Predictive information: From definition to applications to biological systems

Ilya Nemenman

JCSB, Columbia

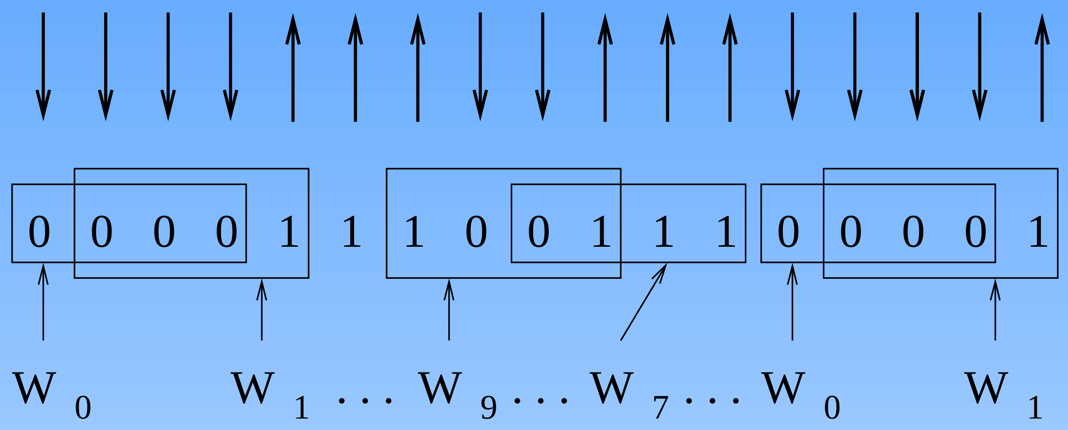
Thanks to: William Bialek, Naftali Tishby

physics/0007070
physics/0103076
q-bio/0402029

Outline

- A curious observation.
- Quantifying predictability and complexity.
- Predictability and optimization in sensory information processing.
- Learning and predictive information.
- Testing models used by animals.
- Bonus material.

Entropy of words in a spin chain



$W_0 = 0\ 0\ 0\ 0$

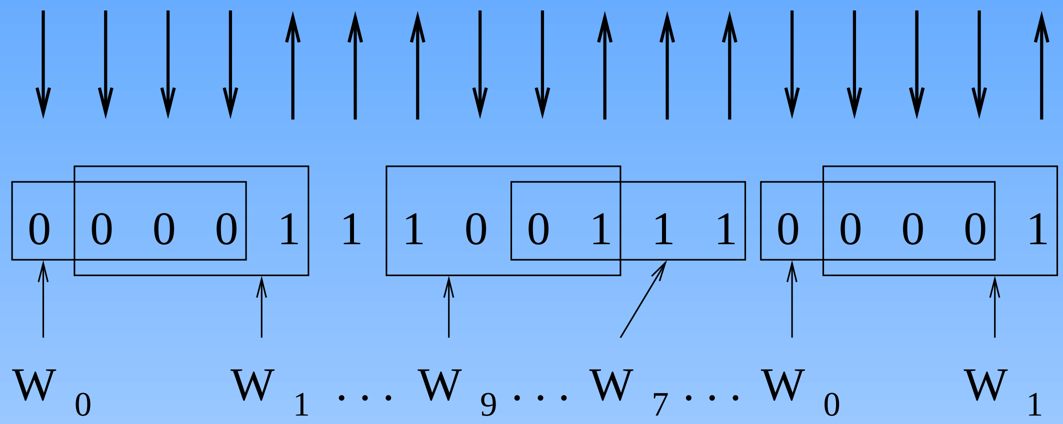
$W_1 = 0\ 0\ 0\ 1$

$W_2 = 0\ 0\ 1\ 0$

$\cdots$

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

Entropy of words in a spin chain



$W_0 = 0\ 0\ 0\ 0$

$W_1 = 0\ 0\ 0\ 1$

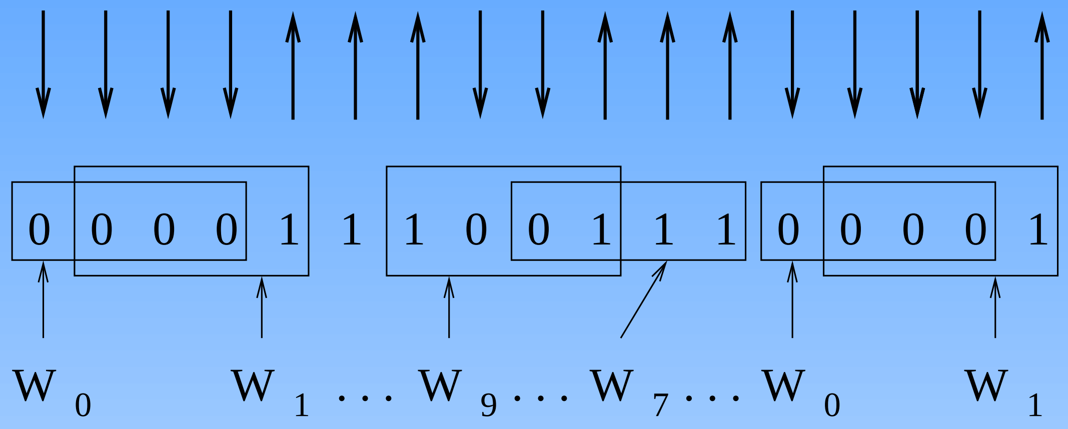
$W_2 = 0\ 0\ 1\ 0$

$\cdots$

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

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

Entropy of words in a spin chain



$W_0 = 0\ 0\ 0\ 0$

$W_1 = 0\ 0\ 0\ 1$

$W_2 = 0\ 0\ 1\ 0$

. . .

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

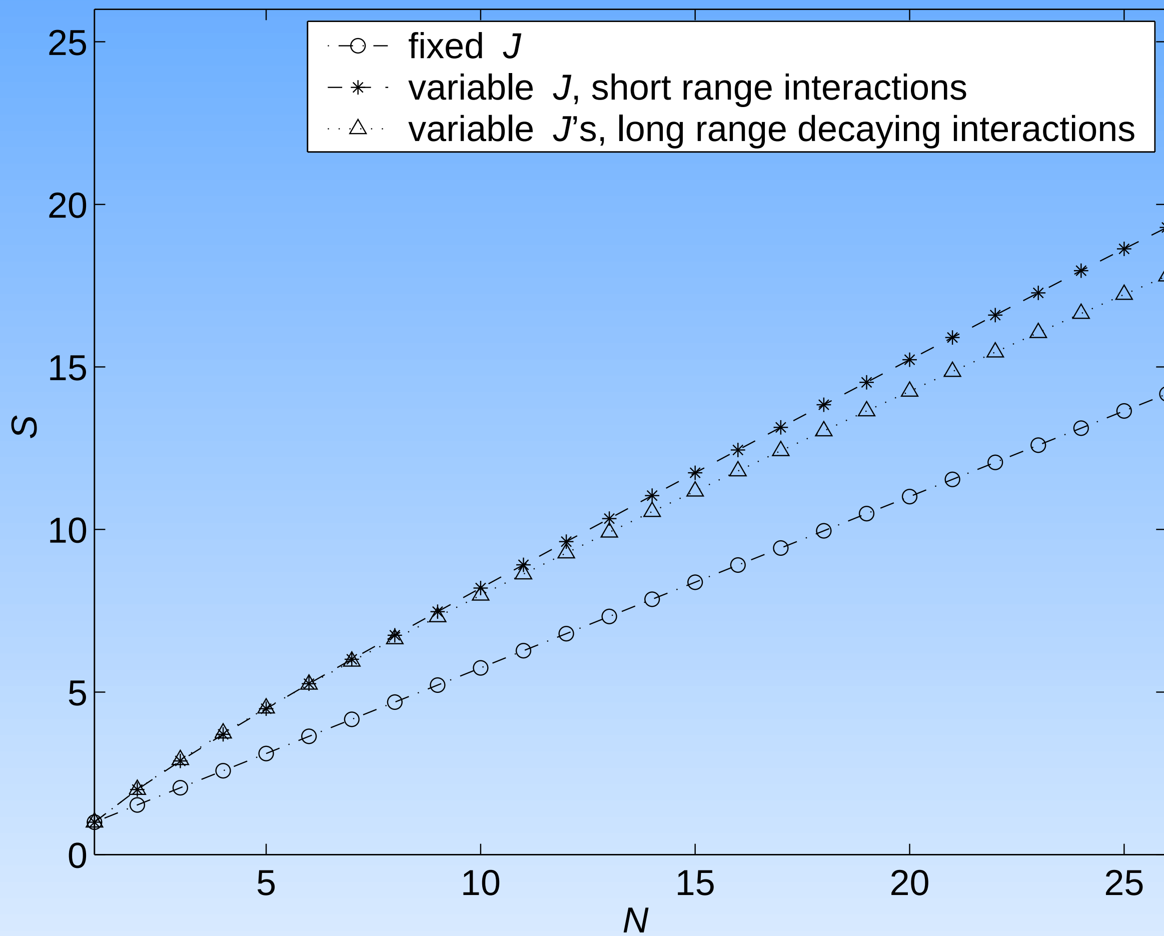
$$S(N) = -\sum_{k=0}^{2^N-1} P_N(W_k)\log_2 P_N(W_k)$$

