www.sciencemag.org/content/full/342/6158/604/suppl/DC1

Supplementary Material for

Parameter Space Compression Underlies Emergent Theories and Predictive Models

Benjamin B. Machta, Ricky Chachra, Mark K. Transtrum, James P. Sethna*

*Corresponding author. E-mail: sethna@lassp.cornell.edu

Published 1 November 2013, *Science* **342**, 604 (2013)
DOI: 10.1126/science.1238723

This PDF file includes:

Supplementary text

Figs. S1 and S2

References (*27–42*)

Supplementary Text

1 Introduction

This document contains relevant background and computational details to accompany the main text. Section 2 provides a pedagogical overview of the information theoretic tools used to quantify distinguishability. Section 3 discusses an application of this formalism to a model of stochastic motion providing details of the calculation underlying fig. 2 in the main text. In addition, it gives an asymptotic analysis of the scaling of the FIM's eigenvalues in the limit of late time observations. Sections 4–7 discuss the Ising model. Section 4 defines the 13-parameter Ising model introduced in the main text. Section 5 outlines the numerical techniques for measuring the FIM, and gives a scaling argument that explains its spectrum before coarsening. Section 6 extends this analysis to the coarsened case. Section 7 gives details on the Monte Carlo techniques used, with emphasis on this paper's implementation of the 'Compatible Monte Carlo' first used in (*24*).

This document also includes a supplement to fig. 1 from the main text (fig. S1). This figure provides additional examples of seemingly disparate models whose FIM's, when fit to their natural data, exhibit a characteristic sloppy spectrum. These include a differential equation model of a circadian rhythm where parameters describe reaction rates and saturation coefficients (*11, 29*), a variational wave-function problem where parameters are Jastrow factors that specify the ground-state wave function (*12*), and a relaxation oscillator where parameters generalize the van der Pol system describing oscillations between two almost stable states (*28*). In addition we include a model from engineering, for fine-tuning the beam of a particle accelerator (the Energy Recovery Linac) where parameters describe the positions of magnets used to shape the beam. This last model was implemented using TAO (the Tool for Accelerator Optics (*30*)) and we thank Georg Hoffstaetter, David Sagan, and Christopher Mayes for the use of their code.

2 Information geometry and the Fisher Information matrix

This section gives an overview of the information theoretic approach used throughout (*16, 31, 32*) motivated by the following questions: how different are two probability distributions, $P_1(x)$ and $P_2(x)$, and what is an appropriate measure of distance between them? Can one test the hypothesis that a set of independent data points $\{x_1, x_2, ..., x_N\}$ (unbeknownst to us generated by P_1) was instead generated by P_2? The probability that P_1 would have generated the data is given by its likelihood:

$$P_1(\{x_1, x_2, ...x_N\}) = \prod_i P_1(x_i) = \exp\left(\sum_i \log P_1(x_i)\right) \tag{S1}$$

To determine which of two candidate models more probably generated this sequence of data, one considers the log likelihood ratio:

$$\lambda(\{x_1, x_2, ...x_N\}) = \log\left(\frac{P_1(\{x_1, x_2, ...x_N\})}{P_2(\{x_1, x_2, ...x_N\})}\right) \tag{S2}$$

If λ is large and positive (negative) than the data suggests P_1 (P_2). Alternatively, if λ is close to zero than either model could be valid and the data is inconclusive. How much data is needed before one should expect that one model distinguishes itself? In a given distribution, λ is a stochastic variable. However, one can define the expectation value for $\lambda(x)$ in distribution P_1, giving the log likelihood per sample that an ensemble drawn from P_1 could have instead been drawn from P_2.

This defines the Kullback-Liebler Divergence, D_{KL}, a statistical measure of how distinguishable P_1 is from P_2 from its data x (*32, 33*)

$$D_{KL}(P_1||P_2) = \sum_x P_1(x)\lambda(x) = \sum_x P_1(x)\log\left(\frac{P_1(x)}{P_2(x)}\right). \tag{S3}$$

Because D_{KL} does not necessarily satisfy $D_{KL}(P_1||P_2) = D_{KL}(P_2||P_1)$, it is not a mathematically proper distance metric[1]. However, D_{KL} becomes symmetric for two 'nearby' models. For a continuously parameterized set of models P_θ where θ is a set of N parameters θ^μ, the infinitesimal D_{KL} between models P_θ and $P_{\theta+\Delta\theta}$ is[2]

$$D_{KL}(P_\theta, P_{\theta+\Delta\theta}) = g_{\mu\nu}\Delta\theta^\mu\Delta\theta^\nu + \mathcal{O}\Delta\theta^3, \tag{S4}$$

where $g_{\mu\nu}$ is the Fisher Information Matrix (FIM), given by [3]

$$g_{\mu\nu}(P_\theta) = -\sum_x P_\theta(x)\frac{\partial}{\partial\theta^\mu}\frac{\partial}{\partial\theta^\nu}\log P_\theta(x). \tag{S5}$$

The quadratic form of the KL-divergence at short distances motivates using the FIM as a metric on parameter space. The FIM is symmetric, positive-definite, and transforms like a covariant rank-2 tensor under parameter transformations, endowing it with all the properties of a Riemannian metric, the study of which is known as information geometry (*16*). In fact, the FIM is the *unique* natural Riemannian metric that is consistent with the additional structure that each point specifies a probability distribution[4].

[1] A distance measure should also satisfy some sort of generalized triangle inequality- at the very least $D(A,B)+D(B,C) \geq D(A,C)$ which is also not necessarily satisfied here.

[2] It is an interesting exercise to show that there is no term linear in $\Delta\theta$. The crucial step uses that P_θ is a probability distributions so that $\partial_\mu \sum_x P_\theta(x) = 0$.

[3] Although the KL-divergence is a common measure of statistical distinguishability among probability distributions, it is not unique. In fact it is a member of a broader class of divergences known as the f-divergences, which take the form $D_f(P_1,P_2) = \sum_x P_1(x) f\left(\frac{P_2(x)}{P_1(x)}\right)$ for some function $f(t)$ that is convex and satisfies $f(1) = 0$. The KL-divergence therefore corresponds to the choice $f(t) = -\log(t)$. Other common choices are $f(t) = 2(1-\sqrt{t})$, corresponding to the Hellinger Distance, and $f(t) = |t-1|$, corresponding to the total variation distance. For our purposes, this distinction is unimportant: the Fisher Information is the lowest order contribution to any f-divergence for infinitesimally separated probability distributions.

[4] Riemannian metrics have more structure than other metric spaces since the metric tensor defines an inner product on the tangent space at each point on the manifold. The FIM is the only inner product that is invariant under specific probabilistically important mappings. The basic argument considers partitions on the domain of the probability distribution, known as Markov mappings. Requiring that the inner product be invariant under these mappings is a rigid constraint that is only satisfied by the FIM (*34, 35*)

Information geometry provides a framework for understanding more generalized Bayesian inference. It gives an immediate derivation of Jeffreys' 'uninformative' prior (*36*): the invariant volume element in any Riemannian geometry is given by $\sqrt{\det(g)}d\theta^1 d\theta^2 \dots d\theta^N$. In a Bayesian inference scheme, choosing a prior on parameter space equal to $\sqrt{\det(g)}/Z$ ensures that model predictions are reparameterization invariant. The normalization constant, Z, is the invariant volume of the manifold that quantifies the amount of information expected to be gained from a single measurement of x.

The FIM is well defined for any models that predict stochastic data. The next sub-sections derive the form of the FIM for two special cases used in this work, the case of Gaussian models, and the case of exponential families familiar from statistical physics. The similarity of parameter space structure in these seemingly very different classes of models suggests that it is not an artifact of the particular choice of stochastic model employed.

2.1 The metric of a Gaussian model

Nonlinear least squares models output a vector of data, y_0^i (for $1 < i < M$), that is generated assuming that the observations y^i are normally distributed with widths σ^i around prediction $\vec{y_0}(\theta)$. The fitting 'cost' or sum of squared residuals is proportional to the negative log likelihood (plus a constant), hence the probability distribution of data is

$$P_\theta(\vec{y}) \sim \exp\left(-\sum_i (y^i - y_0^i(\theta))^2/2\sigma^{i2}\right). \tag{S6}$$

Defining the Jacobian between parameters and scaled data as

$$J_{i\mu} = \frac{1}{\sigma^i}\frac{\partial y_0^i(\theta)}{\partial \theta^\mu}, \tag{S7}$$

the Fisher Information Matrix for least squares problems is given by[5] (*13, 14*)

$$g_{\mu\nu} = \sum_i J_{i\mu} J_{i\nu}. \tag{S8}$$

The Euclidean distance between nearby points in prediction space

$$\begin{aligned} \sum (\Delta y_i)^2 &= \sum_{i,\mu,\nu} \left(\frac{\partial y^i}{\partial \theta^\mu} \Delta\theta^\mu \frac{\partial y^i}{\partial \theta^\nu} \Delta\theta^\nu \right) \\ &= g_{\mu\nu} \Delta\theta^\mu \Delta\theta^\nu \end{aligned} \tag{S9}$$

is the metric tensor contracted with corresponding displacements $\Delta\theta^\mu$ in parameter space. Thus the FIM has a geometric interpretation: distance is locally the same as that measured by embedding the model in the space of scaled data according to the mapping $y_0(\theta)$ (it is *induced* by the Euclidian metric in data space). This metric was shown to be sloppy in seventeen models from the systems biology literature (*11*) and in several other contexts. See Fig. S1 and (*12*).

