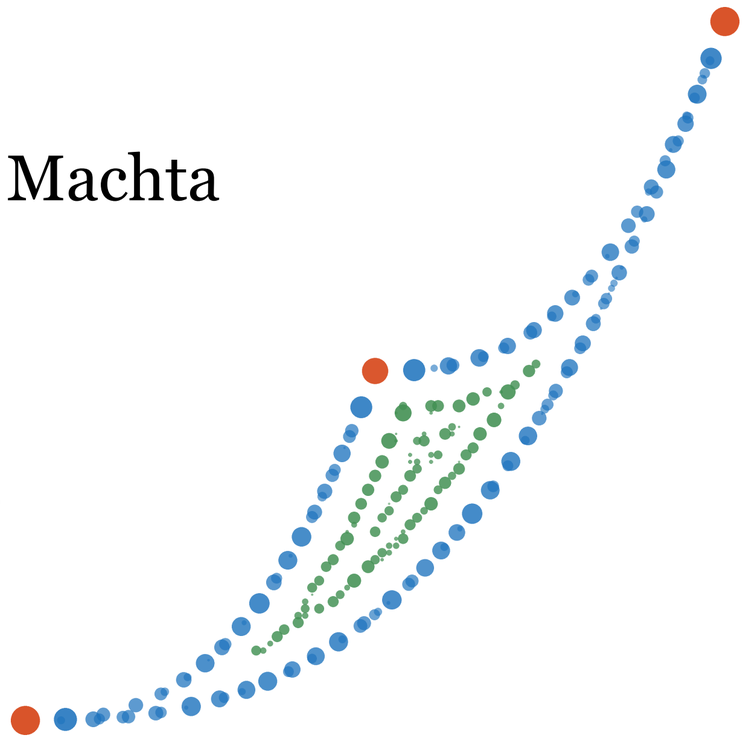
A Recipe for Simple Models

Michael Abbott
UJ

with Henry Mattingly, Mark Transtrum, Ben Machta
Princeton[2] + Brigham Young

arxiv: 1705.01166
“Rational Ignorance: Simpler Models
Learn More from Finite Data”



Background

Entropy is a measure of how much we don't know:

$$S = -\sum_i p_i \log p_i$$

Shannon, 1948

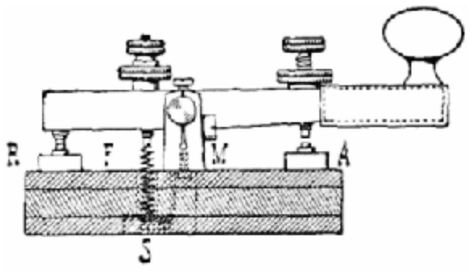
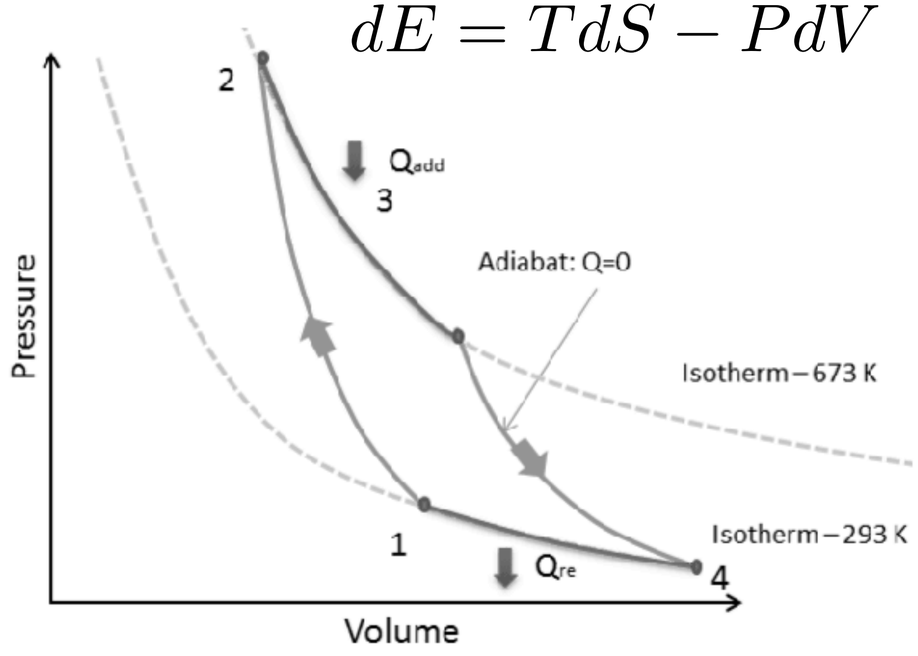
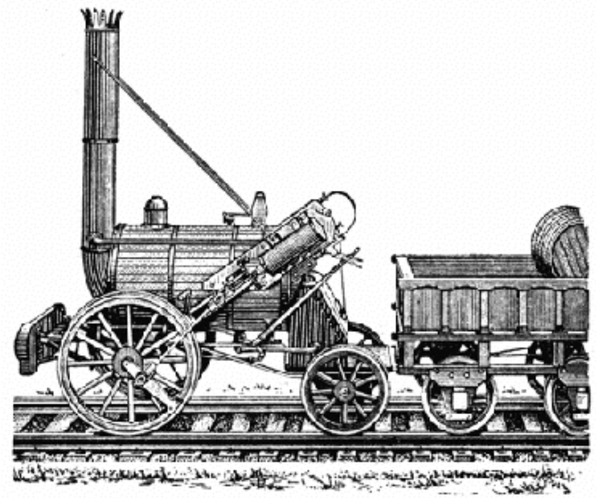


Fig. 6.

But this isn't how it was first used:

$$dE = TdS - PdV$$





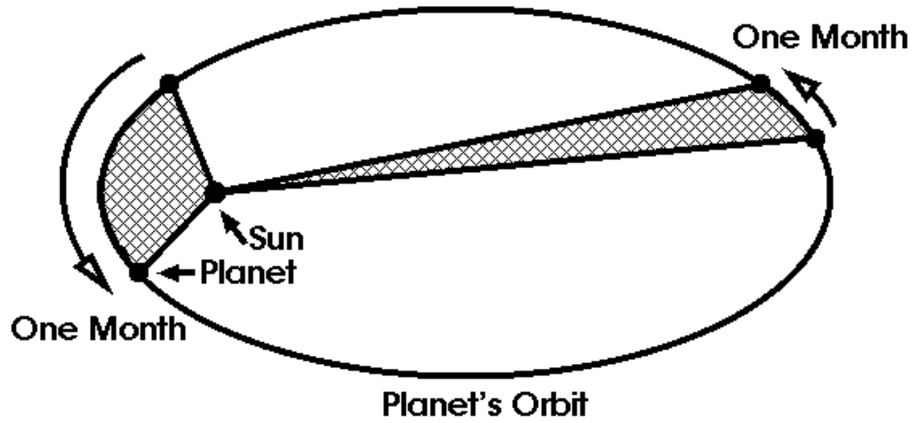
Carnot, 1828

Connection, one you know about atoms: $S = k_B \log \Omega = -k_B \sum_i p_i \log p_i$

(k_B is there because units of temperature and energy were set without knowing this)

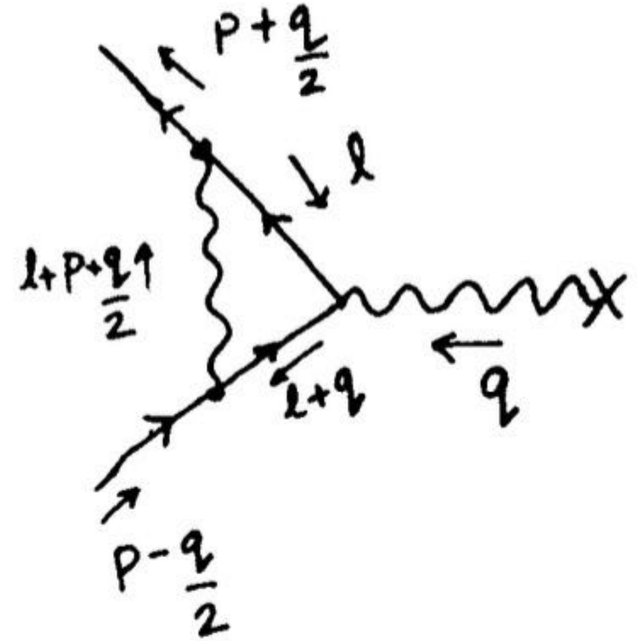
Gibbs, Maxwell, Boltzmann

All of physics is like this:



- Newton didn't need to know what Jupiter was made of (or whether there was life on Mars) only their masses.

- Navier-Stokes equation is the same for water, olive oil, and honey — only density & viscosity matter.

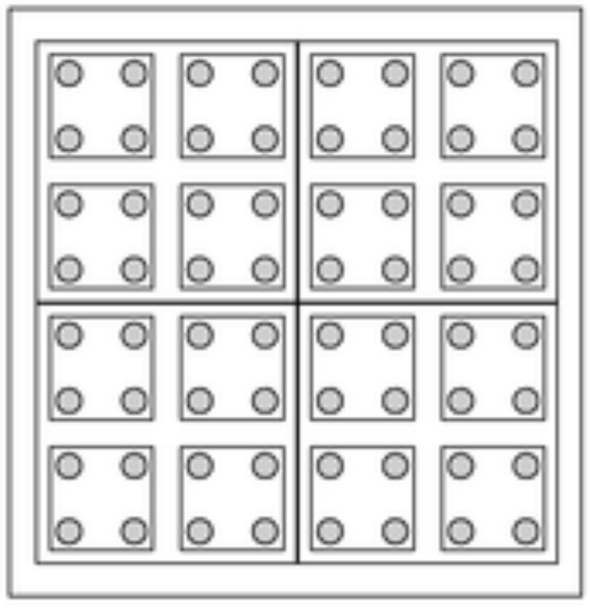


- Feynman didn't need to know about nuclear forces (or the Higgs boson!) to understand QED — only m, e, α

We can describe macroscopic features of many systems while ignorant of many microscopic details — they are irrelevant.

