Journal of Statistical Physics (2019) 176:214–227
https://doi.org/10.1007/s10955-019-02296-2

A Scaling Law From Discrete to Continuous Solutions of Channel Capacity Problems in the Low-Noise Limit

Michael C. Abbott[1,2] · **Benjamin B. Machta**[3,4]

Received: 24 May 2018 / Accepted: 12 April 2019 / Published online: 30 April 2019

Abstract
An analog communication channel typically achieves its full capacity when the distribution of inputs is discrete, composed of just K symbols, such as voltage levels or wavelengths. As the effective noise level goes to zero, for example by sending the same message multiple times, it is known that optimal codes become continuous. Here we derive a scaling law for the optimal number of symbols in this limit, finding a novel rational scaling exponent. The number of symbols in the optimal code grows as $\log K \sim I^{4/3}$, where the channel capacity I increases with decreasing noise. The same scaling applies to other problems equivalent to maximizing channel capacity over a continuous distribution.

1 Introduction

Information theory is concerned with communication in the presence of noise [1]. A noisy channel may be described by the probability distribution $p(x|\theta)$ over received messages $x \in X$, for a given signal $\theta \in \Theta$. The mutual information between input and output $I(X;\Theta)$ depends not only on the channel, but also on the distribution of input signals $p(\theta)$. A choice of communication code implies a choice of this input distribution, and we are interested in $p_\star(\theta)$ which maximizes I, which is then precisely the channel capacity.

Some channels are used repeatedly to send independent signals, as for example in telecommunications. One surprising feature of the optimal distribution $p_\star(\theta)$ in this context is that it is usually discrete: Even when θ may take on a continuum of values, the optimal code uses only finite number K of discrete symbols. In a sense the best code is digital, even though the channel is analog.

✉ Michael C. Abbott
michael.abbott@wigner.mta.hu

Benjamin B. Machta
benjamin.machta@yale.edu

1 Institute of Physics, Jagiellonian University, Ulica Łojasiewicza 11, 30-348 Krakôw, Poland

2 Holographic QFT Group, MTA Wigner Research Centre for Physics, Konkoly-Thege Miklós u. 29-33, 1121 Budapest, Hungary

3 Lewis-Sigler Institute and Department of Physics, Princeton University, Princeton, NJ 08544, USA

4 Department of Physics and Systems Biology Institute, Yale University, New Haven, CT 06520, USA

The opposite limit of the same problem has also been studied, where effectively a single signal θ is sent a large number number of times, generating independent outputs x_i, $i = 1, 2, \ldots m$. In this case we maximize $I(X^m, \Theta)$ over $p(\theta)$ with the goal of transmitting θ to high accuracy. The natural example here is not telecommunications, but instead comes from viewing a scientific experiment as a channel from the parameters θ in a theory, via some noisy measurements, to recorded data x_i. Lindley [2] argued that the channel capacity or mutual information may then be viewed as the natural summary of how much knowledge we will gain. The distribution $p(\theta)$ is then a Bayesian prior, and Bernardo [3,4] argued that the optimal $p_\star(\theta)$ provides a natural choice of uninformative prior. This situation is usually studied in the limit $m \to \infty$, where they named this a reference prior. Unlike the case $m = 1$ above, in this limit the prior is typically continuous, and in fact usually agrees with the better-known Jeffreys prior [5].

The result we report here is a novel scaling law, describing this approach to a continuum. If K is the optimal number of discrete states, the form of our law plotted in Fig. 1 is this:

$$I(X; \Theta) \sim \zeta \log K \quad \text{when } K \to \infty, \qquad \zeta = 3/4. \tag{1}$$

Slope $\zeta = 1$ on this figure represents the absolute bound $I(X; \Theta) \leq \log K$ on the mutual information, which simply encodes the fact that difference between certainty and complete ignorance among $K = 2^n$ outcomes is exactly n bits of information.

While the motivations above come from various fields, our derivation of this law (1) is very much in the tradition of physics. In Sect. 3 we study a field theory for the local number density of delta functions, $\rho(\theta)$. The maximization gives us an equation of motion for this density, and solving this, we find that the average $\rho_0 = \int \rho(\theta)\, d\theta/L$ behaves as

$$\rho_0 = \frac{K}{L} \sim L^{-1+1/\zeta} = L^{1/3} \quad \text{when } L \to \infty, \qquad \zeta = 3/4. \tag{2}$$

Here L is a proper length $L = \int \sqrt{g_{\theta\theta}\, d\theta^2}$, measured with respect to the natural measure on Θ induced by $p(x^m|\theta)$, namely the Fisher information metric (4). At large L, this length is proportional to the number of distinguishable outcomes, thus the information grows as $I(X; \Theta) \sim \log L$. Then the scaling law (1) above reads $K^\zeta \sim L$, equivalent to (2).

We derive this law assuming Gaussian noise, but we believe that it is quite general, because this is (as usual) a good approximation in the limit being taken. Section 4 looks explicitly at another one-dimensional model which displays the same scaling (also plotted in Fig. 1) and then at the generalization to D dimensions. Finally Appendix A looks at a paper from 30 years ago which could have discovered this scaling law. But we begin by motivating in more detail why we are interested in this problem, and this limit.

