Sloppy-Model Universality Class and the Vandermonde Matrix

Joshua J. Waterfall,[1] Fergal P. Casey,[2] Ryan N. Gutenkunst,[1] Kevin S. Brown,[3] Christopher R. Myers,[4] Piet W. Brouwer,[1] Veit Elser,[1] and James P. Sethna[1]

[1]*Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, New York 14853, USA*
[2]*Center for Applied Mathematics, Cornell University, Ithaca, New York 14853, USA*
[3]*Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA*
[4]*Cornell Theory Center, Cornell University, Ithaca, New York 14853, USA*

(Received 15 May 2006; published 12 October 2006)

In a variety of contexts, physicists study complex, nonlinear models with many unknown or tunable parameters to explain experimental data. We explain why such systems so often are *sloppy*: the system behavior depends only on a few "stiff" combinations of the parameters and is unchanged as other "sloppy" parameter combinations vary by orders of magnitude. We observe that the eigenvalue spectra for the sensitivity of sloppy models have a striking, characteristic form with a density of logarithms of eigenvalues which is roughly constant over a large range. We suggest that the common features of sloppy models indicate that they may belong to a common universality class. In particular, we motivate focusing on a Vandermonde ensemble of multiparameter nonlinear models and show in one limit that they exhibit the universal features of sloppy models.

DOI: 10.1103/PhysRevLett.97.150601 PACS numbers: 05.10.−a, 02.30.Zz, 87.15.Aa, 87.16.Xa

Systems with many parameters are often *sloppy*. For practical purposes their behavior depends only on a few stiffly constrained combinations of the parameters; other directions in parameter space can change by orders of magnitude without significantly changing the behavior. Given a suitable cost $C(\mathbf{p})$ measuring the change in system behavior as the parameters $\mathbf{p}$ vary from their original values $\mathbf{p}^{(0)}$ (e.g., a sum of squared residuals), the stiff and sloppy directions can be quantified as eigenvalues and eigenvectors of the Hessian of the cost: $H_{ij} = \partial^2 C/\partial p_i \partial p_j|_{\mathbf{p}^{(0)}}$.

Figure 1 shows the eigenvalues of the cost Hessian for many different systems; those in 1(a)–1(d) and 1(h), are all sloppy. The sensitivity of model behavior to changes along an eigenvector is given by the square root of the eigenvalue—the range in eigenvalues of roughly 1×10^6 for the sloppy models means that one must change parameters along the sloppiest eigendirection a thousand times more than along the stiffest eigendirection in order to change the behavior by the same amount. Although anharmonic effects rapidly become important along sloppy eigendirections, a principal component analysis of a Monte Carlo sampling of low-cost states has a similar spectrum of eigenvalues [1]; the sloppy eigendirections become curved sloppy manifolds in parameter space. Similar sloppy behavior has been demonstrated in 14 systems biology models taken from the literature [2,3], and in three multiparameter interatomic potentials fit to electronic structure data [4]. In these disparate models we see a common, peculiar behavior: the nth stiffest eigendirection is more important than the $(n+1)$th by a roughly constant factor, giving a total range of eigenvalues of typically over a million for any model with more than eight parameters. We call systems exhibiting these characteristic features *sloppy models*.

This sloppiness has a number of important implications. In estimating prediction errors, sloppiness affects both the estimation of statistical errors due to uncertainties in the experimental data [2,3] and allows an estimation of systematic errors due to imperfections in the models (for example, in interatomic potentials [4] and density functional theory [5]). It makes extracting parameter values from fits to sloppy models ill posed [2,6]. Conversely, it is much more efficient to improve the predictivity of a model by fitting parameters to system behavior than by designing experiments that precisely determine the individual parameter values [3]. Sloppy problems are also better approached with optimization algorithms [7,8] (like the Levenberg-Marquardt and Nelder-Mead methods) which can adapt to widely diverging step sizes along different parameter combinations.