Deep understanding of why comes from Kadanoff & Wilson…

Kadanoff, 1966:



1. Consider some 2D spin system, spacing a.
   Interactions governed by some parameters $\theta_1,\theta_2,\theta_3$.

2. Define new "block spins" on a lattice of spacing $2a$.

3. Suppose these have similar equations, but with altered parameters
   $\theta'_1, \theta'_2, \theta'_3$ — Some grow, some shrink.

4. After many repetitions, only some parameters matter — relevant.
   The others you can ignore — irrelevant.

Wilson, 1972:

Quantum field theory was plagued by infinities.
But in certain theories (called renormalizable) these always cancel out.

He showed that these are precisely theories with relevant parameters.

Motivation

Can we do this without scale?

Seek a statistical understanding of what irrelevance means,
i.e. how to select good effective theories.

More familiar concern is overfitting.

Information criteria (AIC, BIC etc.) are expansions around $N=\infty$ ...

$$-\ln P(\{x_i, y_i\}|\text{class}) \approx -\sum_{i=1}^{N} \ln P(x_i) + \frac{1}{2}\chi^2_{\min} + \frac{N}{2}\ln(2\pi\sigma^2) + \frac{K}{2}\ln N + \cdots, \quad (6.296)$$

where the first three terms are $\propto N$, and the omitted terms (including what we have neglected in the saddle point approximation) are constant or decreasing as $N \to \infty$. Again, the negative log probability measures the length of the shortest code for $\{x_i, y_i\}$

Nested Models

Outermost model $p(\theta|x)$ parameters $\theta \in \Theta$, data $x \in X$

Effective model ≡ Subspace of Θ

≡ prior $p(\theta) \neq 0$ only on subspace $p(\vec{\theta}) = p_{\text{rel}}(\theta_1, \theta_2)\ \delta(\theta_3)\delta(\theta_4)$

Bayesian updating respects this:

$$p(\theta|x) = \frac{p(x|\theta)\ p(\theta)}{p(x)} \longleftarrow p(x) = \int d\theta\ p(\theta)\ p(x|\theta)$$

Some Technology

Mutual Information:

$$\mathrm{MI} = S(X) - S(X|\Theta)$$

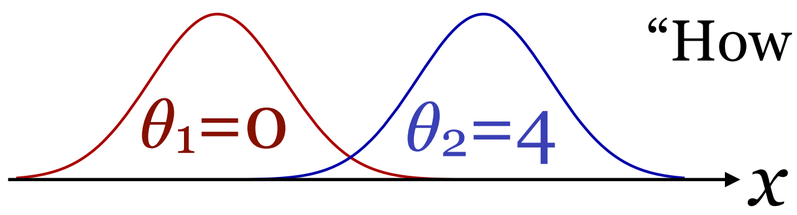
$$= \int d\theta\, p(\theta)\, f_{\mathrm{KL}}(\theta)$$

$$f_{\mathrm{KL}}(\theta) = D_{\mathrm{KL}}\left(p(x|\theta)\middle\|p(x)\right) = \int dx\, p(x|\theta)\, \log\frac{p(x|\theta)}{p(x)}$$

"How many bits do we learn about x by knowing θ?" (or the converse)

Fisher Metric:

$$g_{\mu\nu}(\vec{\theta}) = \int dx\, p(x|\vec{\theta})\, \partial_\mu \log p(x|\vec{\theta})\, \partial_\nu \log p(x|\vec{\theta})$$



"How many standard deviations away from $p(x|\theta_1)$ is $p(x|\theta_2)$?"

Jeffreys Prior:

$$p_{\mathrm{J}}(\vec{\theta}) \propto \sqrt{\det g_{\mu\nu}}$$

"What's the simplest invariant density from this metric?"

Maximising Mutual Information

$$\mathrm{MI} = S(X) - S(X|\Theta)$$

… makes for the most informative model?
… selects the least informative prior?

Properties:

i. Usually discrete: $p_\star(\vec{\theta}) = \sum_{i=1}^{N} \lambda_i \delta(\vec{\theta} - \vec{\theta}_i)$

ii. Approaches Jeffreys' if there is lots of data (or many i.i.d. repetitions)

iii. If there is not much data, the weight is largely on the boundaries of the manifold Θ.

Statisticians:

- Bernando 1979 maximised MI, took limit… and got Jeffreys.
- Berger, Bernardo, Mendoza 1988 saw discreteness, and didn't like it.

Engineers:

- Färber 1967, Smith 1971, Fix 1978 found discreteness — MI exactly equivalent to maximising channel capacity.
- Economists cite the Engineers: Sims et al, Rational Inattention, 2006

Biologists:

- Mayer, Balasubramanian, Mora, Walczak 2014 in different maximisation problems.

Example 1 — Unfair Coin

Probability of heads is θ.

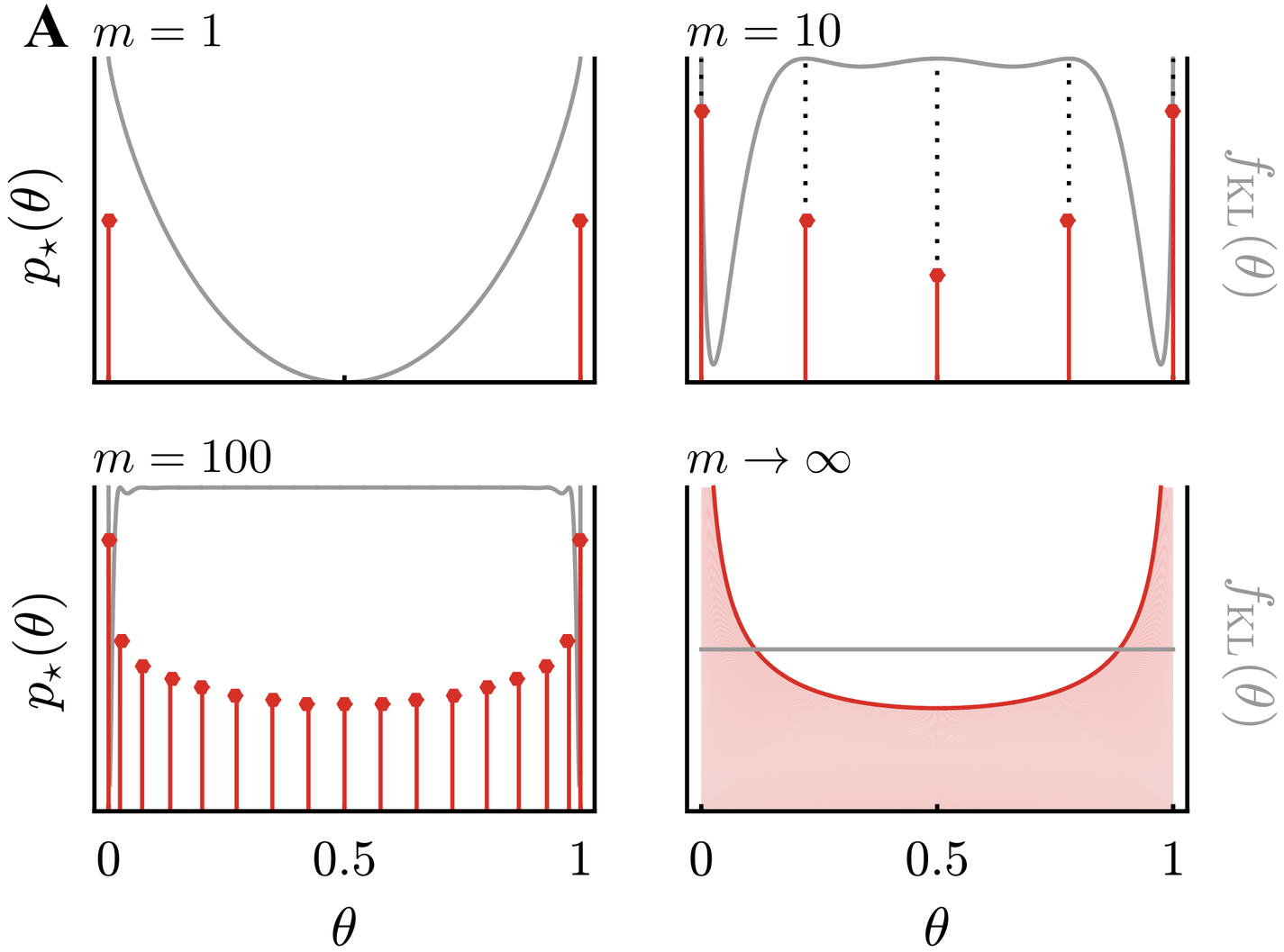
Data is $x_j \in \{0, 1\}$ for $j = 1 \ldots m$
or just the number of heads, y

Model is

$$p(\vec{x}|\theta) = \prod_{j=1}^{m} |x_j - \theta| = \theta^y (1-\theta)^{m-y}$$