2.2 The metric of a Statistical Mechanical Model

Exponential models familiar from statistical mechanics are defined by a parameter set θ dependent Hamiltonian H that assigns an energy to every possible configuration x. Each parameter θ^μ controls the relative weighting of some function of the configuration, $\Phi_\mu(x)$, which together define the probability distribution on configurations through the following (with temperature

[5] This assumes that the uncertainty σ^i does not depend on the parameters, and that errors are diagonal. Both of these assumptions seem reasonable for a wide class of models if measurement error dominates. The more general case is still tractable, but less transparent

and Boltzmann's constant set to 1)

$$\begin{aligned} P(x|\theta) &= \exp(-H_\theta(x))/Z, \\ Z(\theta) &= \exp(-F(\theta)) = \sum_x \exp(-H_\theta(x)), \\ H_\theta(x) &= \sum_\mu \theta^\mu \Phi_\mu(x) \end{aligned} \tag{S10}$$

Here F is the Helmholtz free energy. Many models can be put into this exponential form. For example, the 2d Ising model of section 4 has spins $s_{i,j} = \pm 1$ on a square $L \times L$ lattice with the configuration, $x = \{s_{i,j}\}$, being the state of all spins. The magnetic field, $\theta^0 = h$ multiplies $\Phi_0(\{s_{i,j}\}) = \sum_{i,j} s_{i,j}$, and the nearest neighbor couplings, $\theta^{01} = \theta^{10} = -J$ multiplies $\Phi_1(\{s_{i,j}\}) = \sum_{i,j} s_{i,j}s_{i+1,j} + s_{i,j}s_{i,j+1}$. This form is chosen for convenience in calculating the metric, which is written (*21, 23, 37*)[6]

$$\begin{aligned} g_{\mu\nu} &= \langle -\partial_\mu \partial_\nu \log(P(x)) \rangle \,, \\ &= \langle \partial_\mu \partial_\nu H(x) \rangle + \partial_\mu \partial_\nu \log(z), \\ &= \partial_\mu \partial_\nu \log(z) = -\partial_\mu \partial_\nu F. \end{aligned} \tag{S11}$$

In the last equation we have taken advantage of the fact that the Hamiltonian is linear in parameters θ^μ so that $\langle \partial_\mu \partial_\nu H(x) \rangle = 0$.

[6]Several seemingly reasonable metrics can be defined for systems in statistical mechanics and all give similar results in most circumstances (*23*). Most differences occur either for systems not in a true thermodynamic (N large) limit, or for systems near a critical point. As far as we are aware, Crooks (*21*) was the first to stress that the one used here can be derived from information theoretic principles, perhaps making it the most 'natural' choice. Crooks showed (*21*) that when using this metric 'length' has an interesting connection to dissipation by way of the Jarzynski equality (*38*).

3 A Continuum Limit: Diffusion

In a 'microscopic' model of stochastic motion on a discrete lattice of sites j, parameters θ^μ, for $-N \leq \mu \leq N$ describe the probability that in a discrete time step a particle will transition[7] from site j to site $j + \mu$. Particles are initially at the origin and measurement data consists of the number of particles $\rho_\tau(j)$ at some later time τ.

'Microscopic' measurements of model parameters are taken after starting with the initial probability distribution $\rho_0(j) = \delta_{j,0}$ and observing the new distribution after one time step, $\rho_1(j)$. This distribution is given by

$$\rho_1(j) = \theta^j. \tag{S12}$$

Assuming that the measurement uncertainty of the number of particles at each site is Gaussian, with width[8] $\sigma_{meas} = 1$. The FIM on the parameter space defined in equations S7 and S8 becomes

$$\begin{aligned} J_{i,\mu} &= \partial_\mu \rho_1(i) = \delta_{i,\mu}, \\ g_{\mu\nu} &= \textstyle\sum_i J_{i,\mu} J_{i,\nu}, \\ &= \delta_{\mu\nu}. \end{aligned} \tag{S13}$$

This metric has $2N + 1$ eigenvalues each with value $\lambda = 1$. All of the parameters in this model are measurable with equal accuracy. This of course changes as one examines data that is in the form of densities measured after multiple time steps as discussed next.

[7] If $\sum \theta_\alpha \neq 1$, then particles do not just hop but maybe created or destroyed with net rate $R = \sum \theta_\alpha - 1$. θ^μ then describes the probability that if an isolated particle is at site j at time τ, then one will be at site $j + \mu$ at time $\tau + 1$.

[8] One could carry out a more complicated calculation assuming uncertainty comes from the stochastic nature of the model itself, but with many particles, this approach would yield similar but less transparent results. Changing the measurement uncertainty from 1 to σ_{meas} will multiply all calculated metrics by a trivial factor of $1/\sigma^2_{meas}$ which is omitted for clarity.

3.1 Coarsening the diffusion equation by observing at long times

To calculate the density of particles at position j and time τ, $\rho_\tau(j)$, it is useful to introduce the Fourier transform of the hopping rates, as well as the Fourier transform of the particle density at time τ

$$\begin{aligned} \tilde{\theta}^k &= \sum_{\mu=-N}^{N} e^{-ik\mu}\theta^\mu, \\ \tilde{\rho}_\tau^k &= \sum_{j=-\infty}^{\infty} e^{-ikj}\rho_\tau(j), \\ \rho_\tau(j) &= \tfrac{1}{2\pi}\int_{-\pi}^{\pi} dk e^{ikj}\tilde{\rho}_\tau^k. \end{aligned} \tag{S14}$$

In a time step the density distribution is convolved by the hopping rates. In Fourier space, this is simply written as[9]

$$\tilde{\rho}_\tau^k = \tilde{\theta}^k \tilde{\rho}_{\tau-1}^k. \tag{S15}$$

Initially, all particles are at the origin $\rho_0(j) = \delta_{j,0}$, hence $\tilde{\rho}_0^k \equiv 1$ and

$$\begin{aligned} \tilde{\rho}_\tau^k &= (\tilde{\theta}^k)^\tau, \\ \rho_\tau(j) &= \tfrac{1}{2\pi}\int_{-\pi}^{\pi} dk e^{ikj}(\tilde{\theta}^k)^\tau. \end{aligned} \tag{S16}$$

The Jacobian and metric at time τ can now be written

$$\begin{aligned} J_{j\mu}^\tau &= \partial_\mu \rho_\tau(j) = \tfrac{\tau}{2\pi}\int_{-\pi}^{\pi} dk e^{ik(j-\mu)}(\tilde{\theta}^k)^{\tau-1}, \\ g_{\mu\nu}^\tau &= \tfrac{\tau^2}{2\pi}\int_{-\pi}^{\pi} dk e^{ik(\mu-\nu)}(\tilde{\theta}^k)^{\tau-1}(\tilde{\theta}^{-k})^{\tau-1}. \end{aligned} \tag{S17}$$

Note that the metric now depends on θ. The preceding formulae were used to calculate the sloppy spectrum of the fig. 2. After many steps, the three stiffest eigendirections of $g_{\mu\nu}$ become the three terms in the diffusion equation as discussed next.