Let us begin with the famously ill-posed problem of fitting a sum of exponentials to data [9,10]. Consider a mixture of equal amounts of N radioactive elements whose decay signal is thus the sum of N exponentials with decay rates $(\gamma_1^{(0)}, \ldots, \gamma_N^{(0)})$. We define a cost function for general decay rates as $C = \frac{1}{2}\int_0^\infty [\sum_{i=1}^N \exp(-\gamma_i t) - \sum_{i=1}^N \exp(-\gamma_i^{(0)} t)]^2 d\log t$ (spacing the "data points" equally in log time makes analyzing large ranges of decay constants convenient). Because the decay constants are positive and can have a large range of sizes, we use their logarithms as our parameters ($p_i = \log\gamma_i$), giving model sensitivity to relative changes in the decay rates. The resulting Hessian is $H_{ij}|_{\mathbf{p}^{(0)}} = 2\gamma_i^{(0)}\gamma_j^{(0)}/(\gamma_i^{(0)} + \gamma_j^{(0)})^2$. For the 12 radionuclides described in the caption to Fig. 1(c), the eigenvalues of the Hessian are each separated by nearly one decade; the sloppiest mode has an eigenvalue a factor of 10^{10} smaller (less important) than the stiffest.

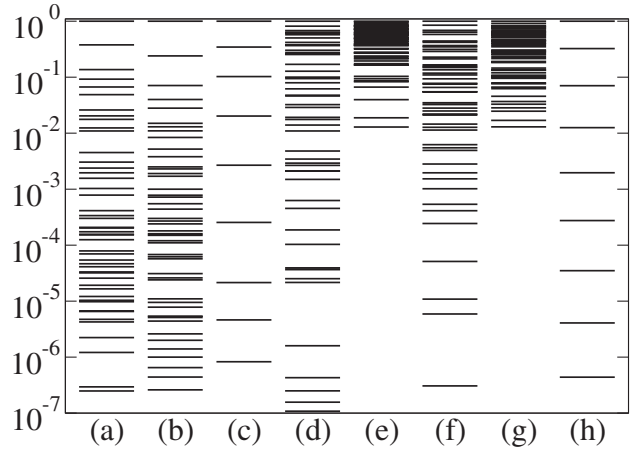

FIG. 1. Eigenvalues giving the stiffness or sloppiness of various models as parameters are varied. Each spectrum has been shifted so that the largest eigenvalue is one. (a) Growth factor signaling model (coupled nonlinear ODEs) for PC12 cells [2], as the 48 parameters (rate and Michaels-Menten constants) are varied. (b) Variational wave function used in quantum Monte Carlo calculations, as the Jastrow parameters (for electron-electron coincidence cusps) are varied, (c) Radioactivity time evolution for a mixture of 12 common radionuclides as the half-lives γ_i are varied. The radionuclides are those available from Perkin and Elmer [19] with half-lives less than 100 days. (Only the first nine eigenvalues are shown.) (d) The same exponential decay model as in (c) with 48 decay constants γ_i randomly spread over a range of e^{50}. (e) One random 48×48 matrix in the Gaussian orthogonal ensemble (GOE) (not sloppy). (f) A product of five random 48×48 matrices, illustrating the random product ensemble (not sloppy, but ill conditioned). (g) A plane in 48 dimensions fit to 68 data points, the same number and data points as for the biology model in column (a) (Wishart statistics, not sloppy). (h) A polynomial fit to data, as the 48 monomial coefficients are varied (the Hilbert matrix [16], sloppy).

This is not the result of an inaccurate mathematical description; it is true for the correct model and parameters with a complete complement of data. The origin of sloppiness is not a simple lack of data where trivial overparametrization leads to unidentifiable parameters.

Unless the individual lifetimes are well separated, the net radiation cannot be used to measure the lifetimes reliably. The difficulty is that the signal is the sum of many functions with similar shapes; one can generate almost identical signals with wildly different values for the parameters. Similarly, the sloppiness in more realistic models is presumably due to the compensation of subsets of parameters with similar effects. If we pick 48 lifetimes whose logarithms are instead uniformly distributed over a range of $2\epsilon = 50$ (largest/smallest $\approx e^{50} \approx 10^{21}$), the density of levels and the variation in spacings between neighboring levels in the new spectrum [Fig. 1(d)] is similar to that of the real-life models in 1(a) and 1(b).

