Predictability, Complexity and Learning

Ilya Nemenman
KITP, UCSB

Joint work with:
William Bialek (Princeton University)
Naftali Tishby (Hebrew University)

`physics/0007070, physics/0103076`

Outline

Outline

- A curious observation.

Outline

- A curious observation.
- Our objectives.

Outline

- A curious observation.
- Our objectives.
- Why and how to use information theory?

Outline

- A curious observation.
- Our objectives.
- Why and how to use information theory?
- A note on ensembles.

Outline

- A curious observation.
- Our objectives.
- Why and how to use information theory?
- A note on ensembles.
- Predictive information for different processes.

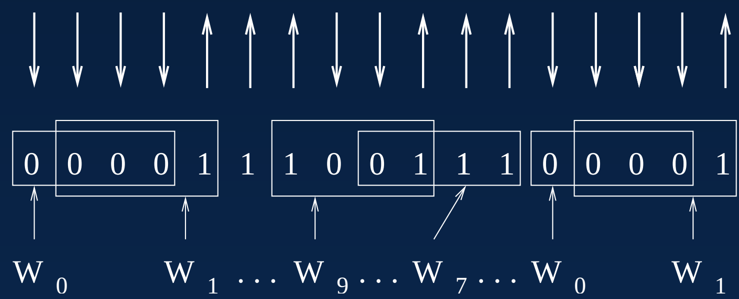
Outline

- A curious observation.
- Our objectives.
- Why and how to use information theory?
- A note on ensembles.
- Predictive information for different processes.
- Unique complexity measure through predictive information.

Outline

- A curious observation.
- Our objectives.
- Why and how to use information theory?
- A note on ensembles.
- Predictive information for different processes.
- Unique complexity measure through predictive information.
- Possible applications.

Entropy of words in a spin chain



$W_0 = 0\ 0\ 0$

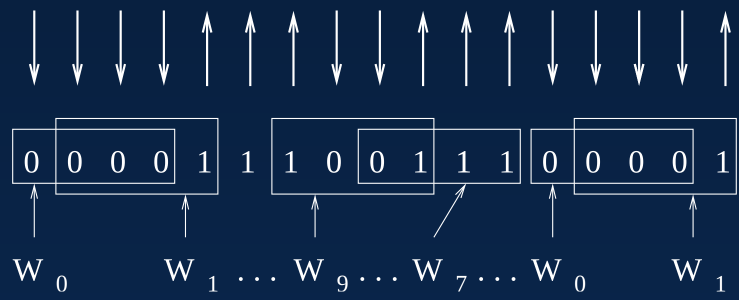
$W_1 = 0\ 0\ 0$

$W_2 = 0\ 0\ 1$

$\cdots$

$W_{15} = 1\ 1\ 1$

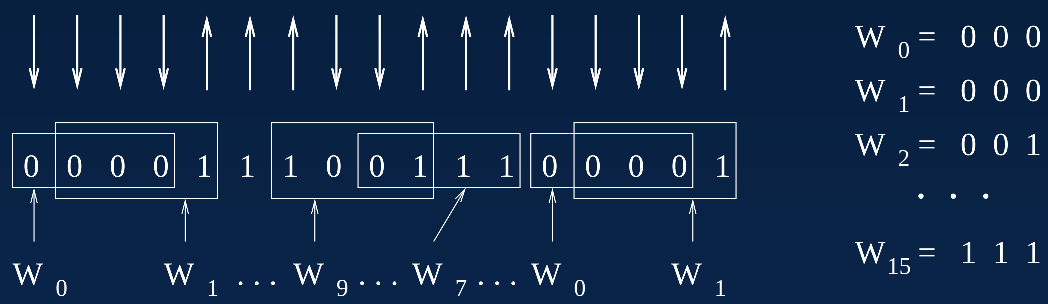
Entropy of words in a spin chain



$W_0 =$ 0 0 0
$W_1 =$ 0 0 0
$W_2 =$ 0 0 1
. . .
$W_{15} =$ 1 1 1

$$S(N) = -\sum_{k=0}^{2^N-1} P_N(W_k)\log_2 P_N(W_k)$$

Entropy of words in a spin chain



$$S(N) = -\sum_{k=0}^{2^N-1} P_N(W_k)\log_2 P_N(W_k)$$

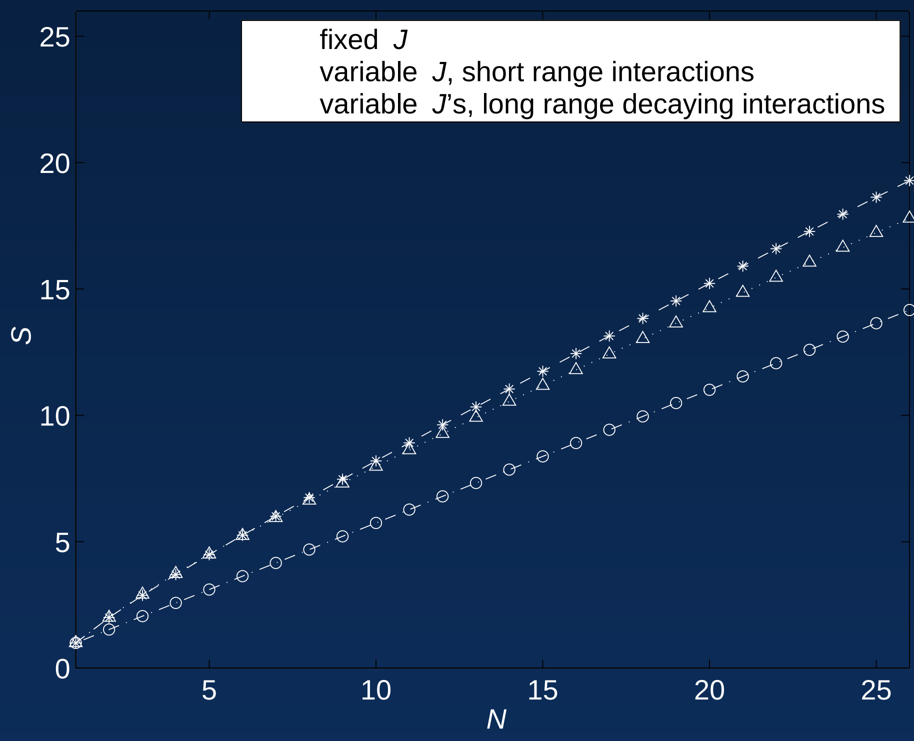
For this chain,

$P(W_0) = P(W_1) = P(W_3) = P(W_7) = P(W_{12}) = P(W_{14}) = 2$, $P(W_8) = P(W_9) = 1$, and all other frequencies (probabilities) are zero. Thus, $S(4) \approx 2.95$ bits.

Entropy of 3 generated chains

- $J_{\mathrm{ij}} = \delta_{\mathrm{i,j+1}}$
- $J_{\mathrm{ij}} = J_0\,\delta_{\mathrm{i,j+1}}$, J_0 is taken at random from $\mathcal{N}(0,1)$ every 400000 spins
- J_{ij} is taken at random from $\mathcal{N}(0,\frac{1}{\mathrm{i-j}})$ every 400000 spins

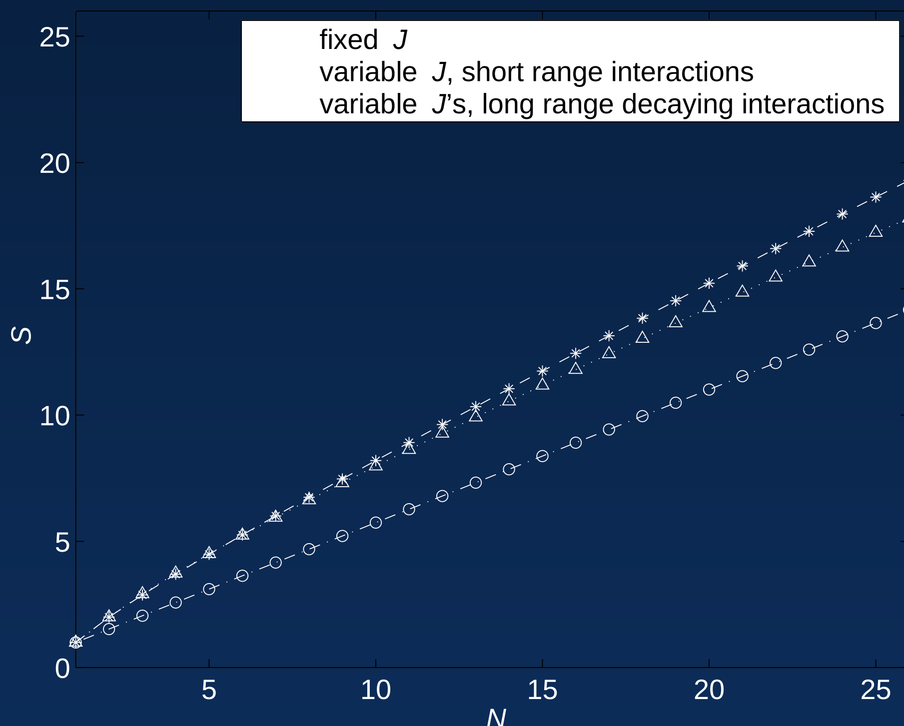
$1 \cdot 10^9$ spins total.



