Complex Systems Methods — 3. Statistical complexity of temporal sequences

Eckehard Olbrich

MPI MiS Leipzig

Potsdam WS 2007/08

Overview

1. Summary Kolmogorov sufficient statistic
2. Intuitive notions of complexity
3. Statistical complexity
   - Entropy convergence and excess entropy
   - Predicitive information
4. Entropy estimation
   - Entropy of a discrete random variable — finite sample corrections

Summary Kolmogorov sufficient statistic

- Kolmogorv complexity

$$K_{\mathcal{U}}(x|l(x)) = \min_{p:\mathcal{U}(p,l(x))=x} l(p)$$

- Kolmogorov structure function, x^n denotes a string of length n

$$K_k(x^n|n) = \min_{\substack{p\,:\,l(p)\leq k \\ \mathcal{U}(p,n)=S \\ x^n\in S\subseteq\{0,1\}^n}} \log|S|$$

- Kolmogorv sufficient statistic: least k such that

$$K_k(x^n|n) + k \leq K(x^n|n) + c \ .$$

- Regularities in x are desribed by describing the **set** S. Given S the string x is random,
- Algorithmic complexity $\equiv$ randomness. But intuitively, complexity measure should quantify structure, not randomness. Thus it is related to an ensemble and not a single object.

Dynamical Systems

all continuous	Partial Differential Equations (PDE's)
discetized space	coupled ordinary differential equations (ODE's)
discretized time	coupled map lattice (CML)
discretized state	cellular automata (CA)

- all dynamical systems either deterministic or stochastic
- Digital computer: Finite state automaton - finite number of discrete states

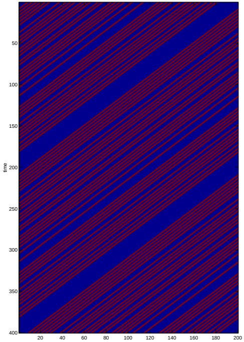
Intuitive notions of complexity — Stephen Wolframs classification of cellular automata

Elementary cellular automata: binary states $\{0, 1\}$. Next neighbour interaction. Rules can be coded by numbers $0, \ldots, 255$:

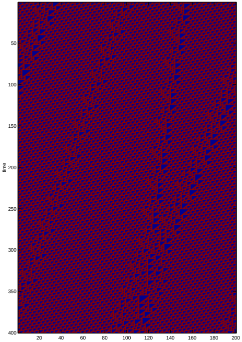
111	110	101	100	011	010	001	000	
0	0	0	1	1	1	1	0	e.g. Rule 30.

- Class 1: Evolution leads to a homogeneous state.
- Class 2: Evolution leads to a set of separated simple stable or periodic structures.
- Class 3: Evolution leads to a chaotic pattern.
- Class 4: Evolution leads to complex localized structures, sometimes long-lived.

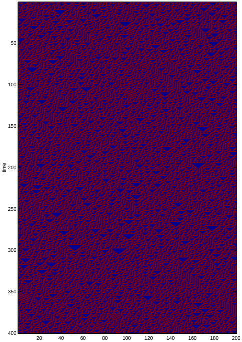
Intuitive notions of complexity — Stephen Wolframs classification of cellular automata



class 2



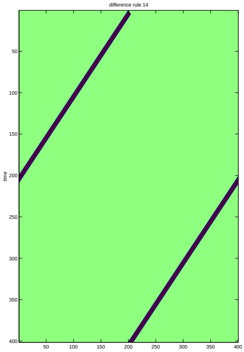
class 4



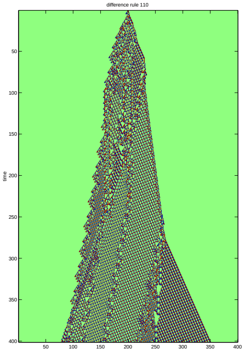
class 3

Intuitive notions of complexity — Stephen Wolframs classification of cellular automata

