Chebyshev Approximation and the Global Geometry of Model Predictions

Katherine N. Quinn,1,* Heather Wilber,2,† Alex Townsend,3,‡ and James P. Sethna1,§

1Physics Department, Cornell University, Ithaca, New York 14853-2501, USA
2Center for Applied Mathematics, Cornell University, Ithaca, New York 14853-3801, USA
3Mathematics Department, Cornell University, Ithaca, New York 14853-4201, USA

(Received 15 September 2018; revised manuscript received 17 February 2019; published 18 April 2019)

Complex nonlinear models are typically ill conditioned or sloppy; their predictions are significantly affected by only a small subset of parameter combinations, and parameters are difficult to reconstruct from model behavior. Despite forming an important universality class and arising frequently in practice when performing a nonlinear fit to data, formal and systematic explanations of sloppiness are lacking. By unifying geometric interpretations of sloppiness with Chebyshev approximation theory, we rigorously explain sloppiness as a consequence of model smoothness. Our approach results in universal bounds on model predictions for classes of smooth models, capturing global geometric features that are intrinsic to their model manifolds, and characterizing a universality class of models. We illustrate this universality using three models from disparate fields (physics, chemistry, biology): exponential curves, reaction rates from an enzyme-catalyzed chemical reaction, and an epidemiology model of an infected population.

DOI: 10.1103/PhysRevLett.122.158302

Complex nonlinear models used to simulate and predict experimentally observed phenomena often exhibit a structural hierarchy: perturbing a few model parameter combinations drastically impacts predictions, whereas most others can vary widely without effect. Such ill-conditioned models are called sloppy. Sloppy models appear to be common, arising in many areas of physics. In critical phenomena, this hierarchy of importance explains the parameter scaling with coarsening for diffusion and the Ising model of magnetism [1]. In accelerator physics, linear combinations of the multitude of tunable beam-line parameters exhibit a geometric hierarchy of importance [2]. Exponential curve fitting, a notoriously ill-conditioned problem, poses a significant challenge, e.g., finding correlators in lattice QCD [3,4]. Sloppy models are not confined to physics, and in fact appear in systems biology [5–7], insect flight [8], power systems [9,10], machine learning [11], and many other areas [12]. Understanding why sloppiness occurs can therefore connect models used across disparate fields.

There are many well-studied cases for insensitivity of model predictions to particular combinations of parameters. Structural identifiability describes systems for which parameters can be analytically exchanged for one another [13,14]. Separation of scales, singular perturbations, and continuum limits can make the behavior at a particular time or distance region depend only on a subset of the underlying parameters [15–17]. Universal critical behavior can yield effective parameter compression on long length scales near continuous transitions [1]. However, these comprehensible sources of sloppiness do not explain the generality of the phenomenon, nor do they offer a rigorous framework by which to quantify the hierarchy of parameter importance. In this Letter, we address the generic sloppiness of multiparameter nonlinear models in the absence of particular mechanisms or small parameters. We unify recently developed geometric descriptions of sloppiness [12] with classical ideas from polynomial approximation theory [18]. We posit that, in many cases, sloppiness is fundamentally linked to the smoothness of the underlying model, and provide a rigorous description of this connection.

The hierarchy of parameter importance that characterizes sloppy models manifests geometrically. Given some model, the space of all possible predictions for all input parameters forms a geometric object known as the model manifold [Fig. 1(a)], whose metric is given by the Fisher information (a measure of the distinguishability between predictions [19], Ch. 2) which sets a lower bound on the possible variance of parameter estimates for an unbiased prior through the Cramér-Rao bound. Studying the geometry of model manifolds yields fruitful information for several reasons: (1) the dominant components reflect emergent behavior of the models (how the microscopic interactions do or do not produce macroscopic behavior [1]), (2) the boundaries represent reduced-model approximations [20], and (3) knowledge of the manifold geometry leads to more efficient data fitting methods [21]. Model manifolds typically form striking hyperribbons [22], so-called because, like ribbons, successive widths follow a geometric decay: they are much longer than they are wide, much wider than they are thick, etc., yielding effective low-dimensional representations. Because directions along the model manifold correspond to specific parameter combinations, there is a direct connection between the hyperribbon nature of
Consider a nonlinear model that depends continuously on $K$ input parameters $\theta = (\theta^1, \ldots, \theta^K)$ to generate predictions $y_\theta(t)$. If we consider the model predictions at $N$ fixed points, $\{t_0, \ldots, t_{N-1}\}$, then our predictions for parameters $\theta$ form an $N$-dimensional vector $Y(\theta) = (y_\theta(t_0), \ldots, y_\theta(t_{N-1})) = (Y_0, \ldots, Y_{N-1})$. We use $\mathcal{Y}$ to represent the model manifold, defined as the space of all possible predictions for all possible parameter combinations [so all allowed $Y(\theta)$]. Specifically, model manifold $\mathcal{Y}$ is a $K$-dimensional surface embedded in an $N$-dimensional prediction space.