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

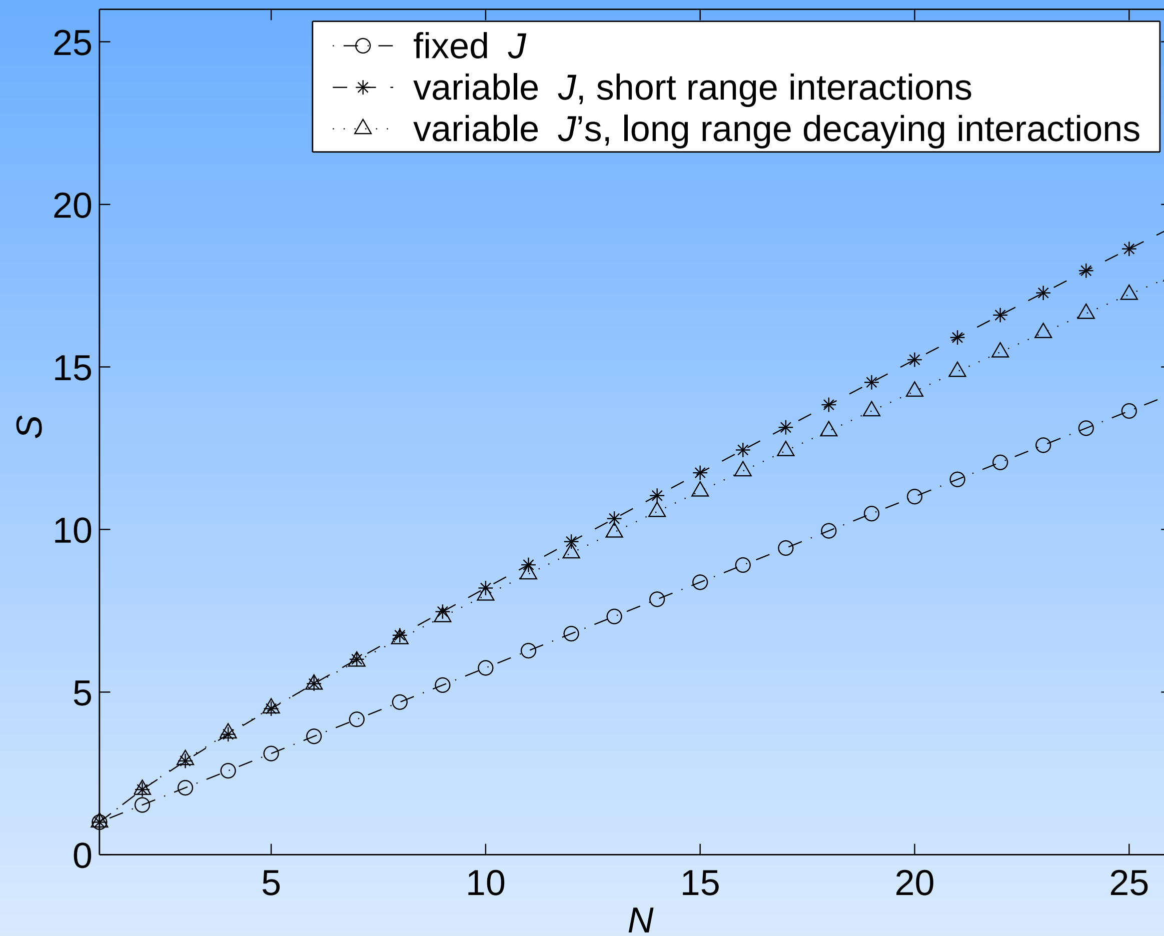
$1 \cdot 10^9$ spins total.



Entropy of 3 generated chains

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

$1 \cdot 10^9$ spins total.

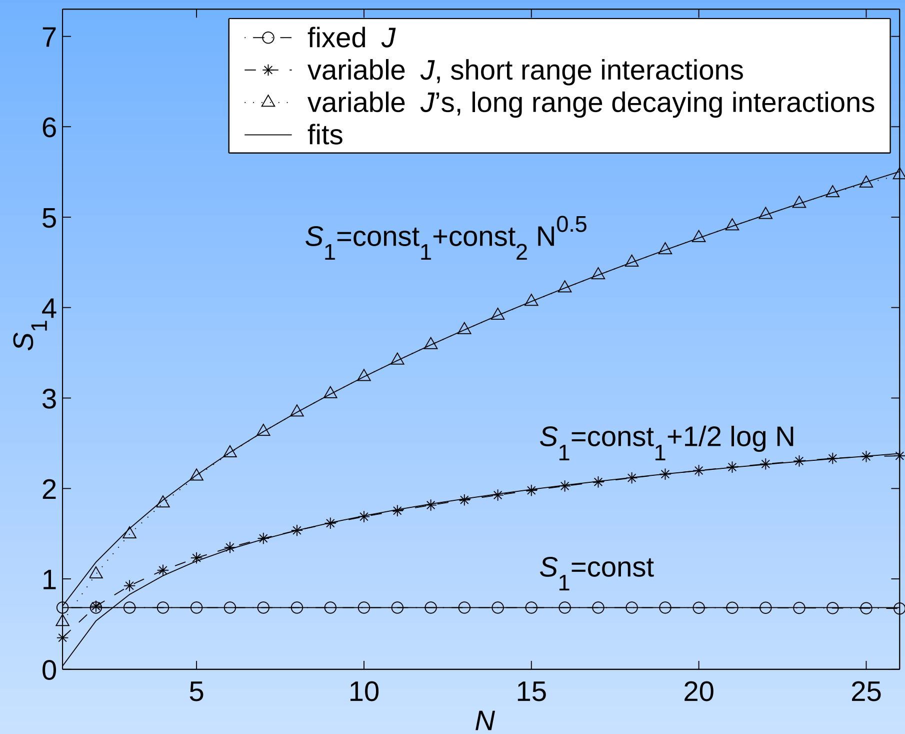


Entropy is extensive!

It shows no distinction between the cases.