[9]This is due to the convolution theorem. For example, see (*39*)

The late time behavior of $g^{\tau}_{\mu\nu}$ is dominated by small k values appearing in the integrand of equation S17. For small values of k

$$\begin{aligned}
\tilde{\theta^k} &= (1+R)(1-ikV-\tfrac{k^2}{2}(D+V^2))+\mathcal{O}(k^3) \\
&= (1+R)\exp(-ikV-D\tfrac{k^2}{2})+\mathcal{O}(k^3), \\
R &= \textstyle\sum_\mu \theta^\mu - 1 \\
V &= \tfrac{1}{1+R}\textstyle\sum_\mu \mu\theta^\mu, \\
D &= \tfrac{1}{1+R}\textstyle\sum_\mu \mu^2\theta^\mu - V^2.
\end{aligned} \tag{S18}$$

In the preceding, note that the first two equations are identical up to second order in k, R is the particle creation rate, V is the drift, and D is the diffusion constant. For the case where the drift $V=0$ and particle creation rate $R=0$, at late times

$$\begin{aligned}
g^{\tau}_{\mu\nu} &\approx \tfrac{\tau^2}{2\pi}\int_{-\infty}^{\infty} dk e^{ik(\mu-\nu)}e^{-D\tau k^2} \\
&\sim \tfrac{\tau^2}{(D\tau)^{1/2}}e^{-(\mu-\nu)^2/4D\tau}.
\end{aligned} \tag{S19}$$

Expanding this in powers of the small parameter $(\mu-\nu)^2/D\tau$ gives

$$\begin{aligned}
g^{\tau}_{\mu\nu} &\sim \tau^2((D\tau)^{-1/2}-(D\tau)^{-3/2}(\mu-\nu)^2/4+\cdots) \\
&= \tau^2\sum_{n=0}^{\infty}\tfrac{(-1)^n(\mu-\nu)^{2n}}{n!(4D\tau)^{n+1/2}}.
\end{aligned} \tag{S20}$$

Each term in the series contributes a single new non-zero eigenvalue which scales like

$$\lambda_n \sim \tau^2\left(\frac{D\tau}{N^2}\right)^{-n-1/2}, n\geq 0. \tag{S21}$$

The corresponding eigenvectors are best understood by considering their projection onto the observables and are proportional to the left singular vectors of J as $v_{L,n} = (1/\lambda_n)J_{i\mu}v^{\mu}_n$.

These are exactly the Hermite polynomials multiplied by a Gaussian with width $2\sigma = \sqrt{D\tau}$. Thus at late times, when the Gaussian goes to a constant in the range $-N$ to N, the stiffest eigendirection is proportional to the non-conservation of particle number $R = \sum_\mu \theta^\mu - 1$, the second measures drift $V = \frac{1}{1+R}\sum_\mu \mu\theta^\mu$, and next is the diffusion constant, D. The next terms are less familiar; those past $n = 2$ never appear in a continuum description, because they are always harder to observe than the diffusion constant by a factor of the ratio of the observation scale ($\sqrt{D\tau}$) to the microscopic scale (N) raised to a negative integer power. It is not possible for the diffusion constant, as defined here, to be zero while any higher cumulants are non-zero, explaining why though drift and the diffusion constant both appear in continuum limits, the physical parameter that describes the third cumulant does not. The next eigendirection measures the skew of the resulting density distribution, while the next one measures the distribution's kurtosis, and so on. It is worth noting that careful observation of a particular θ^μ, somewhat analogous to knowing the bond-angle of a water molecule, would give very little insight on the relevant observables. The exact eigenvalues, measured at steps $\tau = 1$–7 are plotted in fig. 2 of the main text for an $N = 3$ (seven parameter) model where $\theta^\mu = 1/7$ for all μ.

4 A critical point: The Ising model

The 2d square lattice Ising model discussed here has lattice sites $1 < i, j < L$, and degrees of freedom $s_{i,j}$ taking the values of ± 1. The probability of observing a particular configuration on the whole lattice (denoted by $\{s_{i,j}\}$) is defined by a Hamiltonian $H\{s_{i,j}\}$ that assigns each configuration of spins an energy (equation S10). The usual nearest neighbor Ising model has two parameters: a coupling strength J, and a magnetic field h defined through the equation

$$H(\{s_{i,j}\}) = J\sum_{i,j}(s_{ij}s_{ij+1} + s_{ij}s_{i+1j}) + h\sum_{i,j} s_{ij}. \tag{S22}$$

The Ising model discussed in the main text generalizes this to a larger dimensional space of possible models by including in its Hamiltonian the magnetic field θ^0, the usual nearest neighbor coupling term, and 12 other nearby couplings parameterized by $\theta^{\alpha\beta}$. Vertical and horizontal couplings are also allowed to be different. In the form of equation S10

$$\begin{aligned} H(x) = & \sum_{\alpha,\beta} \theta^{\alpha\beta} \Phi_{\alpha\beta}\left(\{s_{i,j}\}\right) + \theta^h \Phi_h\left(\{s_{i,j}\}\right), \\ \Phi_{\alpha\beta}\left(\{s_{i,j}\}\right) = & \sum_{i,j} s_{ij} s_{i+\alpha j+\beta}, \\ \Phi_h\left(\{s_{i,j}\}\right) = & \sum_{i,j} s_{ij}. \end{aligned} \tag{S23}$$

As discussed next, the FIM of this model is calculated along the line through parameter space that describes the usual Ising model ($\theta^{01} = \theta^{10} = J$ and $\theta^{\alpha\beta} = 0$ otherwise) with no magnetic field ($\theta^h = 0$).

5 Measuring the Ising metric

From equation S11, the metric for the generalized Ising model, evaluated at the nearest-neighbor standard zero-field point, can be written in terms of expectation values of observables as follows (except where necessary, the indices $\alpha\beta$ and h are condensed into a single μ)

$$g_{\mu\nu} = \partial_\mu \partial_\nu \log z = \langle \Phi_\mu \Phi_\nu \rangle - \langle \Phi_\mu \rangle \langle \Phi_\nu \rangle . \tag{S24}$$

Furthermore, given a configuration $x = \{s_{i,j}\}$, $\Phi_\mu(x)$ is just a particular two point correlation function (or the total sum of spins for Φ_h) [10]. The Wolff algorithm (*40*) was employed to generate an ensemble of configurations $x_p = \{s_{i,j}\}_p$, for $1 < p < M$, for systems with $L =$

[10] $\Phi_h\left(\{s_{i,j}\}\right) = \sum_{i,j} s_{i,j}$ is efficiently calculated for a given configuration $\{s_{i,j}\}$. $\Phi_{\alpha\beta}\left(\{s_{i,j}\}\right)$ is less trivial: one defines the translated lattice $s'_{i,j}(\alpha,\beta) = s_{i+\alpha,j+\beta}$, in terms of which we write $\Phi_{\alpha\beta}\left(\{s_{i,j}\}\right) = \sum_{i,j} s_{i,j} s'_{i,j}(\alpha,\beta)$.

64 to estimate the distribution defined in equation S24. (Results were checked against exact enumeration of all possible states on lattices up to $L = 4$.) Thus, for an ensemble of M lattice configurations x_i

$$g_{\mu\nu} = \frac{1}{M^2 - M} \sum_{p,q=1,p\neq q}^{M} \Phi_\mu(x_p)\Phi_\nu(x_p) - \Phi_\mu(x_q)\Phi_\nu(x_p). \tag{S25}$$

The results are plotted in supplementary Fig. S2. Away from the critical point in the high temperature phase (small βJ), the results seem somewhat analogous to those found for the diffusion equation viewed at its microscopic scale. All of the parameter eigendirections that control two spin couplings ($\theta^{\alpha\beta}$) are roughly of similar distinguishability. However, as the critical point is approached, the system becomes extremely sensitive both to θ^h and to a certain combination of the $\theta^{\alpha\beta}$ parameters. This divergence has been previously shown for the continuum Ising universality class (*23*) and for the nearest neighbor Ising model (*41*). As discussed in the next section, these two metric eigenvalues diverge with the scaling of the susceptibility ($\chi \sim \xi^{7/4}$, whose eigenvector is θ^h) and specific heat ($C \sim \log(\xi)$, whose eigenvector is a combination of $\theta^{\alpha\beta}$ proportional to the gradient of the critical temperature, $\frac{\partial T_c}{\partial \theta^{\alpha\beta}}$). From an information theoretic point of view, these two parameter combinations seem to become particularly easy to measure near the critical point because the system's behavior becomes extremely sensitive to changes in field and temperature. The behavior of these two eigenvalues seems to have no parallel in the diffusion equation viewed at its microscopic scale.