To bound the model manifold $\mathcal{Y}$ and study its geometry, we consider polynomial approximations of model $y_\theta$. Without loss of generality, we shift and rescale the points so that $\{t_k\}_{k=0}^{N-1} \subset [-1, 1]$. Let $\{\phi_j\}_{j=0}^\infty$ be a complete polynomial basis, and suppose that model $y_\theta(t)$ is decomposed into this basis: $y_\theta(t) = \sum_{j=0}^\infty \beta_j(\theta)\phi_j(t)$. Let $p_{N-1}(t; \theta)$ be the truncated series representing the polynomial approximation to model $y_\theta(t)$. Note that the truncation is set by the number of sampled points, $N$. We can view the coefficients $(\beta_0(\theta), \ldots, \beta_{N-1}(\theta))$ as a set of $N$ parameters. Now, let $P(\theta) = (p_{N-1}(t_0), \ldots, p_{N-1}(t_{N-1})) = (P_0, \ldots, P_{N-1})$ define the polynomial manifold $\mathcal{P}$. Thus, we have model manifold $\mathcal{Y}$ and a polynomial manifold $\mathcal{P}$.

By definition, $P(\theta) = X\mathbf{b}$, where $X_{ij} = \phi_{j-1}(t_{i-1})$ and $\mathbf{b} = (b_0, \ldots, b_{N-1})^T$. Here, $X$ forms a linear map from the space of polynomial coefficients to the space of possible predictions, and is determined by the chosen polynomial basis and fixed points $t_i$. The singular values of $X$ can be used to understand the hyperribbon structure of the polynomial manifold $\mathcal{P}$. Suppose, for example, that $||\mathbf{b}||_2 < r$, so that the coefficient space is bounded in $S$, an $n$-sphere of radius $r$. The action of $X$ on $S$ distorts it into a hyperellipsoid $H_p$. If $\ell_j(H_p)$ is the diameter of the $j$th cross section of hyperellipsoid $H_p$, then

$$\ell_j(H_p) = 2r\sigma_j(X),$$

(1)

where $\sigma_j(X)$ are the ordered singular values of $X$. When $X$ has rapidly decaying singular values, $H_p$ has a hyperribbon structure because there is a strict hierarchy in successive widths. Accounting for the polynomial approximation error $||y_\theta - p_{N-1}||_\infty$, where $|| \cdot ||_\infty$ is the $L^\infty$ norm on $[-1, 1]$, we can define a hyperellipsoid $H_{\rho}$ that must enclose model manifold $\mathcal{Y}$, where the cross-sectional widths are given by...
\[ \ell_j(H_Y) = 2r \sigma_j(X) + 2\|y - p_{N-1}\|_\infty. \]  

In this way, we find that any model manifold \( Y \) is bounded within a hyperribbon whenever \( \sigma_j(X) \) decays geometrically and \( \|y - p_{N-1}\|_\infty \) is small enough. A fundamental question is whether it matters which polynomial basis or which set of time points are chosen to define \( H_p \) and \( H_Y \). The hyperribbon structure of \( Y \), of course, does not depend on our representation of \( y_0 \), but rather on intrinsic properties of the model, such as its smoothness. For example, if for every \( t_0 \in [-1, 1] \), the Taylor expansion of \( y_0 \) at \( t_0 \) has a large enough radius of convergence, any sequence of polynomial interpolants with \( N \) distinct interpolating points converges to \( y_0 \) at a geometric rate with \( N \) [18]. This fact underpins the qualitative observation in Refs. [12,22] that certain analytic models have manifolds bounded within hyperribbons. Here we make that observation rigorous. We consider two such choices. First, we choose our basis functions \( \{\phi_j\}_{j=0}^\infty \) as the Chebyshev polynomials. Truncated Chebyshev expansions converge to \( y_0 \) at an asymptotically optimal rate for polynomial approximation [18]. As we show below, this rate controls the magnitude of \( \sigma_j(X) \) in Eq. (2), and can be used to explicitly bound the cross sectional widths of \( H_Y \). Our bounds deliver an outright description of a hyperribbon that must contain \( Y \).

