Entropy and Shannon information

Entropy and Shannon information

For a random variable X with distribution $p(x)$, the **entropy** is

$$H[X] = - \Sigma_x \, p(x) \log_2 p(x)$$

Information is defined as

$$I[X] = - \log_2 p(x)$$

Mutual information

Typically, "information" = *mutual information*:

how much knowing the value of one random variable r (the response) reduces uncertainty about another random variable s (the stimulus).

Variability in response is due both to different **stimuli** and to **noise**.
How much response variability is "useful", i.e. can represent different messages, depends on the noise. Noise can be specific to a given stimulus.

Mutual information

Information quantifies how *independent* r and s are:

$$I(S;R) = D_{KL}\,[P(R,S),\, P(R)P(S)]$$

Alternatively:

$$I(S;R) = H[R] - \Sigma_s\, P(s)\, H[R|s]\ .$$

Mutual information

Mutual information is the difference between
the total response entropy and the mean noise entropy:

$$I(S;R) = H[R] - \Sigma_s P(s)\, H[R|s)]\ .$$

→ Need to know the conditional distribution P(s|r) or P(r|s).

Take a particular stimulus $s{=}s_0$ and repeat many times to obtain $P(r|s_0)$.
Compute variability due to noise: *noise entropy*

Mutual information

Information is symmetric in r and s

Extremes:

1. response is unrelated to stimulus: $p[r|s]$ = ?, MI = ?

2. response is perfectly predicted by stimulus: $p[r|s]$ = ?

Simple example

r_+ encodes stimulus +, r_- encodes stimulus -

but with a probability of error:

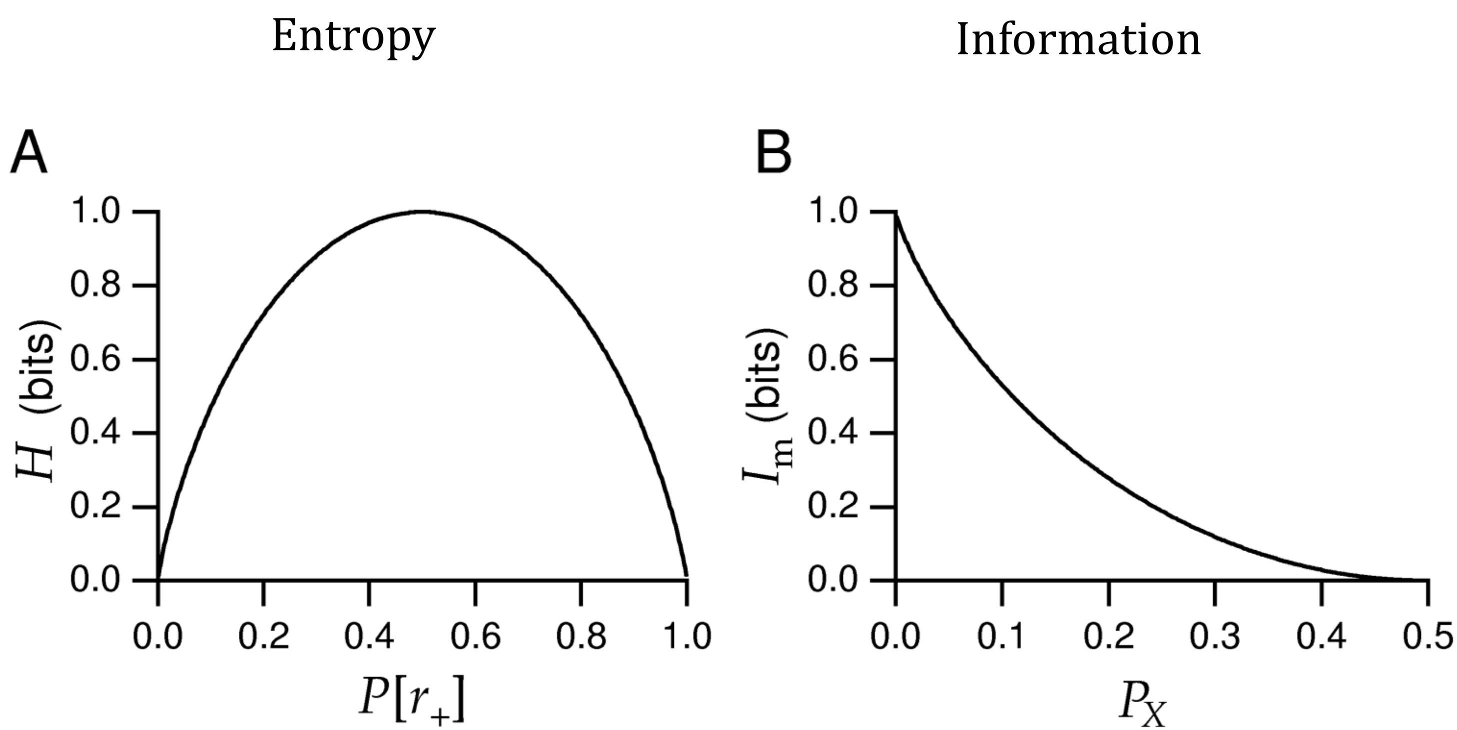
$P(r_+|+) = 1- p$

$P(r_-|-) = 1- p$

What is the response entropy H[r]?

What is the noise entropy?

Entropy and Shannon information



$$H[r] = -p_+ \log p_+ - (1-p_+)\log(1-p_+)$$

When $p_+ = ½$,

$$H[r|s] = -p \log p - (1-p)\log(1-p)$$

Noise limits information

Channel capacity

A communication channel $S \rightarrow R$ is defined by $P(R|S)$

$$I(S;R) = \Sigma_{s,r} P(s)\, P(r|s) \log[\, P(r|s)/P(r) \,]$$

The **channel capacity** gives an upper bound on transmission through the channel:

$$C(R|S) = \sup I(S;R)$$

Source coding theorem

Perfect decodability through the channel:

$$T \xrightarrow{\text{encode}} S \xrightarrow{\text{transmit}} R \xrightarrow{\text{decode}} T'$$

If the entropy of T is less than the channel capacity, then T’ can be perfectly decoded to recover T.

Data processing inequality

Transform S by some function F(S):

$$R \xrightarrow{\text{encode}} S \xrightarrow{\text{transmit}} F(S)$$

The transformed variable F(S) cannot contain more information about R than S.

Calculating information in spike trains

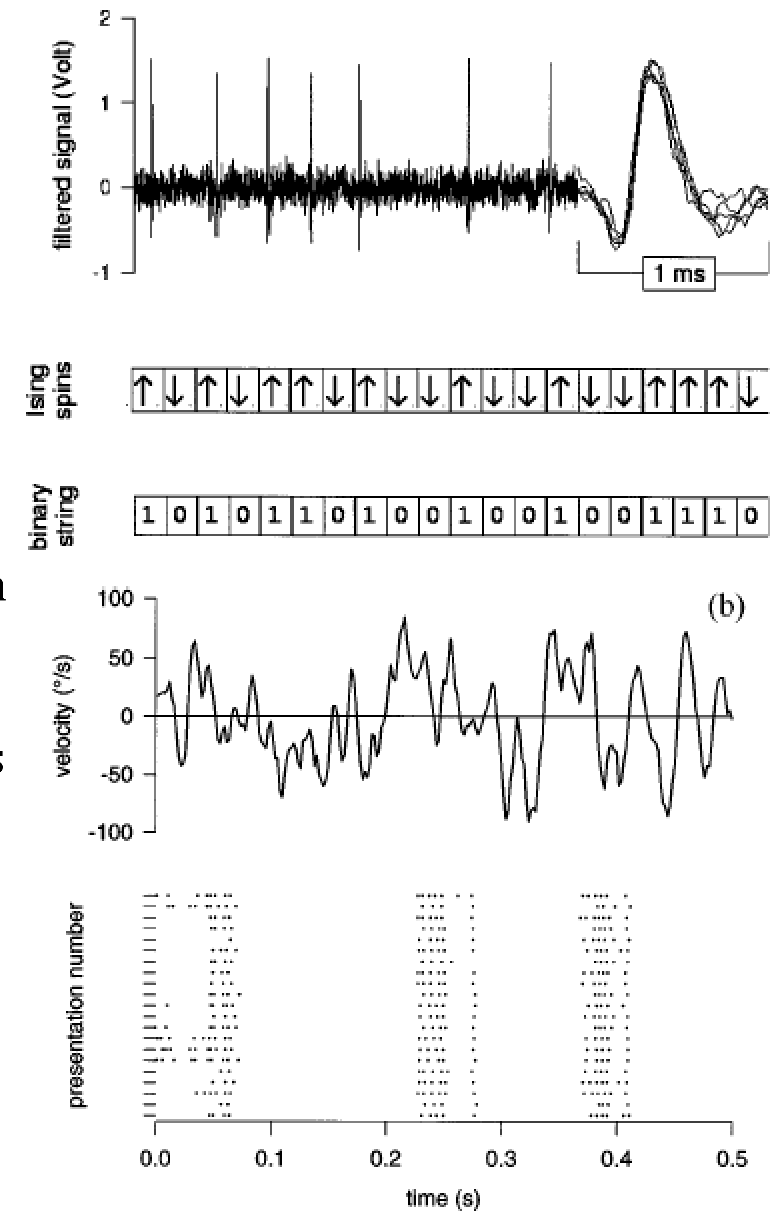
How can one compute the entropy and information of spike trains?

