geodesic-following algorithm[11,18] to ascertain the global structure of the sloppy subspace. The information provided by fitting these $m$ parameters is often enough to make sharp predictions in similar experimental contexts. Note that this is not true for all possible experiments: in general, the relevant information depends on the questions being asked.

The extremely large uncertainties in parameter estimates in sloppy models led to the suggestion that accurate parameter estimates might not be possible.[12] However, advances in the theory of experimental design have suggested that such estimates might be feasible after all,[61–64] although requiring considerable experimental effort.[65] The perspective provided by sloppy model analysis provides at least two alternatives to this method of operation.

First, in spite of the large number of parameters, complex biological systems typically exhibit simple behavior that requires only a few parameters to describe, analogous to how the diffusion equation can describe microscopically diverse processes. Attempting to accurately infer all of the parameters in a complex biological model[66] is analogous to learning all of the mechanical and electrical properties of water molecules in order to accurately predict a diffusion constant. It would involve considerable effort (measuring *all* the microscopic parameters accurately[12]), while the diffusion constant can be easily measured using collective experiments and used to determine the result of any other collective experiment.

Second, in many biological systems, there is considerable uncertainty about the microscopic structure of the system. Sloppiness suggests that an effective model that is microscopically inaccurate may still be insightful and predictive in spite of getting many specific details wrong. For example, a hopping model for thermal conductivity would be "wrong" even though it gives the right thermal diffusion equation. "Wrong" models can provide key insights into the system level behavior because they share important features with the true system. In such a scenario, it is the flexibility provided by large uncertainties in the parameters that allows the model to be useful. Any attempt to infer all the microscopic parameters would break the model, preventing it from being able to fit the data.

Indeed, it is difficult to envision a complete microscopic model in systems biology. Any model will have rates and binding affinities that will be altered by the surrounding complex stew of proteins, ions, lipids, and cellular substructures. Is the well-known dependence of a reaction rate on salt concentration (described by an effective Gibbs free energy tracing over the ionic degrees of freedom) qualitatively different from the dependence of an effective reaction rate on cross talk, regulatory mechanisms, or even parallel or competing pathways not incorporated into the model? We are reminded of quantum field theories, where the properties (say) of the electron known to quantum chemistry are *renormalized* by electron-hole reactions in the surrounding vacuum which are ignored and ignorable at low energies. Insofar as a model provides both insight and correct predictions within its realm of validity, the fact that its parameters have effective, renormalized values incorporating missing microscopic mechanisms should be expected, not disparaged. We hope to bring some of the perspective

regarding effective theories and relevant degrees of freedom as understood in physics to the field of systems biology, in order to better comprehend complex, heterogeneous processes in living systems.

## VII. MORE GENERAL CONSEQUENCES OF SLOPPINESS

The hyperribbon structures implied by interpolation theory and information geometry in Section III have profound implications. Complex scientific models have predictions that vary in far fewer ways than their complexity would indicate. Multiparameter models have behavior that largely depends upon only a few combinations of microscopic parameters. The high-dimensional results of a system with a large number of control parameters will be well encompassed by a rather flat, low-dimensional manifold of behavior. In this section, we shall speculate about these larger issues, and how they may explain the success of our efforts to organize and understand our environment.

### A. Efficacy of principal component analysis

Principal component analysis, or PCA, has long been an effective tool for data analysis. Given a high-dimensional data set, such as the changes of mRNA levels for thousands of genes under several experimental conditions,[67] PCA can provide a reduced description which often retains most of the variation in the original data set in a few linear *components*. Arranging the data into a matrix $R_{jn} + c_j$ of experiments $n$ and data points $j$ centered at $c_j$, PCA uses the *singular value decomposition* (SVD),

$$R = \sum_k \sigma_k \hat{\mathbf{u}}^{(k)} \otimes \hat{\mathbf{v}}^{(k)}, \tag{15}$$