5.1 Scaling analysis of the eigenvalue spectrum

Monte Carlo results were also analyzed with renormalization group (RG) techniques focusing on the critical region, close to the RG fixed point θ_0. After an RG transformation that reduces lengths by a factor of b, the remaining degrees of freedom are described by an effective theory

with parameters θ' related to the original ones by the relationship $\theta'^{\mu} - \theta_0^{\mu} = T_{\nu}^{\mu}(\theta^{\nu} - \theta_0^{\nu})$ where[11] T has left eigenvectors and eigenvalues given by $\mathbf{e}_{\alpha,\mu}^{L}$ and b^{y_α}. It is convenient to switch to the so-called scaling variables, $u_\alpha = \sum_\mu \mathbf{e}_{\alpha,\mu}^{L}\theta^{\mu}$, which have the property that under a renormalization group transformation

$$u'_\alpha = b^{y_\alpha} u_\alpha. \tag{S26}$$

It is also convenient to separate the free energy into a singular part and an analytic part so that

$$\begin{aligned} F(\theta) = &\quad Af^s(u_\alpha(\theta)) + Af^a(u_\alpha(\theta)), \\ f^s = &\quad u_1^{d/2y_1}\mathcal{U}(r_0, ..., r_\alpha), \\ r_\alpha = &\quad u_\alpha / u_1^{y_\alpha/y_1}. \end{aligned} \tag{S27}$$

Here functions f are free energy densities, A is the system size and f^a and $\mathcal{U}$ are both analytic functions of their arguments. Notice that by construction the variables r do not change under an RG transformation: the rescaling of component variables u_α and u_1 cancel. The FIM can be

[11] $\theta'^{\mu} - \theta_0^{\mu} = T_{\nu}^{\mu}(\theta^{\nu} - \theta_0^{\nu})$ is strictly true only if the parameters span the space of possible Ising Hamiltonians, but the analysis holds for $g_{\mu\nu}$ on the space of the original parameters provided the θ' span all possible models, which is assumed in this analysis. Said differently, there is no need for T to be square, and it is sufficient for the analysis presented above to assume that T is 13 by infinite dimensional.

similarly divided into two pieces

$$
\begin{aligned}
g_{\mu\nu} &= g^s_{\mu\nu} + g^a_{\mu\nu} = -A\partial_\mu\partial_\nu f^s - A\partial_\mu\partial_\nu f^a, \\
\\
g^s_{\mu\nu} &= A\textstyle\sum_{\alpha,\beta}(\frac{\partial r_\alpha}{\partial\theta^\mu}\frac{\partial r_\beta}{\partial\theta^\nu})(\frac{\partial}{\partial r^\alpha}\frac{\partial}{\partial r^\beta}\mathcal{U}) \\
&= A\textstyle\sum_{\alpha,\beta}(\frac{\partial u_\alpha}{\partial\theta^\mu}\frac{\partial u_\beta}{\partial\theta^\nu})u_1^{-(y_\alpha+y_\beta-d)/y_1}(\frac{\partial}{\partial r^\alpha}\frac{\partial}{\partial r^\beta}\mathcal{U}) \\
&= A\sum_{\alpha,\beta}(\frac{\partial u_\alpha}{\partial\theta^\mu}\frac{\partial u_\beta}{\partial\theta^\nu})(\frac{\partial}{\partial r^\alpha}\frac{\partial}{\partial r^\beta}\mathcal{U})\xi^{y_\alpha+y_\beta-d}, \\
\\
g^a_{\mu\nu} &= A\textstyle\sum_{\alpha,\beta}(\frac{\partial u_\alpha}{\partial\theta^\mu}\frac{\partial u_\beta}{\partial\theta^\nu})\frac{\partial}{\partial u_\alpha}\frac{\partial}{\partial u_\beta}f^a
\end{aligned}
\tag{S28}
$$

where ξ is the correlation length, which diverges like u_1^{-1/y_1}. By using the dimensionless r variables for analysis of $g^s_{\mu\nu}$, the singular behavior is isolated and expressed in powers of ξ. Now f^a is by assumption an analytic function at the critical point, and the coordinate changes $u_\alpha(\vec{\theta})$ are analytic there, so $g^a_{\mu\nu}$ will have eigenvalues that are all of the same order of magnitude, given by the area A:

$$\lambda^a_i \sim A. \tag{S29}$$

The singular behavior of $g^s_{\mu\nu}$ as the correlation length $\xi \to \infty$ at the critical point controls its eigenvalues. As shown in Appendix A, its eigenvalues scale as

$$\lambda^s_i \sim A\xi^{2y_i-d}. \tag{S30}$$

Hence the singular piece will dominate wherever $2y_i - d \geq 0$. In the 2d Ising model, this is true for the magnetic field as it becomes the largest eigenvector $e_0 = \theta^h$ (with $y_h = 15/8$) along with $e_1 = \partial_\mu u_1$ whose RG exponent is $y_1 = 1$ (in the latter case $2y_i - d = 0$ so there is a logarithmic divergence, as with the Ising model's specific heat). The remaining eigenvectors of

$g_{\mu\nu}$ are dominated by analytic contributions. These analytic contributions, just as in the diffusion equation viewed at its fundamental scale, cause the corresponding eigenvalues to cluster together at a characteristic scale and not exhibit sloppiness (though not necessarily to be exactly the identity). This analysis agrees with the Monte Carlo results plotted in fig. S2.

6 Measuring the Ising metric after coarsening

The FIM after n steps of coarsening is $g_{\mu\nu} = -\left\langle \partial_\mu \partial_\nu \log\left(P(x^n)\right)\right\rangle$ where $x^n = \left\{s_{i,j}\right\}_{\text{for } \{i,j\} \text{ in level } n}$. The levels are defined as follows: If n is even then $\{i,j\}$ is in level n iff $i/2^{n/2}$ and $j/2^{n/2}$ are both integers. If n is odd than $\{i,j\}$ is in level n if and only if $\{i,j\}$ is in level $n-1$ and $(i+j)/2^{n/2+1}$ is an integer[12]. The mapping to level n from level 0 (giving the configuration of retained subset of spins) is denoted[13] as $x^n = C^n(x)$. It will be useful to write $P(x^n)$ in terms of a restricted partition function

$$\begin{aligned} P(x^n) &= \tilde{Z}(x^n)/Z, \\ \tilde{Z}(x^n) &= \sum_x \exp(-H(x))\delta(C^n(x) = x^n). \end{aligned} \tag{S31}$$

Here $\tilde{Z}(x^n)$ is the coarse-grained partition function conditioned on the sub-lattice at level n taking the value x^n, summing over the remaining degrees of freedom. The expectation value of an operator defined at level 0 over configurations which coarsen to the same configuration x^n will be denoted as

$$\left\{Q\right\}_{x^n} = \frac{\sum_x Q(x)\delta(C^n(x) = x^n)\exp(-H(x))}{\tilde{Z}(x^n)}. \tag{S32}$$

[12]The first level is thus a checkerboard, the second has only even sites, the third has a checkerboard of even sites, etc.

[13]The mapping $C^n(x)$ here simply discards all of the spins that do not remain at level N, leaving an $L/2^{n/2} \times L/2^{n/2}$ square lattice for even N and a rotated 'diamond' lattice for odd N. However, this formalism would also apply to other schemes, such as the commonly used block-spin procedure.

$\tilde{Z}(x^n)$ can be treated like a partition function in the usual way. In particular, it is possible to take parameter derivatives of the log of $\tilde{Z}(x^n)$ yielding familiar equations for cumulants

$$\begin{aligned} -\partial_\mu \log(\tilde{Z}(x^n)) &= \{\Phi^\mu\}_{x^n} \\ \partial_\mu \partial_\nu \log(\tilde{Z}(x^n)) &= \{\Phi^\mu \Phi^\nu\}_{x^n} - \{\Phi^\mu\}_{x^n} \{\Phi^\nu\}_{x^n} . \end{aligned} \tag{S33}$$

The calculation will also use nested brackets wherein an outer triangular bracket refers to an expectation value over microscopic configurations and inner curly brackets denote an expectation value in the set of configurations that coarsen to the same x^n. Importantly, a single curly bracket nested in a triangular bracket does not affect expectation values, as every micro state x appears the same number of times in total. However, the presence of two curly brackets in the same one does. For example:

$$\begin{aligned} \langle \{\Phi^\mu \Phi^\nu\}_{x^n} \rangle &= \langle \Phi^\mu \Phi^\nu \rangle \\ \langle \{\Phi^\mu\}_{x^n} \{\Phi^\nu\}_{x^n} \rangle &\neq \langle \Phi^\mu \Phi^\nu \rangle \end{aligned} \tag{S34}$$

With these the FIM can be written as

$$\begin{aligned} g^n_{\mu\nu} &= -\partial_\mu \partial_\nu \big\langle \log\left(P(x^n)\right) \big\rangle \\ &= \partial_\mu \partial_\nu \log(Z) - \big\langle \partial_\mu \partial_\nu \log(\tilde{Z}(C^n(x))) \big\rangle \\ &= g_{\mu\nu} - \big\langle \big\{\Phi_\mu \Phi_\nu\big\}_{C^n(x)} \big\rangle + \big\langle \big\{\Phi_\mu\big\}_{C^n(x)} \big\{\Phi_\nu\big\}_{C^n(x)} \big\rangle \\ &= \big\langle \big\{\Phi_\mu\big\}_{C^n(x)} \big\{\Phi_\nu\big\}_{C^n(x)} \big\rangle - \big\langle \big\{\Phi_\mu\big\}_{C^n(x)} \big\rangle \big\langle \big\{\Phi_\nu\big\}_{C^n(x)} \big\rangle . \end{aligned} \tag{S35}$$