2 Prior Work

How much does observing data x inform us about the parameters θ of a model? Lindley [2] argued that this question could be formalized by considering the mutual information between the parameters and the expected data, $I(X; \Theta)$. In this framework, before data is seen we have some prior on parameter space $p(\theta)$, and a resulting expectation for the probability of data x given by $p(x) = \int d\theta\ p(\theta)\ p(x|\theta)$. After seeing particular data x, and updating $p(\theta)$ to $p(\theta|x)$ using Bayes' rule, the entropy in parameter space will be reduced from $S(\Theta)$ to $S(\Theta|x)$. Thus on average the final entropy is $S(\Theta|X) = S(\Theta) - I(X; \Theta)$, and we have learned information I. Bernardo and others [6] argued that in the absence of any other

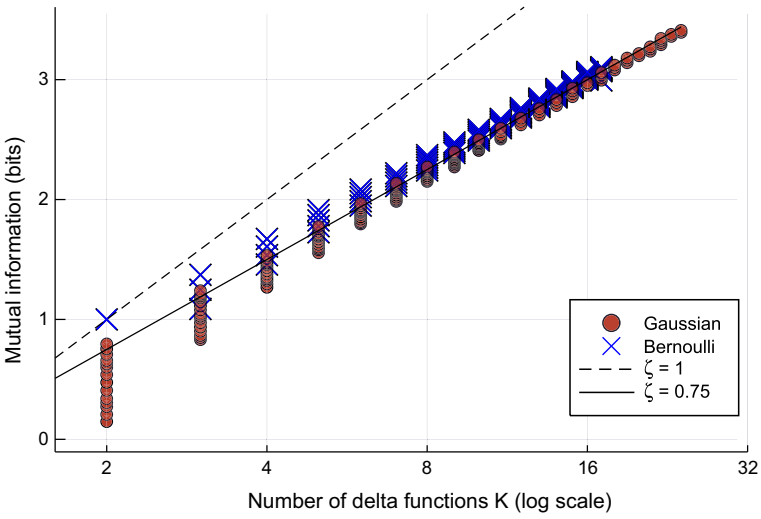

Fig. 1 Scaling law for the number of delta functions as L is increased, plotting $I(X;\Theta)/\log 2$ against K. The lines drawn are $I(X;\Theta) = \zeta \log K$ for $\zeta = 1$ (an absolute bound) and $\zeta = 3/4$ (our scaling law). The blue cross data points are for the Bernoulli model discussed in Sect. 4

knowledge, the prior $p(\theta)$ should be chosen to maximize I, so as to learn as much as possible from the results of an experiment. The statistics community has mostly focused on the limit where data is plentiful—where each experiment is repeated m times, and m goes to infinity. In this limit, the prior which maximizes $I(X^m;\Theta)$ for the aggregate data $(x_1, x_2, \ldots, x_m)$ is known as a reference prior [6]. It usually approaches Jeffreys prior [5], which can also be derived from invariance and geometric considerations, described below.

In a recent paper [7] we argued that the finite data case ($m \neq \infty$) contains surprises which naturally encode a preference for model simplicity. This places model selection and prior selection into the same framework. With finite data, it was long known[1] that the optimal prior $p_\star(\theta)$ is almost always discrete, composed of a finite number of delta functions:

$$p_\star(\theta) = \sum_{a=1}^{K} \lambda_a \, \delta(\theta - \theta_a). \tag{3}$$

The delta functions become more closely spaced as the number of repetitions m increases, with their density approaching the smooth Jeffreys prior as $m \to \infty$, the limit of plentiful data. However, in the data-starved limit, instead this prior has only small number of delta functions, placed as far apart as possible, often at edges of the allowed parameter space. It is the combination of these two limiting behaviors which makes these priors useful for model selection. The typical situation in science is that we have many parameters, of which a few relevant combinations are in the data-rich regime, while many more are in the data-starved regime [16,17]. If we are able to apply the methods of renormalization, then these unmeasurable parameters are precisely the irrelevant directions. We argued that, in general, $p_\star(\theta)$ determines the appropriate model class by placing weight on edges of parameter space

[1] There appear to be several independent discoveries of this fact in the engineering literature [8–10], cited by different groups of later papers. Discreteness was also known in related minimax problems [11–13]. In [7] we reviewed an argument for discreteness from analyticity, and also demonstrated that the algorithm of [14,15], known to be convex, converges to discrete points.

along irrelevant directions, but in the interior along relevant directions. Thus it selects a submanifold of Θ describing an appropriate effective theory, ignoring the irrelevant directions.

The appropriate notion of distance on the parameter manifold Θ should describe how distinguishable the data resulting from different parameter values will be. This is given by the Fisher information metric,

$$ds^2 = \sum_{\mu,\nu=1}^{D} g_{\mu\nu} d\theta^\mu d\theta^\nu, \qquad g_{\mu\nu}(\theta) = \int dx\ p(x|\theta)\ \frac{\partial \log p(x|\theta)}{\partial \theta^\mu}\ \frac{\partial \log p(x|\theta)}{\partial \theta^\nu} \tag{4}$$

which measures distances between points θ in units of standard deviations of $p(x|\theta)$. Such distances are invariant to changes of parameterization, and this invariance is an attractive feature of Jeffreys prior, which is simply the associated volume form, normalized to have total probability 1:

$$p_{\text{J}}(\theta) = \frac{1}{Z}\sqrt{\det g_{\mu\nu}(\theta)}, \qquad Z = \int d\theta \sqrt{\det g_{\mu\nu}(\theta)}.$$

Notice, however, that normalization destroys the natural scale of the metric. Repeating an experiment m times changes the metric $g_{\mu\nu} \to m\ g_{\mu\nu}$, encoding the fact that more data allows us to better distinguish nearby parameter values. But this repetition does not change $p_{\text{J}}(\theta)$. We argued in [7] that this invariance is in fact an unattractive feature of Jeffreys prior. The scale of the metric is what captures the important difference between parameters we can measure well (length $L = \int \sqrt{g_{\theta\theta}\, d\theta^2} \gg 1$ in the Fisher metric) and parameters which we cannot measure at all (length $L \ll 1$).

Instead, the optimal prior $p_\star(\theta)$ has a different invariance, towards the addition of extra irrelevant parameters. Adding extra irrelevant parameters increases the dimensionality of the manifold, and multiplies the volume form by the irrelevant volume. This extra factor, while by definition smaller than 1, can still vary by many orders of magnitude between different points in parameter space (say from 10^{-3} to 10^{-30}). Jeffreys prior, and its implied distribution $p(x)$, is strongly affected by this. It effectively assumes that all parameter directions can be measured, because all are large in the limit $m \to \infty$. But this is not true in most systems of interest to science. By contrast $p_\star(\theta)$ ignores irrelevant directions, giving a distribution $p(x)$ almost unchanged by their addition, or removal.