While a large number of models are sloppy, not all multiparameter models share this quality. The simplest form of multiple linear regression, which is in essence fitting a plane through the origin to a cloud of points, is not sloppy. The Hessian matrix for this type of model is the sample covariance matrix of the data points and is known as a Wishart matrix [11]. The eigenvalues of a Wishart matrix are described by the Marčenko-Pastur distribution [12] and an example is seen in Fig. 1(g). The classic ensembles of random matrix theory [13,14] [Fig. 1(e)] have uniform eigenvalue densities instead of the exponentially large range characteristic of sloppy systems. The ensemble of products of random matrices [15] [Fig. 1(f)] does mimic the exponential spacing of (singular) values but in this case the variance of level spacings is proportional to the mean spacing. Toward the end of this Letter we will see that for sloppy models, in the limit of large spacings, the variance is instead independent of the mean.

Why are so many models sloppy? We can gain insight by considering fitting data for $x \in [0, 1]$ with polynomials. If one considers the polynomials of order N to be sums of monomials, $y(x, \mathbf{p}) = \sum_{i=0}^{N} p_i x^i$, the Hessian is $H_{ij} = 2A_N = \frac{2}{i+j+1}$, the famously ill-conditioned Hilbert matrix [Fig. 1(h)]. Indeed, the coefficients of the monomials are known to be poorly determined in such polynomial fits [7]. Suppose we instead generate the same polynomial fit, but parametrize our polynomial as a sum of the appropriate shifted Legendre polynomials $y(x, \mathbf{p}') = \sum_{i=0}^{N} p_i' L_i(x)$; $L_0 = 1$, $L_1 = \sqrt{3}(2x - 1)$, $L_2 = \sqrt{5}(6x^2 - 6x + 1)$, ... The shifted Legendre polynomials are orthonormal in the L^2 norm on $[0, 1]$, and the Hessian in the $\mathbf{p}'$ basis is the identity matrix. By changing our parametrization from monomial coefficients $\mathbf{p}$ to coefficients $\mathbf{p}'$ in the appropriate orthonormal basis, our sloppiness is completely cured. The sloppiness is due to the fact that the monomial coefficients (natural from many perspectives) are a perverse set of coordinates from the point of view of the behavior of the resulting polynomial. We can quantify this by noting that the transformation S_N from the monomial basis to the orthonormal basis (the coefficients of the shifted Legendre polynomials) has a tiny determinant, and therefore the volume enclosed by the monomial basis vectors shrivels and becomes greatly distorted under the transformation. This determinant can be found by noting that S_N gives a Cholesky decomposition of the Hilbert matrix $A_N = S_N^{\mathsf{T}} S_N$, and thus $\det S_N = \sqrt{\det A_N} = [\prod_{i=1}^{N-1}(i!)]^2 / \sqrt{\prod_{j=1}^{2N-1}(j!)}$ [16]. Physically, the monomials all have roughly the same shape (starting flat near zero, and rising sharply at the end near one), and can be exchanged for one another, while the orthogonal polynomials all have quite distinct shapes. In nonlinear sloppy models the sloppiness is more difficult to remove: (a) the transformation to unsloppy parameters will be nonlinear away from the optimum, often not even single valued, (b) we may not have the insight or the ability to change parametrizations to those natural for fitting purposes, and (c) often the natural parametrization is determined by the science (as in bio-

chemical rate constants, arbitrary linear combinations of which are not biologically motivated).

