Physics/0009032, Physics/0108025
Physics/0007070, Physics/0103076

October 12, 2001
ITP, UCSB

William Bialek, Oya Nemeshman, Nattali Tishby

Predictability, Complexity and Learning
Possible applications.

Unique complexity measure through predictable information.

Predictive information for different processes.

A note on ensembles.

Why and how to use information theory?

Our objectives.

A curious observation.

Outline
Thus, \( S = 2.95 \) bits.

For this chain, we have:

\[
\log \left( \frac{(4 \mathcal{M})^{N_d}}{0} \right) - = (N)S
\]

Entropy of words in a spin chain.
Entropy is extensive. It shows no distinction between the cases.

\[
\begin{align*}
N & \quad 0 \quad 5 \quad 10 \quad 15 \quad 20 \quad 25 \\
S & \quad 0 \quad 5 \quad 10 \quad 15 \quad 20 \quad 25
\end{align*}
\]

\( J \cdot 10^9 \) spins total.

\[
\begin{align*}
ev \ & 400000 \text{ spins} \\
\text{from } N(0, & 1)
\end{align*}
\]

\[
\begin{align*}
\text{every } 400000 \text{ spins} \\
\text{is taken at random}
\end{align*}
\]

\[
\begin{align*}
\text{from } N(0, & 1)
\end{align*}
\]

\[
\begin{align*}
\text{at random from } N(0, & 1)
\end{align*}
\]

\[
\begin{align*}
\text{if } j & = j' \text{ and } i+1
\end{align*}
\]

\[
\begin{align*}
\text{if } j & = j' \text{ and } i+1
\end{align*}
\]

\[
\begin{align*}
\text{if } j & = j' \text{ and } i+1
\end{align*}
\]

\[
\begin{align*}
\text{Entropy of 3 generated chains}
\end{align*}
\]
Complexity of underlying dynamics intuitively increases from \( \log \) to \( \log \) power.

\[ S^1 = \text{const} + \frac{1}{2} \log N \]

Other examples:

- Shows a qualititative distinction between the cases!
- Subextensive component of the entropy

- Theory of phase transitions may not distinguish between the last two cases.
- Entropy density of channel capacity do not distinguish these cases.

(\text{divergent correlation length})

Possibly: some exotic transitions

DNA sequences,

natural texts,

(\text{divergent correlation length})

or at the onset of chaos

systems at phase transitions,

(\text{finite correlation length})

chaotic sequences,

periodic sequences,
to learn

(relationship)
connections between the two (more rules more difficult

(rules describing dynamics) higher complexity

(definition of dynamics) complexity (more

(noise vs. signal)

[learning]

[unifying description of learning] (metric and algorithm/index

Objectives
Solution – Predictability

neeeded for reliable predictions (learning) more features to describe (complexity) more data.

[relations] particular, regular and random sources have low complexity) easier predictions) are generated by more complex sources (more details to predict, not [complexity] high predictability) sources) in any signal are useless since

[unpredictability] nonpredictive features in any intermediate step parameters is only an intermediate step generalize and predict from training examples’ estimation of [learning] we learn (estimate parameters, extrapolate, classify, …) to
\((\mathcal{L})^0S = (\infty, \mathcal{L})^\text{pred}_I \equiv (\mathcal{L})^\text{pred}_I\)

Predicatability is a deviation from extensivity.

Extensive component cancels in predicative information:

\[(\mathcal{L})^0S + \mathcal{L} \cdot 0S = (\mathcal{L})S\]

\[(\mathcal{L} + \mathcal{L})S - (\mathcal{L}S + (\mathcal{L})S) = \]

\[\langle \left[ \frac{(\text{future})^d_{x=0} f_{x} d}{(\text{future past})} \right] \log \rangle = (\mathcal{L}, \mathcal{L})^\text{pred}_I\]

<table>
<thead>
<tr>
<th>Future</th>
<th>Past</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x)</td>
<td>(\mathcal{L})</td>
</tr>
<tr>
<td>(\text{now})</td>
<td>0</td>
</tr>
</tbody>
</table>

\((\text{non-metric, universal way})\)

Quantifying Predicatability
- coding: coding length
- complexity: complexity measures
- learning: universal learning curves

it relates to and generalizes many relevant quantities

\[ 0 = \frac{(L)_{\text{pred}}}{(L)} \lim_{I} \rightarrow \infty \]

\[ 0 = \frac{L}{(L)_{\text{pred}} I} \lim_{I} \rightarrow \infty \]

\[ 0 \geq (L)_{\text{pred}} I \]

\( L \) properties of
be random, but we do not perceive them in this way.