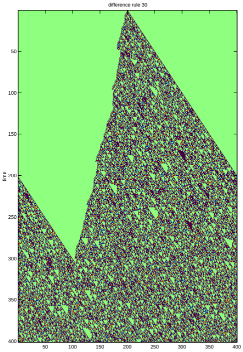
Propagation of a single perturbation



class 2
ordered



class 4
“complex”



class 3
chaotic

Statistical complexity — excess entropy

- $K(x|l(x))$ — minimal length of a program, which produces exactly the string x. Most complex strings are algorithmically random.
- Kolmogorov sufficient statistic: program that describes all regularities in x and computes the set S, which containes all strings with the same regularities as in x, but are otherwise random. Algorithmic complexity can be divided in two parts - regularities and randomness.
- Same idea for a time series (infinite sequence) provided with a stationary distribution:
  - Randomness per symbol is given by the entropy rate

$$h_\infty = \lim_{n\to\infty} h_n \qquad h_n = H(X_0|X_{-1},\ldots,X_{-n})$$

  - Regularities quantified by the excess entropy (Crutchfield) or effective measure complexity (Grassberger)

$$E = \lim_{n\to\infty} E_n \quad \text{with} \quad E_n = (H_n - n \cdot h_n) \qquad \text{and} \quad H_n = H(X_n,\ldots,X_1)$$

The idea which led originally to this complexity measure was to use the converge rate of the conditional entropy to quantify the complexity of a time series:
fast convergence → short memory → low complexity
slow convergence → long memory → high complexity.
Conditional entropies

$$h_n = H(X_0|X_{-1}, \dots, X_{-n}) = H(X_0, X_{-1}, \dots, X_{-n}) - H(X_{-1}, \dots, X_{-n})$$

decrease monotonically. Decay is quantified by the conditional mutual information

$$\delta h_n := h_{n-1} - h_n = MI(X_0 : X_{-n}|X_{-1}, \dots, X_{-n}) .$$

Entropy convergence and excess entropy

Excess entropy:

$$\begin{aligned} E_N &= H(X_1, \dots, X_N) - N h_N \\ &= \sum_{n=0}^{N-1} (h_n - h_N) \qquad \text{using} \quad h_0 = H(X_1) \\ &= \sum_{n=0}^{N-1} \sum_{k=n+1}^{N} \delta h_k \qquad \text{using} \quad h_{n-1} = \delta h_n + h_n \\ &= \sum_{k=1}^{N} k \delta h_k \end{aligned}$$

The slower the convergence of the entropy the larger the excess entropy. If the sequence is Markov' of order m: $\delta h_n = 0$ for $n > m$. Thus $E = E_m$.

Predictive information

Bialek et al. (2000) proposed to quantify the complexity of a time series by the amount of information which the past tells us about the future

$$I_{pred} = MI(\text{past} : \text{future}) = \lim_{n_p, n_f \to \infty} MI(X_{-n_p}, \ldots, X_0 : X_1, \ldots, X_{n_f})$$

and called this *predictive information*.
The predictive information is equal to the excess entropy if the corresponding limits exist:

$$I_{pred} = E$$

Predictive information

$$
\begin{aligned}
I_{pred} &= \lim_{n_p, n_f \to \infty} MI(X_{-n_p}, \dots, X_{-1}, X_0 : X_1, X_2, \dots, X_{n_f}) \\
&= \lim_{n_p, n_f \to \infty} \{ H(X1, \dots, x_{n_f}) + H(X_{-n_p}, \dots, X_0) \\
&\quad - H(X_{-n_p}, \dots, X_0, \dots, X_{n_p}) \} \\
&= \lim_{n_p, n_f \to \infty} \{ E_{n_f} + n_f h_{n_f} + E_{n_p+1} + (n_p + 1) h_{n_p+1} \\
&\quad - E_{n_p+n_f+1} - (n_f + n_p + 1) h_{n_f+n_p+1} \} \\
&= E
\end{aligned}
$$

The excess entropy measures the amount of information which is available from the past for predicting the time series.

Causal states

Causal equivalence: Two past sequences $x_{-\infty}^{0} = x_0, x_{-1}, \ldots$ and $x'^{0}_{-\infty} = x'_0, x'_{-1}, \ldots$ are causal equivalent, if they have the same future, i.e. $p(x_1^\infty | x_{-\infty}^0) = p(x_1^\infty | x'^{0}_{-\infty})$. The equivalence classes are called *causal states*. This notion was introduced by James P. Crutchfield et al. They called the transition graph between the causal states an ϵ-*machine*. In general it seems to be a subclass of hidden Markov models.
Crutchfield and Young (1989): Statistical complexity C_μ as the entropy of the stationary distribution over the states. In general there is

$$C_\mu \geq E \; .$$

Grassberger (1986) called it *set complexity* in the context of regular languages.
In general: The *excess entropy* provides a **lower bound** for the amount of information necessary for an **optimal prediction**.