Entropy of 3 generated chains

- $J_{ij} = \delta_{i,j+1}$
- $J_{ij} = J_0 \, \delta_{i,j+1}$, J_0 is taken at random from $\mathcal{N}(0,1)$ every 400000 spins
- J_{ij} is taken at random from $\mathcal{N}(0, \frac{1}{i-j})$ every 400000 spins

$1 \cdot 10^9$ spins total.

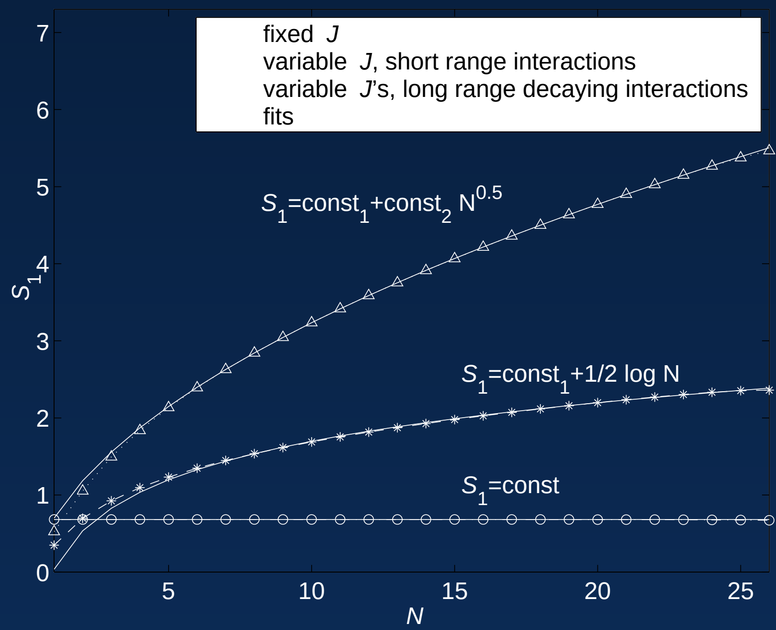


Entropy is extensive!

It shows no distinction between the cases.

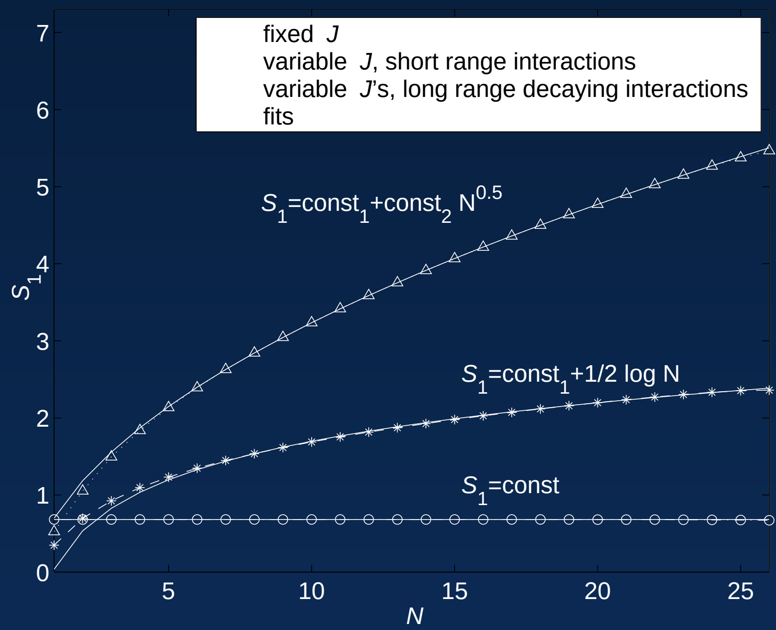
Subextensive component of the entropy

. . . shows a qualitative distinction between the cases!



Subextensive component of the entropy

. . . shows a qualitative distinction between the cases!

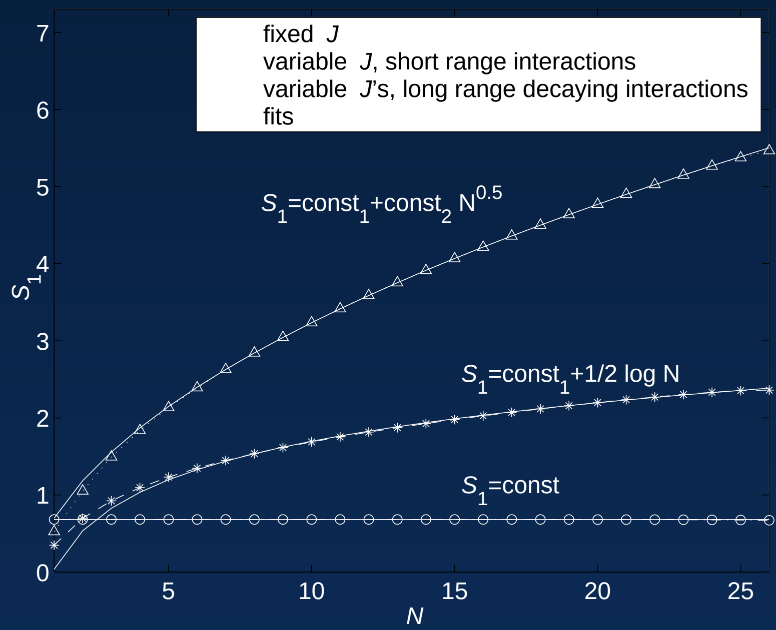


Other examples:

const periodic sequences, chaotic sequences (finite correlation length)

Subextensive component of the entropy

. . . shows a qualitative distinction between the cases!



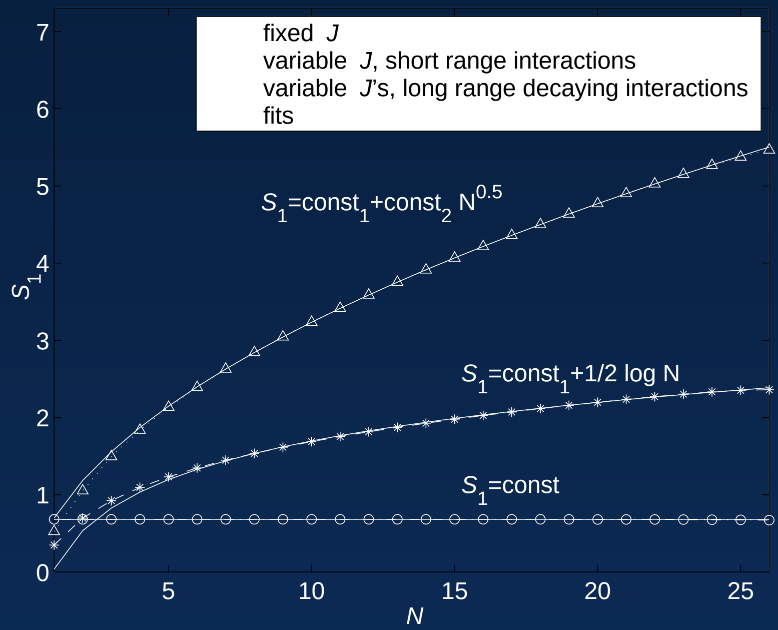
Other examples:

const periodic sequences, chaotic sequences (finite correlation length)

log systems at phase transitions, or at the onset of chaos (divergent correlation length)

Subextensive component of the entropy

. . . shows a qualitative distinction between the cases!