Going from the first to the second line uses equation S31, going from the second to the third uses equation S33 and going from the third to the fourth uses equation S34. The quantity $\big\langle \big\{\Phi_\mu\big\}_{C^n(x)} \big\{\Phi_\nu\big\}_{C^n(x)} \big\rangle$ can be measured by taking each member of an ensemble, x_q, and gen-

erating a sub-ensemble of $x'_{q,r}$ according to the distribution defined by

$$P(x'_{q,r}|x_q) = \frac{\sum_x \exp(-H(x))\delta(C^n(x'_{q,r}) = C^n(x_q))}{\tilde{Z}(C^n(x_q)))}. \tag{S36}$$

Techniques for generating this ensemble, using a form of 'Compatible Monte Carlo' (*24*) are discussed in section 7. From an ensemble of M configurations, with x_q taken from the ensemble of full lattice configurations, and $x_{q,r}$ from the ensemble given by $P(x'_{q,r}|x_q)$ for each x_q, the metric becomes

$$g^n_{\mu\nu} = \frac{1}{(M)(M'^2-M')} \sum_{q,r,s=1 r\neq s}^{q=M\ r,s=M'} \Big(\Phi_\mu(x'_{q,r})\Phi_\nu(x'_{q,s}) - \frac{1}{M-1}\sum_{p=1\ p\neq q}^{M} \Phi_\mu(x'_{q,r})\Phi_\nu(x'_{p,s})\Big). \tag{S37}$$

The results of this Monte Carlo are presented for a 64×64 system at its critical point in fig. 3 of the main text. The analytic corrections to scaling are reduced under coarse-graining, revealing a sloppy spectrum of marginal and irrelevant metric eigenvalues. These irrelevant and marginal eigenvalues continue to behave much as the eigenvalues of the metric in the diffusion equation, becoming progressively less important under coarsening with characteristic eigenvalues. The large eigenvalues are dominated by singular corrections and do not become smaller under coarsening, presumably because they are measured by their collective effects on the large scale behavior measured from large distance correlations.

6.1 Eigenvalue spectrum after coarse-graining

The scaling of the FIM's eigenvalues after coarsening can be estimated by using an RG-like procedure that uses the following steps: (a) discarding the information in certain degrees of freedom, (b) constructing an effective Hamiltonian for the remaining degrees of freedom in a new parameter basis, (c) repeating the analysis for the metric's eigenvalues in the parameter

coordinates of this new effective Hamiltonian, and (d) transforming back into the original coordinates. It is helpful to contrast this approach to a usual RG calculation for a lattice Ising model. In a usual RG calculation, information about certain degrees of freedom is discarded as in (a) and, just as in (b), an effective theory is built that describes the behavior of the remaining degrees of freedom. The approach described below departs from this usual picture in that the goal is not to find this effective theory, but instead to calculate parameter sensitivities of the original microscopic theory. To this end, steps (c) and (d) are added; the effective theory is used only as an intermediate in calculating parameter sensitivities of the original model.

After coarse-graining n times, each observation yields only the spins $\{i, j\}$ remaining at level n, $x^n = \{s_{i,j}\}\Big|_{\{i,j\} \text{ in level n}}$. The probability of a given configuration of these spins x^n can be written in terms of a renormalized model as is typical in RG

$$P(x^n) = \frac{\exp\left(-H^n(x^n)\right)}{Z(A^n, u^n)}, \tag{S38}$$

where H^n is an effective Hamiltonian describing just those spins that are observable after n coarse-graining steps. H^n has new parameters that can be expressed in terms of the scaling variables defined in equation S26 with $u^n_\alpha = b^{y_\alpha n} u_\alpha$. In addition, the area A of the system, in lattice spacings, is reduced to[14] $A^n = b^{-dn} A$, $\partial u^n_\alpha / \partial \theta^\mu = b^{y_\alpha} \partial u_\alpha / \partial \theta^\mu$.

After rescaling, the entropy of the model is smaller by an amount ΔS^n from the original model's entropy. It is customary in RG analysis to subtract this constant from the Hamiltonian, so as to preserve the free energy of the system after rescaling:

$$F^n = F^{n,s} + F^{n,a} + \Delta S^n = F^s + F^a = F \tag{S39}$$

The new model's Hamiltonian is still linear in new parameters, allowing us to use the algebra

[14] here, $b = \sqrt{2}$, $d = 2$

of equation S11 if we remove the constant ΔS from the new Hamiltonian. This would, of course, be an identical model, since the addition of a constant to the free energy does not change any observables. Now expressing the metric for the new observables in terms of the original parameters yields

$$g^n_{\mu\nu}(\theta) = \partial_\mu \partial_\nu (F^{n,s} + F^{n,a}) = \partial_\mu \partial_\nu (F^s + F^a - \Delta S). \tag{S40}$$

Analyzing the singular and analytic contributions to the FIM separately

$$\begin{aligned} g^{s,n}_{\mu\nu} &= \partial_\mu \partial_\nu F^{n,s} = \partial_\mu \partial_\nu F^s = g^s_{\mu\nu}, \\ g^{a,n}_{\mu\nu} &= \partial_\mu \partial_\nu F^{n,a} = b^{-dn} A \partial_\mu \partial_\nu f^{n,a} \\ &= b^{-dn} A \tfrac{\partial u^n_\alpha}{\partial \theta^\mu} \tfrac{\partial u^n_\beta}{\partial \theta^\nu} (\tfrac{\partial}{\partial u^n_\alpha} \tfrac{\partial}{\partial u^n_\beta} f^{n,a}) \\ &= A \textstyle\sum_{\alpha,\beta} b^{(y_\alpha + y_\beta - d)n} (\tfrac{\partial u_\alpha}{\partial \theta^\mu} \tfrac{\partial u_\beta}{\partial \theta^\nu}) (\tfrac{\partial}{\partial u_\alpha} \tfrac{\partial}{\partial u_\beta} f^a) \end{aligned} \tag{S41}$$

The singular piece of the metric is maintained *exactly* because the singular part of the free energy is preserved after an RG step. The implication is that the singular part of the free energy contains long wave-length information. On the other hand, the analytic piece is smaller by $\partial_\mu \partial_\nu \Delta S^n$. The matrix $(\frac{\partial u_\alpha}{\partial \theta^\mu} \frac{\partial u_\beta}{\partial \theta^\nu})(\frac{\partial}{\partial u_\alpha} \frac{\partial}{\partial u_\beta} f^a)$ should be smoothly varying, with n, depending only the u^n which vary only small amount with n near the RG fixed point. Importantly, all of its eigenvalues should continue to take a characteristic value. Thus, after rescaling n times (see equations S29, S30 and Appendix A)

$$\begin{aligned} \lambda^{n,s}_i &\sim A(\xi)^{2y_i - d}, \\ \lambda^{n,a}_i &\sim A b^{n(2y_i - d)}. \end{aligned} \tag{S42}$$

To ensure that the Fisher information is strictly decreasing in every direction upon coarsen-

ing[15], $g^a_{\mu\nu}$ must be negative semidefinite in the subspace of scaling variables where $2y_i - d > 0$. For these relevant directions, $\lambda_i^n \sim A\xi^{2y_i-d} - Ab^{2y_i-d}n$, with $i = 0,\ 1$. Here, the second term only becomes significant when $b^n \sim \xi$ (i.e. when the lattice spacing is comparable to the correlation length). For irrelevant directions, or relevant ones with $0 < 2y_i < d$ (corresponding to $i \geq 2$ in the Ising model), the analytic piece will eventually dominate as the critical point is approached, yielding $\lambda_i \sim Ab^{2y_i-d}$. These results are in quantitative agreement with those plotted in fig. 3 of the main text assuming that the variables project onto irrelevant and marginal scaling variables with leading dimensions of $y = 0$ (blue line in fig. 3 of main text), $y = -2$ (green line in fig. 3 of the main text) and $y = -4$ (purple line in fig. 3 of the main text) consistent with the theoretical predictions for the irrelevant eigenvalue spectrum made in (*42*).

This shrinkage of the FIM is reparameterization invariant in an important way. Although a coordinate system can always be chosen in which the metric is locally the identity, the shrinkage, which can be seen in any coordinate system, quantifies the contraction of the invariant distance between nearby points as observables are coarsened. For example, if we choose a coordinate system in which the metric is the identity when examining microscopic observables, we find that the metric eigenvalues become widely spread after coarsening[16].

It is helpful to contrast the results of this information geometry analysis to those of a more standard RG one. Both can be used to explain the experimental findings of universality: a wide class of microscopic models have identical macroscopic behavior. In an RG picture, one considers a hypothetical large dimensional space of possible Hamiltonians that includes microscopically disparate systems (for example including both ferromagnets and binary fluids). As

[15]In each coarsening step $g^n_{\mu\nu} - g^{n+1}_{\mu\nu}$ must be a positive semidefinite matrix. This is because no parameter combinations can be more measurable from a subset of the data available at level n than from its entirety.