Fisher metric is $g_{\theta\theta} = \dfrac{m}{\theta(1-\theta)}$

$L = \int \sqrt{ds^2} = \pi\sqrt{m}$



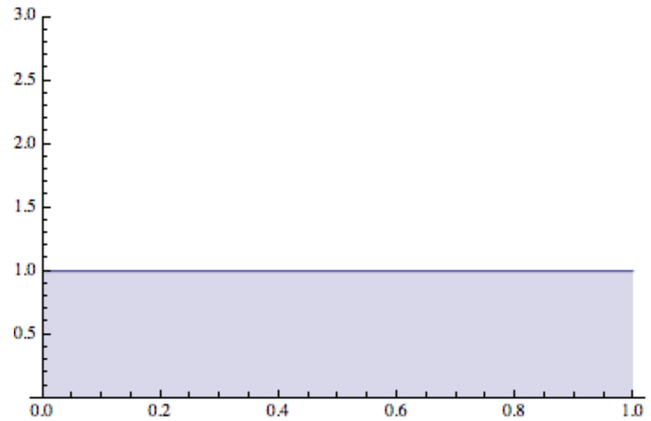
Aside, usual priors are Beta distributions

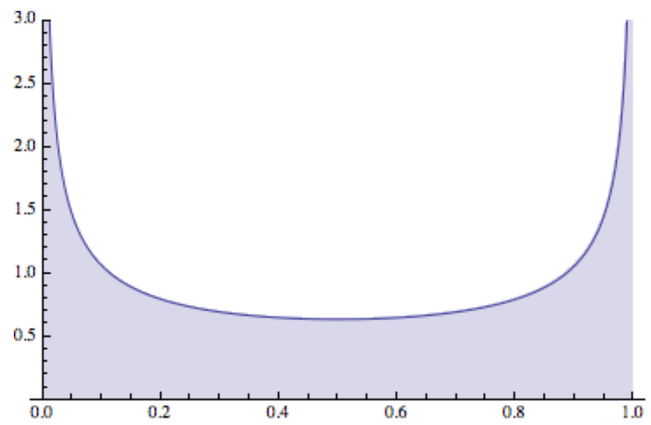
$$p(\theta) = B_{\alpha,\beta}(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

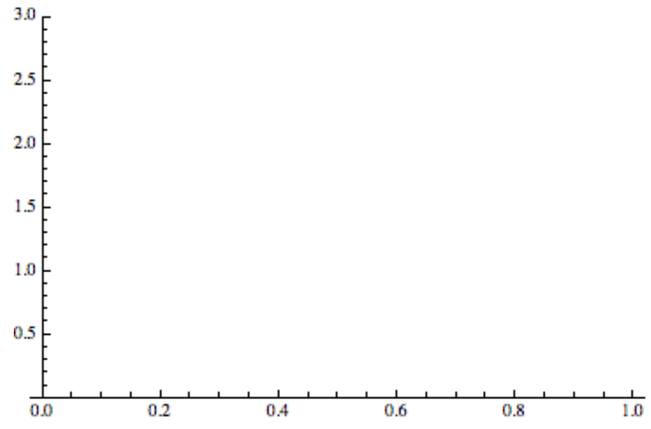
with the nice property that after y heads

$$p(\theta|y) \;=\; B_{\alpha+y,\beta+m-y}(\theta)$$

- Bayes 1794: $B_{1,1}$ = constant
- Jeffreys: $B_{1/2,1/2}$
- Haldane: $B_{0,0}$ improper…







At right, x = (0,1,0,1,0,…)

As m → ∞,
every smooth prior gives the same $p(\theta|x)$

Example 2 — 1D with Gaussian Noise

Estimate position θ from photon arriving at x,
finite resolution σ:

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

Infinite track, finite film?

Easier: finite track, infinite film:

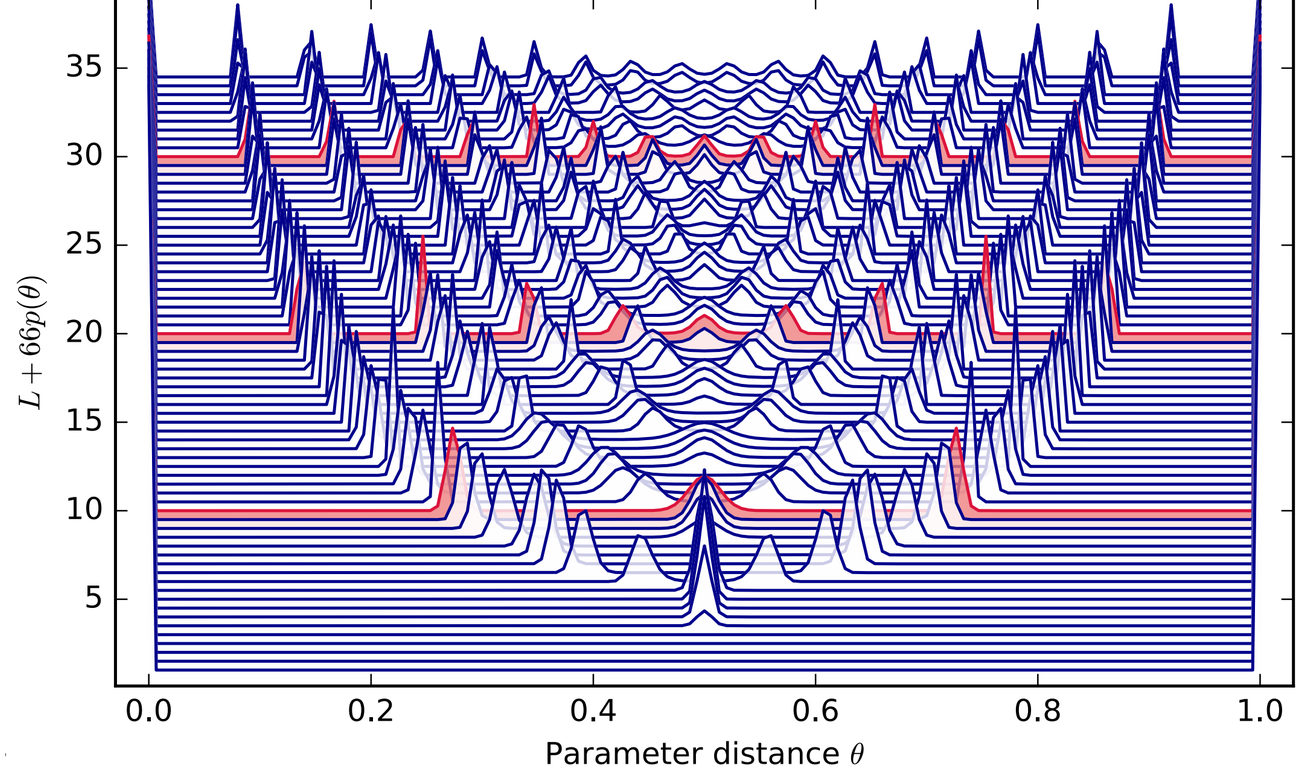
$$\theta \in [0,1], \quad x \in \mathbb{R}$$

$$L = 1/\sigma$$

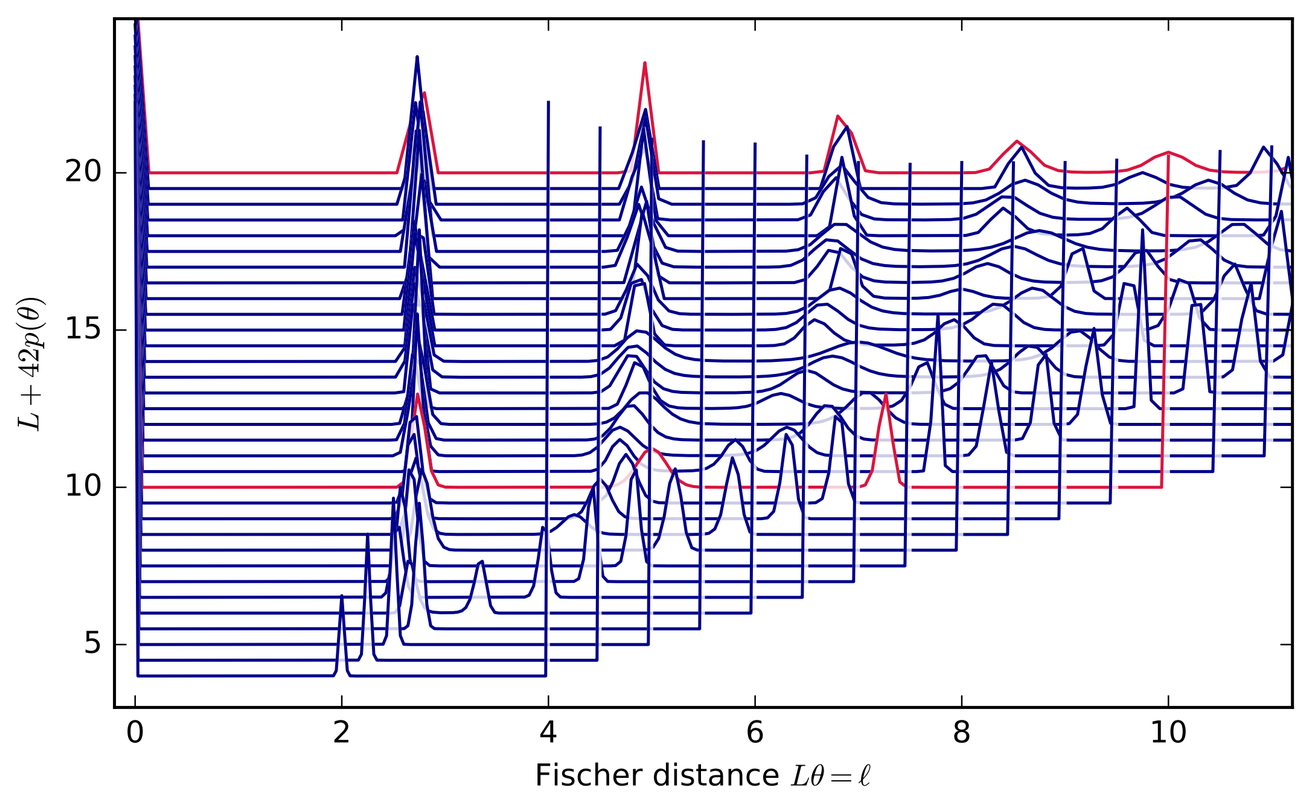
Repetitions = smaller σ.

