Complexity Through Nonextensivity

William Bialek[1], Ilya Nemenman[1], and Naftali Tishby[1,2]

[1]NEC Research Institute, [2]Hebrew University

Presented at the International Workshop on

Frontiers in The Physics of Complex Systems

March 25 - 28, 2001

Dead Sea, Israel

`http://arXiv.org/abs/physics/0007070`

Complexities

- descriptive complexity of single strings – computer science (Kolmogorov complexity, MDL, ...)
- complexity of dynamics (process) – dynamical systems theory (Lyapunov exponents, various entropies, ...)
- complexity of models – learning and statistical inference (Occam factors, MDL, MML, ...)
- complexity (time or space) of problems – computer science

The first three are all *descriptive* complexities, having similar usages, pluses and minuses. One needs a *generalizing* definition.

Descriptive complexities

Usual problems:

What We Want	Problem
complexity $\neq$ randomness	description length $\approx$ entropy $=$ randomness
complexity of dynamics $\approx$ complexity of its output	there can be atypical strings

Intuition: Complexity of a random source and very regular source is low; entropy of their outputs is different. But corrections to the extensivity of the (averaged) entropy are small for both.

Solution (Grassberger 86): We should average over all possible outcomes and focus on subextensive components of entropies!

Different reasoning

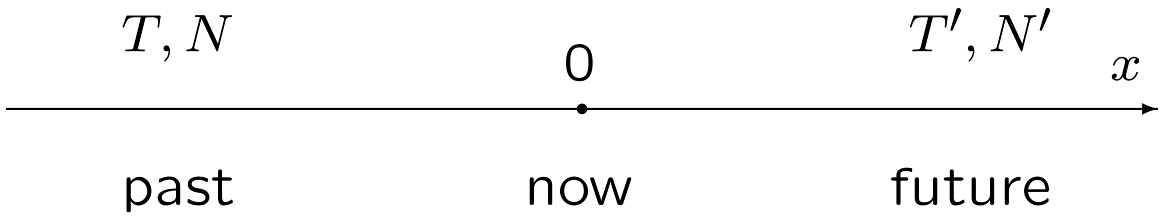
Predicting the future of a sequence:

- we learn (estimate parameters, extrapolate, classify, ...) to *generalize* and *predict* from training examples; estimation of parameters is only an intermediate step
- nonpredictive features in any signal are useless since we observe *now* and react in the *future*
- more features to predict is a problem of higher complexity

Footnote: there's little to predict for both regular and random sequences.

Intuition: only predictive features of signals should be coded; only they are of interest when defining complexity.

Bringing two reasonings together



$$
\begin{aligned}
\mathcal{I}_{\text{pred}}(T, T') &= \left\langle \log_2 \left[\frac{P(x_{\text{future}} | x_{\text{past}})}{P(x_{\text{future}})} \right] \right\rangle \\
&= S(T) + S(T') - S(T + T') \\
S(T) &= \mathcal{S}_0 \cdot T + S_1(T)
\end{aligned}
$$

Thus extensive component cancels in predictive information.

Predictability is nonextensivity!

$$
I_{\text{pred}}(T) \equiv \mathcal{I}_{\text{pred}}(T, \infty) = S_1(T)
$$

Properties of $I_{\text{pred}}(T)$

- $I_{\text{pred}}(T)$ is information, so $I_{\text{pred}}(T) \geq 0$
- $I_{\text{pred}}(T)$ is subextensive, $\lim_{T \to \infty} \frac{I_{\text{pred}}(T)}{T} = 0$
- diminishing returns, $\lim_{T \to \infty} \frac{I_{\text{pred}}(T)}{S(T)} = 0$
- it relates to and generalizes many relevant quantities
  - learning: universal learning curves
  - complexity: complexity measures
  - coding: coding length

How can I_{pred} behave?