[16]Least-Squares models that do not have a concept of coarsening still have a reparameterization invariant manifestation of sloppiness (*13, 14*). These models are typically finite in extent, at least in most directions and contain 'edges' where some metric eigenvalues are zero and where parameters take extreme values (for example a rate constant being either zero or infinity). Although a coordinate change can locally set the metric to the identity, the reparameterization invariant shape of the manifold has a 'hyper-ribbon' structure, with a geometric hierarchy of widths. It is unknown if the Ising model has a similar structure on coarsening.

the renormalization group proceeds the Hamiltonians of their effective models flow towards the same saddle point. The Hamiltonian of this saddle point thus describes the effective interactions of coarsened degrees of freedom. This explains how binary fluids and ferromagnets could have similar effective models for the coarsened observables.

This same hypothetical large dimensional space of Hamiltonians can be considered from an information theory perspective, by adding step (c), calculating the Fisher Information for the effective Hamiltonian and (d), transforming back to microscopic coordinates. Information geometry clarifies that the microscopic Hamiltonians describing binary fluids and ferromagnets produce indistinguishable results for coarsened variables. Although the parameter space distance between microscopic models for binary fluids and ferromagnets is quite large, the 'proper distance' between them defined through the FIM rapidly vanishes upon coarsening. Models for ferromagnets and binary fluids (for which t and h values are identical) differ from each other only along sloppy directions and hence their long-wavelength behaviors become nearly identical. The evolution of the FIM under coarsening tracks the information lost about microscopic details in these physics models. In this information geometry picture of universality, the high dimensional parameter space manifold of systems near Ising critical points collapses onto a two dimensional manifold when its observables are coarsened. This analysis completes what might be seen as a trivial step in RG arguments for universality—demonstrating that nearness in effective model space implies indistinguishability of coarsened observables. The mapping from parameter space distance to metric distance in the space of distinguishability clarifies some confusing points. For example, while FIM distinguishability along relevant parameter directions remain roughly fixed under coarsening, their parameter space distance appears to grow in a usual RG picture. Similarly considering an enlarged hypothetical parameter space likely explains why many models with sloppy FIMs, for example in systems biology, can be predictive even when important components are entirely absent from their microscopic models.

7 Simulation details

As described above, $M = 10,000$–$100,000$ independent members from each ensemble x_p are generated using the standard Wolff algorithm (*40*) implemented on 64×64 periodic square lattices, and are used to calculate the FIM before coarsening.

A variation of the 'Compatible Monte Carlo'[17] method introduced in (*24*) was employed to generate members of the coarse-grained ensemble defined by equation S36. In this method, a Monte Carlo chain is run and any move proposing a switch to a configuration $x'_{p,r}$ for which $C^n(x'_{p,r}) \neq C^n(x_p)$ is rejected. For the mapping $C^n(x_p) = C^n(x_{p,r})$, the simplest implementation equilibrates using Metropolis moves by proposing only the spins not in level n. Additional tricks to speed up convergence are described below.

Consider the task of generating a random member $x'_{p,r}$ for a given x_p at level 1. Because the spins which are free to flip only couple with fixed spins, each one can be chosen independently. As such, choosing each free spin according to its heat bath probability generates an uncorrelated member $x_{p,r}$ of the ensemble defined by x_p in a single step. This idea can be further exploited to exactly calculate the contribution to a metric element at level 1 from a level 0 configuration x. In particular, replacing all of the spins that are not in level 1 with their mean field values defined by $\tilde{s}_{i,j}(x) = \{s_{i,j}\}_{C^n(x)}$ leads to

$$\begin{aligned} \{\Phi_{\alpha\beta}\}_{C^n(x)} &= \sum_{i,j} \tilde{s}_{i,j}(x)\tilde{s}_{i+\alpha,j+\beta}(x), \\ \{\Phi_h\}_{C^n(x)} &= \sum_{i,j} \tilde{s}_{i,j}. \end{aligned} \tag{S43}$$

It is therefore possible to exactly calculate the level 1 quantities $\{\Phi_\mu\}_{C^1(x)}\{\Phi_\nu\}_{C^1(x)}$ for any microscopic configuration x and the corresponding checkerboard configuration $C^1(x)$. The

[17] Ron, Swendsen and Brandt used this technique to generated large equilibrated ensembles close to the critical point, essentially by starting from a small 'coarsened' lattice and iteratively adding layers to generate a large ensemble.

metric at level 1 can now be written

$$g^1_{\mu\nu} = \frac{1}{M^2-M} \sum_{p,q=1,p\neq q}^{M} \Big(\{\Phi_\mu\}_{C^1(x_p)} \{\Phi_\nu\}_{C^1(x_p)} - \{\Phi_\mu\}_{C^1(x_p)} \{\Phi_\nu\}_{C^1(x_q)} \Big). \tag{S44}$$

Beyond level 1 it becomes necessary to use Compatible Monte Carlo. Because of the independence of free spins at level 1, spins at all levels $n \geq 1$ only interact with spins that are already absent at level 1. Therefore, the spins that are free at level 1 (termed the red sites of the checkerboard) are left integrated out. The partition function for a level 1 configuration is most conveniently written in terms of the number of up neighbors, $n^{up}_{i,j}$ that each red site has

$$\begin{aligned} \log \tilde{Z}(C_1(x)) &= \sum_{i,j \text{ not in level 1}} \log\left(z(n^{up}_{i,j})\right), \\ z(n^{up}) &= \cosh\left((\beta J)(2-n^{up})\right), \end{aligned} \tag{S45}$$

Additional spins that are not integrated out at level n are flipped using a heat bath algorithm with the ratio of partition functions in an 'up' vs 'down' configuration used to determine the transition probability. The probability of a spin (at level ≥ 2) transitioning to 'up' after being proposed from the down state is given by $z^{up}_{i,j}/(z^{up}_{i,j} + z^{down}_{i,j})$ with

$$\begin{aligned} z^{up}_{i,j} &= \sum_{\{k,l\} \text{ n.n. of } \{i,j\}} z(n^{up}_{k,l}+1), \\ z^{down}_{i,j} &= \prod_{\{k,l\} \text{ n.n. of } \{i,j\}} z(n^{up}_{k,l}). \end{aligned} \tag{S46}$$

Equilibration is fast as there are effectively no correlations larger than the spacing between fixed spins at level n. This allows generating an ensemble of lattice configurations at level 1, conditioned on the system coarsening to an arbitrary configuration at any level $n > 1$. Equation S37 is thus slightly modified to the following which was used to make fig. 3 for data at level 2 and

higher

$$g^n_{\mu\nu} = \frac{1}{(M)(M'^2-M')} \sum_{q,r,s=1 r\neq s}^{q=M\ r,s=M'} \left(\{\Phi_\mu\}_{c^1(x'_{q,r})} \{\Phi_\nu\}_{c^1(x'_{q,s})} \ -\frac{1}{M-1} \sum_{p=1\ p\neq q}^{M} \{\Phi_\mu\}_{c^1(x'_{q,r})} \{\Phi_\nu\}_{c^1(x'_{p,s})} \right) \tag{S47}$$

Additional Acknowledgement

We thank Yaming Yu for introducing us to Weyl's inequality, and using it to prove the conjecture in reference (*12*). The first steps in Appendix A are an adaptation of his theorem.

Appendix A

Here we discuss the way the eigenvalues of the FIM scale near the Ising critical point, deriving the results quoted in equation S30. Our formula for the FIM is given by

$$\begin{aligned} g^s_{\mu\nu} &= A \sum_{\alpha,\beta} (\tfrac{\partial u^\alpha}{\partial \theta^\mu} \tfrac{\partial u^\beta}{\partial \theta^\nu}) (\tfrac{\partial}{\partial r^\alpha} \tfrac{\partial}{\partial r^\beta} \mathcal{U}) \xi^{y_\alpha + y_\beta - d} \\ &= J^\alpha_\mu \hat{g}^s_{\alpha\beta} J^\beta_\nu \end{aligned} \tag{S48}$$

where $\hat{g}^s_{\alpha\beta}$ is the metric tensor in the scaling variable coordinates $u^\alpha(\vec{\theta})$ for which the renormalization-group flows expand by a factor b^{y_α}, and $J^\beta_\nu = \partial u^\beta / \partial \theta^\nu$ is the Jacobian transforming the natural coordinates θ^ν to the scaling variable coordinates. Our job is to show that the ordered eigenvalues λ^s_i of g^s scale like

$$\lambda^s_i \sim A \xi^{2y_i - d} \tag{S49}$$

(equation S30). To do so, we first demonstrate that the eigenvalues $\hat{\lambda}_i$ of the FIM $\hat{g}^s$ in scaling variable coordinates satisfies this bound, and then show that this scaling is preserved by the transformation J to bare coordinates.