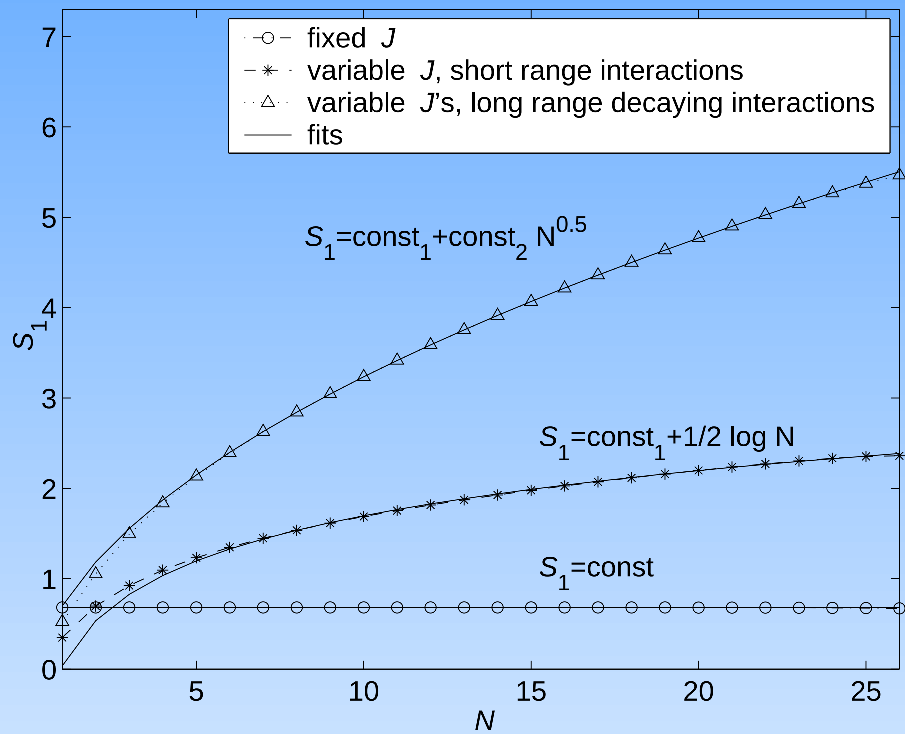
Subextensive component of the entropy

. . . shows a qualitative distinction between the cases!



Subextensive component of the entropy

. . . shows a qualitative distinction between the cases!

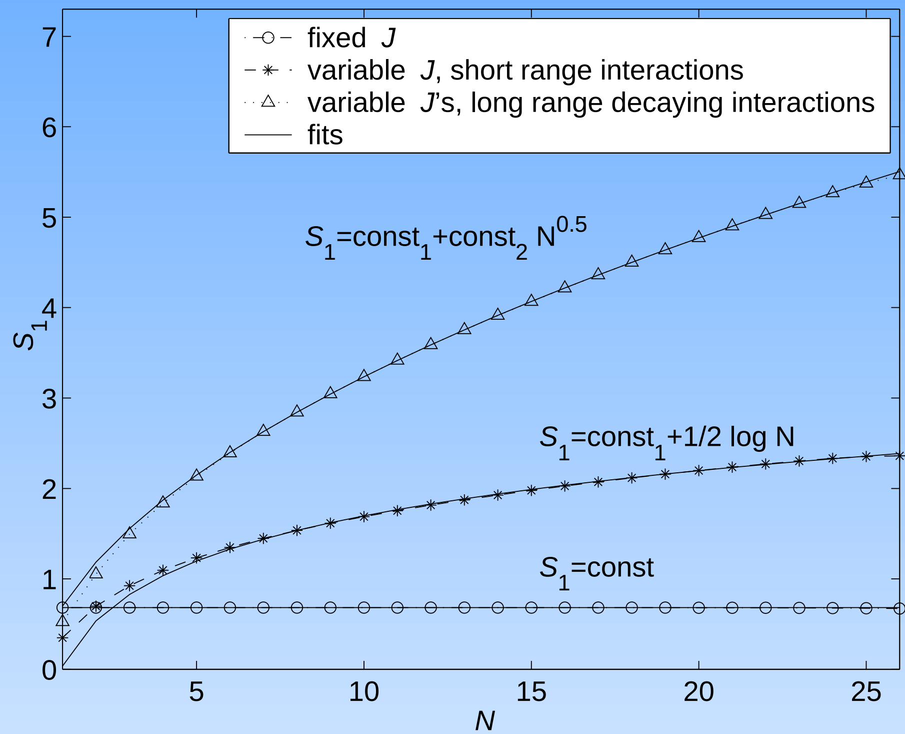


Other examples:

- **const** periodic, fully random, chaotic sequences (finite correlation length)
- **log** systems at phase transitions, or at the onset of chaos (divergent correlation length)
- **power** texts, DNA sequences, (likely) some exotic transitions, (many divergent correlation lengths)

Subextensive component of the entropy

. . . shows a qualitative distinction between the cases!



Other examples:

- **const** periodic, fully random, chaotic sequences (finite correlation length)
- **log** systems at phase transitions, or at the onset of chaos (divergent correlation length)
- **power** texts, DNA sequences, (likely) 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

- unified description of complexity and learning
- make distinction between useful and unusable data
- do this using physical quantities
- understand models used by organisms to represent the world
- understand biological designs by means of optimization principles

Solution – predictability

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

Solution – predictability

- 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*

Solution – predictability

- 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*
- learning happens on all scales in biology
- 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

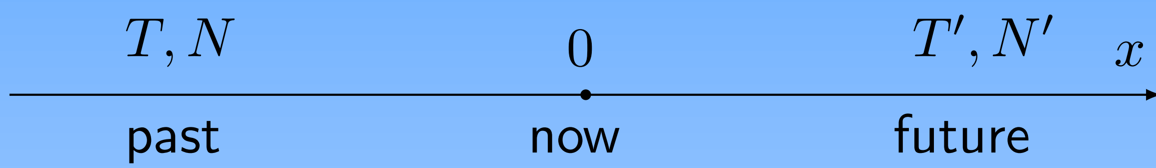
- we learn (estimate parameters, extrapolate, classify, . . .) to *generalize* and *predict* from training examples; estimation of parameters is only an intermediate step
- nonpredictive features in any signal are useless since we observe *now* and react in the *future*
- learning happens on all scales in biology
- 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)
- measuring organisms' learning and prediction performance for signals of different complexity may reveal the underlying models

Solution – predictability

- 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*
- learning happens on all scales in biology
- 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)
- measuring organisms' learning and prediction performance for signals of different complexity may reveal the underlying models
- optimizing predictive information may be the design principle

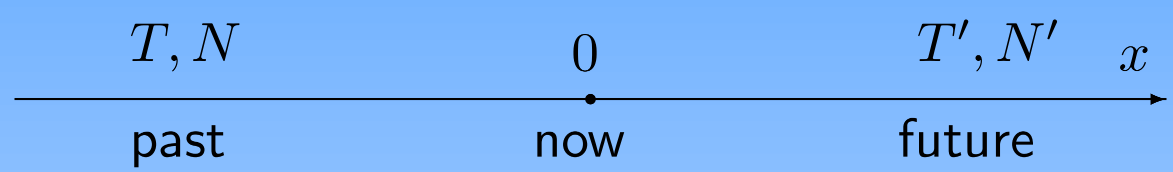
Quantifying predictability

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



Quantifying predictability

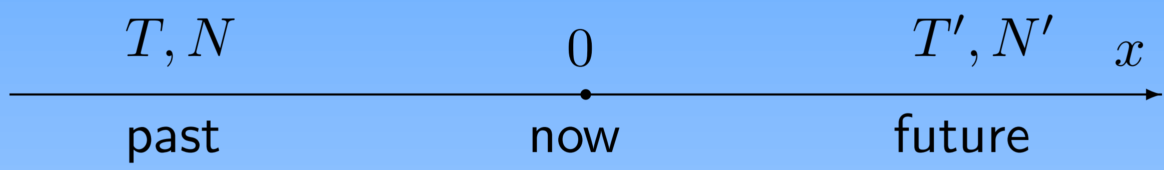
Information theory: non-metric, universal way to quantify learning



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

Quantifying predictability

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



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

Extensive component cancels in predictive information.

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

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

Extensive component cancels in predictive information.

Predictability is a deviation from extensivity!

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

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

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

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

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

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

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
  - complexity: complexity measures
  - coding: model coding length

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 $I_{\rm pred}$ behave?

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

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

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

- learning finite-parameter densities
- well known as $I(N, {\rm parameters}) = I_{\rm pred}(N)$
- physical system at criticality
- (possibly) nonextensive statistics systems

How can $I_{\rm pred}$ behave?

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

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

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

- learning finite-parameter densities
- well known as $I(N, {\rm parameters}) = I_{\rm pred}(N)$
- physical system at criticality
- (possibly) nonextensive statistics systems

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

- learning continuous densities
- language
- some critical phenomena (wetting transitions)
- not well studied

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

- 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 of complexity ...

... are mostly straightforward.

Relations to other definitions of complexity . . .

. . . are mostly straightforward.

For Kolmogorov complexity:

Relations to other definitions of complexity . . .

. . . are mostly straightforward.

For Kolmogorov complexity:

- partition all strings into equivalence classes

Relations to other definitions of complexity ...