Optimum $p(\theta)$ very similar,

in this case found by discretising
and using Blahut-Arimoto algorithm



Spacing is closer in the centre,
and consistent near the ends.



An Algorithm for Computing the Capacity of Arbitrary Discrete Memoryless Channels

SUGURU ARIMOTO

Abstract—A systematic and iterative method of computing the capacity of arbitrary discrete memoryless channels is presented. The algorithm is very simple and involves only logarithms and exponentials in addition to elementary arithmetical operations. It has also the property of monotonic convergence to the capacity. In general, the approximation error is at least inversely proportional to the number of iterations; in certain circumstances, it is exponentially decreasing. Finally, a few inequalities that give upper and lower bounds on the capacity are derived.

I. Introduction

IT IS well known that the capacity of discrete memoryless channels that are symmetric from the input can easily

The mutual information concerning the channel P is defined by

$$I(P;p) = H(p) - H(P;p), \tag{4}$$

where

$$H(p) = \sum_{j=1}^{n} -p_j \log p_j, \tag{5}$$

$$H(P;p) = -\sum_{i=1}^{m} \sum_{j=1}^{n} p(i/j)\, p_j \log \frac{p(i/j)\, p_j}{\sum_{k=1}^{n} p(i/k)\, p_k}, \tag{6}$$

where $p = (p_1, \cdots, p_n) \in \bar{D}^n$ is a probability vector of input symbols. From Shannon's coding theorem [3] the capacity of the memoryless channel P, which will be denoted by $C(P)$, is given by

$$C(P) = \max_{p \in \bar{D}^n} I(P;p). \tag{7}$$

III. Iterative Procedure and Convergence

Based upon the characterization (12) of capacity, we now propose a procedure for capacity evaluation, which comprises the following steps.

i) Initially, choose an arbitrary probability vector $p^1 \in D^n$ (in practice the uniform probability distribution $p_j{}^1 = 1/n$ for all $j = 1, \cdots, n$ is generally suitable). Then, the following two steps are iterated as $t = 1,2, \cdots$.

ii)$_t$ Maximize $H(p^t) - J(P;p^t,\phi)$ with respect to $\phi \in \Phi$ while fixing p^t. According to (11) the maximizing ϕ is

$$\phi^t(j/i) = \frac{p(i/j)\, p_j{}^t}{\sum_{k=1}^{n} p(i/k)\, p_k{}^t}, \tag{16}$$

that is,

$$C(t,t) = \max_{\phi \in \Phi} [H(p^t) - J(P;p^t,\phi)] = H(p^t) - J(P;p^t,\phi^t). \tag{17}$$

iii)$_t$ Maximize $H(p) - J(P;p,\phi^t)$ with respect to $p \in \bar{D}^n$ while fixing ϕ^t. This maximizing probability vector, denoted by p^{t+1}, is given by

$$p_j^{t+1} = \frac{r_j{}^t}{\sum_{k=1}^{n} r_k{}^t},$$

$$r_j{}^t = \exp\left[\sum_{i=1}^{m} p(i/j) \log \phi^t(j/i)\right]. \tag{18}$$

Scaling Law

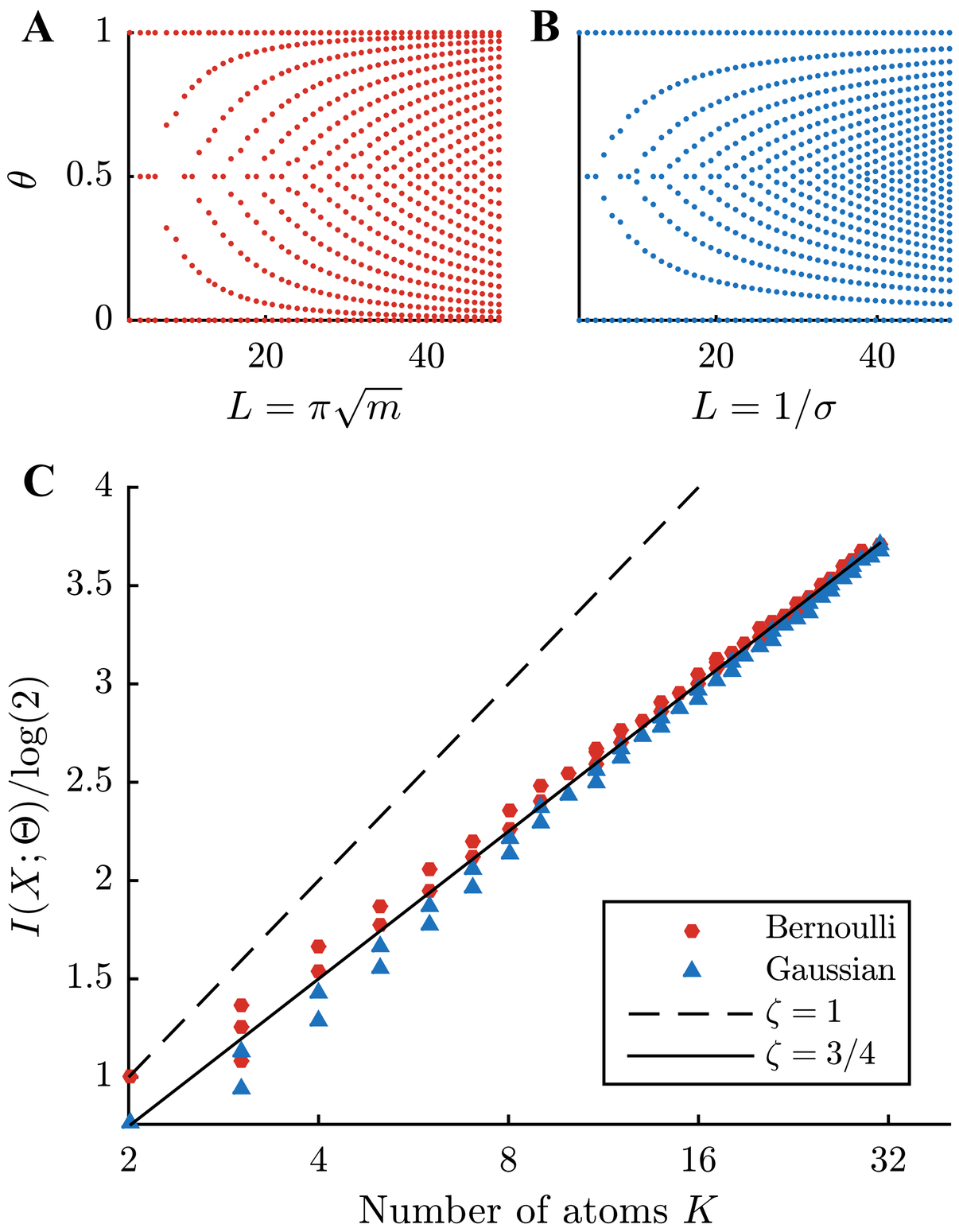
For K delta functions,
bound is MI ≤ log K

We observe MI ~ ζ log K
with ζ = 0.75

Since MI ~ ζ log L when $L \gg 1$,
this implies average spacing goes
like $L/K \sim L^{-1/3} \to 0$.