We also analyze the case where \( \{\phi_j\}_{j=0}^\infty \) are the monomials and \( p_{N-1} \) is the truncated Taylor series expansion of \( y_0 \). In this case, we observe that the numerical computation of \( \sigma_j(X) \) results in excellent practical and universal bounds on the prediction space for large classes of models.

**Chebyshev expansions.**—Suppose that \( y_0 \) has a convergent Chebyshev expansion, so that it is given by \( y_0(t) = \sum_{j=0}^{\infty} c_j(\theta)T_j(t) \), where \( T_j(t) = \cos(j \arccos t) \) is the degree \( j \) Chebyshev polynomial ([18], Ch. 3). We can approximate \( y_0 \) with a degree \( \leq N-1 \) polynomial by truncating the Chebyshev series after \( N \) terms:

\[ p_{N-1}(t; \theta) = \sum_{j=0}^{N-1} c_j(\theta)T_j(t). \]  

Truncated Chebyshev expansions have near-best global approximation properties, and explicit bounds on \( \|y_0 - p_{N-1}\|_\infty \) are known when \( y_0 \) is sufficiently smooth.

We first consider the case where \( y_0 \) is analytic in an open neighborhood of \([-1, 1]\). Such a region contains a Bernstein ellipse \( E_\rho \), defined as the image of the circle \( |z| = \rho \) under the Joukowsky mapping \((z + z^{-1})/2\). It has foci at \( \pm 1 \), and the lengths of its semimajor and semiminor axes sum to \( \rho \). The polynomial in Eq. (3) converges to \( y_0 \) as \( N \to \infty \) at a rate determined by \( \rho \).

**Theorem 1.** Let \( M > 0 \) and \( \rho > 1 \) be constants and suppose that \( y_0(t) \), \( t \in [-1, 1] \), is analytically continuable to the region enclosed by the Bernstein ellipse \( E_\rho \), with \( |y_0| \leq M \) in \( E_\rho \), uniformly in \( \theta \). Let \( p_{N-1}(t; \theta) \) be as in Eq. (3). Then,

\[ (i) \quad \|y_0 - p_{N-1}\|_\infty \leq \frac{2M \rho^{N+1}}{\rho - 1}, \]

\[ (ii) \quad |c_0| \leq M, \quad |c_j(\theta)| \leq 2M \rho^{-j}, \quad j \geq 1. \]  

**Proof.**—For a proof, see Theorem 8.2 in Ref. [18]. To exploit the decay of the coefficients in Eq. (5), we define modified coefficients \( c_j = \rho^j c_j \). We then have that polynomial predictions \( P(\theta) = Xe \), where \( X = JD, J_{ij} = T_{j-i}(t_{i-1}), \) and \( D \) is diagonal with entries \( D_{jj} = \rho^{-j} \). By Eq. (5), we have that \( \|\|_2 < 4M \sqrt{4N - 3} \). This implies that the polynomial manifold \( \mathcal{P} \) is bound in a hyperellipsoid \( H_p \). By Eq. (1), we have that \( \ell_j(H_p) = 8M \sqrt{4N - 3} \sigma_j(X) \). To bound \( \sigma_j(X) \) explicitly, we first prove a conjecture proposed in Ref. [23]:

**Theorem 2.** Let \( S \in \mathbb{R}^{N \times N} \) be symmetric and positive definite. Let \( E \in \mathbb{R}^{N \times N} \) be diagonal with \( E_{ii} = e^{i-1} \) and \( 0 < e < 1 \). If \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \) are the ordered eigenvalues of \( E \), then \( \lambda_{m+1} = O(e^m) \). Specifically,

\[ \lambda_{m+1} \leq \frac{e^{2m}}{1 - e^{-2}} \max_{1 \leq j, k \leq N} |S_{jk}|, \quad 1 \leq m \leq N - 1. \]  

**Proof.**—[24] Consider the rank \( m \) matrix

\[ S_m = (S; : , :)(S; 1 , m, 1; m^\top)(S; 1, m, :), \]  

where \( 1 \leq m \leq N - 1 \), and the notation \( M(:, 1; m) \) denotes the submatrix of \( M \) consisting of its first \( m \) columns. Clearly, \( S_m \) is well defined because \( S(1; m, 1; m) \) is a principal minor of a positive definite matrix and is therefore invertible. Moreover, it can be verified that \( (S - S_m)_{jk} = 0 \) for \( 1 \leq j, k \leq m \).