$$R_{jn} = \sum_k \sigma_k \hat{u}_j^{(k)} \hat{v}_n^{(k)}, \tag{16}$$

to write $R$ as the sum of outer products of orthonormal vectors $\hat{\mathbf{u}}^{(k)}$ in data space and $\hat{\mathbf{v}}^{(k)}$ in the space of experiments. Here, $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ are non-negative "strengths" of the different components $k$. These singular values can be viewed as a generalization of eigenvalues for non-square, non-symmetric matrices. The $\hat{\mathbf{u}}^{(k)}$ for small $k$ describe the "long axes" of the data, viewed as a cloud of points $\mathbf{R}_n$ in data space; $\sigma_k$ is the RMS extent of the cloud in direction $\hat{\mathbf{u}}^{(k)}$. The utility of PCA stems from the fact that in many circumstances only a few components $k$ are needed to provide an accurate reconstruction of the original data. Just as our sloppy eigenvalues converge geometrically to zero, the singular values $\sigma_k$ often rapidly vanish. It is straightforward to show that the truncated SVD keeping only the first, largest $K$ components is an optimal approximation to the data, with total least square error bounded by $\sum_{K+1} \sigma_k^2$. These largest singular components often have physical or biological interpretations—sometimes mundane but useful (e.g., which machine was used to take the data), sometimes scientifically central.

Why does nature often demand so few components to describe large dimensional data sets? Sloppiness provides a new perspective. If the data result from (say) a biological system whose behavior is described by a sloppy model $\mathbf{y}(\boldsymbol{\theta})$, and if the different experiments are sampling different parameter sets $\boldsymbol{\theta}_n$, then the data will be points $R_{jn} + c_j = y_j(\boldsymbol{\theta}_n)$ on the model manifold. Insofar as the model manifold has the hyperribbon structure we predict, it has only a few long axes (corresponding to stiff directions) and it is extrinsically very flat along these axes[11] (Fig. 18). Here, each $y_j - c_j$, being a difference between a data point and the center of the data, will be nearly a linear sum of a small number $K$ of long directions of the model manifold, and the RMS spread along this $k$th direction will be bounded by the width of the model manifold in that direction, plus a small correction for the curvature. As *any* cloud of experimental points must be bounded by the model manifold, the high singular values will be bounded by the hierarchy of widths of the hyperribbon. Hence, our arguments for the hyperribbon structure of the model manifold in multiparameter models provide a fundamental explanation for the success of PCA for these systems.

### B. Efficacy of Levenberg-Marquardt—Improved algorithms

The Levenberg-Marquardt algorithm[28,68,69] is one of the standard algorithms for least squares minimization. Its broad utility can be explained through the lens of sloppy models and geometric insights lead to natural improvements. Minimizing a linear approximation to a nonlinear model with a constraint on the step size,

$$\min_{\delta\theta} |y(\theta_0) + J\delta\theta - y_0|^2, \quad |\delta\theta| \leq \Delta, \tag{17}$$

leads to the iterative algorithm,

$$\delta\theta = -\left(J^T J + \lambda\right)^{-1} J^T \left(y(\theta_0) - y_0\right), \tag{18}$$

where $\lambda$ is a Lagrange multiplier. The FIM ($J^T J$) for a typical sloppy model is extremely ill-conditioned. However, the dampened scaling matrix $J^T J + \lambda$ will have no eigenvalues smaller than $\lambda$. By tuning $\lambda$, the algorithm is able to explicitly control the effects of sloppiness. Furthermore, since the eigenvalues of $J^T J$ are roughly log-linear, $\lambda$ need not be finely tuned to be effective. By slowly decreasing $\lambda$, the algorithm fits the key features of the data first (i.e., the stiffest directions), followed by successive refining approximations (i.e., progressively more sloppy components). The algorithm may still converge slowly as it navigates the extremely narrow canyons of the cost surface (see Figure 7) or fail completely if it becomes trapped near the boundary of the model manifold.[11,18]

Information geometry provides a remarkable new perspective on the Levenberg-Marquardt algorithm. The move $\delta\theta$ for $\lambda = 0$ is the direction in parameter space corresponding to the steepest descent direction in data space; for $\lambda \neq 0$, the move is the steepest descent direction on the *model graph*.[11,18] The fact that the model graph is extrinsically rather flat turns the narrow optimization valleys in parameter space into nearly concentric hyperspheres in data space—explaining the power of the method. Levenberg-Marquardt takes steps along straight lines in parameter space; to take full advantage of the flatness of the model manifold, it should ideally move along geodesics.