Other examples:

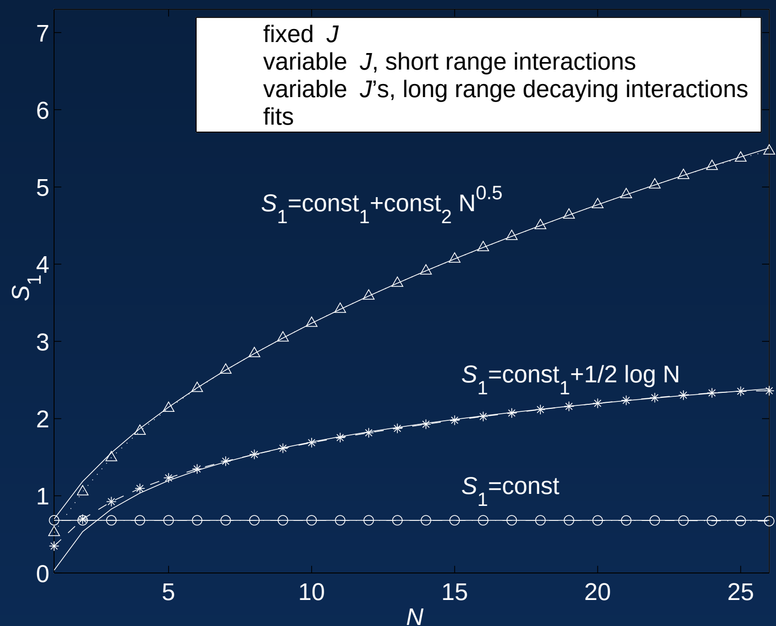
const periodic sequences, chaotic sequences (finite correlation length)

log systems at phase transitions, or at the onset of chaos (divergent correlation length)

power natural texts, DNA sequences, (possibly) some exotic transitions, (many divergent correlation lengths)

Subextensive component of the entropy

. . . shows a qualitative distinction between the cases!



Other examples:

const periodic sequences, chaotic sequences (finite correlation length)

log systems at phase transitions, or at the onset of chaos (divergent correlation length)

power natural texts, DNA sequences, (possibly) some exotic transitions, (many divergent correlation lengths)

- Entropy density or channel capacity do not distinguish these cases.
- Theory of phase transitions may not distinguish between the last two cases.
- Complexity of underlying dynamics intuitively increases from **const** to **power**.

Objectives

Objectives

learning unified description of learning (metric and algorithm independent)

Objectives

learning unified description of learning (metric and algorithm independent)

usability making distinction between useful and unusable data (noise vs. signal)

Objectives

learning unified description of learning (metric and algorithm independent)

usability making distinction between useful and unusable data (noise vs. signal)

complexity universal definition of dynamics' complexity (more rules describing dynamics $\Leftrightarrow$ higher complexity)

Objectives

learning unified description of learning (metric and algorithm independent)

usability making distinction between useful and unusable data (noise vs. signal)

complexity universal definition of dynamics' complexity (more rules describing dynamics $\Leftrightarrow$ higher complexity)

relations connection between the two (more rules $\Leftrightarrow$ more difficult to learn)

Solution – predictability

Solution – predictability

learning we learn (estimate parameters, extrapolate, classify, . . .) to *generalize* and *predict* from training examples; estimation of parameters is only an intermediate step

Solution – predictability

learning we learn (estimate parameters, extrapolate, classify, . . .) to *generalize* and *predict* from training examples; estimation of parameters is only an intermediate step

usability nonpredictive features in any signal are useless since we observe *now* and react in the *future*

Solution – predictability

learning we learn (estimate parameters, extrapolate, classify, . . .) to *generalize* and *predict* from training examples; estimation of parameters is only an intermediate step

usability nonpredictive features in any signal are useless since we observe *now* and react in the *future*

complexity high predictability sources (more details to predict, not easier predictions) are generated by more complex sources (in particular, regular and random sources have low complexity)

Solution – predictability

learning we learn (estimate parameters, extrapolate, classify, . . .) to *generalize* and *predict* from training examples; estimation of parameters is only an intermediate step

usability nonpredictive features in any signal are useless since we observe *now* and react in the *future*

complexity high predictability sources (more details to predict, not easier predictions) are generated by more complex sources (in particular, regular and random sources have low complexity)

relations more features to describe (complexity) $\Leftrightarrow$ more data needed for reliable predictions (learning)

Quantifying predictability

Information theory: non-metric, universal way to quantify learning

Quantifying predictability

Information theory: non-metric, universal way to quantify learning

Quantifying predictability

Information theory: non-metric, universal way to quantify learning

T, N — 0 — T', N' — x

past — now — future

$$\mathcal{I}_{\rm pred}(T,T') = \left\langle \log_2 \left[\frac{P(x_{\rm future}|x_{\rm past})}{P(x_{\rm future})} \right] \right\rangle$$

$$= S(T) + S(T') - S(T+T')$$

Quantifying predictability

Information theory: non-metric, universal way to quantify learning

T, N	0	T', N' x
past	now	future

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

Extensive component cancels in predictive information.

Quantifying predictability

Information theory: non-metric, universal way to quantify learning

T, N — 0 — T', N' — x

past — now — future

$$\mathcal{I}_{\rm pred}(T,T') = \left\langle \log_2 \left[\frac{P(x_{\rm future}|x_{\rm past})}{P(x_{\rm future})}\right] \right\rangle$$

$$= S(T) + S(T') - S(T+T')$$

$$S(T) = \mathcal{S}_0 \cdot T + S_1(T)$$

Extensive component cancels in predictive information.

Predictability is a deviation from extensivity!

Quantifying predictability

Information theory: non-metric, universal way to quantify learning

T, N	0	T', N' x
past	now	future

$$\mathcal{I}_{\rm pred}(T,T') = \left\langle \log_2 \left[\frac{P(x_{\rm future}|x_{\rm past})}{P(x_{\rm future})} \right] \right\rangle$$

$$= S(T) + S(T') - S(T+T')$$

$$S(T) = \mathcal{S}_0 \cdot T + S_1(T)$$

Extensive component cancels in predictive information.

Predictability is a deviation from extensivity!

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

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

- $I_{\text{pred}}(T)$ is information, so $I_{\text{pred}}(T) \geq 0$

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

- $I_{\mathrm{pred}}(T)$ is information, so $I_{\mathrm{pred}}(T) \geq 0$
- $I_{\mathrm{pred}}(T)$ is subextensive, $\lim_{T\to\infty} \frac{I_{\mathrm{pred}}(T)}{T} = 0$

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, $\lim_{T\to\infty} \frac{I_{\text{pred}}(T)}{S(T)} = 0$

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, $\lim_{T\to\infty} \frac{I_{\text{pred}}(T)}{S(T)} = 0$
- prediction and postdiction are symmetric

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, $\lim_{T\to\infty} \frac{I_{\text{pred}}(T)}{S(T)} = 0$
- prediction and postdiction are symmetric
- it relates to and generalizes many relevant quantities

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

- $I_{\rm pred}(T)$ is information, so $I_{\rm pred}(T) \geq 0$
- $I_{\rm pred}(T)$ is subextensive, $\lim_{T\to\infty} \frac{I_{\rm pred}(T)}{T} = 0$
- diminishing returns, $\lim_{T\to\infty} \frac{I_{\rm pred}(T)}{S(T)} = 0$
- prediction and postdiction are symmetric
- it relates to and generalizes many relevant quantities
    - learning: universal learning curves

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, $\lim_{T\to\infty} \frac{I_{\text{pred}}(T)}{S(T)} = 0$
- prediction and postdiction are symmetric
- it relates to and generalizes many relevant quantities
    - learning: universal learning curves
    - complexity: complexity measures

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, $\lim_{T\to\infty} \frac{I_{\text{pred}}(T)}{S(T)} = 0$
- prediction and postdiction are symmetric
- it relates to and generalizes many relevant quantities
    - learning: universal learning curves
    - complexity: complexity measures
    - coding: coding length

Grassberger vs. Kolmogorov

Average or typical vs. particular cases

Grassberger vs. Kolmogorov

Average or typical vs. particular cases

- nothing to learn (predict, encode, describe) for only one string

Grassberger vs. Kolmogorov

Average or typical vs. particular cases

- nothing to learn (predict, encode, describe) for only one string
- atypical data is possible

Grassberger vs. Kolmogorov

Average or typical vs. particular cases

- nothing to learn (predict, encode, describe) for only one string
- atypical data is possible

Complexity (learning properties) is an ensemble (averaged) quantity, even if the ensemble is only implicit.