Entropy:

Discretize the spike train into binary words *w* with letter size Δt, length T. This takes into account correlations between spikes on timescales TΔt.

Compute $p_i = p(w_i)$, then the naïve entropy is

$$S_{\rm naive}(T, \Delta\tau; \text{size}) = -\sum_i \tilde{p}_i \log_2 \tilde{p}_i \; ;$$



Strong et al., 1997; Panzeri et al.

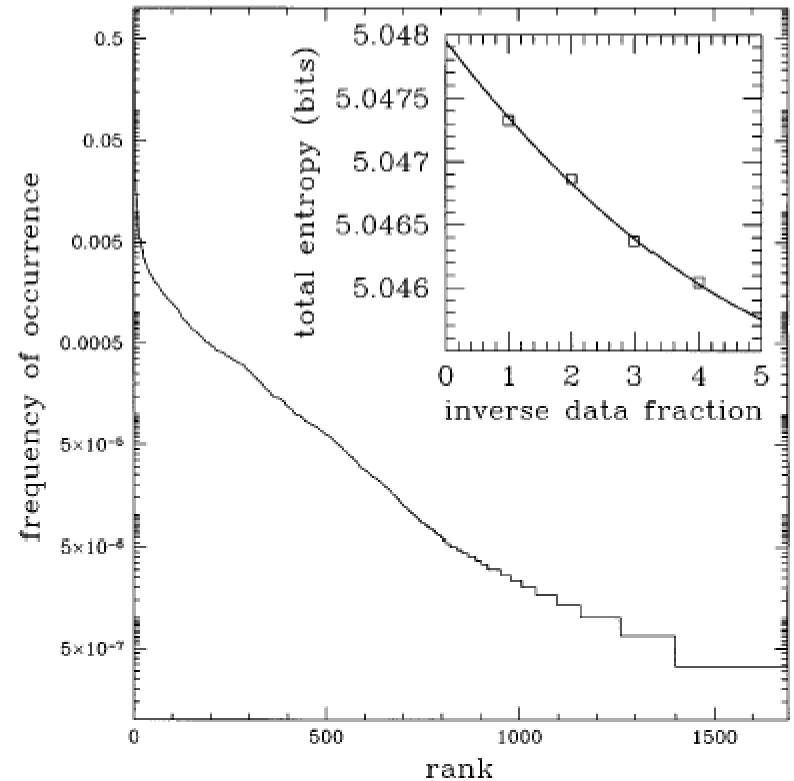
Calculating information in spike trains

Many information calculations are limited by sampling: hard to determine P(w) and P(w|s)

Systematic bias from undersampling.

Correction for finite size effects:

$$S_{\mathrm{naive}}(T, \Delta\tau; \mathrm{size}) = S(T, \Delta\tau) + \frac{S_1(T, \Delta\tau)}{\mathrm{size}} + \frac{S_2(T, \Delta\tau)}{\mathrm{size}^2}.$$



Strong et al., 1997

Calculating information in spike trains

Information : difference between the variability driven by stimuli and that due to noise.

Take a stimulus sequence *s* and repeat many times.