As it happens, the leading term in the geodesic equation is numerically cheap to calculate, providing a "geodesic acceleration" correction to the Levenberg-Marquardt algorithm which greatly improves the performance and reliability of the algorithm.[70,71]

### C. Evolution is enabled

Besides practical consequences for parameter estimation of biochemical networks (Section VI), sloppiness has potential implications for biology and evolution. Specifically, the fact that biological systems often achieve remarkable robustness to environmental perturbations may be less mysterious when taking into account the vastness of sloppy subspaces. For instance, the circadian rhythm in cyanobacteria, controlled by the dynamics of phosphorylation of three interacting Kai proteins, seems remarkable in that it maintains a 24-h cycle over a range of temperature over which kinetic rates in the system are expected to double. Yet the degree of sloppiness in the system suggests that evolution may have to tune only a few stiff parameter directions to get the desired behavior at any given temperature and perhaps only one extra parameter direction to make that behavior robust to temperature variation.[72] Extended, high-dimensional neutral spaces have been identified as a central element underlying robustness and evolvability in living systems,[73] and sloppy parameter spaces play a similar role: a population with individuals spread throughout a sloppy subspace can more easily reach a broader range of phenotypic changes, such that the population is simultaneously highly robust and highly evolvable.[72]

### D. Pattern recognition as low-dimensional representation

The pattern recognition methods we use to comprehend the world around us are clearly low-dimensional representations. Cartoons embody this: we can recognize and appreciate faces, motion, objects, and animals depicted with a few pen strokes. In principle, one could distinguish different people by patterns of scars, fingerprints, or retinal patterns, but our brains instead process subtle overall features. Caricatures in particular build on this low-dimensional representation—exaggerating unusual features of the ears or nose of a celebrity makes them more recognizable, placing them farther along the relevant axes of some model manifold of facial features. Archetypal analysis,[74] a branch of machine learning, analyzes data sets with a matrix factorization similar to PCA, but expressing data points as *convex* sums of features that are not constrained to be orthogonal. In addition, the features must be convex combinations of data points. Archetypal analysis applied to suitably processed facial image data allows faces to be decomposed into strikingly meaningful characteristic features.[74–76] One might speculate, for example, that our facial structures are determined by the effects of genetic and environmental control parameters $\boldsymbol{\theta}$ and that the resulting model manifold of faces has a hyperribbon structure, explaining the success of the linear, low-dimensional archetypal analysis methods and perhaps also the success of our biological pattern recognition skills.

### E. Big data are reducible

Machine learning methods that search for patterns in enormous data sets are a growing feature of our information economy. These methods at root discover low-dimensional representations of the high-dimensional data set. Some tasks, such as the methods used to win the Netflix challenge[77] of predicting what movie users will like, explicitly focus on dimensionality reduction through the use of methods such as PCA. More complex problems, such as digital image recognition, often make use of artificial neural networks, such as stacked denoising autoencoders.[78] Consider the problem of recognizing handwritten digits (the MNIST database). Neural networks are used to fit the data, with parameters $\theta_\alpha$ giving the outputs of the digital neurons, and the model $\mathbf{y}(\boldsymbol{\theta})$ producing a digital image that is optimized to best represent the written digits. The training of these networks focuses on simultaneously developing a model manifold flexible enough to closely mimic the data set of digits and of developing a mapping $\tilde{\mathbf{y}}^{-1}(\mathbf{d})$ from the original data $\mathbf{d}$ depicting the digit to neural outputs $\boldsymbol{\theta} = \tilde{\mathbf{y}}^{-1}(\mathbf{d})$ close to the best fit. Neural networks starting with high-dimensional data routinely distill the data into a much smaller, more comprehensible set of neural outputs $\boldsymbol{\theta}$—which are then used to classify or reconstruct the original data. The neural network thus forms a reduced-dimensional model manifold in the space of possible images. One might guess that a successful neural network would have a hyperribbon structure—since an $N$-neuron network does almost as well as an $N+1$ neuron network, one would imagine that the $(N+1)$th direction would be thin. Initial explorations of a stacked denoising autoencoder trained on the MNIST digit data by Hayden *et al.*[79] instead shows that the neural network forms a roughly equiaxed structure—but that the reconstructed digits form a lower-dimensional structure on the boundary.