Since \( E \) is positive definite and rank(\( S_m \)) = \( m \), we know that \( \lambda_{m+1} \leq \|E(S - S_m)E\|_2 \), where \( \| \cdot \|_2 \) denotes the spectral matrix norm [25]. Using \( \| \cdot \|_F \) to denote the Frobenius norm, we have

\[ \lambda_{m+1}^2 \leq \|E(S - S_m)E\|^2_F \leq \|E(S - S_m)E\|^2 \]

\[ = \sum_{j=m+1}^N \sum_{k=m+1}^N e^{2(j-1)+2(k-1)} |S_{jk} - (S_m)_{jk}|^2 \]

\[ \leq \frac{e^{4m}}{(1 - e^{-2})^2} \max_{1 \leq j, k \leq N} |S_{jk} - (S_m)_{jk}|^2 \]

\[ \leq \frac{e^{4m}}{(1 - e^{-2})^2} \max_{1 \leq j, k \leq N} |S_{jk}|^2, \]

where the last inequality comes from the fact that the block \( S(m+1; N, m+1; N) - S_0(m+1; N, m+1; N) \) is the Schur complement of \( S(1; m, 1; m) \) in \( S [25] \).}

Applying Theorem 2 to \( X^T X = JD^T JD \), we have that for \( j > 1 \), \( \sigma_j(X) \leq \sqrt{Np^{j-2}j/\sqrt{p^2 - 1}} \), where we have used the fact that \( |T_k(t)| \leq 1 \) for \( k \geq 0 \) and \(-1 \leq t \leq 1 \). It
follows from Eqs. (2) and (4) that predictions for $y_\theta(t)$ are bounded by a hyperellipsoid $H_Y$, with

$$\ell_j(H_Y) \leq \frac{2M\sqrt{4N^2 - 3N}\rho^{-j+2}}{\sqrt{\rho^2 - 1}} + 4M\rho^{-N+1} \rho - 1,$$

for $2 \leq j \leq N$, i.e.,

$$\ell_j(H_Y) = O(\rho^{-j} + \rho^{-N}).$$

These bounds indicate that the hyperribbon structure of $H_Y$ is controlled by $\rho$, a parameter characterizing the analyticity of the model. As $\rho$ becomes larger, bounds on the widths of the successive cross sections of $H_Y$ must decay more rapidly: in principle, $H_Y$ becomes successively thinner and more ribbonlike.

When $y_\theta$ is not analytic on an open neighborhood of $[-1, 1]$, the decay rate of $\sigma_j(JD)$ is instead controlled by the smoothness of $y_\theta$ on $[-1, 1]$. Furthermore, when we consider models with two experimental conditions (for instance, time and temperature) these bounds can be extended to the two-dimensional case. We provide more discussion of nonanalytic and two-dimensional cases in the Supplemental Material [26].

Taylor expansions.—The degree $N - 1$ truncated Taylor polynomial of $y_\theta$ is $p_{N-1}(t) = \sum_{k=0}^{N-1} a_k(\theta)(t - t_0)^k$, where $a_k(\theta) = y^{(k)}_\theta(t_0)/k!$. We describe the analyticity of $y_\theta$ using the following condition: for all $N \geq 1$,

$$\sum_{k=0}^{N-1} \left(\frac{R^k}{k!} \frac{d^k y_\theta(t)}{dt^k}\right)^2 < C^2 N,$$

where $C > 0, R > 1$ are constants in $\theta$. A straightforward but tedious calculation outlined in the Supplemental Material [26] shows that the lengths of the resulting hyperellipsoid are given by

$$\ell_j(H_R) \leq \frac{2CN}{\sqrt{R^2 - 1}} R^{-j+2}.$$
third and fourth author’s research areas, and Peter Lepage for his expertise and insight into the connection with lattice QCD. K. N. Q. received support from a Natural Sciences and Engineering Research Council of Canada (NSERC) Fellowship, and J. P. S. and K. N. Q. were supported by the National Science Foundation (NSF) through Grant No. DMR-1719490. A. T. was supported by NSF Grant No. DMS-1818757, and H. W. was supported by NSF Grant No. DGE-1650441.


[24] Previous proofs with weaker bounds were provided through private communications with Ari Turner and Yaming Yu.


[26] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.122.158302 for bounds on nonanalytical models, numerical results for high-dimensional manifolds, proof for the truncated Taylor series, extensions to two-dimensional experimental conditions, and a full description of how the visualizations of the model manifolds were achieved, which includes Refs. [18,27–30].