Excess entropy — some examples

- Sequence with period n: $h_m = 0$ for $m \geq n$, in particular $h_\infty = 0$. The entropy $H_{m \geq n} = \log n$ remains constant for $m \geq n$, thus $E = \log n$. All periodic sequences of the same length have the same complexity according to this measure. (Alternative: transient information introduced by Crutchfield and Feldman).
- Markov chain: $h_\infty = h_1$. $E = \delta h_1 = h_0 - h_1 = MI(X_{n+1} : X_n)$.
- Chaotic maps: Usually exponential decay of h_n — finite E.
- Feigenbaum point: $h_\infty = 0$, but $h_n \propto 1/n$, thus E diverges.

Complexity for the period doubling route to chaos

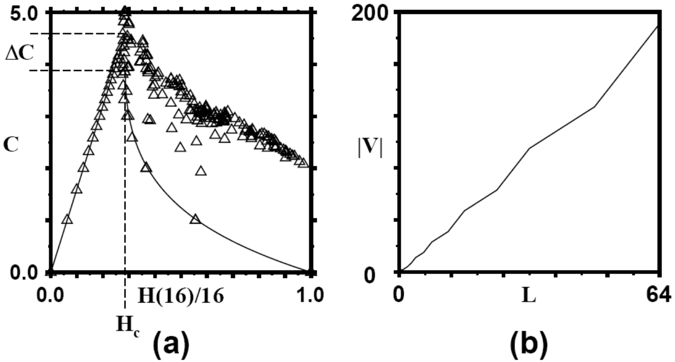


Figure 6 (a) Statistical complexity C_μ versus specific entropy $H(L)/L$ for the period-doubling route to chaos. Triangles denote estimated $(C_\mu, H(L)/L)$ at 193 values of the logistic map nonlinearity parameter. ϵ-machines were reconstructed using a subsequence length of $L = 16$. The heavy solid lines overlaying some of this empirical data are the analytical curves derived for C_0 versus $H_0(L)/L$. (After [24].) (b) At one of the critical parameter values of the period-doubling cascade in the logistic map the number $\|\mathbf{V}\|$ of inferred states grows without bound. Here $r = r_c \approx 3.5699456718695445\ldots$ and the sequence length ranges up to $L = 64$ where $\|\mathbf{V}\| = 196$ states are found. It can be shown, and can be inferred from the figure, that the per symbol density of states $\|\mathbf{V}(L)\|/L$ does not have a limiting value as $L \to \infty$. (After [56].)

From Crutchfield/Young 1990. Computation at the onset of chaos.

Excess entropy and attractor dimension

Deterministic system with continous state observables: The entropy $H_m(\epsilon) = H(X_1, \ldots, X_m; \epsilon)$ scales with respect to the resolution ϵ:

m $<$ D

$$H_m(\epsilon) = H_m^c - m \log \epsilon + \mathcal{O}(\epsilon)$$

m $>$ D

$$H_m(\epsilon) = const - D \log \epsilon + \mathcal{O}(\epsilon)$$

Because $h_\infty = h_{KS}$ does not depend on ϵ we have

$$E \propto -D \log \epsilon \ .$$

The better one knows the initial conditions of the system the better the system can be predicted.

References

- Peter Grassberger, Toward a quantitative theory of self-generated complexity, International Journal of Theoretical Physics, 25 (1986), 907-938.
- James P. Crutchfield and David P. Feldman, Regularities unseen, randomness observed: Levels of entropy convergence Chaos, 13 (2003), 25-54.
- William Bialek, Ilya Nemenman and Naftali Tishby, Predictability, complexity, and learning. Neural Computation, 13 (2001), 2409-2463.
- Remo Badii and Antonio Politi, Complexity — Hierarchical structures and scaling in physics, Cambridge University Press, 1997.

Entropy of a discrete randomm variable — finite sample corrections

Reference: P. Grassberger, Entropy Estimates from Insufficient Samplings, arXiv:physics/0307138v1

- N data points randomly and *independently* distributed on M boxes. The number n_i of points in each box is a random variable with an expectation value $z_i := E[n_i] = p_i N$. Their distribution is binomial

$$P(n_i; p_i, N) = \binom{N}{n_i} p_i^{n_i} (1 - p_i)^{N - n_i}$$

- For $p_i \ll 1 \quad \forall i$ the n_i can be assumed to be Poisson distributed

$$P(n_i; z_i) = \frac{z_i^{n_i}}{n_i!} e^{-z_i}$$

Entropy of a discrete random variable — finite sample corrections

- Entropy

$$H = -\sum_{i=1}^{M} p_i \log p_i = \ln N - \frac{1}{N}\sum_{i=1}^{M} z_i \ln z_i$$