Example: all pictures can be described using only a few symbols. In this case, the ensemble is not necessary for the description of the picture.

Complexity (learning properties) is an ensemble (averaged) quantity that is possible with atypical data.

Nothing to learn (predict, encode, describe) for only one string.

Grassberger vs. Kolmogorov
The ghost of Bayes

Large $N$ expansion around maximum likelihood value is almost always valid

\[
\frac{(B) \int d(B | X) d + (A) \int d(A | X) d}{(X | A) \partial B(A | X) d \int (A) \int d} = \frac{(X) d}{(A) \int d(A | X) d} = (X | A) d
\]
learning continuous densities

Learning more features as $N$ grows

$\lim_{N \to \infty} \frac{\text{pred}\_I}{N} \approx \text{const} \times \log_2 N$

precise learning of a fixed set of

fully stochastic (Markov) processes

simply predictable (periodic, constant, etc.) processes

$\lim_{N \to \infty} \text{pred}\_I \approx \text{const}$

not well studied

Learning continuous densities

How can $\text{pred}\_I$ behave?
\[\begin{align*}
\log_2 (\{ x \})_D \sum_{I=1}^N \log_2 (\{ x \})_D N_{x'} \ldots \cdot x' \int - = (N)S \equiv (N_{x'} \ldots x'_I, x'_I)P \\
(x|!x) \otimes \prod_N (x)_D x_P \int = (N_{x'} \ldots x'_I, x'_I)P \\
(x|!x) \otimes \prod_N = (x|N_{x'} \ldots x'_I, x'_I)_P \\
\text{random samples from the distribution} \quad \text{prior distribution or parameters} \\
\text{dimensionality of } a \text{ may be infinite} \\
\text{unknown parameters } a \\
\text{probability density function for } x \text{ parameterized by } a \\
\text{specific examples: problem setup}
\end{align*}\]
\[ \left[ \mathcal{N}_{3N} - \mathfrak{E}(\varnothing) \mathbb{D} \chi \mathbb{P} \int \right] \mathcal{N} \left[ (\varnothing) \mathbb{D} \chi \mathbb{P} \int - = (N) S \right] \]

This separates into the extensive and the subextensive terms.

\[ \left\{ \left[ (\varnothing | x) \mathbb{D} \mathcal{N} \times \sum_{N} (\varnothing) \mathbb{D} \chi \mathbb{P} \int (\varnothing | x) \mathbb{D} \mathcal{N} \times \sum_{N} (\varnothing) \mathbb{D} \chi \mathbb{P} \int \right] \right\} (\varnothing) \mathbb{D} \chi \mathbb{P} \int - = (N) S \]
\[
\sum_{N} \left( \langle \Psi | H | \Psi \rangle - \varepsilon_N \langle \Psi | \Phi \rangle \right) = \left[ \sum_{N} \log_2 \left( \frac{\langle \Psi | \Phi \rangle}{\langle \Psi \Phi | \Phi \rangle} \right) \right] - \left[ \sum_{N} \frac{1}{N} \right] \equiv \langle \Psi | \Phi \rangle \tag{15}
\]

**Annealed approximation**

Under some (known) conditions we may have
Thus learning is annealing at decreasing temperature; the density could be very different for different targets:

\[
1 = (x) d^x K p \int = (x : a)^d dp \int \\
[(x | x) K \Lambda a - a] g(x) d^x K p \int = (x : a)^d \\
[a N - ] e x p (x : a)^d dp \int = (N : x) Z
\]

We can rewrite the partition function as:

Density of states
Caution: Speed of approach to this asymptotics is rarely investigated.

\[ N \frac{\log{\frac{x}{p}}}{p} \approx (\text{a})^S \]

\[
\frac{z/(z-p)A(z-I)}{z/|z|} \frac{(\textbf{x}^T \textbf{X})^{-1}}{|x|} (\textbf{v})_d \approx (\textbf{v} : 0 \leftarrow A)^d
\]

\[
\cdots + (v_u - v_u)^d \sum_{i=1}^{\infty} z_i \approx (\textbf{v} || \textbf{v})_{KL}
\]