We make use of Weyl's inequality for matrix eigenvalues, which implies that if B and M

are real, symmetric matrices and $B - M$ is nonnegative definite, then each ordered eigenvalue of B is greater than or equal to the corresponding one of M. Let us write

$$\hat{g}^s_{\alpha\beta} = A\xi^{-d}\left(\frac{\partial}{\partial r^\alpha}\frac{\partial}{\partial r^\beta}\mathcal{U}\right)\xi^{y_\alpha+y_\beta} = A\xi^{-d}EME \tag{S50}$$

where $M_{\alpha\beta} = \partial^2\mathcal{U}/\partial r^\alpha \partial r^\beta$ and $E_{\sigma\rho} = \delta_{\sigma\rho}\xi^{y_\sigma}$. This form of $\hat{g}^s$ is similar to that of matrices studied in (*12*).

Let C be the maximum eigenvalue of M, and let $B_{\alpha\beta} = C\delta_{\alpha\beta}$, so in particular $B - M$ is nonnegative definite, and hence $W^T(B - M)W \geq 0$ for any vector W.

Conclusion: $(A\,\xi^{-d}EBE - \hat{g}^s)$ is nonnegative definite, and thus $\hat{g}^s$ has sorted eigenvalues $\hat{\lambda}_i \leq CA\xi^{2y_i-d}$.

Argument: Because $\hat{g}^s = A\xi^{-d}EME$, for any vector V,

$$V^T(A\xi^{-d}EBE - \hat{g}^s)V = V^T(A\xi^{-d}E(B-M)E)V = A\xi^{-d}W^T(B-M)E)W \geq 0, \tag{S51}$$

where $W = EV = VE$. Since $B_{\alpha\beta} = C\delta_{\alpha\beta}$ and $E_{\alpha\beta} = \xi^{y_\alpha}\delta_{\alpha\beta}$ are diagonal, the sorted eigenvalues of $A\xi^{-d}EBE$ are just $CA\xi^{2y_i-d}$, which by Weyl's inequality bound the sorted eigenvalues of $\hat{g}^s$.

We now need to transform from the scaling coordinates u^α to the original coordinates θ^ν. The mapping from scaling variable to bare coordinates is non-orthogonal. Let the eigenvector of $\hat{g}^s$ corresponding to $\hat{\lambda}_i$ be $\hat{v}_i$. Each scaling-coordinate eigenvector transforms to a vector in parameter space,

$$V^i_\mu = \sum_\alpha \hat{v}^\alpha_i J^\alpha_\mu = \sum_\alpha \hat{v}^\alpha_i \frac{\partial u^\alpha}{\partial \theta^\mu}. \tag{S52}$$

The V^is are neither orthogonal nor normalized. The metric in parameter space can be written

as:

$$g^s_{\mu\nu} = \sum_{i=1}^{\infty} \hat{\lambda}_i V^i_\mu V^i_\nu \tag{S53}$$

Conclusion: The sorted eigenvalues of g^s, the FIM matrix in the original coordinates, scale as $\lambda_i \sim A\xi^{2y_i-d}$.

Argument: Consider the truncated version of this matrix formed by adding just the first N contributions:

$$g^{s,N}_{\mu\nu} = \sum_{i=1}^{N} \hat{\lambda}_i V^i_\mu V^i_\nu. \tag{S54}$$

It is positive semidefinite, with rank N. Also, $g^{s,N+1} - g^{s,N}$ is nonnegative definite, so Weyl's inequality tells us that the sorted eigenvalues of $g^{s,N+1}$ are each greater than or equal to those of $g^{s,N}$, $\lambda^{i,N+1} > \lambda^{i,N}$. As traces of matrices sum, we also have that $\sum_i \lambda^{i,N+1} - \lambda^{i,N} = \hat{\lambda}_{N+1}|V^{N+1}|^2$. This implies that all eigenvalues must increase, with none increasing by more than $\hat{\lambda}_{N+1}|V^{N+1}|^2$. As the mapping from parameter space to scaling variables is analytic at the critical point, the normalization factor $|V|^2$ is order one (does not diverge as $\xi \to \infty$). Hence the eigenvalue λ_i in parameter space is a sum of positive terms $\sim \hat{\lambda}_j$ for $j \geq i$. Since by the Lemma $\hat{\lambda}_j \leq CA\xi^{2y_j-d}$, as $\xi \to \infty$ the dominant term will be $\hat{\lambda}_i$, so $\lambda_i \sim A\xi^{2y_i-d}$.

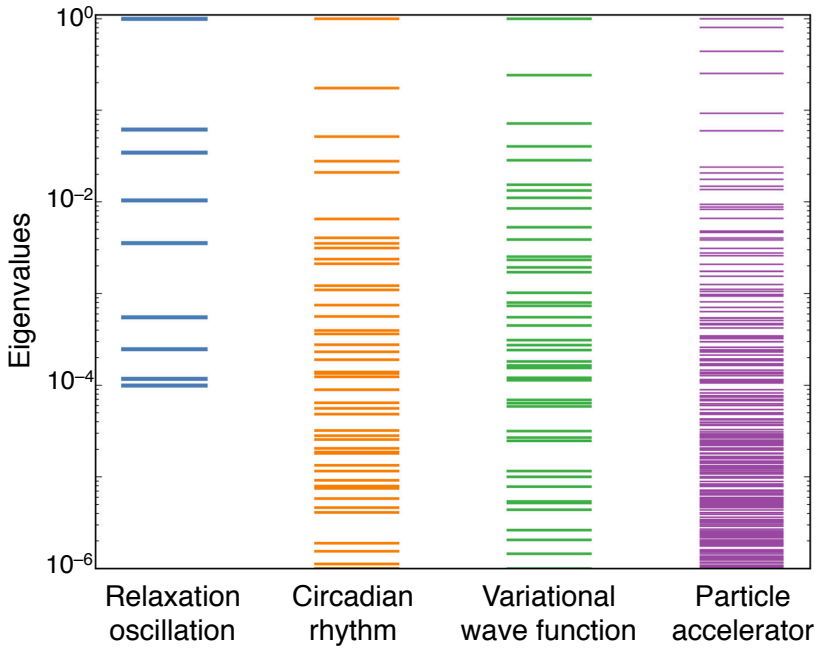

Fig. S1. Normalized eigenvalues of the Fisher Information Matrix (FIM) for four models. The 'Relaxation oscillation' model is a modified Van der Pol system taken from (*28*). Eigenvalues of the genetic network describing 'Circadian rhythm' model (*29*) are calculated in (*11*). 'Variational wave function' eigenvalues are taken from Quantum Monte Carlo simulations as Jastrow parameters are varied (*12*). 'Particle accelerator' is a model of beam shape simulated using the Tool for Accelerator Optics (*30*) as discussed briefly in this supplement. Only the first six decades for each set are shown. Additional examples are in fig. 1.

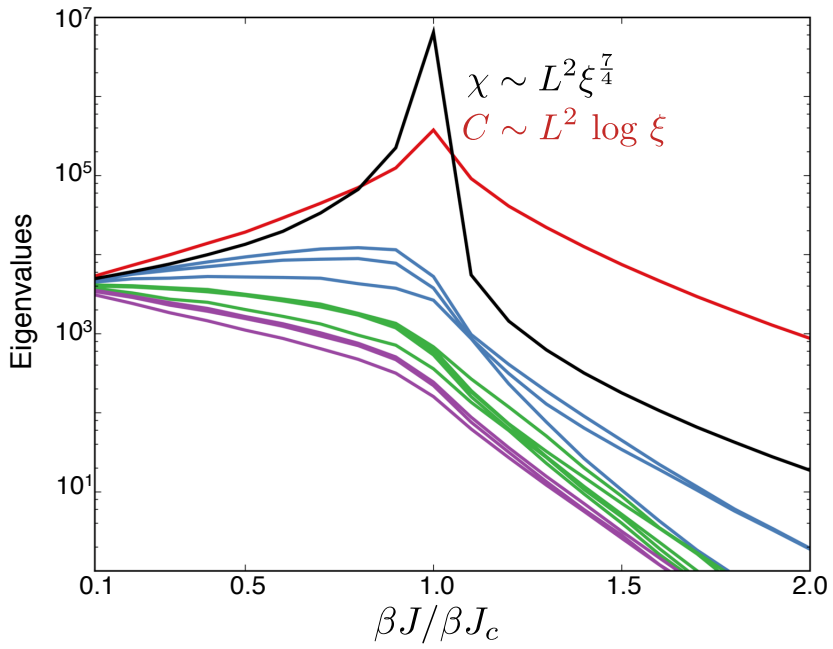

Fig. S2. Eigenvalues of the FIM versus J/J_c. The enlarged 13 parameter Ising model of size $L = 64$ is described in the text. Magnetic field h is taken to be zero. Two eigenvalues become large near the critical point, each diverging with characteristic exponents describing the divergence of the susceptibility and specific heat respectively. The other eigenvalues vary smoothly as the critical point is crossed. Furthermore they take a characteristic scale determined by the system size and are not widely distributed in log. (In the phase separated region, $\beta J > \beta J_c$ we use the connected correlation function in calculating g_{00}. This corresponds to calculating eigenvalues in 'infinitesimal field'. It allows calculation of the FIM in the phase but arbitrarily close to the phase boundary at which there is a net spontaneous magnetization. Without this the FIM would have one spuriously large eigenvalue, quantifying the large symmetry breaking affect of an arbitrarily small applied field.)