... 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 of complexity ...

... 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 of complexity ...

... are mostly straightforward.

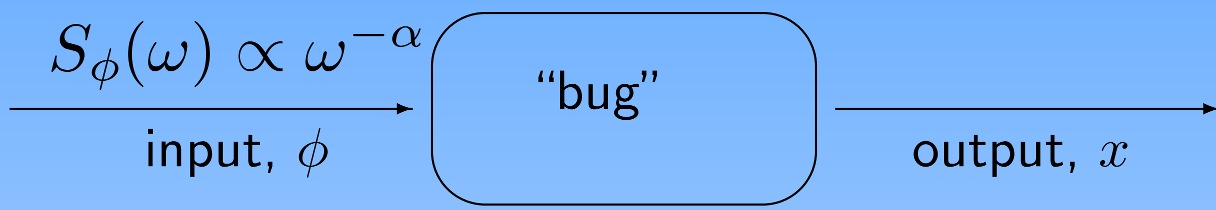
For Kolmogorov complexity:

- 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

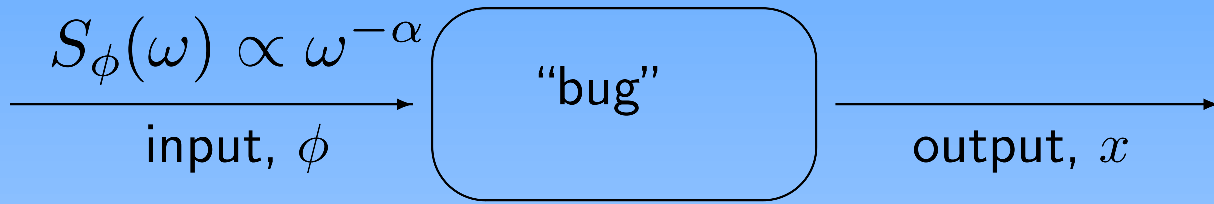
If sufficient statistics exist, then $C_K \approx I_{\rm pred}$. Otherwise $C_K > I_{\rm pred}$.

C_K is unique up to a constant.

$I_{\rm pred}$ optimization in biology

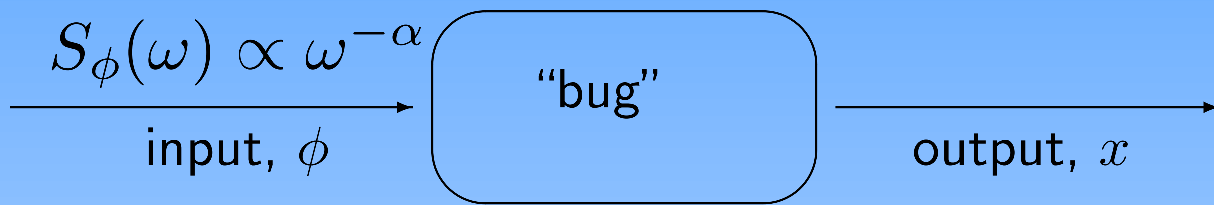


$I_{\rm pred}$ optimization in biology



$$\tau \frac{dx}{dt} = -x + \phi(t) + \eta(t), \quad \langle \eta(t)\eta(0)\rangle = 1/I_0\,\delta(t)$$

$I_{\rm pred}$ optimization in biology

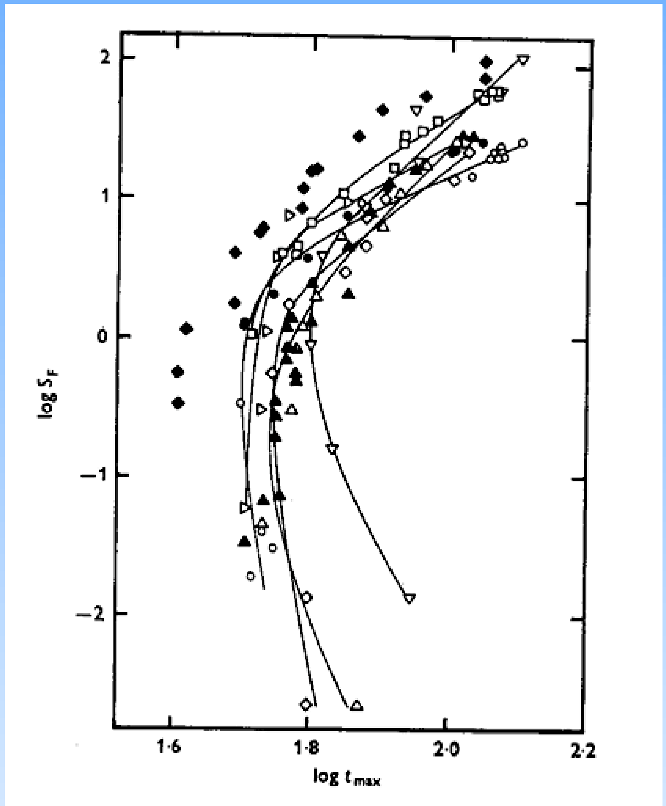


$$\tau \frac{dx}{dt} = -x + \phi(t) + \eta(t), \quad \langle \eta(t)\eta(0)\rangle = 1/I_0\,\delta(t)$$

$$\mathcal{I}([\phi],[x]) = \lim_{T\to\infty} \frac{1}{2T}\int_{-T/2}^{T/2} \frac{d\omega}{2\pi} \log\left(1 + \frac{S_\phi(\omega)}{1/I_0}\right)$$

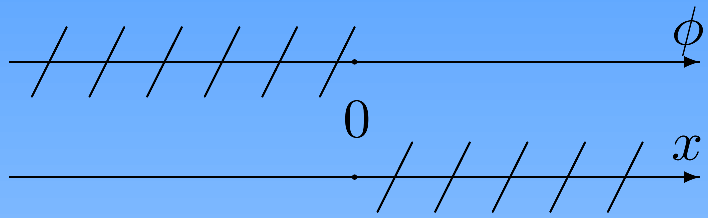
Maximization w.r.t. τ is meaningless.

I_{pred} extraction and maximization

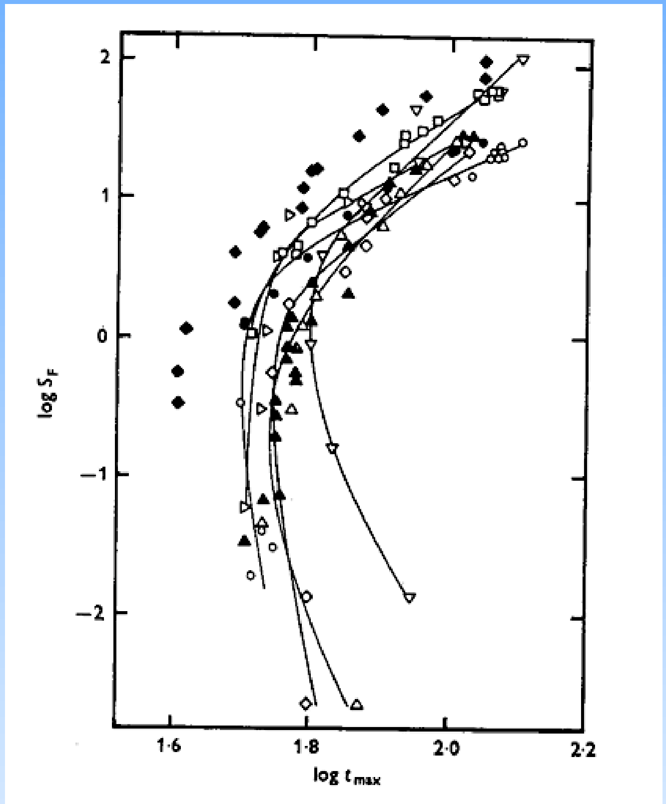


(Baylor and Hodgkin, 1974)

