One parameter is always enough

Steven T. Piantadosi
Department of Brain and Cognitive Sciences
358 Meliora Hall, P.O. Box 270268
University of Rochester
Rochester, NY 14627

We construct an elementary equation with a single real valued parameter that is capable of fitting any "scatter plot" on any number of points to within a fixed precision. Specifically, given given a fixed $\epsilon > 0$, we may construct f_θ so that for any collection of ordered pairs $\{(x_j, y_j)\}_{j=0}^n$ with $n, x_j \in \mathbb{N}$ and $y_j \in (0, 1)$, there exists a $\theta \in [0, 1]$ giving $|f_\theta(x_j) - y_j| < \epsilon$ for all j simultaneously. To achieve this, we apply prior results about the logistic map, an iterated map in dynamical systems theory that can be solved exactly. The existence of an equation f_θ with this property highlights that "parameter counting" fails as a measure of model complexity when the class of models under consideration is only slightly broad.

The mathematician John von Neumann famously admonished that with four free parameters he could make an elephant, and with five he could make it wiggle its trunk [1]. Indeed, the number of free parameters is often taken as a proxy of model complexity intuitively as well as in quantitative model comparison measures like AIC [2] and BIC [3]. While these measures can be shown to be statistically principled or optimal for certain classes of models [4], they are often used to evaluate arbitrary models by practitioners in a given field. The aim of this short note is to show that, in fact, very simple, elementary models exist that are capable of fitting arbitrarily many points to an arbitrary precision using only a single real-valued parameter θ. This is not always due to severe pathologies—one such model, studied here, is infinitely continuously differentiable as a function of θ. The existence of this model has implications for statistical model comparison, and shows that great care must be taken in machine learning efforts to discover equations from data [5–7] since some simple models can fit any data set arbitrarily well.

We will consider the simple setting of a "scatter plot" of x-values at natural numbers $0, 1, 2, 3, \ldots, n$ and y-values in $(0, 1)$. We show how to construct an elementary function f_θ that, as θ varies, may fit any collection of ordered pairs $\{(x_i, y_j)\}_{j=0}^n$ to an arbitrary precision $\epsilon > 0$. The existence of a solution for $x = 0, 1, 2, \ldots, n$ implies the existence of solution for any subset of these integers. The general approach taken here will be to first find an initial condition of a chaotic dynamical system whose orbit comes close to values related to each y_j. Then, an exact solution

of this dynamical system yields an equation $y = f_\theta(x)$ that, as x varies, recovers the system's dynamics with initial condition θ. This approach is related to attempts to encode computations into chaotic dynamical systems [8, 9]. The techniques deployed here are not novel mathematically, but this lesson from dynamical systems theory has not been explicitly articulated in the literature on statistics and model comparison.

We will make use of the logistic map $m(z) = 4z(1-z)$ whose iterated application can be solved exactly [10] for a given initial value θ as

$$m^k(\theta) = \sin^2\left[2^k \arcsin\sqrt{\theta}\right]. \tag{1}$$

This solution follows from the double angle identity,

$$\begin{aligned} m(\sin^2(z)) &= 4\sin^2(z)(1-\sin^2(z)) \\ &= 4\sin^2(z)\cos^2(z) \\ &= \sin(2z)^2 \end{aligned} \tag{2}$$

and the requirement that $m^0(\theta) = \theta$. The map m is chaotic [11] and it is well-established that m may be viewed as a shift map on θ through its conjugacy via $\varphi(z) = \sin^2(2\pi z)$ to the Bernoulli map,

$$S(z) = \begin{cases} 2z & \text{if } 0 < z < \dfrac{1}{2} \\ 2z-1 & \text{if } \dfrac{1}{2} \leq z < 1. \end{cases} \tag{3}$$

S has the effect of removing the first bit of a binary expansion $0.z_1z_2z_3\cdots$ of z, so that

$$S(0.z_1z_2z_3\cdots) = 0.z_2z_3\cdots. \tag{4}$$

This property of S means that we may construct a point $\omega \in (0,1)$ whose orbit under S will bring it arbitrarily close to each member of any collection of points. Specifically, let us fix $\epsilon > 0$ and choose $r \in \mathbb{N}$ so that $2^{-r} < \epsilon/2$. We will define $y'_j = \varphi^{-1}(y_j)$ and denote the binary expansion of y'_j as $0.y'_{j1}y'_{j2}y'_{j3}\cdots$. Define the parameter value $\omega \in (0,1)$ by concatenating the first r binary digits of each y'_j,

$$\omega = 0.y'_{11}y'_{12}\cdots y'_{1r}y'_{21}y'_{22}\cdots y'_{2r}\cdots y'_{n1}y'_{n2}\cdots y'_{nr}. \tag{5}$$