### F. Science is possible

In fields like ecology, systems biology, and macroeconomics, grossly simplified models capture important features of the behavior of incredibly complex interacting systems. If what everyone ate for breakfast was crucial in determining the economic productivity each day, and breakfast eating habits were themselves not comprehensible, macroeconomics would be doomed as a subject. We argue that adding more complexity to a model produces diminishing returns in fidelity, *because the model predictions have an underlying hyperribbon structure.*

### G. Different models can describe the same behavior

We are told that science works by creating theories and testing rival theories with experiments to determine which is wrong. A more nuanced view allows for effective theories of limited validity—Newton was not wrong and Einstein right, Newton's theory is valid when velocities are slow compared to the speed of light. In more complex environments, several theoretical descriptions can cast useful light onto the same phenomena ("soft" and "hard" order parameters for magnets and liquid

crystals[80] (Chap. 9)). Also, in fields like economics and systems biology, all descriptions are doomed to neglect pathways or behavior without the justification of a small parameter. So as long as these models are capable of capturing the "long axes" of the model manifold in the data space of known behavior, and are successful at predicting the behavior in the larger data space of experiments of interest, one must view them as successful. Many such models will in general exist—certainly reduced models extracted systematically from a microscopic model (Section IV), but other models as well. Naturally, one should design experiments that test the limits of these models and cleanly discriminate between rival models. Our information geometry methods could be useful in the design of experiments distinguishing rival models; current methods that linearize about expected behavior[81] could be replaced by geometric methods that allow for large uncertainties in model parameters corresponding to nearly indistinguishable model predictions.

### H. Why is the world comprehensible?

Surely the reason that handwritten digits have a hyperribbon structure—that we do not use random dot patterns to write numbers—is partially related to the way our brain is wired. We recognize cartoons easily; therefore, the information in our handwriting is encapsulated in cartoon-like subrepresentations. But physics presumably has low-dimensional representations (Section V) independently of the way our brain works. The continuum limit describes our world perturbatively in the inverse length and time scales of the observation; the renormalization group in addition perturbs in the distance to the critical point. Why is science successful in other fields, systems biology and macroeconomics, for example? Is it a selection effect—do we choose to study subjects where our brains see patterns (low-dimensional representations), and then describe those patterns using theories with hyperribbon structures? Or are there deep underpinning structures (evolution, game theory) that guide the behavior into comprehensible patterns? A cellular control circuit where hundreds of parameters all individually control important, different aspects of the behavior would be incomprehensible without full microscopic information, discouraging us from trying to model it. On the other hand, it would seem challenging for such a circuit to arise under Darwinian evolution. Perhaps, modularity and comprehensibility themselves are the result of evolution.[82–85]

Philosophers of science have long noticed and speculated about the "unreasonable effectiveness of mathematics"[13] in constructing a hierarchy of physical theories.[14] Although the final explanation for this mystery remains elusive, sloppiness provides a new way to frame the question using the language of information geometry. The problem is thus translated from philosophical speculation into one that can be studied systematically and rigorously. Indeed, a partial explanation has been proposed based on interpolation theory. While interpolation arguments cannot explain the hyperribbon structure in all contexts, they suggest that a comprehensive explanation might be possible and what such an explanation would look like.