$\lim_{N\to\infty} I_{\text{pred}} = \text{const}$ no long-range structure

- simply predictable (periodic, constant, etc.) processes
- fully stochastic (Markov) processes

$\lim_{N\to\infty} I_{\text{pred}} = \text{const} \times \log_2 N$ precise learning of a fixed set of parameters

- learning finite-parameter densities (functions)
- dynamics with divergent correlation times
- analyzed as $I(N, \text{parameters}) = I_{\text{pred}}(N)$

$\lim_{N\to\infty} I_{\text{pred}} = \text{const} \times N^{\xi}$ $0 < \xi < 1$ learning more features as N grows

- learning nonparametric densities (functions) with smoothness constraints
- some cellular automata
- natural languages
- not well studied

Density of states

For a stochastic process described by an unknown model $\bar{\boldsymbol{\alpha}}$ taken at random from $\mathcal{P}(\boldsymbol{\alpha})$ the randomness (disorder) due to $\vec{x}_i$ is often unimportant and behavior of S_1 is governed to the leading order only by the model family properties:

$$\begin{aligned}
S_1(N) &= \left\langle \log \int dD \rho(D; \bar{\boldsymbol{\alpha}}) \exp[-ND] \right\rangle_{\bar{\boldsymbol{\alpha}}} + O(N^0) \\
\rho(D; \bar{\boldsymbol{\alpha}}) &= \int d^K \alpha \mathcal{P}(\boldsymbol{\alpha}) \delta[D - D_{\mathsf{KL}}(\bar{\boldsymbol{\alpha}} || \boldsymbol{\alpha})] \\
D_{\mathsf{KL}}(\bar{\boldsymbol{\alpha}} || \boldsymbol{\alpha}) &= \int d\vec{x} Q(\vec{x} | \bar{\boldsymbol{\alpha}}) \log \frac{Q(\vec{x} | \bar{\boldsymbol{\alpha}})}{Q(\vec{x} | \boldsymbol{\alpha})}
\end{aligned}$$

Then predictive properties depend on $D \to 0$ behavior of the density.

Power–law density function

The exponent is equivalent to the dimensionality in statistical systems.

$$\begin{aligned} \rho(D \to 0; \bar{\boldsymbol{\alpha}}) &\approx A(\bar{\boldsymbol{\alpha}}) D^{(d-2)/2} \quad \Rightarrow \\ S_1 &\approx \frac{d}{2} \log_2 N \end{aligned}$$

- well studied case;
- happens for most finite parameter models (including Markov chains) in learning, phase transitions, dynamical systems at the onset of chaos;
- speed of approach to this asymptotics is rarely investigated.

Essential zero in the density function

As $d \to \infty$ we may imagine the following behavior

$$\begin{aligned} \rho(D \to 0; \bar{\boldsymbol{\alpha}}) &\approx A(\bar{\boldsymbol{\alpha}}) \exp\left[-\frac{B(\bar{\boldsymbol{\alpha}})}{D^\mu}\right], \quad \mu > 0 \quad \Rightarrow \\ S_1(N) &\sim N^{\mu/(\mu+1)} \end{aligned}$$

- not well studied case;
- as $\mu \to \infty$, $S_1(N)$ grows and then vanishes to the leading order when it becomes extensive;
- observed when longer sequences allow progressively more detailed description of the underlying dynamics (natural languages, some dynamical systems, nonparametric learning, finite parameter learning models with increasing number of parameters $K \sim N^{\mu/(\mu+1)}$.

I_{pred} as a unique measure of complexity

Complexity measure must be:

- some kind of entropy (we proclaim Shannon's postulates):
  - monotonic in N for N equally likely signals,
  - additive for statistically independent signals,
  - a weighted sum of measure at branching points if measuring a leaf on a tree;
- reparameterization, quantization invariant $\Rightarrow$ subextensive;
- insensitive to invertible temporally local transformations (e. g., $x_k \to x_k + \xi x_{k-1}$—measuring device with inertia);

The divergent subextensive term measures complexity uniquely!

Relations to other definitions ...

... are mostly straightforward.

For Kolmogorov complexity:

- partition all strings into equivalence classes;
- define Kolmogorov complexity $C_E(s)$ of a sequence s with respect to the partition as a length of the shortest program that can generate a sequence from the class s belongs to;
- equivalence = indistinguishable conditional distributions of futures;

Result: If sufficient statistics exist, then $C_E \approx I_{\text{pred}}$. Otherwise $C_E > I_{\text{pred}}$. (Relates to TMC and *statistical complexity*). C_E is unique up to a constant.

What's next?

- separating predictive information from non–predictive using the 'relevant information' technique;
- reflection to physics — finding order parameters for phase transitions using behavior of the predictive information;
- reflection to biology — is predictive information maximization a guiding principle for animal behavior? how complex are the models we use in learning?
- reflection to dynamical systems theory — what is the predictive information and complexity of various systems? of natural languages?
- reflection to statistics — nonparametric extensions of MDL (predictive information *is* a property of the data, not of the model [*N,B NIPS-2000*]).