Analytically...

$$\mathcal{H} = e^{-(2\pi)^2\rho^2}\left[\rho^4(\rho')^2 + 1\right]$$



Example 3 — Radioactivity

Suppose we wish to determine the composition of an unknown radioactive source: as data we have n_t Geiger counter clicks at various times t (all together $\vec{n}$), and as parameters we have A_i activity units of isotope i with half life T_i (or rather, decay constant $k_i = \log 2/T_i$). The forward model is of course a Poisson distribution at each time:

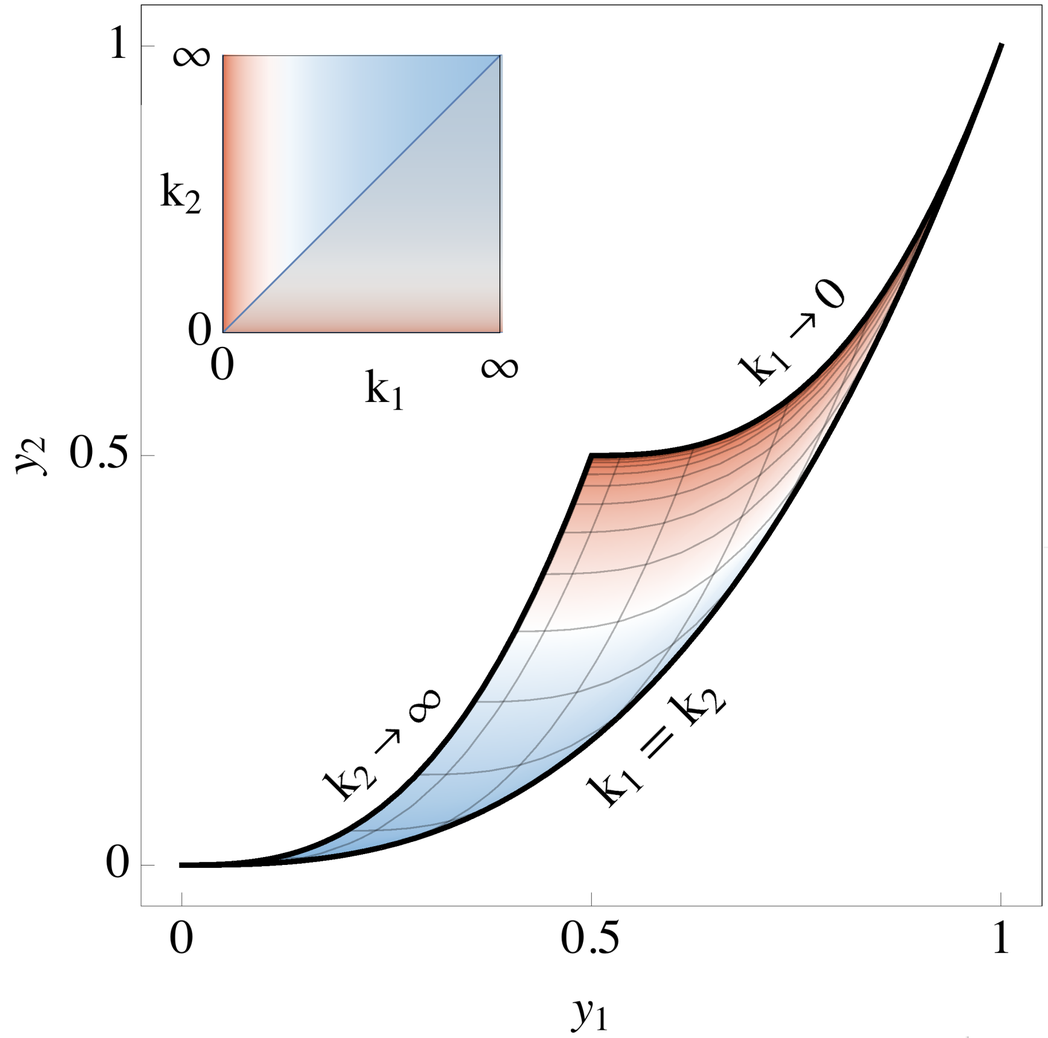
$$p(n_t|\vec{\theta}) = \frac{e^{-y_t} y_t^{n_t}}{n_t!}, \qquad y_t = \sum_i A_i e^{-k_i t}.$$

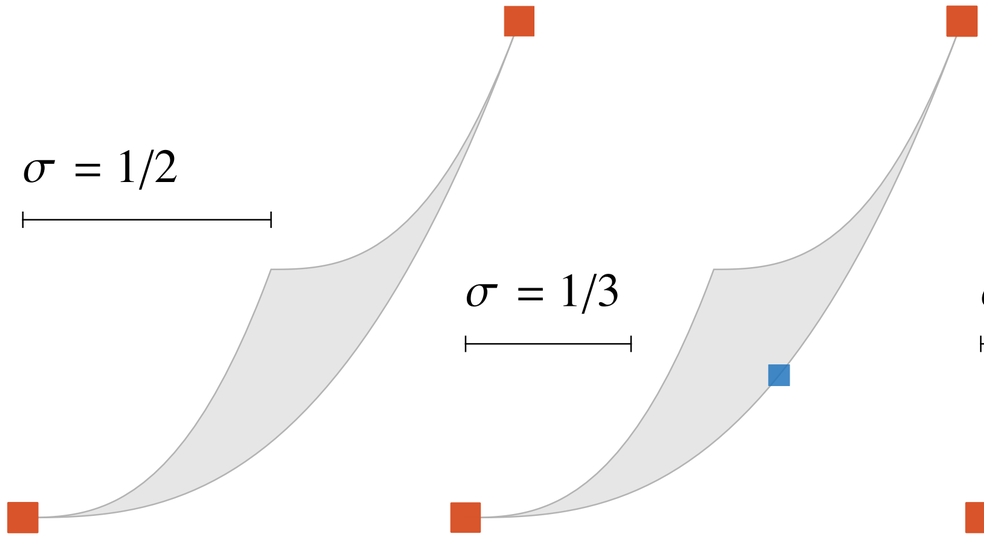
If we measure the number of counts at only two times t_1 and t_2, and approximate $p(n_t|\vec{\theta})$ by a Gaussian with $\sigma_t = \sqrt{y_t}$, then we have almost a two-dimensional version of example (9) above:

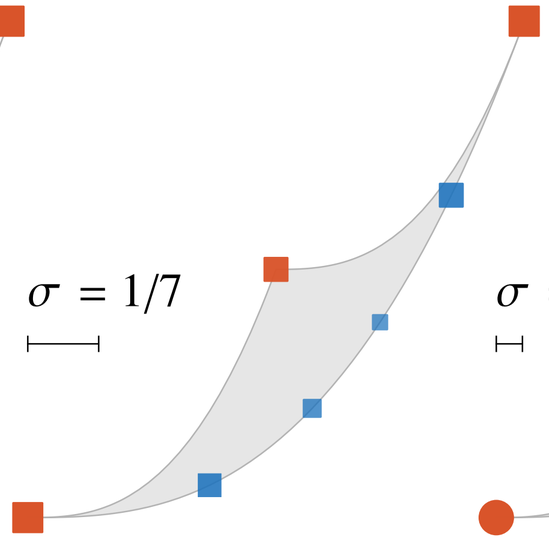
$$p(\vec{n}|\vec{\theta}) \propto e^{-(n_1-y_1)^2/2\sigma_1}\, e^{-(n_2-y_2)^2/2\sigma_2}. \qquad (11)$$

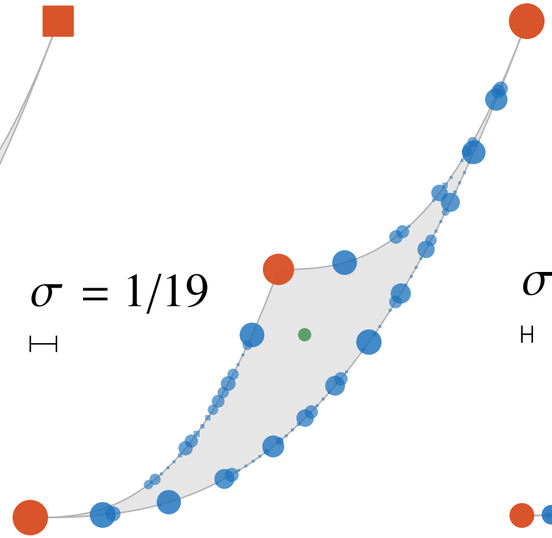
The centre of the distribution $\vec{y} = (y_1, y_2)$ plays the role of θ above. However not all points (y_1, y_2) are allowed, and the boundaries of the allowed region will determine the behavior.

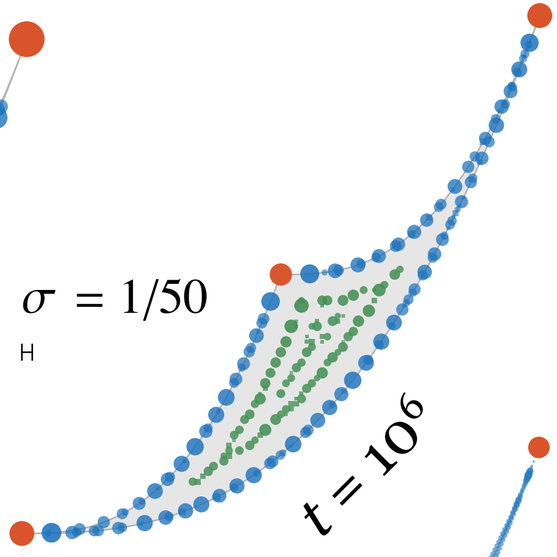
We can see the essential behavior with just two isotopes in fixed quantities $y_t = \frac{1}{2}(e^{-k_1 t} + e^{-k_2 t})$ and fix $\sigma_1 = \sigma_2 = \sigma$ and $t_2 = 2t_1$. While the parameter