Grassberger vs. Kolmogorov

Average or typical vs. particular cases

- nothing to learn (predict, encode, describe) for only one string
- atypical data is possible

Complexity (learning properties) is an ensemble (averaged) quantity, even if the ensemble is only implicit.

Example: all pictures can be random, but we do not perceive them this way.

The ghost of Bayes

Model family (ensemble) A	Model family (ensemble) B
$Q_A(x_1 \dots x_N \vert \boldsymbol{\alpha})$, $\mathcal{P}_A(\boldsymbol{\alpha})$, $Pr(A)$	$Q_B(x_1 \dots x_N \vert \boldsymbol{\beta})$, $\mathcal{P}_B(\boldsymbol{\beta})$, $Pr(B)$

The ghost of Bayes

Model family (ensemble) A

$Q_A(x_1 \ldots x_N|\boldsymbol{\alpha})$, $\mathcal{P}_A(\boldsymbol{\alpha})$, $Pr(A)$

Model family (ensemble) B

$Q_B(x_1 \ldots x_N|\boldsymbol{\beta})$, $\mathcal{P}_B(\boldsymbol{\beta})$, $Pr(B)$

is $X = \{x_1 \ldots x_N\}$ from A or B?

The ghost of Bayes

Model family (ensemble) A

$Q_A(x_1 \ldots x_N|\boldsymbol{\alpha})$, $\mathcal{P}_A(\boldsymbol{\alpha})$, $Pr(A)$

Model family (ensemble) B

$Q_B(x_1 \ldots x_N|\boldsymbol{\beta})$, $\mathcal{P}_B(\boldsymbol{\beta})$, $Pr(B)$

is $X = \{x_1 \ldots x_N\}$ from A or B?

$$P(A|X) = \frac{P(X|A)Pr(A)}{P(X)} = \frac{Pr(A)\int d\boldsymbol{\alpha}\mathcal{P}_A(\boldsymbol{\alpha})Q_A(X|\boldsymbol{\alpha})}{P(X|A)Pr(A) + P(X|B)Pr(B)}$$

The ghost of Bayes

Model family (ensemble) A
$Q_A(x_1 \dots x_N|\boldsymbol{\alpha})$, $\mathcal{P}_A(\boldsymbol{\alpha})$, $Pr(A)$

Model family (ensemble) B
$Q_B(x_1 \dots x_N|\boldsymbol{\beta})$, $\mathcal{P}_B(\boldsymbol{\beta})$, $Pr(B)$

is $X = \{x_1 \dots x_N\}$ from A or B?

$$P(A|X) = \frac{P(X|A)Pr(A)}{P(X)} = \frac{Pr(A)\int d\boldsymbol{\alpha}\mathcal{P}_A(\boldsymbol{\alpha})Q_A(X|\boldsymbol{\alpha})}{P(X|A)Pr(A) + P(X|B)Pr(B)}$$

Large N expansion around maximum likelihood value is almost always valid

$$\log P(A|X) \rightarrow \sum_i \underbrace{\log Q_A(X|\boldsymbol{\alpha}_{\mathrm{ML}})} - \underbrace{\frac{1}{2}\log\det\frac{\partial^2 \log Q_A(X|\boldsymbol{\alpha}_{\mathrm{ML}})}{\partial\alpha_a\partial\alpha_b}} + \dots$$

The ghost of Bayes

Model family (ensemble) A

$Q_A(x_1 \dots x_N|\boldsymbol{\alpha})$, $\mathcal{P}_A(\boldsymbol{\alpha})$, $Pr(A)$

Model family (ensemble) B

$Q_B(x_1 \dots x_N|\boldsymbol{\beta})$, $\mathcal{P}_B(\boldsymbol{\beta})$, $Pr(B)$

is $X = \{x_1 \dots x_N\}$ from A or B?

$$P(A|X) = \frac{P(X|A)Pr(A)}{P(X)} = \frac{Pr(A)\int d\boldsymbol{\alpha}\mathcal{P}_A(\boldsymbol{\alpha})Q_A(X|\boldsymbol{\alpha})}{P(X|A)Pr(A) + P(X|B)Pr(B)}$$

Large N expansion around maximum likelihood value is almost always valid

$$\log P(A|X) \to \sum_i \underbrace{\log Q_A(X|\boldsymbol{\alpha}_{\mathrm{ML}})}_{\text{goodness of fit}} - \underbrace{\frac{1}{2}\log\det\frac{\partial^2 \log Q_A(X|\boldsymbol{\alpha}_{\mathrm{ML}})}{\partial\alpha_a\partial\alpha_b}}_{\text{generalization error, fluctuations, complexity}} + \dots$$

How can I_{pred} behave?

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

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

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
- well known as $I(N, \text{parameters}) = I_{\text{pred}}(N)$

How can I_{pred} behave?

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

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

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

- learning finite-parameter densities
- well known as $I(N, \mathrm{parameters}) = I_{\mathrm{pred}}(N)$

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

- learning continuous densities
- not well studied

Specific examples: problem setup

Specific examples: problem setup

$Q(\vec{x}\|\boldsymbol{\alpha})$	p. d. f. for $\vec{x}$ parameterized by unknown parameters $\boldsymbol{\alpha}$
$\dim \boldsymbol{\alpha} = K$	dimensionality of $\boldsymbol{\alpha}$, may be infinite
$\mathcal{P}(\boldsymbol{\alpha})$	prior distribution of parameters
$\vec{x}_1 \cdots \vec{x}_{\mathrm{N}}$	random samples from the distribution

Specific examples: problem setup

$Q(\vec{x}|\boldsymbol{\alpha})$ p. d. f. for $\vec{x}$ parameterized by unknown parameters $\boldsymbol{\alpha}$

$\dim \boldsymbol{\alpha} = K$ dimensionality of $\boldsymbol{\alpha}$, may be infinite

$\mathcal{P}(\boldsymbol{\alpha})$ prior distribution of parameters

$\vec{x}_1 \cdots \vec{x}_{\mathrm{N}}$ random samples from the distribution

$$P(\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_{\mathrm{N}}|\boldsymbol{\alpha}) = \prod_{\mathrm{i}=1}^{\mathrm{N}} Q(\vec{x}_{\mathrm{i}}|\boldsymbol{\alpha})$$

Specific examples: problem setup

$Q(\vec{x}|\boldsymbol{\alpha})$ p. d. f. for $\vec{x}$ parameterized by unknown parameters $\boldsymbol{\alpha}$

$\dim \boldsymbol{\alpha} = K$ dimensionality of $\boldsymbol{\alpha}$, may be infinite

$\mathcal{P}(\boldsymbol{\alpha})$ prior distribution of parameters

$\vec{x}_1 \cdots \vec{x}_{\mathrm{N}}$ random samples from the distribution

$$
\begin{aligned}
P(\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_{\mathrm{N}}|\boldsymbol{\alpha}) &= \prod_{\mathrm{i}=1}^{\mathrm{N}} Q(\vec{x}_{\mathrm{i}}|\boldsymbol{\alpha}) \\
P(\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_{\mathrm{N}}) &= \int d^K\alpha \mathcal{P}(\boldsymbol{\alpha}) \prod_{\mathrm{i}=1}^{\mathrm{N}} Q(\vec{x}_{\mathrm{i}}|\boldsymbol{\alpha})
\end{aligned}
$$