## VIII. CONCLUSION

What began as a rather pragmatic exercise in parameter fitting has blossomed into an enterprise that stretches across the landscape of science. The work described here has both methodological implications for the development and validation of scientific models (in the areas of optimization, machine learning, and model reduction) and philosophical implications for how we reason about the world around us. By investigating and characterizing in detail the geometric and topological structures underlying scientific models, this work connects bottom-up descriptions of complex processes with top-down inferences drawn from data, paving the way for emergent theories in physics, biology, and beyond.

## ACKNOWLEDGMENTS

We would like to thank Alex Alemi, Phil Burnham, Colin Clement, Josh Fass, Ryan Gutenkunst, Lorien Hayden, Lei Huang, Jaron Kent-Dobias, Ben Nicholson, and Hao Shi for their assistance and insights. This work was supported in part by NSF DMR 1312160 (J.P.S.), NSF IOS 1127017 (C.R.M.), the John Templeton Foundation through a grant to SFI to study complexity (B.C.D.), the U.S. Army Research Laboratory and the U.S. Army Research Office under Contract No. W911NF-13-1-0340 (B.C.D.), and a Lewis-Sigler Fellowship (B.B.M.).

[1]F. Dyson, Nature **427**, 297 (2004).
[2]J. Ditley, B. Mayer, and L. Loew, Biophys. J. **104**, 520 (2013).
[3]K. S. Brown and J. P. Sethna, Phys. Rev. E **68**, 021904 (2003).
[4]K. S. Brown, C. C. Hill, G. A. Calero, C. R. Myers, K. H. Lee, J. P. Sethna, and R. A. Cerione, Phys. Biol. **1**, 184 (2004).
[5]J. J. Waterfall, F. P. Casey, R. N. Gutenkunst, K. S. Brown, C. R. Myers, P. W. Brouwer, V. Elser, and J. P. Sethna, Phys. Rev. Lett. **97**, 150601 (2006).
[6]S. L. Frederiksen, K. W. Jacobsen, K. S. Brown, and J. P. Sethna, Phys. Rev. Lett. **93**, 216401 (2004).
[7]R. Gutenkunst, "Sloppiness, modeling, and evolution in biochemical networks," Ph.D. thesis, Cornell University, 2007, http://ecommons.library.cornell.edu/handle/1813/8206.
[8]G. J. Berman and Z. J. Wang, J. Fluid Mech. **582**, 153 (2007).
[9]B. B. Machta, R. Chachra, M. Transtrum, and J. P. Sethna, Science **342**, 604 (2013).
[10]A. Ruhe, SIAM J. Sci. Stat. Comput. **1**, 481 (1980).
[11]M. K. Transtrum, B. B. Machta, and J. P. Sethna, Phys. Rev. E **83**, 036701 (2011).
[12]R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna, PLoS Comput. Biol. **3**, 1871 (2007).
[13]E. P. Wigner, Commun. Pure Appl. Math. **13**, 1 (1960).
[14]P. W. Anderson *et al.*, Science **177**, 393 (1972).
[15]S. Amari and H. Nagaoka, *Methods of Information Geometry*, Translations of Mathematical Monographs (American Mathematical Society, 2000).
[16]B. Averick, R. Carter, J. More, and G. Xue, Preprint MCS-P153-0694, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois (1992).
[17]J. Kowalik and J. Morrison, Math. Biosci. **2**, 57 (1968).
[18]M. K. Transtrum, B. B. Machta, and J. P. Sethna, Phys. Rev. Lett. **104**, 060201 (2010).
[19]A. Atkinson, A. Donev, and R. Tobias, *Optimum Experimental Designs, with SAS* (Oxford University Press, UK, 2007).
[20]J. M. Bernardo and A. F. M. Smith, *Bayesian Theory* (John Wiley & Sons, 2009), Vol. 405.
[21]S. Kullback, *Information Theory and Statistics* (Courier Corporation, 1997).
[22]E. Beale, J. R. Stat. Soc. Ser. B (Methodological) **22**, 41 (1960).
[23]D. M. Bates and D. G. Watts, J. R. Stat. Soc. Ser. B (Methodological) **42**, 1 (1980).
[24]S.-i. Amari, *Differential-Geometrical Methods in Statistics* (Springer, 1985).
[25]S.-i. Amari, O. E. Barndorff-Nielsen, R. Kass, S. Lauritzen, and C. Rao, Lecture Notes-Monograph Series, i (1987).