References and Notes

1. E. P. Wigner, The unreasonable effectiveness of mathematics in the natural sciences. *Commun. Pure Appl. Math.* **13**, 1–14 (1960). doi:10.1002/cpa.3160130102

2. P. W. Anderson, More is different. *Science* **177**, 393–396 (1972). doi:10.1126/science.177.4047.393

3. U. Alon, Simplicity in biology. *Nature* **446**, 497 (2007). doi:10.1038/446497a Medline

4. G. J. Stephens, B. Johnson-Kerner, W. Bialek, W. S. Ryu, Dimensionality and dynamics in the behavior of *C. elegans*. *PLOS Comput. Biol.* **4**, e1000028 (2008). doi:10.1371/journal.pcbi.1000028 Medline

5. G. J. Stephens, L. C. Osborne, W. Bialek, Searching for simplicity in the analysis of neurons and behavior. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 15565–15571 (2011). doi:10.1073/pnas.1010868108 Medline

6. T. D. Sanger, Human arm movements described by a low-dimensional superposition of principal components. *J. Neurosci.* **20**, 1066–1072 (2000). Medline

7. F. Corson, E. D. Siggia, Geometry, epistasis, and developmental patterning. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 5568–5575 (2012). doi:10.1073/pnas.1201505109 Medline

8. M. Schmidt, H. Lipson, Distilling free-form natural laws from experimental data. *Science* **324**, 81–85 (2009). doi:10.1126/science.1165893

9. T. Mora, W. Bialek, Are biological systems poised at criticality? *J. Stat. Phys.* **144**, 268–302 (2011). doi:10.1007/s10955-011-0229-4

10. K. S. Brown, C. C. Hill, G. A. Calero, C. R. Myers, K. H. Lee, J. P. Sethna, R. A. Cerione, The statistical mechanics of complex signaling networks: Nerve growth factor signaling. *Phys. Biol.* **1**, 184–195 (2004). doi:10.1088/1478-3967/1/3/006 Medline

11. R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, J. P. Sethna, Universally sloppy parameter sensitivities in systems biology models. *PLOS Comput. Biol.* **3**, 1871–1878 (2007). doi:10.1371/journal.pcbi.0030189 Medline

12. J. J. Waterfall, F. P. Casey, R. N. Gutenkunst, K. S. Brown, C. R. Myers, P. W. Brouwer, V. Elser, J. P. Sethna, Sloppy-model universality class and the Vandermonde matrix. *Phys. Rev. Lett.* **97**, 150601 (2006). doi:10.1103/PhysRevLett.97.150601 Medline

13. M. K. Transtrum, B. B. Machta, J. P. Sethna, Why are nonlinear fits to data so challenging? *Phys. Rev. Lett.* **104**, 060201 (2010). doi:10.1103/PhysRevLett.104.060201 Medline

14. M. K. Transtrum, B. B. Machta, J. P. Sethna, Geometry of nonlinear least squares with applications to sloppy models and optimization. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **83**, 036701 (2011). doi:10.1103/PhysRevE.83.036701 Medline

15. Materials and methods are available as supplementary materials on *Science* Online.

16. S. Amari, H. Nagaoka, *Methods of Information Geometry*, (American Mathematical Society, Providence, RI, 2000).

17. I. J. Myung, V. Balasubramanian, M. A. Pitt, Counting probability distributions: Differential geometry and model selection. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11170–11175 (2000). doi:10.1073/pnas.170283897 Medline

18. V. Balasubramanian, Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Comput.* **9**, 349–368 (1997). doi:10.1162/neco.1997.9.2.349

19. P. M. Chaikin, T. C. Lubensky, *Principles of Condensed Matter Physics* (Cambridge Univ. Press, Cambridge, 1995).

20. S. L. Veatch, O. Soubias, S. L. Keller, K. Gawrisch, Critical fluctuations in domain-forming lipid mixtures. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 17650–17655 (2007). doi:10.1073/pnas.0703513104 Medline

21. G. E. Crooks, Measuring thermodynamic length. *Phys. Rev. Lett.* **99**, 100602 (2007). doi:10.1103/PhysRevLett.99.100602 Medline

22. J. Cardy, *Scaling and Renormalization in Statistical Physics* (Cambridge Univ. Press, Cambridge, 1996).

23. G. Ruppeiner, Riemanniam geometry in thermodynamic fluctuation theory. *Rev. Mod. Phys.* **67**, 605–659 (1995). doi:10.1103/RevModPhys.67.605

24. D. Ron, R. H. Swendsen, A. Brandt, Inverse Monte Carlo renormalization group transformations for critical phenomena. *Phys. Rev. Lett.* **89**, 275701 (2002). doi:10.1103/PhysRevLett.89.275701 Medline

25. F. P. Casey, D. Baird, Q. Feng, R. N. Gutenkunst, J. J. Waterfall, C. R. Myers, K. S. Brown, R. A. Cerione, J. P. Sethna, Optimal experimental design in an epidermal growth factor receptor signalling and down-regulation model. *IET Syst. Biol.* **1**, 190–202 (2007). doi:10.1049/iet-syb:20060065 Medline

26. J. F. Apgar, D. K. Witmer, F. M. White, B. Tidor, Sloppy models, parameter uncertainty, and the role of experimental design. *Mol. Biosyst.* **6**, 1890–1900 (2010). doi:10.1039/b918098b Medline

27. In (*13*, *14*), we used interpolation theorems and information geometry to show that multiparameter models fit to collective data have model manifolds that form hyper-ribbons in data space with geometrically spaced widths.

28. R. Chachra, M. K. Transtrum, J. P. Sethna, Structural susceptibility and separation of time scales in the van der Pol oscillator. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **86**, 026712 (2012). doi:10.1103/PhysRevE.86.026712 Medline

29. J. C. Locke, P. O. Westermark, A. Kramer, H. Herzel, Global parameter search reveals design principles of the mammalian circadian clock. *BMC Syst. Biol.* **2**, 22 (2008). doi:10.1186/1752-0509-2-22 Medline

30. D. Sagan, J. Smith, The TAO accelerator simulation program. *Proceedings of the Particle Accelerator Conference* (2005), pp. 4159–4161.

31. C. E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948). doi:10.1002/j.1538-7305.1948.tb01338.x

32. T. Cover, J. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).

33. S. Kullback, R. Leibler, On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951). doi:10.1214/aoms/1177729694

34. N. N. Čencov, *Statistical Decision Rules and Optimal Inference* (American Mathematical Society, Providence, RI, 1981).

35. L. Campbell, An extended Čencov characterization of the information metric. *Proc. Am. Math. Soc.* **98**, 135 (1986).

36. H. Jeffreys, An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. A Math. Phys. Sci.* **186**, 453–461 (1946). doi:10.1098/rspa.1946.0056 Medline

37. M. Prokopenko, J. T. Lizier, O. Obst, X. R. Wang, Relating Fisher information to order parameters. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **84**, 041116 (2011). doi:10.1103/PhysRevE.84.041116 Medline

38. C. Jarzynski, Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* **78**, 2690–2693 (1997). doi:10.1103/PhysRevLett.78.2690

39. G. Arfken, H. Weber, *Mathematical Methods for Physicists* (Academic Press, New York, 2001).

40. U. Wolff, Collective Monte Carlo updating for spin systems. *Phys. Rev. Lett.* **62**, 361–364 (1989). doi:10.1103/PhysRevLett.62.361 Medline

41. D. C. Brody, A. Ritz, Information geometry of finite Ising models. *J. Geom. Phys.* **47**, 207–220 (2003). doi:10.1016/S0393-0440(02)00190-0

42. M. Caselle, M. Hasenbusch, A. Pelissetto, E. Vicari, Irrelevant operators in the twodimensional Ising model. *J. Phys. A* **35**, 4861–4888 (2002). doi:10.1088/0305-4470/35/23/305