Specific examples: problem setup

$Q(\vec{x}|\boldsymbol{\alpha})$ p. d. f. for $\vec{x}$ parameterized by unknown parameters $\boldsymbol{\alpha}$

$\dim \boldsymbol{\alpha} = K$ dimensionality of $\boldsymbol{\alpha}$, may be infinite

$\mathcal{P}(\boldsymbol{\alpha})$ prior distribution of parameters

$\vec{x}_1 \cdots \vec{x}_{\rm N}$ random samples from the distribution

$$
\begin{aligned}
P(\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_{\rm N}|\boldsymbol{\alpha}) &= \prod_{\rm i=1}^{\rm N} Q(\vec{x}_{\rm i}|\boldsymbol{\alpha}) \\
P(\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_{\rm N}) &= \int d^K\alpha \mathcal{P}(\boldsymbol{\alpha}) \prod_{\rm i=1}^{\rm N} Q(\vec{x}_{\rm i}|\boldsymbol{\alpha}) \\
S(\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_{\rm N}) &\equiv S(N) \\
&= -\int d\vec{x}_1 \cdots d\vec{x}_{\rm N}\, P(\{\vec{x}_{\rm i}\}) \log_2 P(\{\vec{x}_{\rm i}\})
\end{aligned}
$$

Separating the extensive term

$$S(N) \quad = \quad -\int d^K\bar{\boldsymbol{\alpha}}\mathcal{P}(\bar{\boldsymbol{\alpha}})\left\{d^N\vec{x}\prod_{\rm j=1}^{\rm N}Q(\vec{x}_{\rm j}|\bar{\boldsymbol{\alpha}})\log_2\int d^K\alpha\mathcal{P}(\boldsymbol{\alpha})\prod_{\rm i=1}^{\rm N}Q(\vec{x}_{\rm i}|\boldsymbol{\alpha})\right\}$$

Separating the extensive term

$$
\begin{aligned}
S(N) &= -\int d^K\bar{\boldsymbol{\alpha}}\mathcal{P}(\bar{\boldsymbol{\alpha}})\left\{d^N\vec{x}\prod_{\mathrm{j}=1}^{\mathrm{N}}Q(\vec{x}_{\mathrm{j}}|\bar{\boldsymbol{\alpha}})\log_2\int d^K\alpha\mathcal{P}(\boldsymbol{\alpha})\prod_{\mathrm{i}=1}^{\mathrm{N}}Q(\vec{x}_{\mathrm{i}}|\boldsymbol{\alpha})\right\} \\
&= -\int d^K\bar{\boldsymbol{\alpha}}\mathcal{P}(\bar{\boldsymbol{\alpha}})\left\{d^N\vec{x}\prod_{\mathrm{j}=1}^{\mathrm{N}}Q(\vec{x}_{\mathrm{j}}|\bar{\boldsymbol{\alpha}})\right. \\
&\qquad\left.\times\log_2\prod_{\mathrm{j}=1}^{\mathrm{N}}Q(\vec{x}_{\mathrm{j}}|\bar{\boldsymbol{\alpha}})\int d^K\alpha\mathcal{P}(\boldsymbol{\alpha})\overbrace{\prod_{\mathrm{i}=1}^{\mathrm{N}}\left[\frac{Q(\vec{x}_{\mathrm{i}}|\boldsymbol{\alpha})}{Q(\vec{x}_{\mathrm{i}}|\bar{\boldsymbol{\alpha}})}\right]}^{\exp\left[-N\mathcal{E}_N(\boldsymbol{\alpha},\bar{\boldsymbol{\alpha}};\{\vec{x}_{\mathrm{i}}\})\right]}\right\}
\end{aligned}
$$

Separating the extensive term

$$
\begin{aligned}
S(N) &= -\int d^K \bar{\boldsymbol{\alpha}} \mathcal{P}(\bar{\boldsymbol{\alpha}}) \left\{ d^N \vec{x} \prod_{\mathrm{j}=1}^{\mathrm{N}} Q(\vec{x}_{\mathrm{j}}|\bar{\boldsymbol{\alpha}}) \log_2 \int d^K \alpha \mathcal{P}(\boldsymbol{\alpha}) \prod_{\mathrm{i}=1}^{\mathrm{N}} Q(\vec{x}_{\mathrm{i}}|\boldsymbol{\alpha}) \right\} \\
&= -\int d^K \bar{\boldsymbol{\alpha}} \mathcal{P}(\bar{\boldsymbol{\alpha}}) \left\{ d^N \vec{x} \prod_{\mathrm{j}=1}^{\mathrm{N}} Q(\vec{x}_{\mathrm{j}}|\bar{\boldsymbol{\alpha}}) \right. \\
&\qquad \left. \times \log_2 \prod_{\mathrm{j}=1}^{\mathrm{N}} Q(\vec{x}_{\mathrm{j}}|\bar{\boldsymbol{\alpha}}) \int d^K \alpha \mathcal{P}(\boldsymbol{\alpha}) \overbrace{\prod_{\mathrm{i}=1}^{\mathrm{N}} \left[\frac{Q(\vec{x}_{\mathrm{i}}|\boldsymbol{\alpha})}{Q(\vec{x}_{\mathrm{i}}|\bar{\boldsymbol{\alpha}})} \right]}^{\exp\left[-N\mathcal{E}_N(\boldsymbol{\alpha},\bar{\boldsymbol{\alpha}};\{\vec{x}_{\mathrm{i}}\})\right]} \right\}
\end{aligned}
$$

This separates $S(N)$ into the extensive and the subextensive terms

$$
\begin{aligned}
\mathcal{S}_0 &= \int d^K \alpha \mathcal{P}(\boldsymbol{\alpha}) \left[-\int d\vec{x} Q(\vec{x}|\boldsymbol{\alpha}) \log_2 Q(\vec{x}|\boldsymbol{\alpha}) \right], \\
S_1(N) &= -\int d^K \bar{\alpha}\, d^N \vec{x}_{\mathrm{i}} \mathcal{P}(\bar{\boldsymbol{\alpha}}) \log_2 \left[\int d^K \alpha P(\boldsymbol{\alpha}) \mathrm{e}^{-N\mathcal{E}_N} \right]
\end{aligned}
$$

Annealed approximation

Under some (known) conditions we may have

$$\psi(\boldsymbol{\alpha}, \bar{\boldsymbol{\alpha}}; \{x_{\rm i}\}) \quad \equiv \quad \underbrace{\mathcal{E}_N(\boldsymbol{\alpha}, \bar{\boldsymbol{\alpha}}; \{\vec{x}_{\rm i}\})} \quad - \quad \underbrace{D_{\rm KL}(\bar{\boldsymbol{\alpha}}||\boldsymbol{\alpha})}$$

Annealed approximation

Under some (known) conditions we may have

$$\psi(\boldsymbol{\alpha}, \bar{\boldsymbol{\alpha}}; \{x_{\rm i}\}) \quad \equiv \quad \underbrace{\mathcal{E}_N(\boldsymbol{\alpha}, \bar{\boldsymbol{\alpha}}; \{\vec{x}_{\rm i}\})}_{\text{quenched energy}} \quad - \quad \underbrace{D_{\rm KL}(\bar{\boldsymbol{\alpha}}||\boldsymbol{\alpha})}_{\text{annealed energy}}$$

Annealed approximation

Under some (known) conditions we may have

$$
\begin{aligned}
\psi(\boldsymbol{\alpha}, \bar{\boldsymbol{\alpha}}; \{x_{\rm i}\}) &\equiv \underbrace{\mathcal{E}_N(\boldsymbol{\alpha}, \bar{\boldsymbol{\alpha}}; \{\vec{x}_{\rm i}\})}_{\text{quenched energy}} - \underbrace{D_{\rm KL}(\bar{\boldsymbol{\alpha}}||\boldsymbol{\alpha})}_{\text{annealed energy}} \\
&\equiv -\frac{1}{N}\sum_{\rm i=1}^{\rm N} \ln\left[\frac{Q(\vec{x}_{\rm i}|\boldsymbol{\alpha})}{Q(\vec{x}_{\rm i}|\bar{\boldsymbol{\alpha}})}\right] + \int d\vec{x} Q(\vec{x}|\bar{\boldsymbol{\alpha}}) \ln\left[\frac{Q(\vec{x}|\boldsymbol{\alpha})}{Q(\vec{x}|\bar{\boldsymbol{\alpha}})}\right] \\
&\overset{\sim}{\rightarrow} 0
\end{aligned}
$$