Bayesian priors have been a contentious subject from the beginning, with arguments foreshadowing some of the later debates about wavefunctions. Uninformative priors such as Jeffreys prior, which depend on the likelihood function $p(x|\theta)$ describing a particular experiment, fit into the so-called objective Bayesian viewpoint. This is often contrasted with a subjective viewpoint, in which the prior represents our state of knowledge, and we incorporate every possible new experimental result by updating it: "today's posterior is tomorrow's prior" [18]. Under such a view the discrete $p_\star(\theta)$, which is zero at almost every θ, would be extremely strange. Note however that this subjective viewpoint is already incompatible with Jeffreys prior. If we invent a different experiment to perform tomorrow, we ought to go back and change our prior to the one appropriate for the joint experiment, resulting in an update not described by Bayes' rule [19]. We differ only in that we regard (say) 10^6 repetitions of the same experiment as also being a different experiment, since this is equivalent to obtaining a much higher-resolution instrument.

But the concerns of this paper are different. We are interested in the relevant directions in parameter space, as it is along such directions that we are in the regime in which the scaling law (1) holds. In the derivation of this, in Sect. 3 below, we study one such dimension

in isolation. While we use the notation of the above statistics problem, we stress the result applies equally to problems from other domains, as discussed above.

3 Derivation

This section studies a model in just one dimension, measuring $\theta \in [0, L]$ with Gaussian noise of known variance, thus

$$p(x|\theta) = \tfrac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\theta)^2/2\sigma^2}. \tag{5}$$

We consider only one measurement, $m = 1$, since more repetitions are equivalent to having less noise. It is convenient to choose units in which $\sigma = 1$, so that θ measures proper distance: $g_{\theta\theta} = 1$. Thus L is the length of parameter space, in terms of the Fisher metric. Jeffreys prior is a constant $p_{\mathrm{J}}(\theta) = 1/L$. The optimal prior has K points of mass:

$$p_\star(\theta) = \sum_{a=1}^{K} \lambda_a\, \delta(\theta - \theta_a), \qquad \theta_1 = 0, \quad \theta_K = L, \qquad \sum_{a=1}^{K} \lambda_a = 1.$$

The positions θ_a and weights λ_a should be fixed my maximizing the mutual information, This is symmetric $I(X;\Theta) = I(\Theta; X)$, so can be written

$$I(X;\Theta) = S(X) - S(X|\Theta) \tag{6}$$

where the entropy and relative entropy are

$$S(X) = -\int dx\; p(x) \log p(x), \qquad p(x) = \int d\theta\; p(x|\theta)\, p(\theta)$$
$$S(X|\Theta) = \int d\theta\, p(\theta) \Big[-\int dx\; p(x|\theta) \log p(x|\theta) \Big].$$

For this Gaussian model, the relative entropy $S(X|\Theta) = \frac{1}{2} + \frac{1}{2}\log 2\pi$ is independent of the prior, so it remains only to calculate the entropy $S(X)$.

On an infinite line, the entropy would be maximized by a constant $p(x)$, i.e. a prior with delta functions spaced infinitesimally close together. But on a very short line, we observe that entropy is maximized by placing substantial weight at each end, with a gap before the next delta function. The idea of our calculation is that the behavior on a long but finite line should interpolate between these two regimes. We work out first the cost of a finite density of delta functions, and then the local cost of a spatially varying density, giving us an equation of motion for the optimum $\rho(x)$. By solving this we learn how the density increases as we move away from the boundary. The integral of this density then gives us K with the desired scaling law.

Since the deviations from a constant $p(x)$ will be small, we write

$$p(x) = \frac{1}{L}[1 + w(x)], \qquad \int dx\; w(x) = 0$$

and then expand the entropy in powers of $w(x)$:

$$S(X) = \log L - \frac{1}{2L}\int_0^L dx\; w(x)^2 + \mathcal{O}(w^4)$$
$$= \log L - \frac{1}{2}\sum_k |w_k|^2 + \cdots . \tag{7}$$

Here our convention for Fourier transforms is that

$$w_k = \int_0^L \frac{dx}{L}\, e^{-ikx} w(x), \qquad k \in \frac{2\pi}{L}\mathbb{Z}.$$

Constant Spacing

Consider first the effect of a prior which is a long string of delta functions at constant spacing a, which we assume to be small compared to the standard deviation $\sigma = 1$, which in turn is much less than the length L.[2] This leads to

$$p(x) = \frac{a}{L}\sum_{n\in\mathbb{Z}} \frac{1}{\sqrt{2\pi}} e^{-(x-na)^2/2}. \tag{8}$$

Because this is a convolution of a Dirac comb with a Gaussian kernel, its Fourier transform is simply a product of such pieces. Let us write the transformation of the positions of the sources as follows:

$$c^0(x) = a\sum_{n\in\mathbb{Z}} \delta(x - na), \qquad \Rightarrow \qquad c_k^0 = \frac{a}{L}\sum_{n\in\mathbb{Z}} e^{-ikna} = \sum_{m\in\mathbb{Z}} \delta_{k-m\frac{2\pi}{a}}.$$

The zero-frequency part of p_k is the constant term in $p(x)$, with the rest contributing to $w(x)$:

$$w_k = \sum_{m\neq 0} \delta_{k-m\frac{2\pi}{a}}\, e^{-k^2/2} = \begin{cases} e^{-k^2/2} & k \in \frac{2\pi}{a}\mathbb{Z} \text{ and } k \neq 0 \\ 0 & \text{else.} \end{cases}$$

The lowest-frequency terms at $k = \pm 2\pi/a$ give the leading exponential correction to the entropy:

$$S(X) = \log L - e^{-q^2} + \mathcal{O}(e^{-2q^2}), \qquad q = \frac{2\pi}{a}. \tag{9}$$

As advertised, any nonzero spacing $a > 0$ (i.e. frequency $q < \infty$) reduces the entropy from its maximum.