What causes this even distribution of relative stiffnesses over so many decades of scales? To form strong conclusions about sloppy models we must establish criteria sufficient to exclude the large variety of multiparameter systems that will not be sloppy. First, we specialize to models where the cost is a sum of squared residuals $C(\mathbf{p}) = \sum_m r_m^2$, where the sum may be continuous (e.g., an integral over time) and $r_m = y_m(\mathbf{p}) - d_m$ is the deviation of theory $y(\mathbf{p})$ from the experimental datum d_m; all of our examples of sloppy models are of this type. Second, to avoid including systems where each parameter is the subject of a separate experiment isolating that component, we make the (strong) assumption that all of the residuals $r_m(\mathbf{p})$ depend on the parameters $\mathbf{p}$ in a symmetric fashion (e.g., permuting p leaves r_m unchanged). This allows us to recast the residuals into the basis of power sum polynomials of the parameters, $r_m(\mu_1, \mu_2, \ldots)$, $\mu_k = \sum_{i=1}^N p_i^k$, which can also be viewed as the moments of the parameter distribution. Third, we noticed that in fitting exponentials the compensable nature of different parameters increased when they were restricted to smaller ranges; here we will assume that the parameters are all confined to a small range $p_i \in [\bar{p} \pm \epsilon]$. Thus if we define $\epsilon_i = p_i - \bar{p}$, the residuals $r_m(\mu_1, \mu_2, \ldots)$ can be written as functions of the moments $\mu_k = \sum_{i=1}^N \epsilon_i^k$.

In general, the Hessian is

$$H_{ij} = \sum_m \left(\frac{\partial r_m}{\partial p_i} \frac{\partial r_m}{\partial p_j} + r_m \frac{\partial^2 r_m}{\partial p_i \partial p_j} \right) \tag{1}$$

but for the correct model at the true parameters the cost is zero, so $r_m = 0 \forall m$ and $H = J^\top J$ with the Jacobian

$$J_{mj} = \frac{\partial r_m}{\partial p_j} = \sum_{k=1}^K \frac{\partial r_m}{\partial \mu_k} k \epsilon_j^{k-1} = A_{mk} V_{kj}, \tag{2}$$

where $A_{mk} = \frac{\partial r_m}{\partial \mu_k} k$, $V_{kj} = \epsilon_j^{k-1}$, and K is the maximum degree (possibly ∞) to which we expand in ϵ. Thus $H = J^\top J = V^\top A^\top A V$. Here V, the famous Vandermonde matrix, is the heart of the sloppy-model universality class. Reminiscent of random matrix theory ensembles, we are now interested in the Vandermonde ensemble of Hessians of the form $V^\top A^\top A V$. The Vandermonde matrix is well known primarily because its determinant (for $N = K$) can be expressed analytically, $\det(V) = \prod_{i<j}(\epsilon_i - \epsilon_j)$. As $\epsilon \to 0$ this product is tiny, $\det(V) = \mathcal{O}(\epsilon^{N(N-1)/2})$. While the elements of A do, in general, depend on the parameter values, they either approach a constant or zero in this limit and we can see that the determinant of H, $\det(H) = \det(V)^2 \det(A)^2$ is smaller still. As we saw with the Hilbert matrix and fitting monomials to data, transformation matrices with very small determinants are a signature of sloppy models.

To show that the eigenvalues in our Vandermonde ensemble are evenly spread in logarithm, we will make use of an apparent truth about matrices:

Conjecture 1.—Let $S \in \mathbb{R}^{n \times n}$ be symmetric and positive definite. Let $E \in \mathbb{R}^{n \times n}$ be diagonal with $E_{ii} = \epsilon^{i-1}$ and $0 < \epsilon \ll 1$. Then the mth largest eigenvalue of ESE is $\mathcal{O}(\epsilon^{2(m-1)})$ [less than some constant times $\epsilon^{2(m-1)}$].

We have two reasons to believe this conjecture is true. (1) Treating the off-diagonal components of ESE as a perturbation, the corrections to the mth eigenvalue are of order $\epsilon^{2(m-1)}$ to all orders in perturbation theory, despite the fact that many of the perturbing elements are large compared to the diagonal entries. (2) Extensive numerical tests show an even sharper result: the mth largest eigenvalue, λ_m, is bounded above by the mth largest row sum of EES for all ϵ, where the row sum for row k is $\sum_l \epsilon^{2(k-1)} |S_{kl}|$. This implies that $\lambda_m \leq \|S\|_\infty \epsilon^{2(m-1)}$, and also (for $\epsilon = 1$) implies the remarkable apparent fact that the sorted eigenvalues of any symmetric positive definite matrix are each bounded by their corresponding sorted row sums. □