Annealed approximation

Under some (known) conditions we may have

$$\begin{aligned}
\psi(\boldsymbol{\alpha}, \bar{\boldsymbol{\alpha}}; \{x_{\rm i}\}) &\equiv \underbrace{\mathcal{E}_N(\boldsymbol{\alpha}, \bar{\boldsymbol{\alpha}}; \{\vec{x}_{\rm i}\})}_{\text{quenched energy}} - \underbrace{D_{\rm KL}(\bar{\boldsymbol{\alpha}} || \boldsymbol{\alpha})}_{\text{annealed energy}} \\
&\equiv -\frac{1}{N}\sum_{\rm i=1}^{\rm N} \ln\left[\frac{Q(\vec{x}_{\rm i}|\boldsymbol{\alpha})}{Q(\vec{x}_{\rm i}|\bar{\boldsymbol{\alpha}})}\right] + \int d\vec{x} Q(\vec{x}|\bar{\boldsymbol{\alpha}}) \ln\left[\frac{Q(\vec{x}|\boldsymbol{\alpha})}{Q(\vec{x}|\bar{\boldsymbol{\alpha}})}\right] \\
&\stackrel{\sim}{\rightarrow} 0 \\
S_1(N) &\stackrel{\sim}{\rightarrow} S_1^{(\rm a)}(N) \\
&\equiv -\int d^K \bar{\alpha} \mathcal{P}(\bar{\boldsymbol{\alpha}}) \underbrace{\log_2 \overbrace{\int d^K \alpha P(\boldsymbol{\alpha}) {\rm e}^{-N D_{\rm KL}}}}
\end{aligned}$$

Annealed approximation

Under some (known) conditions we may have

$$
\begin{aligned}
\psi(\boldsymbol{\alpha},\bar{\boldsymbol{\alpha}};\{x_{\rm i}\}) &\equiv \underbrace{\mathcal{E}_N(\boldsymbol{\alpha},\bar{\boldsymbol{\alpha}};\{\vec{x}_{\rm i}\})}_{\text{quenched energy}} - \underbrace{D_{\rm KL}(\bar{\boldsymbol{\alpha}}||\boldsymbol{\alpha})}_{\text{annealed energy}} \\
&\equiv -\frac{1}{N}\sum_{\rm i=1}^{\rm N}\ln\left[\frac{Q(\vec{x}_{\rm i}|\boldsymbol{\alpha})}{Q(\vec{x}_{\rm i}|\bar{\boldsymbol{\alpha}})}\right] + \int d\vec{x}Q(\vec{x}|\bar{\boldsymbol{\alpha}})\ln\left[\frac{Q(\vec{x}|\boldsymbol{\alpha})}{Q(\vec{x}|\bar{\boldsymbol{\alpha}})}\right] \\
&\stackrel{\sim}{\rightarrow} 0 \\
S_1(N) &\stackrel{\sim}{\rightarrow} S_1^{(\rm a)}(N) \\
&\equiv -\int d^K\bar{\alpha}\mathcal{P}(\bar{\boldsymbol{\alpha}})\underbrace{\log_2 \overbrace{\int d^K\alpha P(\boldsymbol{\alpha}){\rm e}^{-ND_{\rm KL}}}^{\text{annealed partition function, } Z(\bar{\boldsymbol{\alpha}};N)}}_{\text{annealed free energy, } F(\bar{\boldsymbol{\alpha}};N)}
\end{aligned}
$$

Density of states

We can rewrite the partition function

$$Z(\bar{\boldsymbol{\alpha}}; N) = \int dD \rho(D; \bar{\boldsymbol{\alpha}}) \exp[-ND]$$

Density of states

We can rewrite the partition function

$$Z(\bar{\boldsymbol{\alpha}}; N) = \int dD \rho(D; \bar{\boldsymbol{\alpha}}) \exp[-ND]$$

$$\rho(D; \bar{\boldsymbol{\alpha}}) = \int d^K \alpha \mathcal{P}(\boldsymbol{\alpha}) \delta[D - D_{\mathrm{KL}}(\bar{\boldsymbol{\alpha}} || \boldsymbol{\alpha})]$$

Density of states

We can rewrite the partition function

$$Z(\bar{\boldsymbol{\alpha}}; N) = \int dD \rho(D; \bar{\boldsymbol{\alpha}}) \exp[-ND]$$

$$\rho(D; \bar{\boldsymbol{\alpha}}) = \int d^K \alpha \mathcal{P}(\boldsymbol{\alpha}) \delta[D - D_{\mathrm{KL}}(\bar{\boldsymbol{\alpha}} || \boldsymbol{\alpha})]$$

$$\int dD \rho(D; \bar{\boldsymbol{\alpha}}) = \int d^K \alpha \mathcal{P}(\boldsymbol{\alpha}) = 1$$

Density of states

We can rewrite the partition function

$$
\begin{aligned}
Z(\bar{\boldsymbol{\alpha}}; N) &= \int dD \rho(D; \bar{\boldsymbol{\alpha}}) \exp[-ND] \\
\rho(D; \bar{\boldsymbol{\alpha}}) &= \int d^K \alpha \mathcal{P}(\boldsymbol{\alpha}) \delta[D - D_{\mathrm{KL}}(\bar{\boldsymbol{\alpha}} || \boldsymbol{\alpha})] \\
\int dD \rho(D; \bar{\boldsymbol{\alpha}}) &= \int d^K \alpha \mathcal{P}(\boldsymbol{\alpha}) = 1
\end{aligned}
$$

The density ρ could be very different for different targets.

Density of states

We can rewrite the partition function

$$
\begin{aligned}
Z(\bar{\boldsymbol{\alpha}}; N) &= \int dD \rho(D; \bar{\boldsymbol{\alpha}}) \exp[-ND] \\
\rho(D; \bar{\boldsymbol{\alpha}}) &= \int d^K \alpha \mathcal{P}(\boldsymbol{\alpha}) \delta[D - D_{\mathrm{KL}}(\bar{\boldsymbol{\alpha}} || \boldsymbol{\alpha})] \\
\int dD \rho(D; \bar{\boldsymbol{\alpha}}) &= \int d^K \alpha \mathcal{P}(\boldsymbol{\alpha}) = 1
\end{aligned}
$$

The density ρ could be very different for different targets.

Thus learning is annealing at decreasing temperature.

Properties of predictive information (and learning) almost always depend on $D = 0$ behavior of the density.

Power–law density function

$$\rho(D \to 0; \bar{\boldsymbol{\alpha}}) \approx A(\bar{\boldsymbol{\alpha}}) D^{(d-2)/2}$$

Power–law density function

$$\rho(D \to 0; \bar{\boldsymbol{\alpha}}) \approx A(\bar{\boldsymbol{\alpha}}) D^{(d-2)/2}$$

Example: *sound* finite parameter models, $\dim \boldsymbol{\alpha} = d$.