I_{pred} extraction and maximization

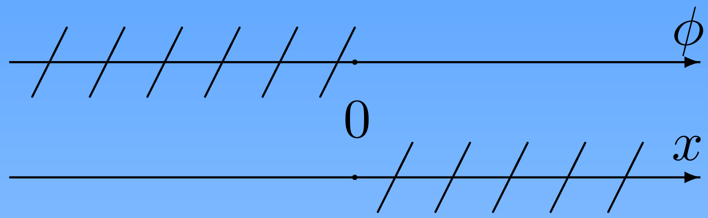


$I([x_{\text{past}}], [\phi_{\text{future}}])$ – too difficult



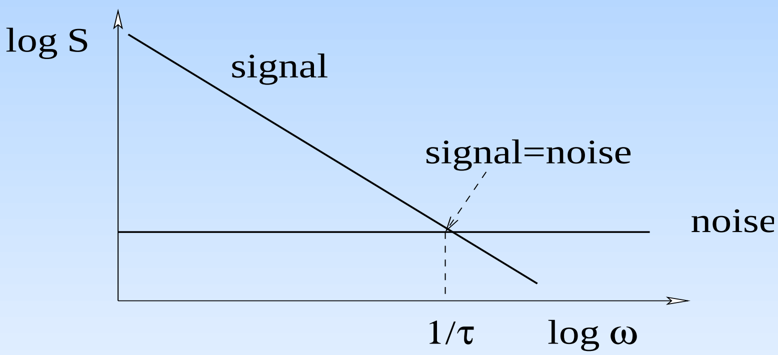
(Baylor and Hodgkin, 1974)

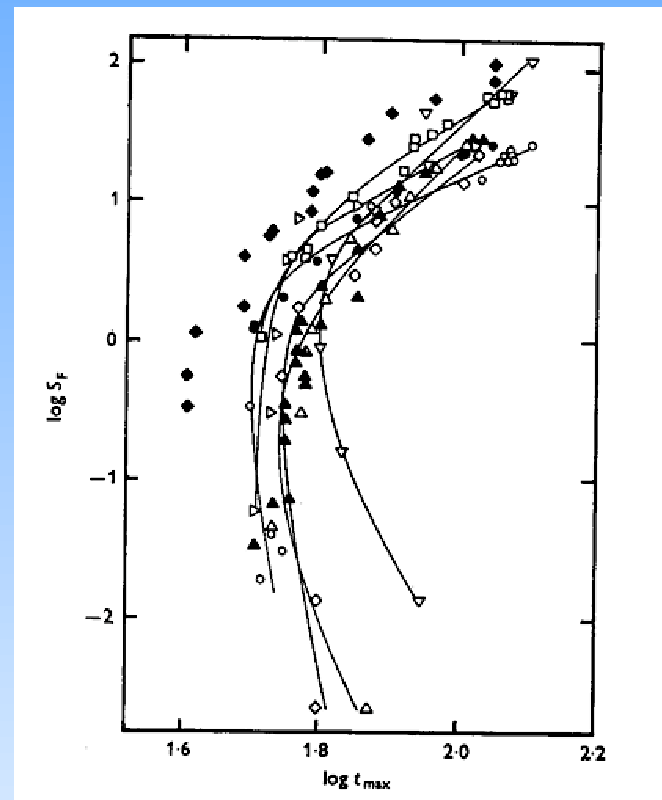
$I_{\rm pred}$ extraction and maximization



$I([x_{\rm past}],[\phi_{\rm future}])$ – too difficult

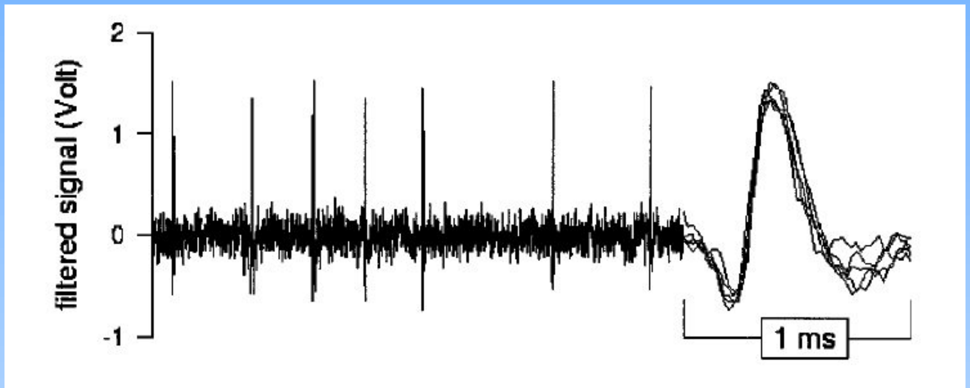
$$I(x_0,\phi_0) = \log \frac{\langle \phi^2 \rangle}{\langle \phi^2 \rangle - \frac{\langle \phi_f^2 \rangle^2}{\langle x^2 \rangle}}$$



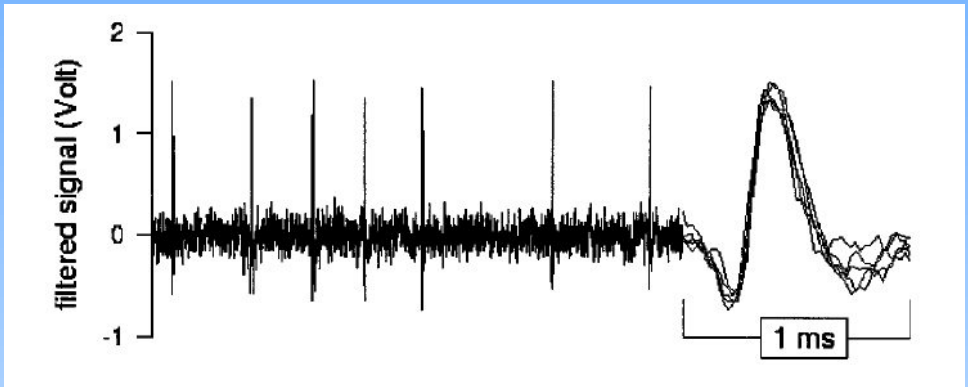


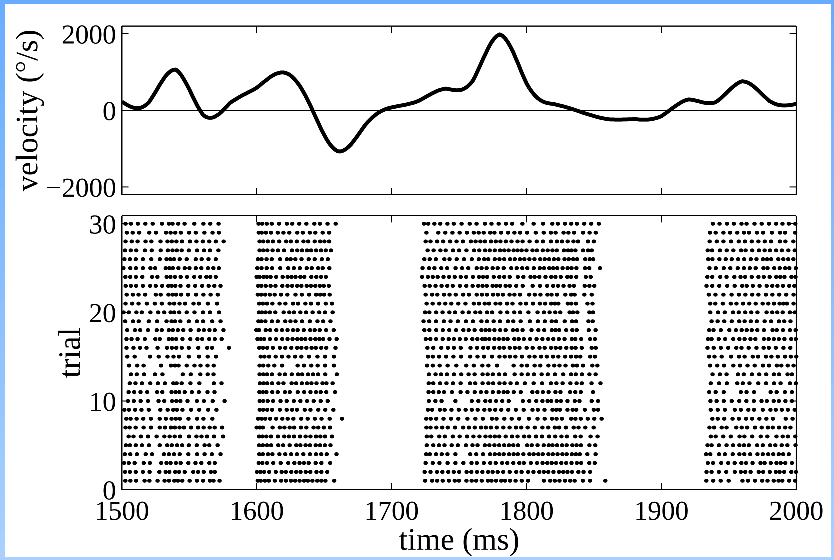
(Baylor and Hodgkin, 1974)