Variable Spacing

Now consider perturbing the positions of the delta functions by a slowly varying function $\Delta(x)$, and multiplying their weights by $1 + h(x)$. We seek a formula for the entropy in terms of $\Delta(x)$, while allowing $h(x)$ to adjust so as to minimize the disturbance. This cannot be done perfectly, as $h(x)$ is only sampled at spacing a, so only contributions at frequencies lower than $q = 2\pi/a$ will be screened. Thus we expect what survives to appear with the same exponential factor as (9). In particular this ensures that at infinite density, no trace of

[2] We summarise all the scales involved in (11), see also Fig. 2.

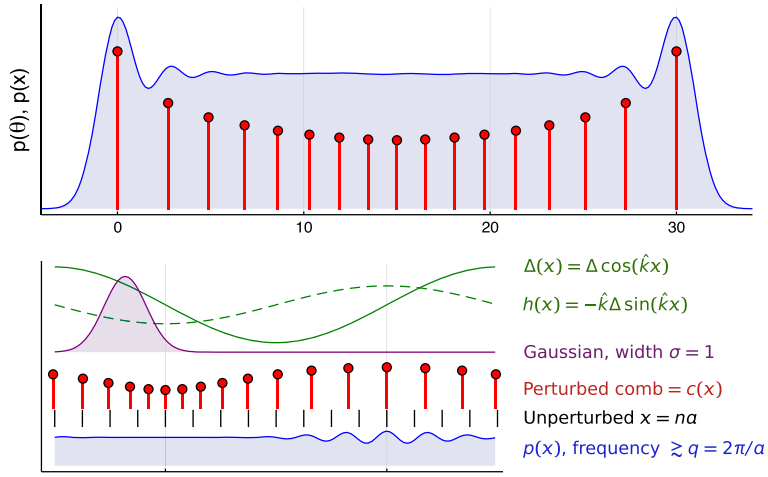

Fig. 2 Above, numerical solution $p_\star(\theta)$ for $L = 30$, in which we observe that as the spacing of the delta functions grows closer together, their weights compensate to leave $p(x)$ almost constant, with deviations $w(x)$ at a wavelength comparable to the spacing. Below, a diagram to show the scales involved when perturbing the positions of the delta functions in our derivation. These are arranged from longest to shortest wavelength, see also (11)

$\Delta(x)$ remains. And that is necessary in order for the limit to agree with Jeffreys prior, which is a constant.

Figure 2 illustrates how the positions and weights of $p_\star(\theta)$ compensate to leave $p(x)$ almost constant in the interior, in a numerical example. Below that it shows how the functions $\Delta(x)$ and $h(x)$ used here mimic this effect.

The comb of delta functions $c^0(x)$ we had above is perturbed to

$$c(x) = [1 + h(x)]\, a \sum_{n\in\mathbb{Z}} \delta\big(x - na - \Delta(na)\big) = [1 + h(x)]\, c^\Delta(x).$$

The effect of $h(x)$ is a convolution in frequency space:

$$c_k = c_k^\Delta + \sum_{k'} h_{k'} c^\Delta_{k-k'}.$$

It will suffice to study $\Delta(x) = \Delta\cos(\hat{k}x)$, i.e. frequencies $\pm\hat{k}$ only: $\Delta_k = \frac{1}{2}\Delta(\delta_{k-\hat{k}} + \delta_{k+\hat{k}})$. The driving frequency is $\hat{k} \ll q$. We can expand in the amplitude Δ to write

$$\begin{aligned} c_k^\Delta &= \frac{a}{L}\sum_n e^{-ik\left(na+\Delta(na)\right)} \\ &= c_k^0 - \frac{ik\Delta}{2}\Big[c^0_{k-\hat{k}} + c^0_{k+\hat{k}}\Big] - \frac{k^2\Delta^2}{8}\Big[2c_k^0 + c^0_{k-2\hat{k}} + c^0_{k+2\hat{k}}\Big] + \mathcal{O}(\Delta^3). \end{aligned}$$

The order Δ term has contributions at $k = \hat{k} \ll q = 2\pi/a$, which can be screened in the full c_k by setting $h_k = +ik\Delta_k$ i.e. $h(x) = -\hat{k}\Delta\sin(\hat{k}x)$. What survives in c_k then are

contributions at $k = 0$, $k = \pm q$ and $k = \pm q \pm \hat{k}$:[3]

$$c_k = \delta_k + \sum_{\pm} \delta_{k\pm q}\Big(1 - \frac{q^2\Delta^2}{4}\Big) + \sum_{\pm}[\delta_{k\mp q-\hat{k}} + \delta_{k\mp q+\hat{k}}]\Big(\pm\frac{i\,q\,\Delta}{2}\Big) + \mathcal{O}(\delta_{k\pm 2q},\,\Delta^2)$$

All but the zero-frequency term are part of $w_k = (c_k - \delta_k)e^{-k^2/2}$, and enter (7) independently, giving this:

$$\begin{aligned} S(X) &= \log L - e^{-q^2}\Big(1 - \frac{q^2\Delta^2}{4}\Big)^2 - \Big[e^{-(q-\hat{k})^2} + e^{-(q+\hat{k})^2}\Big]\Big(\frac{\Delta q}{2}\Big)^2 + \mathcal{O}(\Delta^4) + \mathcal{O}(e^{-2q^2}) \\ &= \log L - e^{-q^2}\Big[1 + \Delta^2\Big(q^4\hat{k}^2 + \frac{1}{3}q^6\hat{k}^4 + \frac{2}{45}q^8\hat{k}^6 + \ldots\Big)\big[1 + \mathcal{O}(1/q^2)\big] + \ldots\Big] + \ldots \end{aligned} \tag{10}$$

As promised, the order Δ^2 term comes with the same overall exponential as in (9) above. Restoring units briefly, the expansion in round brackets makes sense only if $\hat{k}\,q\,\sigma^2 \ll 1$.[4] In terms of length scales this means $\frac{2\pi/\hat{k}}{\sigma} \gg \frac{\sigma}{a}$, or writing all the assumptions made:

$$\underset{\text{comb spacing}}{a = 2\pi/q} \ll \underset{\text{kernel width}}{\sigma = 1} \lll \underset{\text{perturbation wavelength}}{2\pi/\hat{k}} \ll \underset{\text{box size}}{L}. \tag{11}$$