Motivated by numerical evidence that to leading order in ϵ the eigenvectors of the Hessian are the right singular vectors of the Vandermonde matrix, we shall transform into that basis. We first bound the singular values of the Vandermonde matrix. Conveniently, $VV^\top$ has the form necessary for Conjecture 1. The singular values of V are the positive square root of the eigenvalues of $VV^\top$. Factoring the appropriate power of ϵ from each row of the Vandermonde matrix gives $V = EX$ and $VV^\top = EXX^\top E$, where E is the same as in Conjecture 1 and the elements of X are bounded by one. Equating $XX^\top$ with the matrix S in Conjecture 1, we conclude that the eigenvalues of $VV^\top$ scale as $\lambda_m(VV^\top) = \mathcal{O}(\epsilon^{2(m-1)})$ and thus $\sigma_m(V) = \mathcal{O}(\epsilon^{m-1})$.

We now transform the Hessian into this basis, and again use Conjecture 1 to bound its eigenvalues. Starting with the decomposition $H = V^\top A^\top A V$, taking the singular value decomposition of $V = U \Sigma W^\top$, and transforming the Hessian into the basis of the right singular vectors of the V, we have $W^\top H W = \tilde{H} = \Sigma^\top U^\top A^\top A U \Sigma$. We know that $\Sigma_{ii} = \mathcal{O}(\epsilon^{i-1})$. By construction, the elements of A are well behaved as $\epsilon \to 0$ and since U is an orthogonal matrix its elements too cannot diverge in this limit. This means that $\tilde{H}_{ij} = \mathcal{O}(\epsilon^{i+j-2})$. By Conjecture 1 we know that $\lambda_i(\tilde{H}) = \mathcal{O}(\epsilon^{2(i-1)})$ and since $\tilde{H}$ is simply an orthogonal transformation of H, $\lambda_i(H) = \mathcal{O}(\epsilon^{2(i-1)})$. Rigorous universality is only expected as the system size approaches infinity. Empirically we find, from studying a variety of models [3–5] as well as subsystems of models [like PC12 in Fig. 1(a)] [17], that models with more than roughly eight parameters are often recognizably sloppy.

Do these results tell us anything about the statistics of level spacings? Unless two parameters are strictly degenerate or the residuals are independent of a particular moment of the parameter distribution, $\lambda_i = l_i \epsilon^{2(i-1)}$ for some nonzero coefficient l_i. The relative spacing between neighboring eigenvalues is $s_i = \log(\lambda_i/\lambda_{i+1}) = \log(l_i/l_{i+1}) - 2\log\epsilon$. For a fixed model but an ensemble of random

parameters, the distribution of coefficients l_i has a finite width as $\epsilon \to 0$. Therefore, the distribution of s_i over the ensemble, normalized by $2\log\epsilon$ such that the average spacing is unity, goes to one with a width which vanishes as $\epsilon \to 0$. This means that the whole system is becoming not only more sloppy (larger spacing) but it is becoming almost deterministically so (strong level repulsion). Figure 1(c) is a clear depiction of this remarkably strong level repulsion.

What is the link between the Vandermonde ensemble at small ϵ and the behavior of real world sloppy models [Fig. 1 columns (a), (b)] and also the behavior at large ϵ [column (d)]? These systems share the roughly uniform density of log eigenvalues over many decades that is the signature of sloppy models but do not exhibit strong level repulsion. The real world models also do not share the strict requirement that the residuals be perfectly symmetric functions of the parameters. We conjecture that while not all of the parameters are interchangeable in real world sloppy models, there are Vandermonde subsystems lurking below the surface. Thus the fastest decay rates in column (d) constitute one Vandermonde subsystem and the slowest decay rates another. Indeed, the Poisson statistics of level spacings when fitting exponential decays from a wide range [e.g., $2\epsilon = 50$ as in (d)] can be reproduced by superimposing the spectra of several separate experiments, each fitting decays from a narrower range [e.g., $2\epsilon = 3.5$ as in (c)]. Such a decomposition into Vandermonde subsystems is also illustrated by modifying the net radiation model to include the initial amounts of the elements as unknown parameters. Now the parameters clearly separate into two classes—decay rates and initial amounts. Each class alone fits the assumptions of the Vandermonde ensemble, produces rigidly (strong level repulsion) sloppy spectra, and generates nearly equivalent patterns of changes in the residuals. When mixed together however, the fact that parameters from one class cannot compensate for parameters of the other class destroys the correlations between levels and they do not repel each other anymore. Similarly, a full many body wave function in quantum Monte Carlo calculations [18] decomposes into the sloppy space of the Jastrow parameters in Fig. 1(b) and a nonsloppy subspace of the configuration interaction coefficients describing single-particle orbitals.