Due to the construction of ω and the ability to interpret S as removing the leftmost bit, $S^{rj}(\omega)$ agrees with y'_j on its first r bits, so,

$$|S^{rj}(\omega) - y'_j| < 2^{-r} < \epsilon/2 \quad \text{for all } j = 0, 1, 2, \ldots, n. \tag{6}$$

The ability to construct such an orbit relies ultimately on the fact that S is continuous and topologically mixing. Since φ is a homeomorphism between S and m, $S^{rj} = \varphi^{-1} \circ m^{rj} \circ \varphi$. Moreover, φ is Lipschitz continuous and in particular $2\,|x-y| > |\varphi(x) - \varphi(y)|$ for all $x, y \in (0,1)$. Putting these two facts together with (6) yields that for all j,

$$\epsilon > 2\,|S^{rj}(\omega) - y'_j| = 2\,|\varphi^{-1}(m^{rj}(\varphi(\omega)))) - \varphi^{-1}(y_j)| > |m^{rj}(\varphi(\omega)))) - y_j|, \tag{7}$$

where the last inequality follows the Lipschitz condition and application of φ to each term inside the absolute value. This presentation has elided one technical factor, which is that φ is not one-to-one on $(0,1)$ and so φ^{-1} has two possible values. This has the consequence that in (7), $m^{rj}(\varphi(\omega))))$ may be close to *either* y_j or $1 - y_j$, since φ is symmetrical about $1/2$. To address this, we may always use the lower value for φ^{-1} and scale the y_j so that they are always below 0.5. This scaling may then be inverted in the output of the final equation if desired.

Equation (7) shows that m^{rj} will come ϵ close to each of the y_j when started on value $\theta = \varphi(\omega)$. Thus, we may define a single parameter equation,

$$f_\theta(x) = m^{rx}(\theta) = \sin^2\left[2^{rx} \arcsin\sqrt{\theta}\right] \tag{8}$$

where choosing $\theta = \varphi(\omega)$ yields that $|f_\theta(x_j) - y_j| < \epsilon$ for all j. Of course, the y_j were freely chosen, showing that f_θ can approximate any data set $\{(x_j, y_j)\}_{j=0}^n$ as θ varies in $[0,1]$. Note that for a fixed ϵ, the number of data points n that can be fit is not bounded. In addition, this f_θ is continuous and differentiable for fixed r—indeed infinitely continuously differentiable—and so it will satisfy nearly all regularity conditions that would normally weed out such pathological functions in the context of parameter estimation.

To illustrate that (8) can fit an arbitrary data set as θ varies, Figure 1 shows the value $f_\theta(0), f_\theta(1), f_\theta(2), \ldots$, for two values of θ. Each parameter value was created by following the construction above using target y_j chosen at each x value from the black pixels of a line drawing of either an elephant (left) or signature (right). The implementation used the arbitrary precision library mpmath in python [12]. Both are able to be fit well by $f_\theta(x)$ if θ is appropriately tuned. This single parameter model provides a large improvement over the prior state of the art in fitting an elephant [13, 14]. Note that the only shown x values are integers, and between these integers are the rapidly oscillating sin patterns implied by (8).

The equation $f_\theta(x)$ is extraordinarily sensitive to its single parameter θ and in fact will generalize to $x > n$ in ways that depend only on the digits of θ are after the last digit of y'_n. Thus, while fitting the data, generalization behavior is completely determined by the free parameter's less significant