Entropy Density

We can think of this entropy (10) as arising from some density: $S(X) = \int \frac{dx}{L}\mathcal{S}(x) = \mathcal{S}_0$. Our claim is that this density takes the form

$$\mathcal{S}(x) = \log L - L\,e^{-(2\pi)^2\rho(x)^2}\Big[1 + \frac{128\pi^6}{3}\rho(x)^4\rho'(x)^2\Big]. \tag{12}$$

The constant term is clearly fixed by (9). To connect the kinetic term to (10), we need

$$\rho(x) = \frac{1}{a(1+\Delta'(x))} = \frac{1}{a}\big[1 - \Delta'(x) + \Delta'(x)^2 + \mathcal{O}(\Delta^3)\big]$$

thus $\rho'(x) = -\frac{1}{a}\Delta''(x) + \ldots$ and

$$e^{-(2\pi)^2\rho(x)^2} = e^{-q^2}\big[1 + 2q^2\Delta'(x) + 2q^4\Delta'(x)^2 + \mathcal{O}(\Delta^3)\big]\big[1 + \mathcal{O}(1/q^2)\big].$$

Multiplying these pieces, the order Δ^1 term of $\mathcal{S}(x)$ integrates to zero. We can write the order Δ^2 term in terms of Fourier coefficients (using (7), and $\Delta'_k = ik\Delta_k$), and we recover the leading terms in (10). The next term there $q^8\hat{k}^6$ would arise from a term $\rho(x)^6\rho''(x)^2$ in the density, which we neglect.[5]

The Euler-Lagrange equations from (12) read

$$0 = \rho(x)^4\rho''(x) + 2\rho(x)^3\rho'(x)^2 - 4\pi^2\rho(x)^5\rho'(x)^2 + \frac{3}{32\pi^4}\rho(x).$$

[3] The contributions at and $k = \pm q \pm 2\hat{k}$ will only matter at order Δ^4 in $S(X)$.

[4] In footnote 5 we confirm that this indeed holds.

[5] The term $q^8\hat{k}^6$ in $S(X)$ (10) corresponds to a term $\rho^6(\rho'')^2$ in $\mathcal{S}(x)$ (12). This gives a term in the equations of motion (13) going like $x^{9\eta-4}$, which goes to zero as $x \to \infty$ with $\eta = 1/3$. Thus we are justified in dropping this.

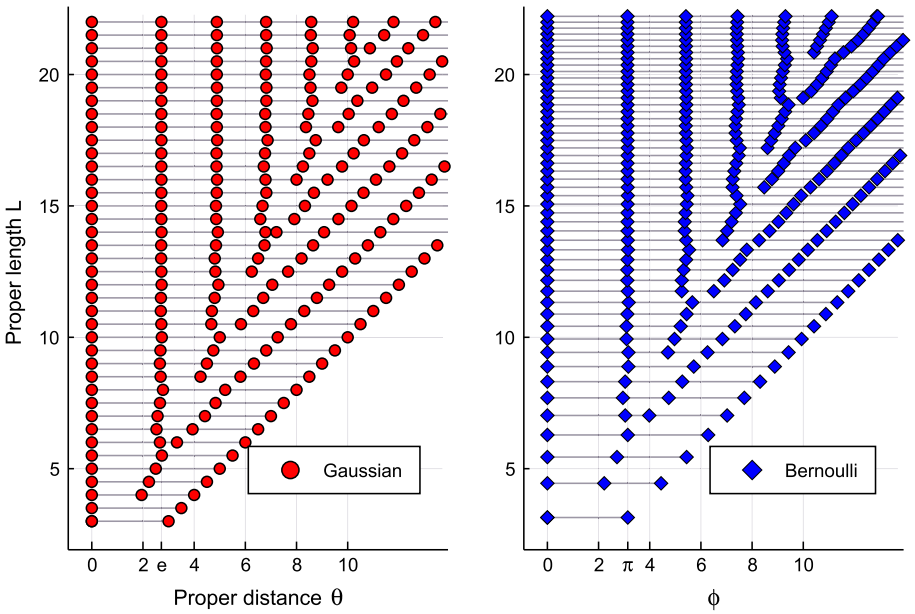

Fig. 3 Positions of delta functions in optimal priors $p_\star(\theta) = \sum_{a=1}^{K} \lambda_a \, \delta(\theta - \theta_a)$, for various values of L. Each horizontal line represents one prior. We observe that the second (and third...) delta functions occur at fixed proper distance from the first, justifying the fixed boundary condition on $\rho(x)$. On the right, we show similar data for the Bernoulli model of Sect. 4 below, in terms of proper distance ϕ. Here $L = \pi\sqrt{m}$ for $m = 1, 2, 3, \ldots, 50$

We are interested in the large-x behavior of a solution with boundary condition at $x = 0$ of $\rho = 1$. Or any constant density, but this value is independent of L because the only interaction is of scale $\sigma = 1$. This is also what we observe numerically, shown in Fig. 3. Making the ansatz $\rho(x) = 1 + x^\eta$ with $\eta > 1$, these four terms scale as

$$x^{5\eta-2}, \quad x^{5\eta-2}, \quad x^{7\eta-2}, \quad x^\eta, \qquad \text{all} \times e^{-x^{2\eta}}, \; x \to \infty. \tag{13}$$

Clearly the first two terms are subleading to the third, and thus the last two terms must cancel each other. We have $7\eta - 2 = \eta$ and thus $\eta = 1/3$. Then the total number of delta functions in length L is

$$K = \int_0^L dx \, \rho(x) \sim L^{4/3} \tag{14}$$

establishing the result (2).