[26]M. K. Murray and J. W. Rice, in *Differential Geometry and Statistics* (CRC Press, 1993), Vol. 48.
[27]S.-i. Amari and H. Nagaoka, in *Methods of Information Geometry* (American Mathematical Society, 2007), Vol. 191.
[28]D. Marquardt, J. Soc. Ind. Appl. Math. **11**, 431 (1963).
[29]D. M. Bates and D. G. Watts, Technometrics **23**, 179 (1981).
[30]A. Caticha and R. Preuss, Phys. Rev. E **70**, 046127 (2004).
[31]M. Girolami and B. Calderhead, J. R. Stat. Soc.: Ser. B (Statistical Methodology) **73**, 123 (2011).
[32]M. Spivak, *A Comprehensive Introduction to Differential Geometry* (Publish or Perish, 1979).
[33]T. Ivancevic, *Applied Differential Geometry: A Modern Introduction* (World Scientific Publishing Co., Inc, 2007).
[34]J. Stoer, R. Bulirsch, W. Gautschi, and C. Witzgall, *Introduction to Numerical Analysis* (Springer-Verlag, 2002).
[35]D. Bates and D. Watts, Ann. Stat. **9**, 1152 (1981).
[36]D. Bates, D. Hamilton, and D. Watts, Commun. Stat.-Simul. Comput. **12**, 469 (1983).
[37]D. Bates and D. Watts, *Nonlinear Regression Analysis and Its Applications* (John Wiley, 1988).
[38]J. Wei and J. C. Kuo, Ind. Eng. Chem. Fundam. **8**, 114 (1969).
[39]J. C. Liao and E. N. Lightfoot, Biotechnol. Bioeng. **31**, 869 (1988).
[40]H. Huang, M. Fairweather, J. Griffiths, A. Tomlin, and R. Brad, Proc. Combust. Inst. **30**, 1309 (2005).
[41]N. Goldenfeld, *Lectures on Phase Transitions and the Renormalization Group* (Addison-Wesley, Advanced Book Program, Reading, 1992).
[42]J. Zinn-Justin, *Phase Transitions and Renormalization Group* (Oxford University Press, 2007).
[43]V. Saksena, J. O'reilly, and P. V. Kokotovic, Automatica **20**, 273 (1984).
[44]P. Kokotovic, H. K. Khali, and J. O'Reilly, in *Singular Perturbation Methods in Control: Analysis and Design* (SIAM, 1999), Vol. 25.
[45]D. Naidu, Dynamics of Continuous Discrete and Impulsive Systems Series B **9**, 233 (2002).
[46]A. C. Antoulas, in *Approximation of Large-Scale Dynamical Systems* (SIAM, 2005), Vol. 6.
[47]C. H. Lee and H. G. Othmer, J. Math. Biol. **60**, 387 (2010).
[48]B. Moore, IEEE Trans. Autom. Control **26**, 17 (1981).
[49]G. Dullerud and F. Paganini, *Course in Robust Control Theory* (Springer-Verlag, New York, 2000).
[50]S. Gugercin and A. C. Antoulas, Int. J. Control **77**, 748 (2004).
[51]K. Zhou, C. D'Souza, and J. R. Cloutier, Syst. Control Lett. **24**, 235 (1995).
[52]L. Li and F. Paganini, Automatica **41**, 145 (2005).
[53]H. Sandberg and R. M. Murray, Optim. Control Appl. Methods **30**, 225 (2009).
[54]J. M. Scherpen, Syst. Control Lett. **21**, 143 (1993).
[55]S. Lall, J. E. Marsden, and S. Glavaški, Int. J. Robust Nonlinear Control **12**, 519 (2002).
[56]A. J. Krener, *Analysis and Design of Nonlinear Control Systems* (Springer, 2008), pp. 41–62.