Example: sound finite parameter models, \( \dim \textbf{a} = p \). Then for \( a \approx 0 \)

\[ \frac{z/(z-p)A(\textbf{v})V}{z/(z-2)^2} \approx (\textbf{v} : 0 \leftarrow A)^d \]

Power-law density function
relationships using order parameters. This is similar to describing physical systems with cor
relations and from intrinsic long-range correlations coming

\[ N \log_2 \frac{2}{\lambda + \lambda} = (N)^T_S \]

\[ N \log_2 \frac{2}{\lambda} = 0 \quad \Rightarrow \ \lambda + 0 SN \leftarrow \]

\[ (\omega|\{x\}) \log_2 (\omega|\{x\}) \bar{X}_NP \int - \equiv [\omega|\{x\}] S \]

Another example: Learning Markov pro

cess with long range intrinsic correlations such that
order when $T_\text{exec}$ becomes extensive
as complexity grows and then vanishes to the leading

\[ N \log \frac{2}{(1+\eta)/\eta N} \sim (N)^{\Theta T_\text{exec}} \not\sim \frac{(1+\eta)/\eta N}{(1+\eta)/\eta [A]} = (\mathcal{O})C \]

As $p \to 0$ we may imagine the following behavior

\[ \text{essential singularity in the density function} \]
crossover from complexity to randomness demonstrates a

heuristic argument for the dimensionality and the smooth-

techniques exponent given by \( \mu \), and the smooth-

\[
(x)_{0}^{\mathcal{N}} \sim (N)_{I}^{1}
\]

time

increasing number of effective parameters (bins) of adaptive

\[
\left( \frac{1}{N} \right) \frac{\mu_{2} \mu_{2}}{N^{1/2}} \approx (N)_{(I)}^{1}
\]

\[
\left( \frac{D}{[\langle x \lambda \rangle]_{d}} \right)^{d} \int_{-\varepsilon}^{\varepsilon} d[(x)_{0}] \mathcal{V} = (\phi_{0}^{+} \leftarrow d)^{d}
\]

\[
\left[ 1 - (x)_{0} \int \frac{0_{1}}{1} \right] \int \left( \frac{\phi_{0}}{\phi_{0}} \right) \frac{x_{1}}{l} \frac{Z_{1}}{1} \int_{-\varepsilon}^{\varepsilon} d[(x)_{0}] = \quad [(x)_{0}]_{d}
\]

Callian, and Strong 1996.

Example of the power-law I

Learning a nonparametric (infinite parameter) density

\( (x)_{0}^{\mathcal{N}} \sim (N)_{I}^{1} \)

\( \mu_{2} \mu_{2} \)
Which complexity do we want to define?

- must be attached to an ensemble, not a single realization
- expressible in conventional physical terms
- usable for Occam–style punishment in statistical inference
- zero for totally random and for easily predictable processes
- computationally or descriptive complexity; thus it must be
- complexity of dynamics that generates a time series (not
Complementarity uniquely
The divergent superextensive term measures

\[ \log p_1(x) = \log p_2(x) + \text{loc.} \text{ oper.} \]

This may present a problem in higher dimensions.

same book in different languages – same universality class
Invariant under invertible temporarily locally transformations •
(tonicity, continuity, additivity)

some kind of entropy (we proclaim Shannon’s postulates: mono-

Complexity measure
Result: If sufficient statistics exist, then $C_k \approx I_{\text{pred}}$. Otherwise, futures equivalence = indistinguishable conditional distributions of that can generate a sequence from the class $s$ belongs to respect to the partition as a length of the shortest program.

- Define Kolmogorov complexity $C_k(s)$ of a sequence $s$ with partition all strings into equivalence classes.
- For Kolmogorov complexity:

... are mostly straightforward.

... Relations to other definitions...
predictive information is a property of the data, not of the
natural language?

reflections to statistics — nonparametric extensions of
MDL

are the models we use in learning?

tion a guiding principle for animal behavior? how complex
reflections to biology — is predictive information maximiza-

reflections to subextensive statistical mechanics

reflections to physics — analysis of phase transitions, conne-

the relevant information, technique

separating predictive information from non-predictive using

What's next?