4 Extensions

The other one-dimensional example studied in [7] was Bernoulli problem, of determining the weighting of an unfair coin given the number of heads seen after m flips:

$$p(x|\theta) = \frac{m!}{x!(m-x)!}\theta^x(1-\theta)^{m-x}, \qquad \theta \in [0,1], \qquad x \in \{0, 1, 2, 3, \ldots, m\}. \tag{15}$$

The Fisher metric here is

$$g_{\theta\theta} = \frac{m}{\theta(1-\theta)} \qquad \Rightarrow \qquad L = \int_0^1 \sqrt{g_{\theta\theta}\, d\theta^2} = \pi\sqrt{m}$$

and we define the proper parameter ϕ by

$$ds^2 = \frac{m\, d\theta^2}{\theta(1-\theta)} = d\phi^2 \qquad \Leftarrow \qquad \theta = \sin^2\left(\frac{\phi}{2\sqrt{m}}\right), \quad \phi \in \left[0, \pi\sqrt{m}\right].$$

The optimal prior found by maximizing the mutual information is again discrete, and when $m \to \infty$ it also obeys the scaling law (1) with the same slope ζ. Numerical data showing this is also plotted in Fig. 1 above. This scaling relies on the behavior far from the ends of the interval, where this binomial distribution can be approximated by a Gaussian:

$$p(x|\theta) \approx \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m\theta)^2/2\sigma^2}, \qquad m \to \infty,\ \theta \text{ finite}, \qquad \sigma^2 = m\,\theta(1-\theta). \tag{16}$$

The agreement of these very different models suggests that the $\zeta = 3/4$ power is in some sense universal, for nonsingular one-dimensional models.

Near to the ends of the interval, we observe in Fig. 3 that first few delta functions again settle down to fixed proper distances. In this regime (16) is not a good approximation, and instead the binomial (15) approaches a Poisson distribution:

$$p(x|\theta) \approx \frac{e^{-\mu}\mu^{-x}}{x!}, \qquad m \to \infty, \quad \mu = m\theta \approx \frac{\phi^2}{2} \text{ finite}. \tag{17}$$

The first few positions and weights are as follows:[6]

$$\begin{aligned} \phi_2 &\approx 3.13 & \lambda_2/\lambda_1 &\approx 0.63 \\ \phi_3 &\approx 5.41 & \lambda_3/\lambda_1 &\approx 0.54 \\ \phi_4 &\approx 7.42 & \lambda_4/\lambda_1 &\approx 0.49\,. \end{aligned} \tag{18}$$

This implies that the second delta function is at mean $\mu \approx 2.47$, skipping the first few integers x.

More Dimensions

Returning to the bulk scaling law, one obvious thing to wonder is whether this extends to more dimensions. The trivial example is to consider the same Gaussian model (8) in D-dimensional cube:

$$p(\mathbf{x}|\boldsymbol{\theta}) \propto e^{-(\mathbf{x}-\boldsymbol{\sigma})^2/2}, \quad \boldsymbol{\theta} \in [0, L]^D,\ \mathbf{x} \in \mathbb{R}^D. \tag{19}$$

This simply factorizes into the same problem in each direction: (6) is the sum of D identical mutual information terms. Thus the optimal prior is simply

$$p_\star(\boldsymbol{\theta}) = \prod_{\mu=1}^{D} p_\star(\theta_\mu) = \sum_{a_1 \ldots a_D = 1}^{K} \lambda_{a_1} \cdots \lambda_{a_D}\, \delta(\theta - \theta_{a_1}) \cdots \delta(\theta - \theta_{a_D})$$

[6] For the Gaussian model (5), the corresponding table reads:

$$\begin{aligned} \theta_2 &\approx 2.718 & \lambda_2/\lambda_1 &\approx 0.672 \\ \theta_3 &\approx 4.889 & \lambda_3/\lambda_1 &\approx 0.582\,. \end{aligned}$$

See also [20] and references therein.

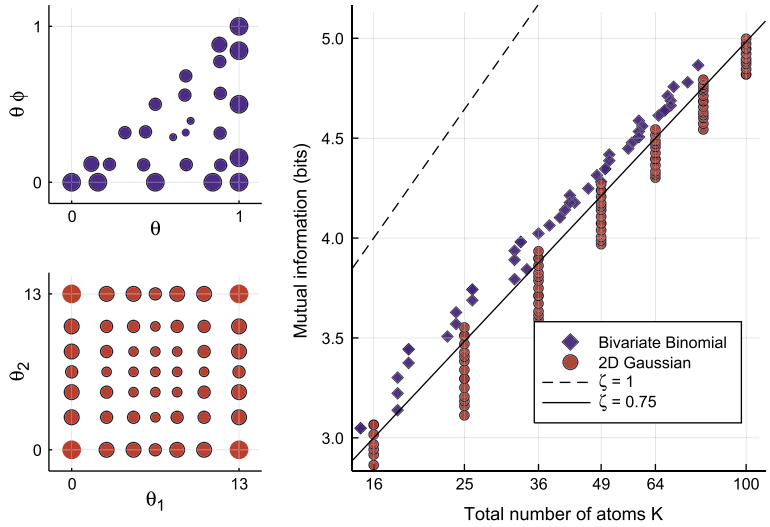

Fig. 4 Scaling law in two dimensions. On the right we plot $I(X;\Theta)/\log 2$ against $\log K_{\text{tot}}$ for the bivariate binomial model (21) and the $D = 2$ Gaussian model (19). On the left we show examples of the priors, for $m = 20$ and $L_1 = L_2 = 13$. For the bivariate binomial the plot axes are $(\theta, \theta\phi)$ so as to respect the topology of the parameter space, but the figure is not isometric to Θ

with the same coefficients as in (3) above. The total number of delta functions is $K_{\text{tot}} = K^D$ which scales as

$$K_{\text{tot}} \sim L^{D/\zeta} = V^{1/\zeta}, \qquad \zeta = 3/4, \quad V \to \infty. \tag{20}$$

We believe that this scaling law is also generic, provided the large-volume limit is taken such that all directions expand together. If the scaling arises from repeating an experiment m times, then this will always be true as all directions grow as $\sqrt{m}$.