Power–law density function

$$\rho(D \to 0; \bar{\boldsymbol{\alpha}}) \approx A(\bar{\boldsymbol{\alpha}}) D^{(d-2)/2}$$

Example: *sound* finite parameter models, $\dim \boldsymbol{\alpha} = d$.

$$D_{\mathrm{KL}}(\bar{\boldsymbol{\alpha}} || \boldsymbol{\alpha}) \quad \xrightarrow{\boldsymbol{\alpha} \to \bar{\boldsymbol{\alpha}}} \quad \frac{1}{2} \sum_{\mu\nu} (\bar{\alpha}_\mu - \alpha_\mu) \mathcal{F}_{\mu\nu} (\bar{\alpha}_\nu - \alpha_\nu) + \cdots$$

Power–law density function

$$\rho(D \to 0; \bar{\boldsymbol{\alpha}}) \approx A(\bar{\boldsymbol{\alpha}}) D^{(d-2)/2}$$

Example: *sound* finite parameter models, $\dim \boldsymbol{\alpha} = d$.

$$D_{\mathrm{KL}}(\bar{\boldsymbol{\alpha}} || \boldsymbol{\alpha}) \quad \xrightarrow{\boldsymbol{\alpha} \to \bar{\boldsymbol{\alpha}}} \quad \frac{1}{2} \sum_{\mu\nu} (\bar{\alpha}_\mu - \alpha_\mu) \mathcal{F}_{\mu\nu} (\bar{\alpha}_\nu - \alpha_\nu) + \cdots$$

$$\rho(D; \bar{\boldsymbol{\alpha}}) \quad \xrightarrow{D \to 0} \quad \mathcal{P}(\bar{\boldsymbol{\alpha}}) \frac{2\pi^{d/2}}{\Gamma(d/2)} \left(\det \mathcal{F}\right)^{-1/2} D^{(d-2)/2}$$

Power–law density function

$$\rho(D \to 0; \bar{\boldsymbol{\alpha}}) \approx A(\bar{\boldsymbol{\alpha}}) D^{(d-2)/2}$$

Example: *sound* finite parameter models, $\dim \boldsymbol{\alpha} = d$.

$$D_{\mathrm{KL}}(\bar{\boldsymbol{\alpha}} || \boldsymbol{\alpha}) \quad \xrightarrow{\boldsymbol{\alpha} \to \bar{\boldsymbol{\alpha}}} \quad \frac{1}{2} \sum_{\mu\nu} (\bar{\alpha}_\mu - \alpha_\mu) \mathcal{F}_{\mu\nu} (\bar{\alpha}_\nu - \alpha_\nu) + \cdots$$

$$\rho(D; \bar{\boldsymbol{\alpha}}) \quad \xrightarrow{D \to 0} \quad \mathcal{P}(\bar{\boldsymbol{\alpha}}) \frac{2\pi^{d/2}}{\Gamma(d/2)} (\det \mathcal{F})^{-1/2} D^{(d-2)/2}$$

$$S_1^{(\mathrm{a})} \quad \approx \quad \frac{d}{2} \log_2 N$$

Power–law density function

$$\rho(D \to 0; \bar{\boldsymbol{\alpha}}) \approx A(\bar{\boldsymbol{\alpha}}) D^{(d-2)/2}$$

Example: *sound* finite parameter models, $\dim \boldsymbol{\alpha} = d$.

$$D_{\mathrm{KL}}(\bar{\boldsymbol{\alpha}}||\boldsymbol{\alpha}) \quad \xrightarrow{\boldsymbol{\alpha}\to\bar{\boldsymbol{\alpha}}} \quad \frac{1}{2}\sum_{\mu\nu}(\bar{\alpha}_\mu - \alpha_\mu)\mathcal{F}_{\mu\nu}(\bar{\alpha}_\nu - \alpha_\nu) + \cdots$$

$$\rho(D; \bar{\boldsymbol{\alpha}}) \quad \xrightarrow{D\to 0} \quad \mathcal{P}(\bar{\boldsymbol{\alpha}})\frac{2\pi^{d/2}}{\Gamma(d/2)}(\det \mathcal{F})^{-1/2} D^{(d-2)/2}$$

$$S_1^{(\mathrm{a})} \quad \approx \quad \frac{d}{2}\log_2 N$$

Speed of approach to this asymptotics is rarely investigated.

Another example

Learning $Q(\vec{x}_1 \cdots \vec{x}_{\rm N}|\boldsymbol{\alpha})$, a finite parameter Markov process with long range intrinsic correlations such that

$$\begin{aligned} S\left[\{\vec{x}_{\rm i}\}|\boldsymbol{\alpha}\right] &\equiv -\int d^N\vec{x}\, Q(\{\vec{x}_{\rm i}\}|\boldsymbol{\alpha}) \log_2 Q(\{\vec{x}_{\rm i}\}|\boldsymbol{\alpha}) \\ &\to N\mathcal{S}_0 + \mathcal{S}_0^*; \qquad \mathcal{S}_0^* = \frac{K'}{2}\log_2 N \end{aligned}$$

Another example

Learning $Q(\vec{x}_1 \cdots \vec{x}_{\rm N}|\boldsymbol{\alpha})$, a finite parameter Markov process with long range intrinsic correlations such that

$$
\begin{aligned}
S\left[\{\vec{x}_{\rm i}\}|\boldsymbol{\alpha}\right] \;&\equiv\; -\int d^N\vec{x}\, Q(\{\vec{x}_{\rm i}\}|\boldsymbol{\alpha}) \log_2 Q(\{\vec{x}_{\rm i}\}|\boldsymbol{\alpha}) \\
&\rightarrow\; N\mathcal{S}_0 + \mathcal{S}_0^*; \qquad \mathcal{S}_0^* = \frac{K'}{2}\log_2 N \\
S_1^{\rm (a)}(N) \;&\approx\; \frac{K+K'}{2}\log_2 N
\end{aligned}
$$

Another example

Learning $Q(\vec{x}_1 \cdots \vec{x}_{\mathrm{N}}|\boldsymbol{\alpha})$, a finite parameter Markov process with long range intrinsic correlations such that

$$
\begin{aligned}
S\left[\{\vec{x}_{\mathrm{i}}\}|\boldsymbol{\alpha}\right] &\equiv -\int d^N\vec{x}\, Q(\{\vec{x}_{\mathrm{i}}\}|\boldsymbol{\alpha}) \log_2 Q(\{\vec{x}_{\mathrm{i}}\}|\boldsymbol{\alpha}) \\
&\rightarrow N\mathcal{S}_0 + \mathcal{S}_0^*; \qquad \mathcal{S}_0^* = \frac{K'}{2}\log_2 N \\
S_1^{(\mathrm{a})}(N) &\approx \frac{K+K'}{2}\log_2 N
\end{aligned}
$$

Predictive information does not distinguish predictability coming from unknown parameters and from intrinsic long–range correlations.

This is similar to describing physical systems with correlations using order parameters.

Essential singularity in the density

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

$$\rho(D \to 0; \bar{\boldsymbol{\alpha}}) \approx A(\bar{\boldsymbol{\alpha}}) \exp\left[-\frac{B(\bar{\boldsymbol{\alpha}})}{D^{\mu}}\right], \qquad \mu > 0$$

Essential singularity in the density

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

$$\rho(D \to 0; \bar{\boldsymbol{\alpha}}) \approx A(\bar{\boldsymbol{\alpha}}) \exp\left[-\frac{B(\bar{\boldsymbol{\alpha}})}{D^\mu}\right], \quad \mu > 0$$

$$C(\bar{\boldsymbol{\alpha}}) = [B(\bar{\boldsymbol{\alpha}})]^{1/(\mu+1)} \left(\frac{1}{\mu^{\mu/(\mu+1)}} + \mu^{1/(\mu+1)} \right)$$

$$S_1^{(\mathrm{a})}(N) \approx \frac{1}{\ln 2} \langle C(\bar{\boldsymbol{\alpha}}) \rangle_{\bar{\boldsymbol{\alpha}}} N^{\mu/(\mu+1)}$$

Essential singularity in the density

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

$$\rho(D \to 0; \bar{\boldsymbol{\alpha}}) \approx A(\bar{\boldsymbol{\alpha}}) \exp\left[-\frac{B(\bar{\boldsymbol{\alpha}})}{D^\mu}\right], \qquad \mu > 0$$

$$C(\bar{\boldsymbol{\alpha}}) = [B(\bar{\boldsymbol{\alpha}})]^{1/(\mu+1)} \left(\frac{1}{\mu^{\mu/(\mu+1)}} + \mu^{1/(\mu+1)}\right)$$

$$S_1^{(\mathrm{a})}(N) \approx \frac{1}{\ln 2} \langle C(\bar{\boldsymbol{\alpha}}) \rangle_{\bar{\boldsymbol{\alpha}}} N^{\mu/(\mu+1)}$$

- finite parameter model with increasing number of parameters $K \sim N^{\mu/(\mu+1)}$; $S_1(N) \sim N^{\mu/\mu+1}$, not $S_1(N) \sim \frac{N^{\mu/\mu+1}}{2} \log N$