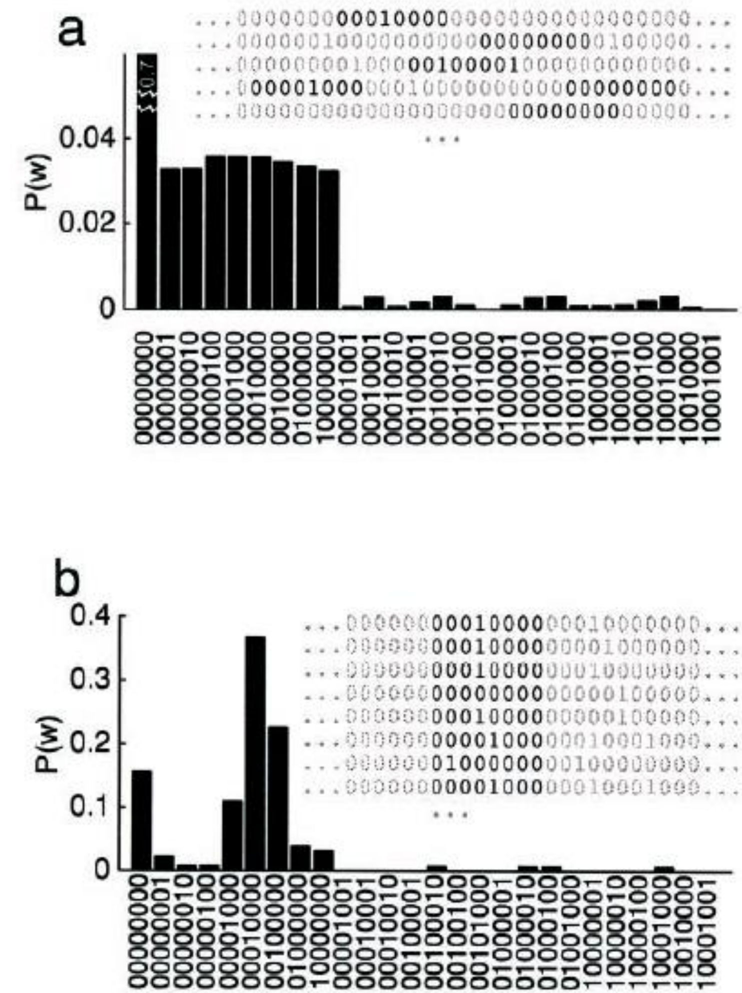
For each time in the repeated stimulus, get a set of words P(*w*|*s*(t)).

Average over *s* → average over time:

$$H_{noise} = \langle H[P(w|s_i)] \rangle_i.$$

Choose length of repeated sequence long enough to sample the noise entropy adequately.

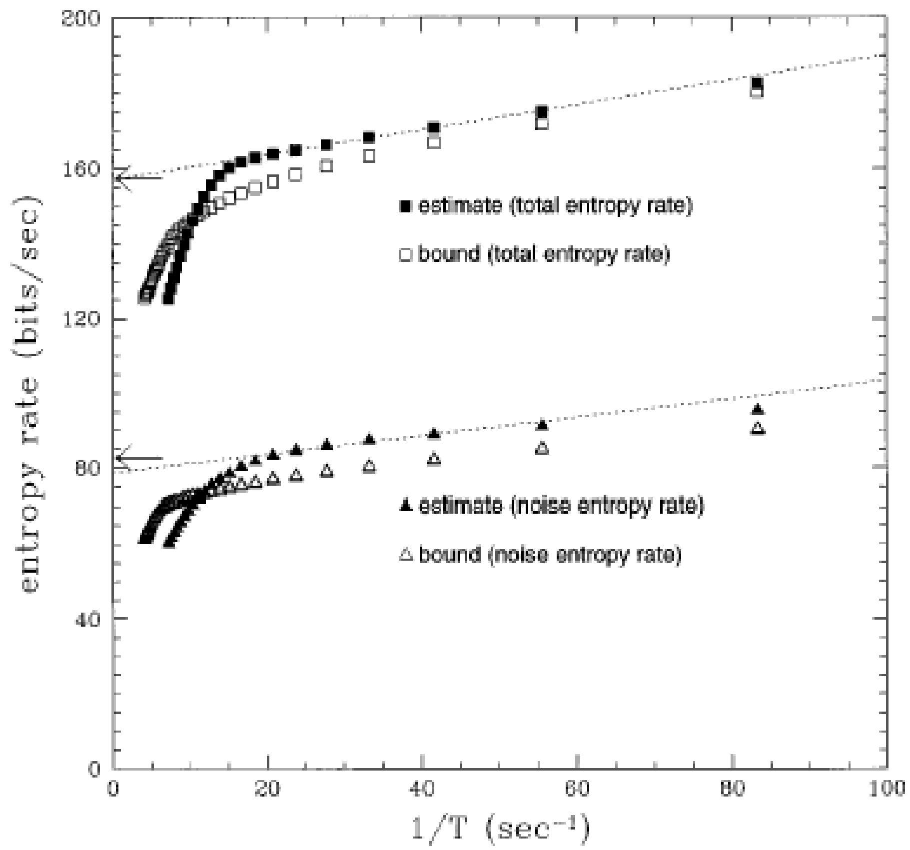
Finally, do as a function of word length T and extrapolate to infinite T.