To check this in a less trivial example, we consider now the bivariate binomial problem studied by [21]. We have two unfair coins whose weights we wish to determine, but we flip the second coin only when the first coin comes up heads. After m throws of the first coin, the model is

$$p(x, y|\theta, \phi) = \binom{m}{x}\theta^x(1-\theta)^{m-x}\binom{x}{y}\phi^y(1-\phi)^{x-y}. \tag{21}$$

with $\theta, \phi \in [0, 1]$ and $0 \le y \le x \le m \in \mathbb{Z}$. The Fisher information metric here is

$$ds^2 = \frac{m}{\theta(1-\theta)}d\theta^2 + \frac{m\,\theta}{\phi(1-\phi)}d\phi^2$$

which implies

$$V = \int_0^1 d\theta \int_0^1 d\phi \,\sqrt{\det g_{\mu\nu}} = 2\pi m$$

and $p_{\text{J}}(\theta, \phi) = \frac{1}{2\pi}[(1-\theta)\phi(1-\phi)]^{-1/2}$. Topologically the parameter space is a triangle, since at $\theta = 0$ the ϕ edge is of zero length. The other three sides are each of length $\pi\sqrt{m}$, and so will all grow in proportion as $m \to \infty$.

We can find the optimal priors for this numerically.[7] In Fig. 4 we see that the mutual information obeys the same law as (1) above: $I(X;\Theta) \sim \zeta \log K$ with $\zeta \approx 0.75$. Since the

[7] See the appendix of [7] for a discussion of the algorithms used here, and [22] for software.

Fisher volume is proportional to the number of distinguishable states $I(X;\Theta) \sim \log V$, this implies (20).

Finally, suppose that instead of a square (or an equilateral triangle), a two-dimensional Θ has one direction much longer than the other:

$$L_1 = a_1\sqrt{m}, \qquad L_2 = a_2\sqrt{m}, \qquad a_1 \ll a_2.$$

Then as we increase m we will pass through three regimes, according to how many of the lengths are long enough to be in the scaling regime:

$$\begin{aligned} L_1, L_2 \lesssim 1 \quad &: \quad K \text{ constant} \\ L_1 \lesssim 1 \ll L_2 \quad &: \quad K \sim L^{1/\zeta} \propto m^{1/2\zeta} = m^{2/3} \\ 1 \ll L_1, L_2 \quad &: \quad K \sim L^{2/\zeta} = V^{1/\zeta} \propto m^{1/\zeta} = m^{4/3}. \end{aligned} \tag{22}$$

The last regime is the one we discussed above. When plotting K against $\log m$ (or $\log L$), we expect to see a line with a series of straight segments, and an increase in slope every time another dimension becomes relevant.[8]

5 Conclusion

The fact that $\zeta < 1$ is important for the qualitative behavior of the priors studied in [7]. This is what ensures that the number of delta functions $K \sim L^{1/\zeta}$ grows faster than the Fisher length of parameter space L, ensuring that discreteness washes out in the asymptotic limit $L \to \infty$. Parameters which we can measure with good accuracy are in this limit. For such a parameter, the posterior $p(\theta|x)$, which is also discrete, has substantial weight on an increasing number of points, and in this sense approaches a continuous description.

In Sect. 4 we also studied some generalizations beyond what we did in [7]. Very near to the end of a long parameter, the discreteness does not wash out as $L \to \infty$, and we wrote down its proper position for Gaussian and Poisson models. And we observed that this scaling law holds in any number of dimensions, if stated in terms of the mutual information (1). But stated in terms of the length L, it gives a slope which depends on the number of large dimensions (22), and hence has phase transitions as more parameters become relevant.

While our motivation here was finding optimal priors, our conclusions apply to a much larger class of problems, including the maximization of channel capacity over a continuous input distribution [14,15,24–26], which is formally equivalent to what we did above. This problem is where mutual information was first discussed [1], and discreteness was first seen in this context [8–10]. This maximization is also equivalent to a minimax optimization problem [27], and discreteness was known in other minimax problems slightly earlier [11–13]; see [3,4] for other work in statistics. More recently discreteness has been employed in economics [28,29], and is seen in various systems optimized by evolution [30–32]. This scaling law should apply to all of these examples, when interpolating between the coarse discreteness at small L and the continuum $L \to \infty$.

Acknowledgements We thank Henry Mattingly and Mark Transtrum for collaboration on [7]. This work was performed in part at Aspen Center for Physics, which is supported by National Science Foundation grant PHY-1607611; M.C.A.'s visit was supported by a grant from the Simons Foundation. M.C.A. was supported by NCN grant 2012/06/A/ST2/00396, and a Wigner Fellowship. B.B.M. was supported by a Lewis-Sigler

[8] These transitions are what LaMont and Wiggins call high-temperature freeze-out [23].

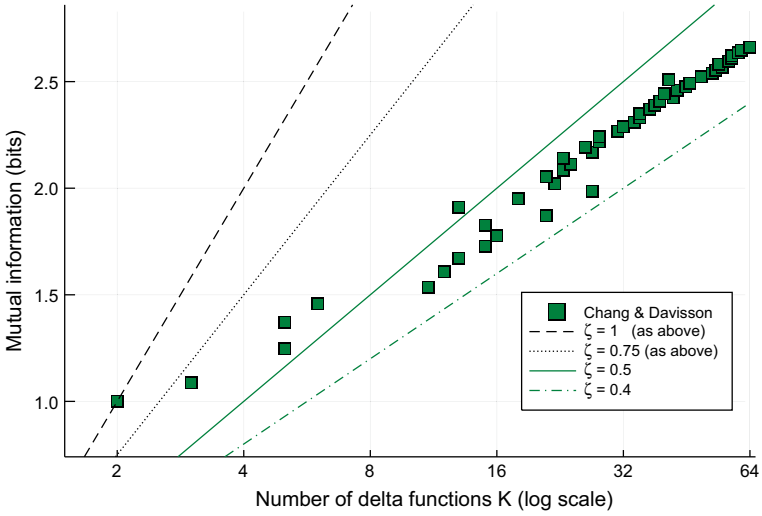

Fig. 5 Approximate scaling law (23) seen in the data from Table 1 of [24]

Fellowship and by NSF PHY 0957573. Open access funding provided by MTA Wigner Research Centre for Physics (MTA Wigner FK, MTA EK).