Essential singularity in the density

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

$$\rho(D \to 0; \bar{\boldsymbol{\alpha}}) \approx A(\bar{\boldsymbol{\alpha}}) \exp\left[-\frac{B(\bar{\boldsymbol{\alpha}})}{D^{\mu}}\right], \quad \mu > 0$$

$$C(\bar{\boldsymbol{\alpha}}) = [B(\bar{\boldsymbol{\alpha}})]^{1/(\mu+1)} \left(\frac{1}{\mu^{\mu/(\mu+1)}} + \mu^{1/(\mu+1)}\right)$$

$$S_1^{(\mathrm{a})}(N) \approx \frac{1}{\ln 2} \langle C(\bar{\boldsymbol{\alpha}}) \rangle_{\bar{\boldsymbol{\alpha}}} N^{\mu/(\mu+1)}$$

- finite parameter model with increasing number of parameters $K \sim N^{\mu/(\mu+1)}$; $S_1(N) \sim N^{\mu/\mu+1}$, not $S_1(N) \sim \frac{N^{\mu/\mu+1}}{2} \log N$
- as $\mu \to \infty$ complexity grows and then vanishes to the leading order when $S_1^{(\mathrm{a})}$ becomes extensive

Example of the power–law I_{pred}

Learning a nonparameteric (infinite parameter) density $Q(x) = 1/l_0 \mathrm{e}^{-\phi(x)}$, $x \in [0, L]$, with some smoothness constraints (Bialek, Callan, and Strong 1996).

$$\mathcal{P}[\phi(x)] = \frac{1}{\mathcal{Z}} \exp\left[-\frac{l}{2}\int dx \left(\frac{\partial\phi}{\partial x}\right)^2\right] \delta\left[\frac{1}{l_0}\int dx\, \mathrm{e}^{-\phi(x)} - 1\right]$$

Example of the power–law I_{pred}

Learning a nonparameteric (infinite parameter) density $Q(x) = 1/l_0 \mathrm{e}^{-\phi(x)}$, $x \in [0, L]$, with some smoothness constraints (Bialek, Callan, and Strong 1996).

$$
\begin{aligned}
\mathcal{P}[\phi(x)] &= \frac{1}{\mathcal{Z}} \exp\left[-\frac{l}{2}\int dx \left(\frac{\partial\phi}{\partial x}\right)^2\right] \delta\left[\frac{1}{l_0}\int dx\, \mathrm{e}^{-\phi(x)} - 1\right] \\
\rho(D \to 0; \bar{\phi}) &= A[\bar{\phi}(x)] D^{-3/2} \exp\left(-\frac{B[\bar{\phi}(x)]}{D}\right)
\end{aligned}
$$

Example of the power–law $I_{\rm pred}$

Learning a nonparameteric (infinite parameter) density $Q(x) = 1/l_0 \mathrm{e}^{-\phi(x)}$, $x \in [0, L]$, with some smoothness constraints (Bialek, Callan, and Strong 1996).

$$\mathcal{P}[\phi(x)] = \frac{1}{\mathcal{Z}} \exp\left[-\frac{l}{2}\int dx \left(\frac{\partial\phi}{\partial x}\right)^2\right] \delta\left[\frac{1}{l_0}\int dx\, \mathrm{e}^{-\phi(x)} - 1\right]$$

$$\rho(D \to 0; \bar\phi) = A[\bar\phi(x)] D^{-3/2} \exp\left(-\frac{B[\bar\phi(x)]}{D}\right)$$

$$S_1^{(\mathrm{a})}(N) \approx \frac{1}{2\ln 2}\sqrt{N}\left(\frac{L}{l}\right)^{1/2}$$

Power–law density example: continuation

- increasing number of 'effective parameters' (bins) of adaptive size $\sim \sqrt{l/NQ(x)}$

Power–law density example: continuation

- increasing number of 'effective parameters' (bins) of adaptive size $\sim \sqrt{l/NQ(x)}$
- heuristic arguments for the dimensionality ζ and the smoothness exponent η give $S_1(N) \sim N^{\zeta/2\eta}$ — demonstrates a crossover from complexity to randomness

Which complexity do we want to define?

Which complexity do we want to define?

- complexity of dynamics that generates a time series (not computational or descriptive complexity); thus it must be zero for totally random and for easily predictable processes

Which complexity do we want to define?

- complexity of dynamics that generates a time series (not computational or descriptive complexity); thus it must be zero for totally random and for easily predictable processes
- usable for Occam–style punishment in statistical inference

Which complexity do we want to define?

- complexity of dynamics that generates a time series (not computational or descriptive complexity); thus it must be zero for totally random and for easily predictable processes
- usable for Occam–style punishment in statistical inference
- expressible in conventional physical terms

Which complexity do we want to define?

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

Complexity measure

Complexity measure

- some kind of entropy (we proclaim Shannon's postulates: monotonicity, continuity, additivity)

Complexity measure

- some kind of entropy (we proclaim Shannon's postulates: monotonicity, continuity, additivity)
- invariant under invertible temporally local transformations

  ($x_k \to x_k + \xi x_{k-1}$: measuring device with inertia, article with misprints, same book in different languages – same universality class)

$$\log P_1(x) = \log P_2(x) + \text{loc. oper.} \;\Rightarrow\; C[P_1(x)] = C[P_2(x)]$$

This may present a problem in higher dimensions.

Complexity measure

- some kind of entropy (we proclaim Shannon's postulates: monotonicity, continuity, additivity)
- invariant under invertible temporally local transformations

($x_k \to x_k + \xi x_{k-1}$: measuring device with inertia, article with misprints, same book in different languages – same universality class)

$$\log P_1(x) = \log P_2(x) + \text{loc. oper.} \;\Rightarrow\; C[P_1(x)] = C[P_2(x)]$$

This may present a problem in higher dimensions.

The divergent subextensive term measures complexity uniquely!

Relations to other definitions . . .

. . . are mostly straightforward.

Relations to other definitions . . .

. . . are mostly straightforward.

For Kolmogorov complexity:

Relations to other definitions . . .

. . . are mostly straightforward.

For Kolmogorov complexity:

- partition all strings into equivalence classes

Relations to other definitions . . .

. . . are mostly straightforward.

For Kolmogorov complexity:

- partition all strings into equivalence classes
- define Kolmogorov complexity $C_K(s)$ 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

Relations to other definitions . . .

. . . are mostly straightforward.

For Kolmogorov complexity:

- partition all strings into equivalence classes
- define Kolmogorov complexity $C_K(s)$ 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

Relations to other definitions . . .

. . . are mostly straightforward.

For Kolmogorov complexity:

- partition all strings into equivalence classes
- define Kolmogorov complexity $C_K(s)$ 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

If sufficient statistics exist, then $C_K \approx I_{\rm pred}$. Otherwise $C_K > I_{\rm pred}$.

C_K is unique up to a constant.

What's next?

What's next?

extraction separating predictive information from non–predictive using the 'relevant information' technique

What's next?

extraction separating predictive information from non–predictive using the 'relevant information' technique

physics of phase transitions, connection to subextensive statistical mechanics

What's next?

extraction separating predictive information from non–predictive using the 'relevant information' technique

physics of phase transitions, connection to subextensive statistical mechanics

statistics extensions of MDL (predictive information *is* a property of the data, not of the model)

What's next?

extraction separating predictive information from non–predictive using the 'relevant information' technique

physics of phase transitions, connection to subextensive statistical mechanics

statistics extensions of MDL (predictive information *is* a property of the data, not of the model)

learning unification of approaches: Bayesian, SRM, MDL, Cucker-Smale. . .

Continuation: What's next?

Continuation: What's next?

neuro- and cognitive sciences is predictive information maximization a guiding principle for animal behavior? how complex are the models we use in learning?

Continuation: What's next?

neuro- and cognitive sciences is predictive information maximization a guiding principle for animal behavior? how complex are the models we use in learning?

bioinformatics what is predictive information of natural symbolic sequences? (DNA, languages, spike trains) can we use changes in predictability for data partitioning? for model building?

Continuation: What's next?

neuro- and cognitive sciences is predictive information maximization a guiding principle for animal behavior? how complex are the models we use in learning?

bioinformatics what is predictive information of natural symbolic sequences? (DNA, languages, spike trains) can we use changes in predictability for data partitioning? for model building?

dynamical systems theory what is predictive information and complexity of various systems?