[57]B. C. Daniels and I. Nemenman, e-print arXiv:1404.6283 [q-bio.QM] (2014).
[58]B. C. Daniels and I. Nemenman, PLoS One **10**, e0119821 (2015).
[59]M. K. Transtrum and P. Qiu, Phys. Rev. Lett. **113**, 098701 (2014).
[60]M. K. Transtrum, G. Hart, and P. Qiu, preprint arXiv:1409.6203 (2014).
[61]J. F. Apgar, D. K. Witmer, F. M. White, and B. Tidor, Mol. BioSyst. **6**, 1890 (2010).
[62]M. Vilela, S. Vinga, M. A. Maia, E. O. Voit, and J. S. Almeida, BMC Syst. Biol. **3**, 47 (2009).
[63]K. Erguler and M. P. H. Stumpf, Mol. BioSyst. **7**, 1593 (2011).
[64]M. Transtrum and P. Qiu, BMC Bioinf. **13**, 181 (2012).
[65]R. Chachra, M. K. Transtrum, and J. P. Sethna, Mol. BioSyst. **7**, 2522 (2011).
[66]E. Lee, A. Salic, R. Kruger, R. Heinrich, and M. W. Kirschner, PLoS Biol. **1**, e10 (2008).
[67]M. Ringner, Nat. Biotechnol. **26**, 303 (2008).
[68]K. Levenberg, Q. Appl. Math. **2**, 164 (1944).
[69]W. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing* (Cambridge University Press, 2007).
[70]M. K. Transtrum and J. P. Sethna (manuscript in revision), http://arxiv.org/abs/1201.5885.
[71]M. Transtrum and J. P. Sethna, "geodesiclm," http://sourceforge.net/projects/geodesiclm/ (2011).
[72]B. C. Daniels, Y. J. Chen, J. P. Sethna, R. N. Gutenkunst, and C. R. Myers, Curr. Opin. Biotechnol. **19**, 389 (2008).
[73]A. Wagner, *Robustness and Evolvability in Living Systems* (Princeton University Press, 2005).
[74]A. Cutler and L. Breiman, Technometrics **36**, 338 (1994).
[75]M. Mørup and L. K. Hansen, Neurocomputing **80**, 54 (2012), Special Issue on Machine Learning for Signal Processing 2010.
[76]C. Thurau, K. Kersting, and C. Bauckhage, in *ICDM '09. Ninth IEEE International Conference on Data Mining, 2009* (CPS: Conference Publishing Services, Los Alamitos, CA, 2009), pp. 523–532.
[77]Y. Koren, R. Bell, and C. Volinsky, Computer **42**, 30 (2009).
[78]P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, in *Proceedings of the 25th International Conference on Machine Learning, ICML '08* (ACM, New York, NY, USA, 2008), pp. 1096–1103.
[79]L. X. Hayden, A. A. Alemi, and J. P. Sethna, "Information geometry of neural networks" (unpublished).
[80]J. P. Sethna, *Statistical Mechanics: Entropy, Order Parameters, and Complexity* (Oxford University Press, Oxford, 2006), http://www.physics.cornell.edu/sethna/StatMech/.
[81]F. P. Casey, D. Baird, Q. Feng, R. N. Gutenkunst, J. J. Waterfall, C. R. Myers, K. S. Brown, R. A. Cerione, and J. P. Sethna, IET Syst. Biol. **1**, 190 (2007).
[82]M. Kirschner and J. Gerhart, Proc. Natl. Acad. Sci. U. S. A. **95**, 8420 (1998).
[83]L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, Nature **402**, C47 (1999).
[84]N. Kashtan and U. Alon, Proc. Natl. Acad. Sci. U. S. A. **102**, 13773 (2005).
[85]J. Clune, J.-B. Mouret, and H. Lipson, Proc. R. Soc. B **280**, 20122863 (2013).