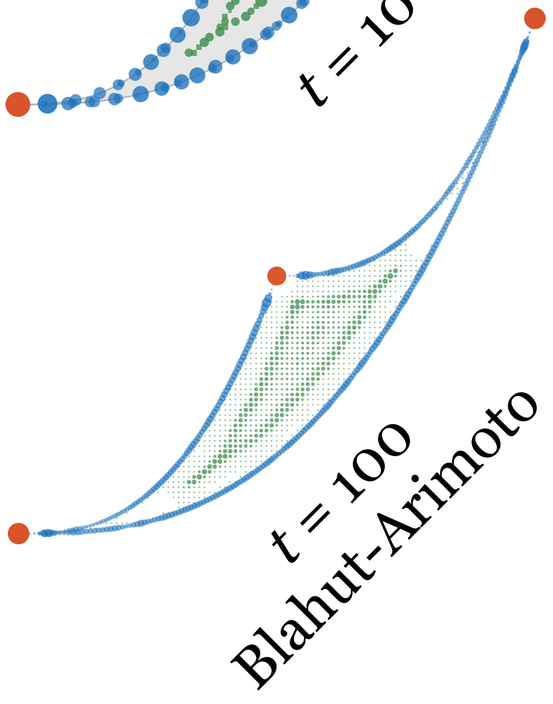






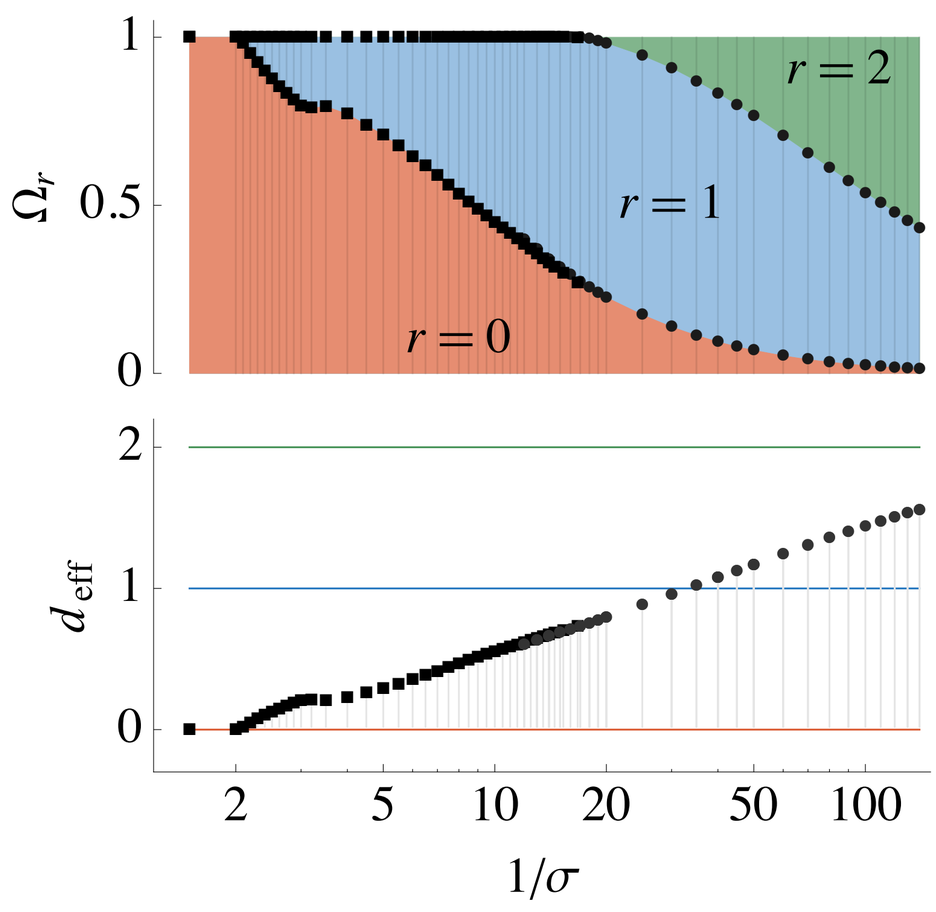






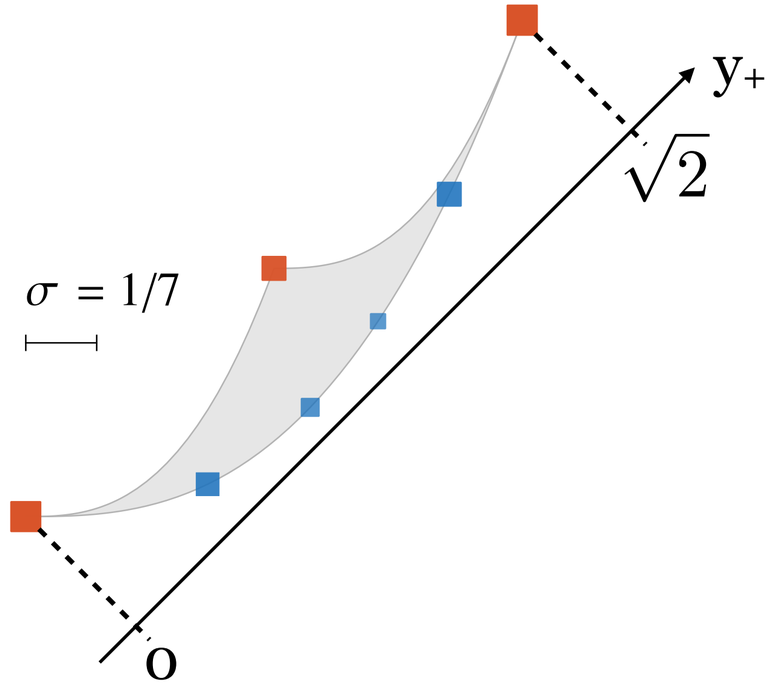
As noise is decreased, gradual progression from zero- to two-dimensional effective model.

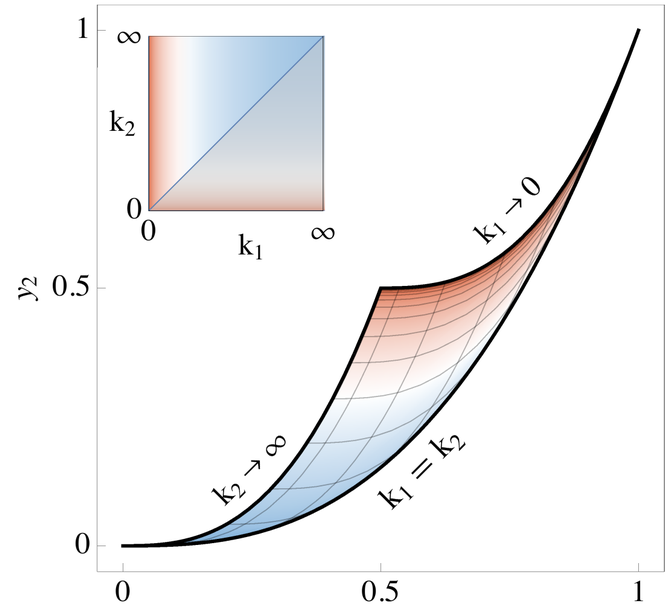
$$d_{\text{eff}} = \sum_{r=0}^{2} r\,\Omega_r$$



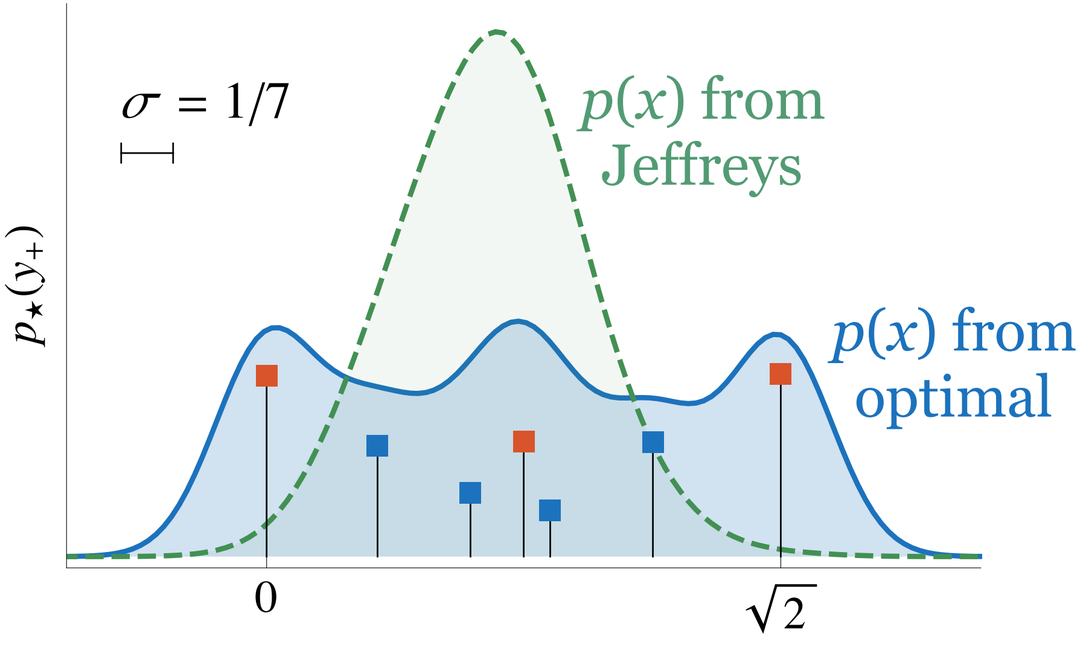
At medium values…

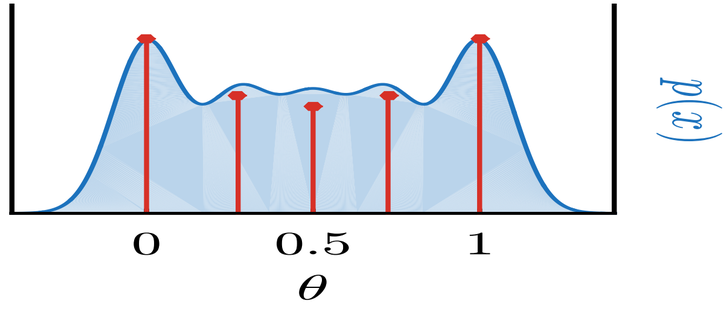
Plot the data from this, compared to Jeffreys:





This edge is k_1=k_2





1D Gaussian model (from before)

Manifold View

Typical situation is a “hyper-ribbon”
a few long directions, many short directions.