Reinagel and Reid, ‘00

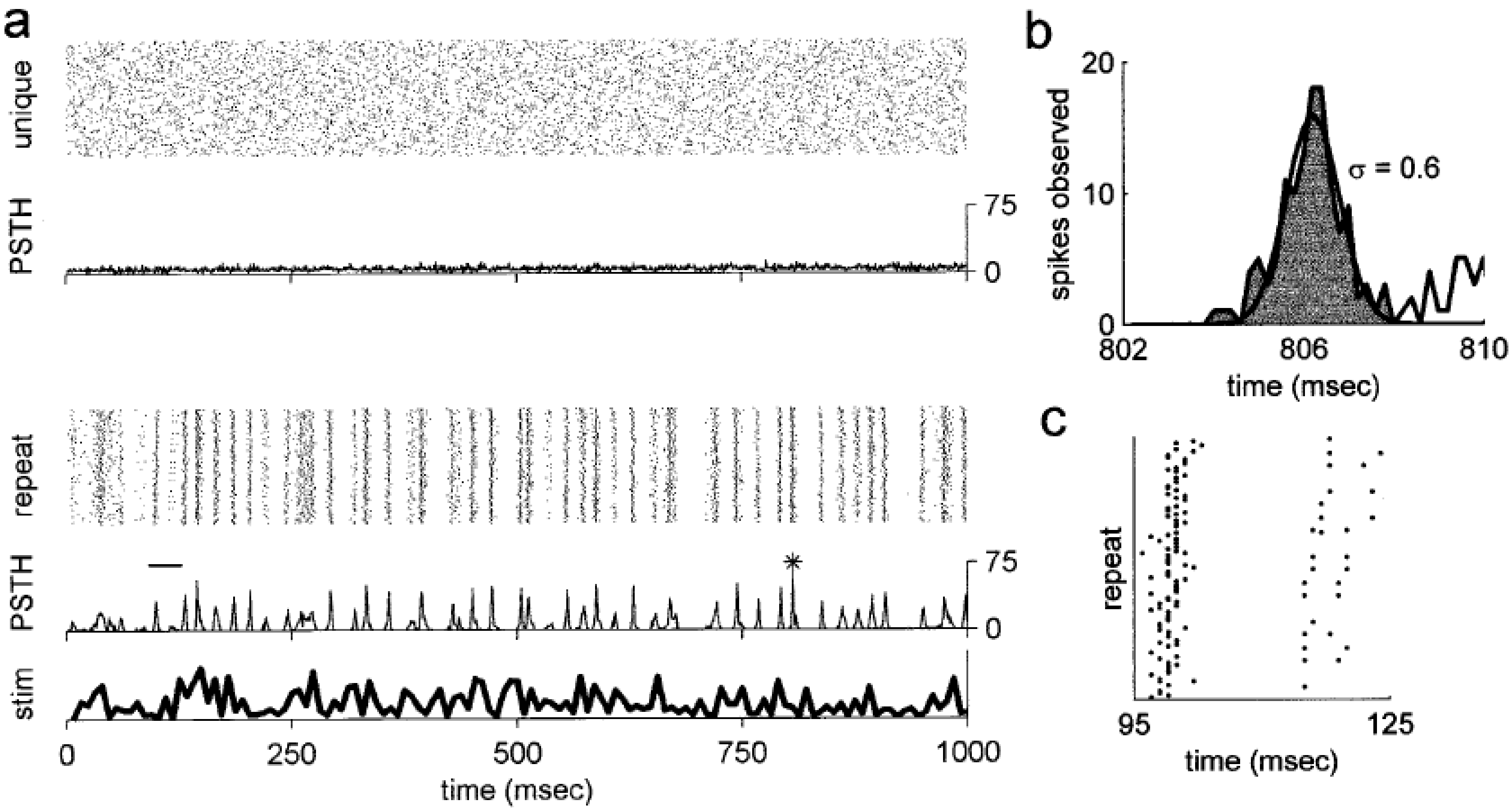
Calculating information in spike trains

Fly H1:
obtain information rate of
~80 bits/sec or 1-2 bits/spike.