- Naive estimator

$$\hat{H}_{naive} = \ln N - \frac{1}{N}\sum_{i=1}^{M} n_i \ln n_i$$

- In general the estimator is biased, i.e.

$$\Delta H := E[\hat{H}] - H \neq 0 \,. \qquad \Delta H_{naive} < 0 \,.$$

Entropy of a discrete random variable — finite sample corrections

- In the limit of large N and M each contribution $z_i \ln z_i$ will be statistically independent and can be estimated as function of n_i:

$$z_i \ln z_i \approx z_i \hat{\ln} z_i = n_i \phi(n_i) \qquad E[z_i \hat{\ln} z_i] = \sum_{n_i=1}^{\infty} n_i \phi(n_i) P(n_i; z_i) \; .$$

  Implizit assumption: $n_i = 0$ gives no information about p_i.

- Estimator

$$\hat{H}_\phi = \ln N - \frac{1}{N} \sum_{i=1}^{M} n_i \phi(n_i)$$

Finite sample corrections — Grassbergers Result 1

- Grassbergers result:

$$E[n\psi(n)] = z \ln z + zE_1(z)$$

with the digamma function

$$\psi(x) = \frac{d \ln \Gamma(x)}{dx} \quad \text{and} \quad E_1 = \Gamma(0, x) = \int_1^\infty \frac{e^{-xt}}{t} dt \ .$$

- For large z $zE_1(z) \approx e^{-z}$, thus neglecting this term gives the estimator

$$\hat{H}_\psi = \ln N - \frac{1}{N} \sum_{i=1}^{M} n_i \psi(n_i) \ .$$

Finite sample corrections — Comparison with other results

- Approximations for the digamma function leads to known estimators:
  1. $\psi(x) \approx \ln x$: naive estimator
  2. $\psi(x) \approx \ln x - 1/(2x)$: Miller correction
  3. $\psi(x) < \ln x - 1/(2x) < \ln x$ leads to $\hat{H}_\psi > \hat{H}_{Miller} > \hat{H}_{naive}$
- Grassbergers best estimator:

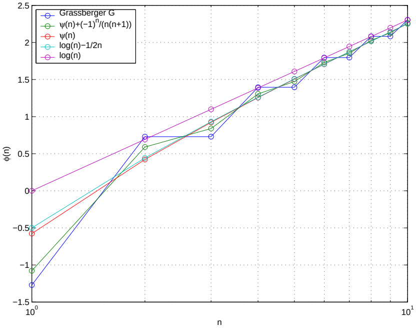
$$\hat{H}_G = \ln N - \frac{1}{N} \sum_{i=1}^{M} n_i G_{n_i}$$

with

$$G_n = \psi(n) + (-1)^n \sum_{k=1}^{\infty} \frac{1}{(n+2k)(n+2k+1)}$$

or $G_1 = -\gamma - \ln 2 \quad G_2 = 2 - \gamma - \ln 2 \quad G_{2n+1} = G_{2n}$ and $G_{2n+2} = G_{2n} + \frac{2}{2n+1}$ for $n \geq 1$.

Finite sample corrections — Comparison between different corrections

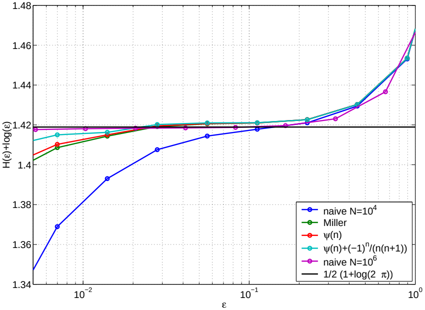


Note that $\psi(x) + \frac{(-1)^n}{n(n+1)}$ corresponds to an approximation of

$$G_n = \psi(n) + (-1)^n \sum_{k=1}^{\infty} \frac{1}{(n+2k)(n+2k+1)}$$

with considering only the $k = 0$ term in the sum.

Test — Entropy of a Gaussian distribution



Differential entropy of a Gaussian distribution

$$H^C_{Gauss} = \frac{1}{2}(1 + \log(2\pi\sigma^2)) = H(\epsilon) + log(\epsilon) + \mathcal{O}(\epsilon)$$