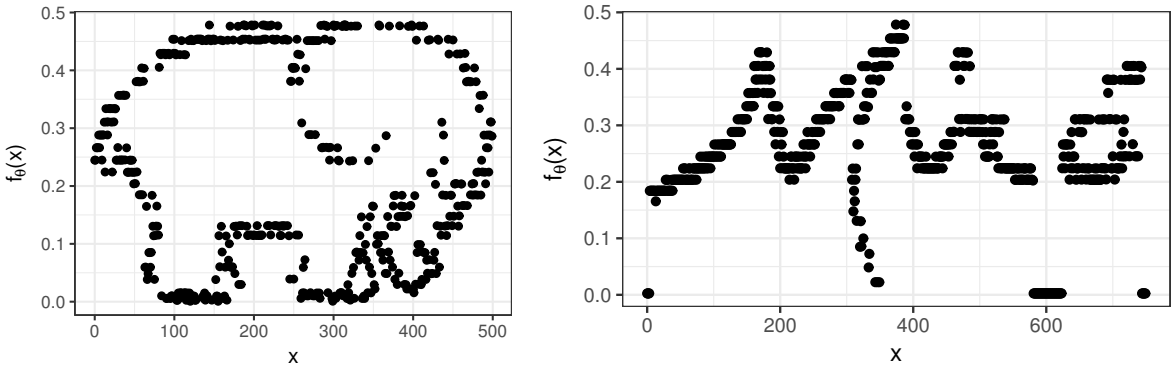

FIG. 1: A scatter plot of f_θ for $\theta = 0.244684726673474545822754065 6\cdots$ plotted at integer x values, showing that a single parameter can fit an elephant (left). The same model run with parameter $\theta = 0.00242654180550004019353876 20\cdots$ showing a fit of a scatter plot to Joan Miró's signature (right). Both use $r = 8$ and require hundreds to thousands of digits of precision in θ.

digits. This implies that there can be no guarantees about the performance of f_θ in extrapolation, despite its good fit. Thus, the construction shows that *even a single parameter* can overfit the data, and therefore it is not always preferable to use a model with fewer parameters. This fact is related to the observation, in a setting of classification, that $f(x) = \sin(x)$ has an infinite VC-dimension [15].

The existence of such a simple equation with such freedom in behavior illustrates a more basic problem that model complexity cannot be determined by counting parameters. More generally, uncritical use of a "parameter counting" approach ignores the fact that a single real-valued parameter potentially contains an unboundedly large amount of information since a real number requires an infinite number of bits to specify. Indeed, the set of real numbers that can even be described with finitely many bits (e.g. by a Turing machine) is countable and thus has measure zero. Given the existence of injective maps between $\mathbb{R}^n$ and $\mathbb{R}$ [16], the number of parameters in a model cannot be a meaningful measure of its complexity once the class of models is large enough to implement these maps and effectively decode one single number into many. However, such embeddings are not continuous nor likely constructible as an ordinary looking equation that a scientist is likely to encounter.

The example provided in this paper shows that the infinite amount of information in a real valued parameter can be *decoded* quite simply, using just sin and exponentiation. The existence of such a simple yet problematic equation implies that attempts both at broad model comparison and automatic discovery of equations from data may often be ill-posed. Quantitatively, parameter-counting methods should be dispreferred relative to model comparisons based on measures that incorporate

the precision required of real-valued parameters, including Minimum Description Length [17]. The result also emphasizes the importance of constraints on scientific theories that are enforced independently from the measured data set, with a focus on careful a priori consideration of the class of models that should be compared [4].

[1] F. Dyson, "A meeting with Enrico Fermi," *Nature*, vol. 427, no. 6972, p. 297, 2004.

[2] H. Akaike, "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on*, vol. 19, no. 6, pp. 716–723, 1974.

[3] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[4] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, 2003.

[5] P. Langley, G. L. Bradshaw, and H. A. Simon, "BACON.5: The discovery of conservation laws," in *IJCAI*, vol. 81, pp. 121–126, 1981.

[6] J. Koza, *Genetic programming: on the programming of computers by means of natural selection*. Cambridge, MA: MIT Press, 1992.

[7] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," *Science*, vol. 324, no. 5923, pp. 81–85, 2009.

[8] S. Sinha and W. L. Ditto, "Dynamics based computation," *Physical Review Letters*, vol. 81, no. 10, p. 2156, 1998.

[9] T. Munakata, S. Sinha, and W. L. Ditto, "Chaos computing: implementation of fundamental logical gates by chaotic elements," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 49, no. 11, pp. 1629–1633, 2002.

[10] E. Schröder, "Ueber iterirte Functionen," *Mathematische Annalen*, vol. 3, no. 2, pp. 296–322, 1870.

[11] R. Devaney, *An introduction to Chaotic Dynamical Systems*. Westview Press, 2008.

[12] F. Johansson, "mpmath: a Python library for arbitrary-precision floating-point arithmetic," 2013.

[13] J. Wei, "Least square fitting of an elephant," *Chemtech*, vol. 5, no. 2, pp. 128–129, 1975.

[14] J. Mayer, K. Khairy, and J. Howard, "Drawing an elephant with four complex parameters," *American Journal of Physics*, vol. 78, no. 6, pp. 648–649, 2010.

[15] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, vol. 1. Springer series in statistics New York, 2001.

[16] G. Peano, "Sur une courbe, qui remplit toute une aire plane," *Mathematische Annalen*, vol. 36, no. 1, pp. 157–160, 1890.

[17] P. D. Grünwald, *The Minimum Description Length Principle*. MIT press, 2007.