Data: eig($g_{\mu\nu}$) for many systems:

Maximising MI seems to “quantise” the manifold:

Many points along directions $L \gg 1$

Almost ignores directions $L \ll 1$

Repeating m times expands all directions,
so $m \to \infty$ erases this distinction.

$$p^m(\vec{x}|\theta) = \prod_{i=1}^{m} p(x_i|\theta)$$
$$\implies \quad g^m_{\mu\nu} = m\, g_{\mu\nu}$$

But in the real world, data is finite.
(LHC: 10^{18} bits?)

Manifold Boundaries

Restricting to $\partial\Theta$ is a known way of generating simpler models:

- Analytically recover known approximations
- Numerically find few-parameter models (using Fisher metric to find closest boundary…)

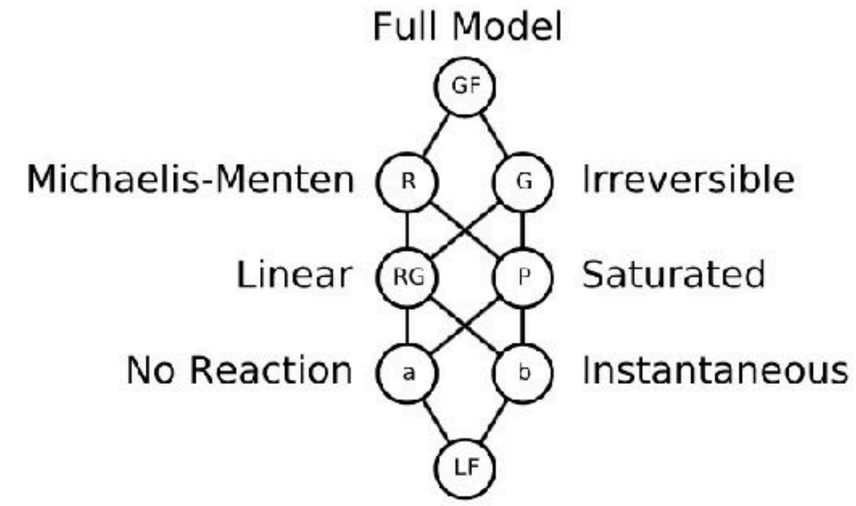
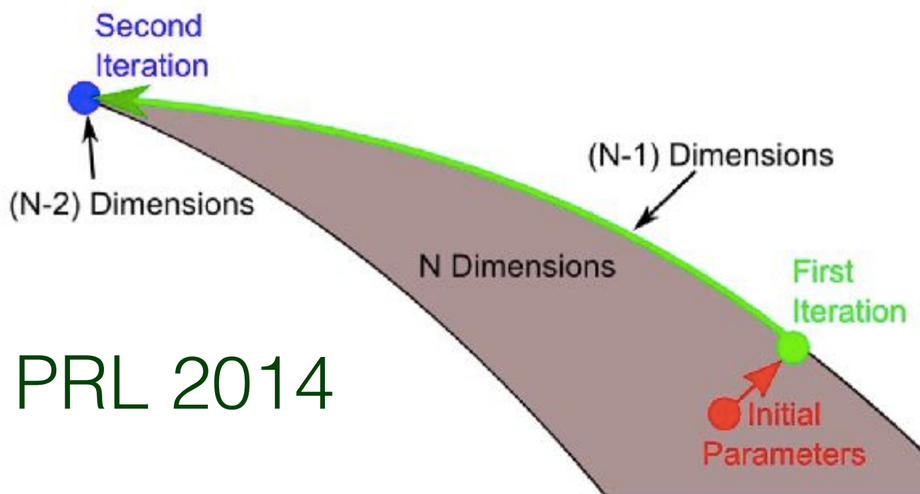
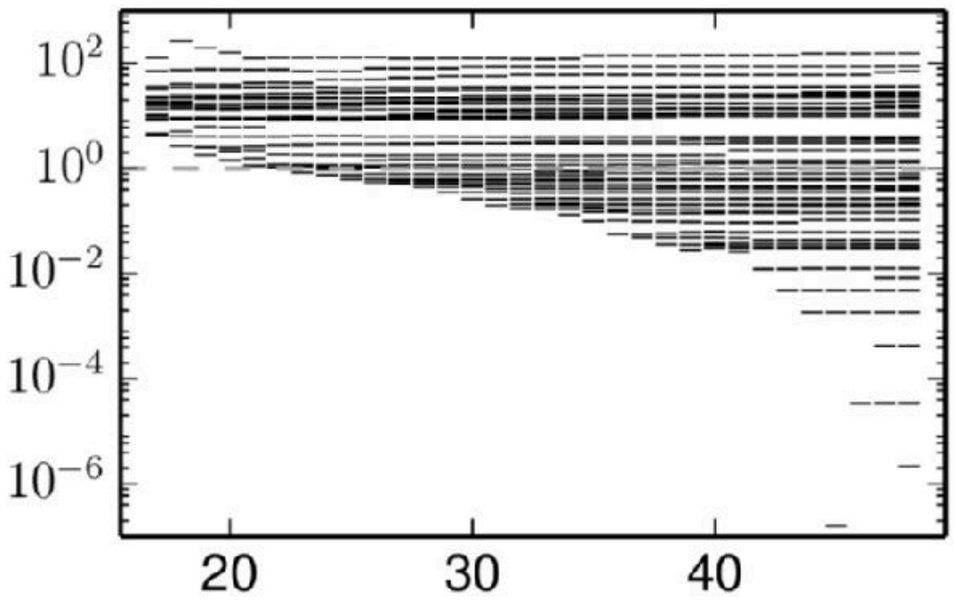


FIG. 16: **Model Hierarchy for coarsened Enzyme-Substrate Model.** The nodes of the Hasse diagram for





Transtrum & Qiu, PRL 2014

important in making model behavior realistic. One appeal of the rational inattention idea (that is, of modeling agents as finite-capacity channels) is that it can, in principle, explain the observed patterns of inertial and random behavior by a mechanism with much fewer free parameters. Another is that it fits well with intuition; most people every day encounter, or could very easily encounter, much more information than is, in principle, relevant to their economic behavior or that they actually respond to. The notion that this is because there are limits to "attention," and that such limits might behave like finite Shannon capacity, is intuitively appealing.

While a traditional optimizing economic agent will respond precisely and continuously to information in the environment (for example, prices), a rationally inattentive agent will respond imprecisely. He will react in discrete jumps to signals that a fully rational agent would respond to continuously; or he will react somewhat randomly. As the capacity constraint slackens, the capacity-constrained agent's behavior approximates that of the fully optimizing agent, but with a tight capacity constraint his behavior will be much more weakly correlated with external information than the behavior of a fully optimizing agent would be.

An optimizing capacity-constrained agent

Economics Literature

They want to explain sticky prices.

In frictionless classical economics, an agent updates its behaviour continuously. This requires infinite bandwidth.

"Rational Inattention" optimises with a finite cost of information:

$$\min_{p(x,y)} \left[\left\langle U(x,y) \right\rangle - \lambda I(X;Y) \right] \qquad \text{with given } p(y) \text{ and } \lambda$$

Sims, 2006
Jung, Kim, Matějka, Sims

Biology Literature

Every signalling pathway is
a communication channel.

Here is one conveying position
along a fly embryo:

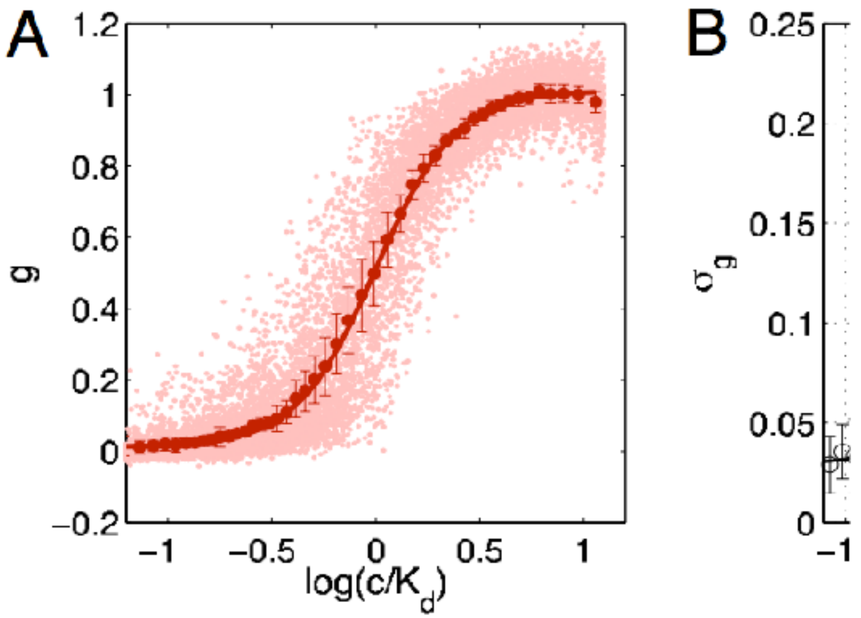


Fig. 2. The Bcd/Hb input/output relationship

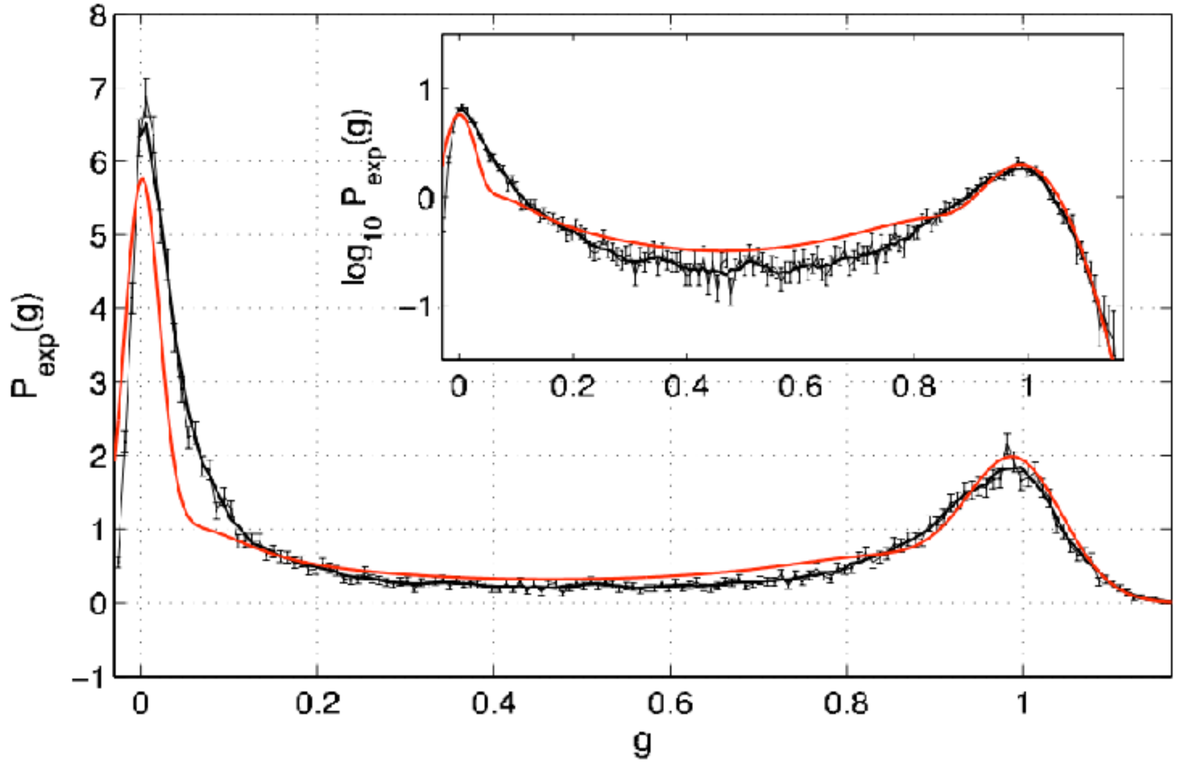


Fig. 4. The measured (black) and optimal (red) distributions of Hunchback expression levels. The measured distribution is estimated from data of ref. 11 by making a histogram of the g values for each data point in Fig. 2. The optimal solution corresponds to the capacity of $I_{opt}(c; g) = 1.7$ bits. The same plot is

Tkačik, Callan & Bialek, 2008

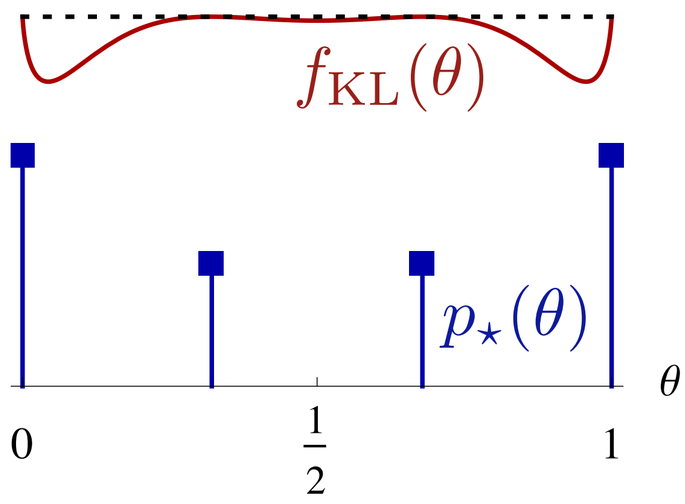
Curve fitting?

Consider $p(\vec{x}|\vec{\theta}) \propto \prod_i e^{-(x_i - g(z_i))/2\sigma^2}$

with noisy measurements of $g(z) = \sum_{n=1}^{D} \theta_n z^n$

For D>2 you need a smarter algorithm...

MCMC using $f_{\mathrm{KL}}(\theta)$? Lafferty & Wasserman 2001, Dauwels 2005



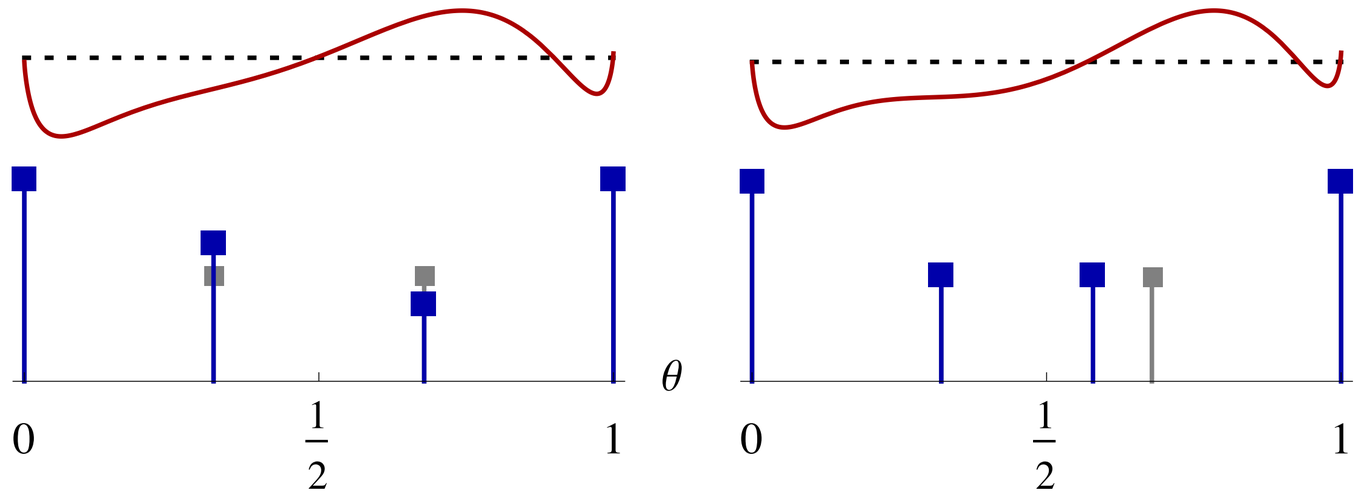
Recall unfair coin:

$f_{\mathrm{KL}}(\theta)$ = MI at atoms, maximum.

Away from optimal prior:

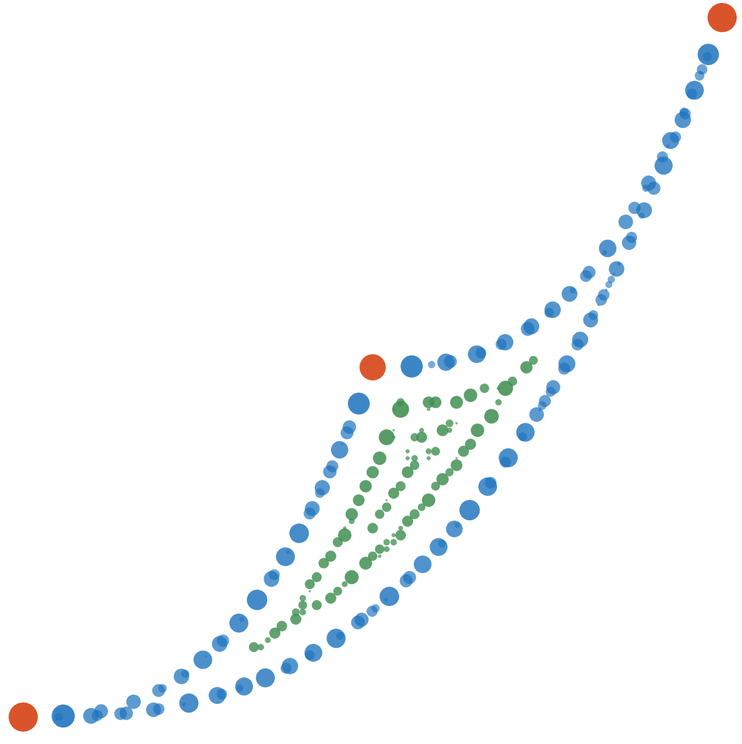
I hear you are interested in functions like

$$g(z) = \sum_n \theta_{2n} \max(z - \theta_{2n+1}, 0)$$



Summary

- Simplicity from irrelevance ≠ avoiding overfitting
- Describe model selection as part of prior choice
- Maximising MI chooses well,
  provided $m < \infty$!
- Along the way, scaling law $\zeta = 3/4$



Papers

- “Rational Ignorance…”, 1705.01166
- Machta et. al. “Parameter Space Compression…” Science, 2013 [1303.6738]

Thank you!