These results motivate algorithms for the decomposition of real world sloppy models into rigidly sloppy Vandermonde subspaces whose components are effectively redundant. Such a decomposition would be useful for three separate reasons: (a) explaining why a particular model is sloppy overall, (b) suggesting routes for model reduction and coarse graining by subsuming degrees of freedom within Vandermonde systems, and (c) prescribing changes in parameters to alter specific aspects of model behavior.

Complex models from a wide array of scientific fields are *sloppy*: they each have an exponentially large range of sensitivities to changes in underlying parameter values. This occurs because the parameters natural for experimental manipulation or human description are often a severe distortion of the basis natural for describing system behavior. Far from being a deficiency, sloppiness is in fact a saving grace of complex models—provided the right combinations of parameters are known, they provide nontrivial and well-constrained predictions despite surprisingly unconstrained parameters overall. The origins and implications of sloppiness in its various incarnations offers new, fundamental insights into complex systems.

The authors thank R. Cerione, P. Flaherty, D. Schneider, and C. Umrigar for helpful discussions. We also thank C. Umrigar for the variational quantum Monte Carlo eigenvalues. J. J. W. was supported by the DOE Computational Science Graduate program. R. N. G. was supported by an NIH Molecular Biophysics Training Grant, No. T32-GM-08267. C. R. M. acknowledges support from USDA-ARS Project No. 1907-21000-017-05. This work was supported by NSF Grant No. DMR-0218475.

[1] K. Brown and J. Sethna, Phys. Rev. E **68**, 021904 (2003).
[2] K. Brown, C. Hill, G. Calero, C. Myers, K. Lee, J. Sethna, and R. Cerione, Phys. Med. Biol. **1**, 184 (2004).
[3] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna (to be published).
[4] S. Frederiksen, K. Jacobsen, K. Brown, and J. Sethna, Phys. Rev. Lett. **93**, 165501 (2004).
[5] J. Mortensen, K. Kaasbjerg, S. Frederiksen, J. Norskov, J. Sethna, and K. Jacobsen, Phys. Rev. Lett. **95**, 216401 (2005).
[6] K. Gadkar, J. Varner, and F. Doyle III, IEEE Systems Biology **2**, 17 (2005).
[7] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C* (Cambridge University Press, Cambridge, England, 1996), 2nd ed.
[8] F. P. Casey, J. J. Waterfall, R. N. Gutenkunst, C. R. Myers, and J. P. Sethna (to be published).
[9] C. Lanczos, *Applied Analysis* (Prentice-Hall, Englewood Cliffs, NJ, 1956).
[10] A. van den Bos and J. H. Swarte, IEEE Transactions on Signal Processing **41**, 313 (1993).
[11] J. Wishart, Biometrika **20**, 32 (1928).
[12] V. Marčenko and L. Pastur, Math USSR Sbornik **1**, 457 (1967).
[13] M. L. Mehta, *Random Matrices* (Academic, New York, 1991).
[14] T. Guhr, A. Müller-Groeling, and H. A. Weidenmüller, Phys. Rep. **299**, 189 (1998).
[15] C. Beenakker, Rev. Mod. Phys. **69**, 731 (1997).
[16] D. Hilbert, Acta Math. Acad. Sci. Hung. **18**, 155 (1894).
[17] J. J. Waterfall, Ph.D. thesis, Cornell University (to be published).
[18] *Quantum Monte Carlo Methods in Physics and Chemistry*, edited by P. Nightengale and C. Umrigar (Kluwer, Dordrecht, 1999).
[19] Radioactive Decay Tables (2004), http://las.perkinelmer.com/content/TotallyRAD/pages/decaytables.html.