Fly H1 predictive information

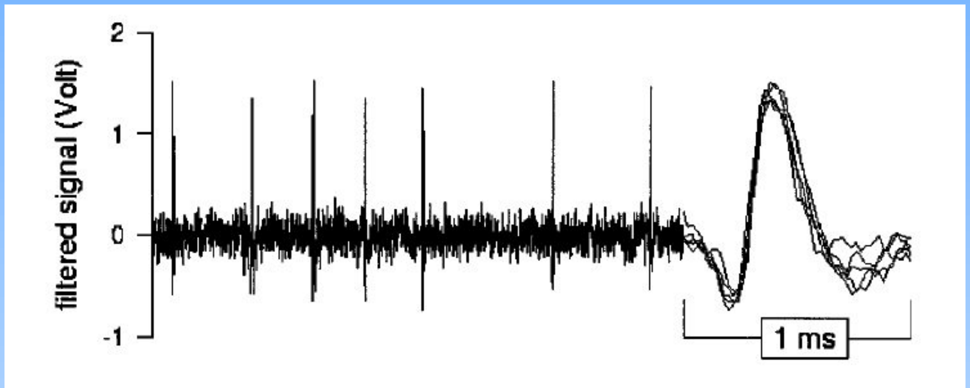


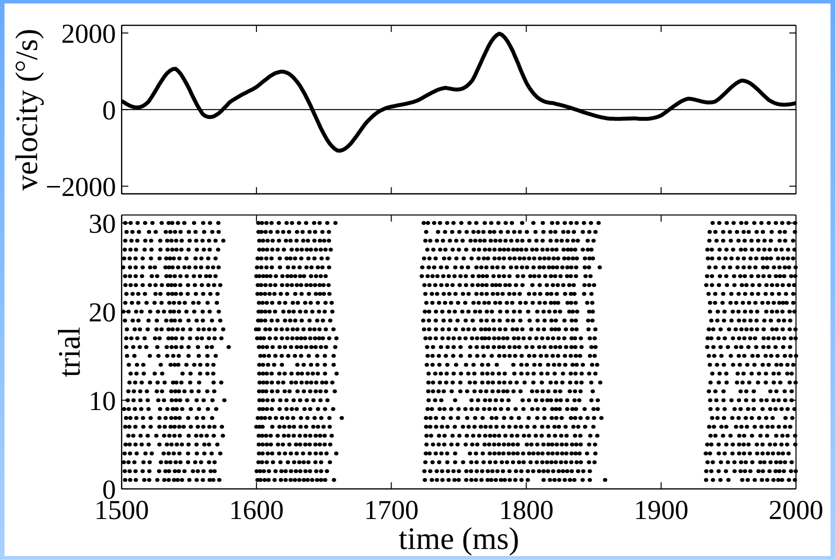
Fly H1 predictive information





Fly H1 predictive information





Estimate $I(\text{spikes}_{\text{past}}, v_{\text{future}})$.

Close to maximum!

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

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

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

Density of states

$$\mathcal{E}_N \equiv \frac{1}{N}\sum_i \log\left[\frac{Q(\vec{x}_\mathrm{i}|\bar{\boldsymbol{\alpha}})}{Q(\vec{x}_\mathrm{i}|\boldsymbol{\alpha})}\right] \xrightarrow{\text{anneal}} \epsilon = \int d\vec{x}\, Q(\vec{x}|\bar{\boldsymbol{\alpha}}) \log\frac{Q(\vec{x}|\bar{\boldsymbol{\alpha}})}{Q(\vec{x}|\boldsymbol{\alpha})}$$

Density of states

$$
\begin{aligned}
\mathcal{E}_N &\equiv \frac{1}{N}\sum_i \log\left[\frac{Q(\vec{x}_{\rm i}|\bar{\boldsymbol{\alpha}})}{Q(\vec{x}_{\rm i}|\boldsymbol{\alpha})}\right] \xrightarrow{\text{anneal}} \epsilon = \int d\vec{x}\, Q(\vec{x}|\bar{\boldsymbol{\alpha}}) \log\frac{Q(\vec{x}|\bar{\boldsymbol{\alpha}})}{Q(\vec{x}|\boldsymbol{\alpha})} \\
Z(\bar{\boldsymbol{\alpha}}; N) &= \int d\epsilon\, \rho(\epsilon; \bar{\boldsymbol{\alpha}}) \exp[-N\epsilon] \\
\rho(\epsilon; \bar{\boldsymbol{\alpha}}) &= \int d^K\alpha\, \mathcal{P}(\boldsymbol{\alpha}) \delta[\epsilon - D_{\rm KL}(\bar{\boldsymbol{\alpha}}||\boldsymbol{\alpha})]
\end{aligned}
$$

Density of states

$$\mathcal{E}_N \equiv \frac{1}{N}\sum_i \log\left[\frac{Q(\vec{x}_{\rm i}|\bar{\boldsymbol{\alpha}})}{Q(\vec{x}_{\rm i}|\boldsymbol{\alpha})}\right] \xrightarrow{\text{anneal}} \epsilon = \int d\vec{x}\, Q(\vec{x}|\bar{\boldsymbol{\alpha}}) \log \frac{Q(\vec{x}|\bar{\boldsymbol{\alpha}})}{Q(\vec{x}|\boldsymbol{\alpha})}$$

$$Z(\bar{\boldsymbol{\alpha}}; N) = \int d\epsilon\, \rho(\epsilon; \bar{\boldsymbol{\alpha}}) \exp[-N\epsilon]$$

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

$$\int d\epsilon\, \rho(\epsilon; \bar{\boldsymbol{\alpha}}) = \int d^K\alpha\, \mathcal{P}(\boldsymbol{\alpha}) = 1 \quad \text{annealing works!}$$

Density of states

$$\mathcal{E}_N \equiv \frac{1}{N}\sum_i \log\left[\frac{Q(\vec{x}_{\rm i}|\bar{\boldsymbol{\alpha}})}{Q(\vec{x}_{\rm i}|\boldsymbol{\alpha})}\right] \xrightarrow{\text{anneal}} \epsilon = \int d\vec{x}\, Q(\vec{x}|\bar{\boldsymbol{\alpha}}) \log\frac{Q(\vec{x}|\bar{\boldsymbol{\alpha}})}{Q(\vec{x}|\boldsymbol{\alpha})}$$

$$Z(\bar{\boldsymbol{\alpha}}; N) = \int d\epsilon\, \rho(\epsilon; \bar{\boldsymbol{\alpha}}) \exp[-N\epsilon]$$

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

$$\int d\epsilon\, \rho(\epsilon; \bar{\boldsymbol{\alpha}}) = \int d^K\alpha\, \mathcal{P}(\boldsymbol{\alpha}) = 1 \quad \text{annealing works!}$$

Annealed approximation (almost) always works.
Learning is annealing at decreasing temperature.
Nonzero $\rho \Longrightarrow$ consistency in learning.

Density at $\epsilon \to 0$, $I_{\rm pred}$, and learning

Occam factor, generalization error, prediction error, fluctuation determinant:

$$\mathcal{D}(\bar{\boldsymbol{\alpha}}; N) \approx -\log \int d\epsilon\, \rho(\epsilon; \bar{\boldsymbol{\alpha}}) \mathrm{e}^{-N\epsilon}$$

Density at $\epsilon \to 0$, I_{pred}, and learning

Occam factor, generalization error, prediction error, fluctuation determinant:

$$\mathcal{D}(\bar{\boldsymbol{\alpha}}; N) \approx -\log \int d\epsilon\, \rho(\epsilon; \bar{\boldsymbol{\alpha}}) \mathrm{e}^{-N\epsilon}$$

Predictive information:

$$I_{\mathrm{pred}}(N) \approx \int d^K \bar{\alpha}\, \mathcal{P}(\bar{\boldsymbol{\alpha}})\, \mathcal{D}(\bar{\boldsymbol{\alpha}}, N)$$

Density at $\epsilon \to 0$, $I_{\rm pred}$, and learning

Occam factor, generalization error, prediction error, fluctuation determinant:

$$\mathcal{D}(\bar{\boldsymbol{\alpha}}; N) \approx -\log \int d\epsilon\, \rho(\epsilon; \bar{\boldsymbol{\alpha}}) \mathrm{e}^{-N\epsilon}$$

Predictive information:

$$I_{\rm pred}(N) \approx \int d^K \bar{\alpha}\, \mathcal{P}(\bar{\boldsymbol{\alpha}})\, \mathcal{D}(\bar{\boldsymbol{\alpha}}, N)$$

Universal learning curves:

$$\begin{aligned} \Lambda(\bar{\boldsymbol{\alpha}}; N) &\equiv D_{\rm KL}(\bar{\boldsymbol{\alpha}} || \boldsymbol{\alpha}_{\rm est}) \approx \frac{d\mathcal{D}(\bar{\boldsymbol{\alpha}}; N)}{dN} \\ \Lambda(N) &\equiv \int d\bar{\boldsymbol{\alpha}}\, \mathcal{P}(\bar{\boldsymbol{\alpha}}) \Lambda(\bar{\boldsymbol{\alpha}}; N) \approx \frac{dI_{\rm pred}}{dN} \end{aligned}$$

Finite number of states and finite $I_{\rm pred}$

$$\rho(\epsilon; a_i) \quad = \quad \sum_{j=1}^{M} \mathcal{P}_j \, \delta(d_{ij} - \epsilon)$$

Finite number of states and finite $I_{\rm pred}$

$$\begin{aligned}
\rho(\epsilon; a_i) &= \sum_{j=1}^{M} \mathcal{P}_j \, \delta(d_{ij} - \epsilon) \\
\mathcal{D}(a_i; N) &= c_1 - c_2 \, \exp[-N c_3] \\
\Lambda(a_i; N) &\approx c_2 c_3 \exp[-N c_3]
\end{aligned}$$

$I_{\rm pred}$ saturates as $N \to \infty$

Power–law density function

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

Power–law density function

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

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

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

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

Power–law density function

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

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

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

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

Speed of approach to this asymptotics is rarely investigated.