A Chang and Davisson

Chang and Davisson [24] came close to finding this scaling law. In our notation, they maximized $I(X;\Theta)$ for various L, and noted the number of delta functions K' their algorithm used to achieve good enough accuracy. But this is not quite the optimal K: they tend to use too many delta functions, in a way which varies with L. This results in a slightly different law:

$$I(X;\Theta) \sim \zeta' \log K', \qquad \zeta' \lesssim 0.5\,. \tag{23}$$

Their paper does not mention this, nor attempt to plot this data as in Fig. 5.

References

1. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 623–656 (1948)
2. Lindley, D.V.: On a measure of the information provided by an experiment. Ann. Math. Stat. **27**, 986–1005 (1956)
3. Berger, J.O., Bernardo, J.M., Mendoza, M.: On Priors that Maximize Expected Information, Recent Developments in Statistics and Their Applications , pp. 1–20 (1988). https://www.uv.es/~bernardo/1989Seoul.pdf
4. Zhang, Z.: Discrete Noninformative Priors. PhD thesis, Yale University (1994). [UMI 9523257]
5. Clarke, B.S., Barron, A.R.: Jeffreys' prior is asymptotically least favorable under entropy risk. J. Stat. Plan. Inference **41**, 37–60 (1994)

6. Bernardo, J.M.: Reference posterior distributions for Bayesian inference. J. R. Stat. Soc. B **41**, 113–147 (1979). http://www.uv.es/~bernardo/1979JRSSB.pdf
7. Mattingly, H.H., Transtrum, M.K., Abbott, M.C., Machta, B.B.: Maximizing the information learned from finite data selects a simple model. PNAS. **115**, 1760–1765 (2018). arXiv:1705.01166
8. Färber, G.: Die Kanalkapazität allgemeiner Übertragunskanäle bei begrenztem Signalwertbereich beliebigen Signalübertragungszeiten sowie beliebiger Störung. Arch. Elektr. Übertr. **21**, 565–574 (1967)
9. Smith, J.G.: The information capacity of amplitude-and variance-constrained scalar gaussian channels. Inf. Control **18**, 203–219 (1971)
10. Fix, S.L.: Rate distortion functions for squared error distortion measures, Proc. 16th Annu. Allerton Conf. Commun. Control Comput. 704–711 (1978)
11. Ghosh, M.N.: Uniform approximation of minimax point estimates. Ann. Math. Stat. **35**, 1031–1047 (1964)
12. Casella, G., Strawderman, W.E.: Estimating a bounded normal mean. Ann. Stat. **9**, 870–878 (1981)
13. Feldman, I.: Constrained minimax estimation of the mean of the normal distribution with known variance. Ann. Stat. **19**, 2259–2265 (1991)
14. Arimoto, S.: An algorithm for computing the capacity of arbitrary discrete memoryless channels. IEEE Trans. Inf. Theory **18**, 14–20 (1972)
15. Blahut, R.: Computation of channel capacity and rate-distortion functions. IEEE Trans. Inf. Theory **18**, 460–473 (1972)
16. Gutenkunst, R.N., Waterfall, J.J., Casey, F.P., Brown, K.S., Myers, C.R., Sethna, J.P.: Universally sloppy parameter sensitivities in systems biology models. PLoS Comp. Biol. **3**, e189 (2007)
17. Machta, B.B., Chachra, R., Transtrum, M.K., Sethna, J.P.: Parameter space compression underlies emergent theories and predictive models. Science **342**, 604–607 (2013). [arXiv:1303.6738]
18. Lindley, D.V.: Bayesian Statistics, A Review. SIAM, Philadelphia (1972)
19. Lewis, N.: Combining independent Bayesian posteriors into a confidence distribution, with application to estimating climate sensitivity. J. Stat. Plan. Inference **195**, 80–92 (2018)
20. Amir, A., Lemeshko, M., Tokieda, T.: Surprises in numerical expressions of physical constants. Am. Math. Mon. **123**, 609–612 (2016). [arXiv:1603.00299]
21. Polson, N., Wasserman, L.: Prior distributions for the bivariate binomial. Biometrika **77**, 901–904 (1990)
22. Abbott, M.C.: Rational Ignorance, https://github.com/mcabbott/RationalIgnorance.jl (2017)
23. LaMont, C.H., Wiggins, P.A.: A correspondence between thermodynamics and inference, arXiv:1706.01428
24. Chang, C.-I., Davisson, L.D.: On calculating the capacity of an infinite-input finite (infinite)-output channel. IEEE Trans. Inf. Theory **34**, 1004–1010 (1988)
25. Lafferty, J., Wasserman, L.: Iterative Markov chain Monte Carlo computation of reference priors and minimax risk, Proc. 17th conf. Uncert. AI 293–300 (2001) arXiv:1301.2286
26. Dauwels, J.: Numerical computation of the capacity of continuous memoryless channels. In: Proc. 26th Symp. Inf. Theory Benelux (2005). http://www.dauwels.com/files/memoryless.pdf
27. Haussler, D.: A general minimax result for relative entropy. IEEE Trans. Inf. Theory **43**, 1276–1280 (1997)
28. Sims, C.A.: Rational inattention: beyond the linear-quadratic case. Am. Econ. Rev. **96**, 158–163 (2006)
29. Jung, J., Kim, J., Matejka, F., Sims, C.A.: Discrete Actions in Information-Constrained Decision Problems. http://www.princeton.edu/~sims/#RIDiscrete (2015)
30. Balasubramanian, V., Sterling, P.: Receptive fields and functional architecture in the retina. J. Physiol. **587**, 2753–2767 (2009)
31. Mayer, A., Balasubramanian, V., Mora, T., Walczak, A.M.: How a well-adapted immune system is organized. PNAS **112**, 5950–5955 (2015). [arXiv:1407.6888]
32. Sharpee, T.O.: Optimizing neural information capacity through discretization. Neuron **94**, 954–960 (2017)