Another example

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

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

Another example

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

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

Another example

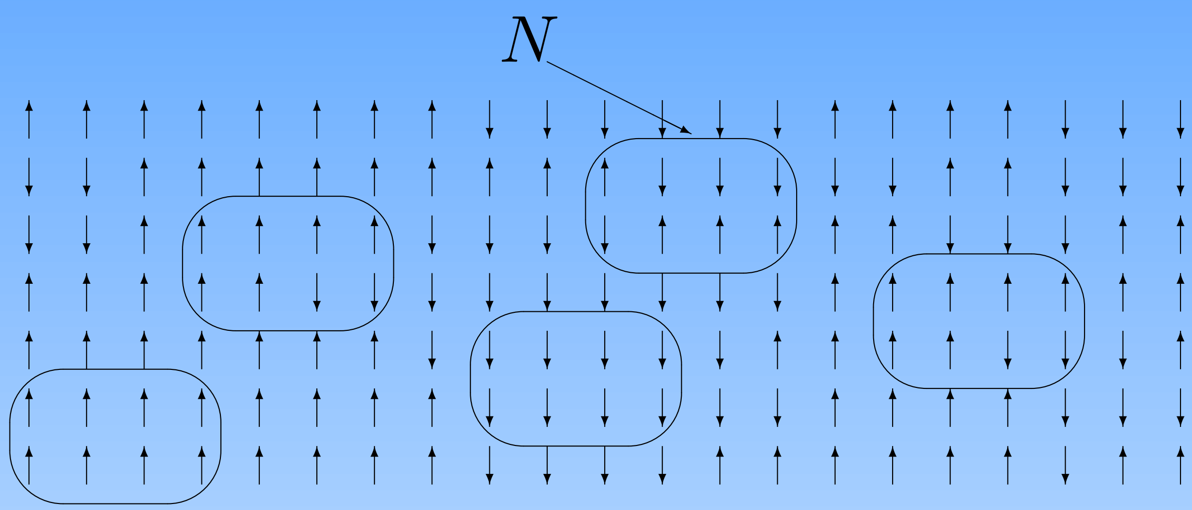
Learning $Q(\vec{x}_1 \cdots \vec{x}_{\rm N}|\boldsymbol{\alpha})$, a finite parameter Markov process with long range intrinsic correlations such that

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

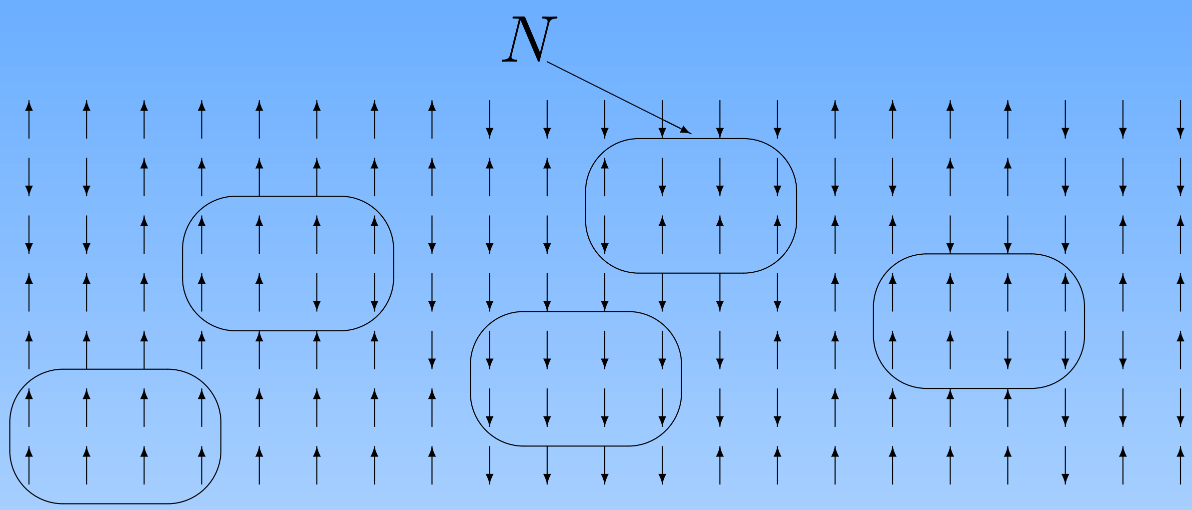
Do not distinguish predictability from unknown parameters and from intrinsic correlations.

In physics similar to: order parameters $\Longleftrightarrow$ interactions.

RG, not finite size scaling!



RG, not finite size scaling!



$$S(N) = S(\text{block}) + S(\text{spin}|\text{block})$$

Scaling fields carry information across.

Is $I_{\text{pred}} = f(\text{scaling exponents}) \log N$?

Essential singularity in the density

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

$$S_1^{(\mathrm{a})}(N) \propto N^{\mu/(\mu+1)}$$

Essential singularity in the density

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

$$S_1^{(\mathrm{a})}(N) \propto N^{\mu/(\mu+1)}$$

- 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 $I_{\rm pred}$

Learning a smooth nonparameteric density $Q(x) = 1/l_0 \mathrm{e}^{-\phi(x)}$, $x \in [0, L]$ (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 $I_{\rm pred}$

Learning a smooth nonparameteric density $Q(x) = 1/l_0 {\rm e}^{-\phi(x)}$, $x \in [0, L]$ (Bialek, Callan, and Strong 1996).

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

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

$$S_1^{(\rm a)}(N) \propto \sqrt{N}\left(\frac{L}{l}\right)^{1/2}$$

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

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

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

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

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

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

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

Learning a smooth nonparameteric density $Q(x) = 1/l_0 {\rm e}^{-\phi(x)}$, $x \in [0, L]$ (Bialek, Callan, and Strong 1996).

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

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

$$S_1^{(\rm a)}(N) \propto \sqrt{N}\left(\frac{L}{l}\right)^{1/2}$$

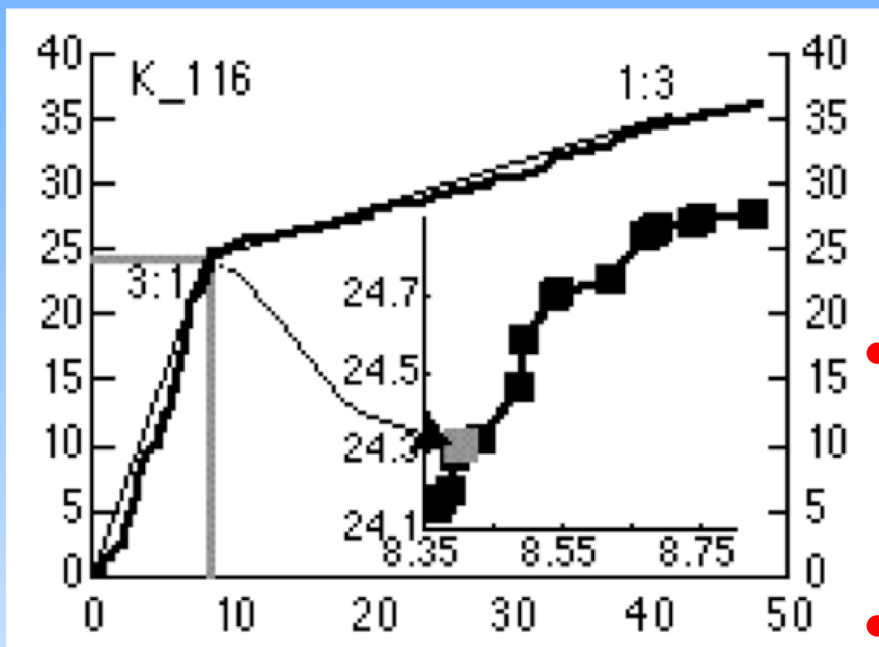
- increasing number of "effective parameters" (bins) of adaptive size $\sim \sqrt{l/NQ(x)}$
- heuristic arguments for the dimensionality ζ and the smoothness exponent η give $S_1(N) \sim N^{\zeta/2\eta}$ — demonstrates a crossover from complexity to randomness

Which model is being used?

for QFT or nested asymptotics kicks in fast

- asymptotic decay rate should signify the model
-

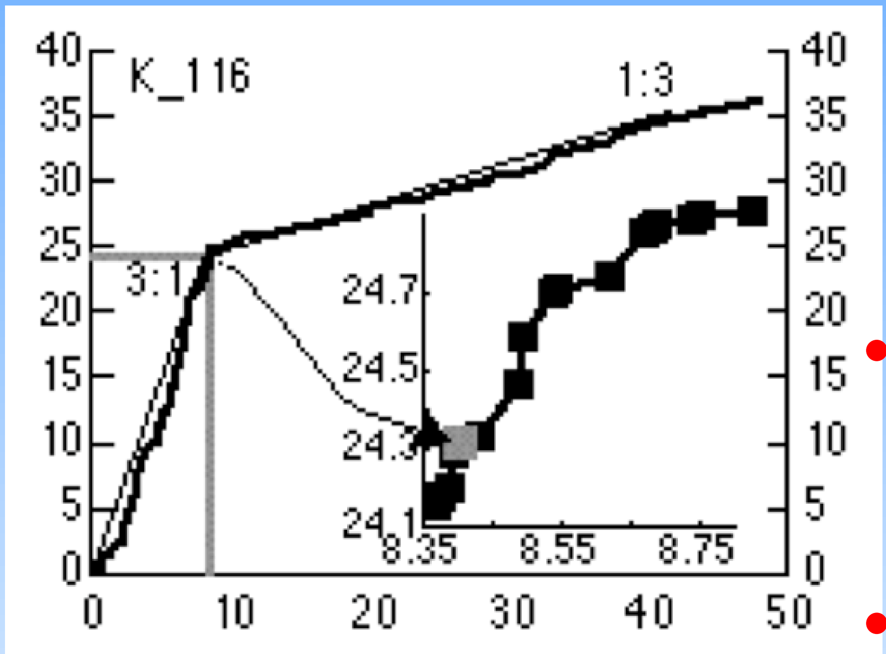
Which model is being used?



(Gallistel et al., 2001)

- for QFT or nested asymptotics kicks in fast
- asymptotic decay rate should signify the model
- decay rate too fast to observe
- noisy learning

Which model is being used?

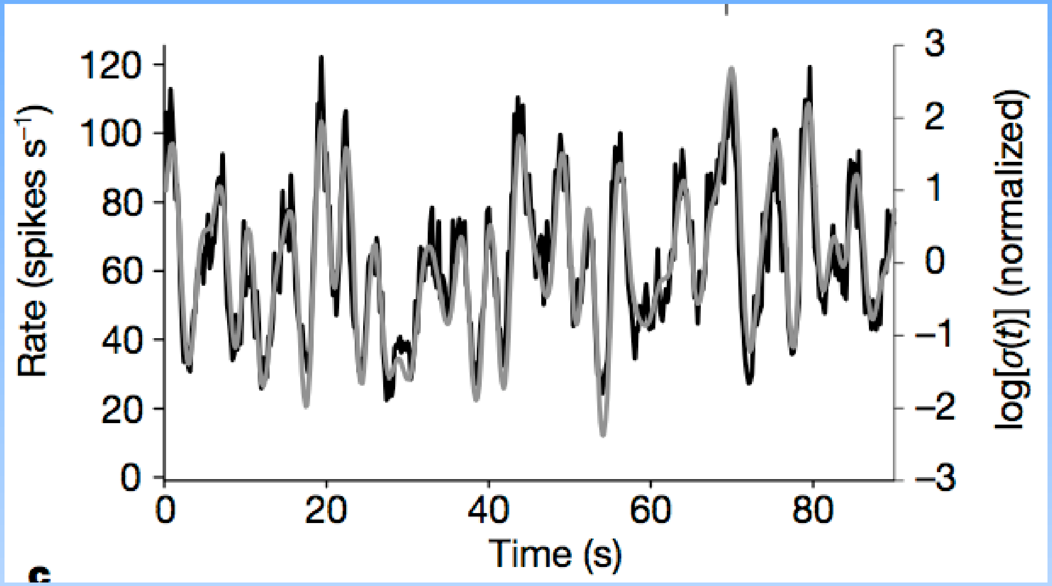


(Gallistel et al., 2001)

- for QFT or nested asymptotics kicks in fast
- asymptotic decay rate should signify the model
- decay rate too fast to observe
- noisy learning

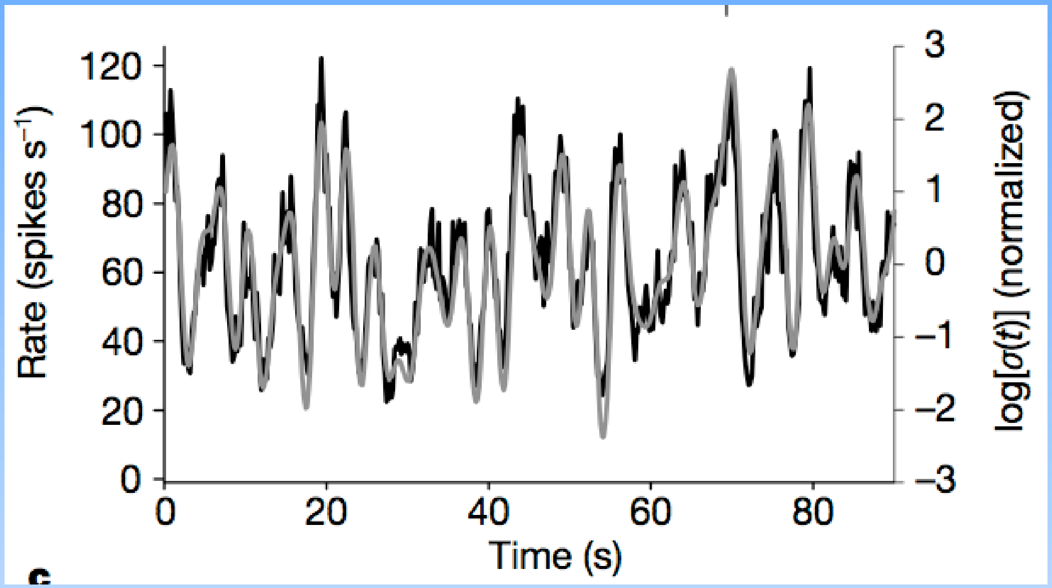
maybe FDT? $\frac{\partial \Lambda}{\partial N} = -\zeta_N \Lambda^\nu$

Fluctuations (drifting target) and dissipation (learning curve)



(Fairhall et al., 2001)

Fluctuations (drifting target) and dissipation (learning curve)



(Fairhall et al., 2001)

$$\Delta_{\rm rms} = \left\{ \nu^{1/\nu} \frac{\Gamma\left(\frac{3}{2\nu}\right)}{\Gamma\left(\frac{1}{2\nu}\right)} \right\}^{1/2} \left(\frac{\Omega}{\zeta}\right)^{1/(2\nu)}$$

What's next?

extraction separating predictive information from non–predictive using the *Information Bottleneck* technique

physics of phase transitions, connection to subextensive statistical mechanics

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

biology what is predictive information of natural symbolic sequences (DNA, languages, spike trains)? animal behavior? can we understand molecular circuits in terms of learning (extracting $I_{\rm pred}$?

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