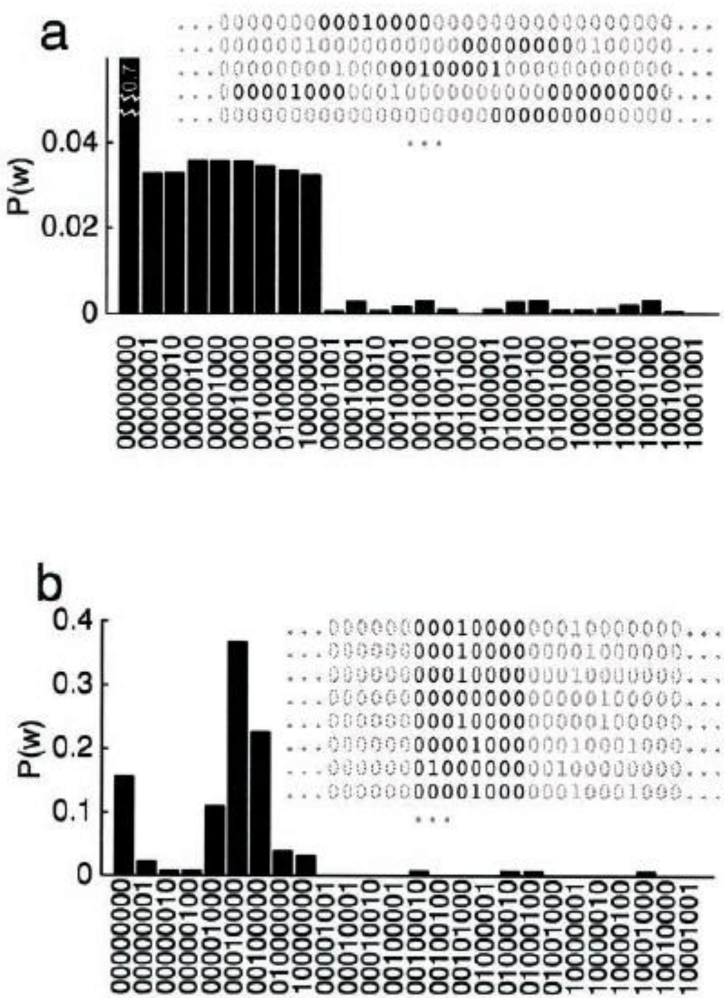


Calculating information in the LGN

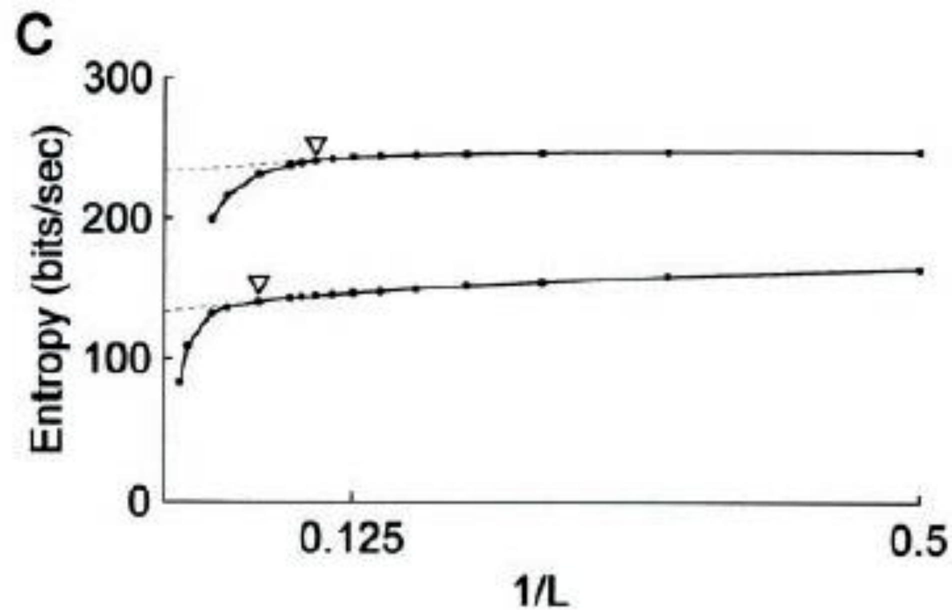
Another example: temporal coding in the LGN (Reinagel and Reid ‘00)



Calculating information in the LGN

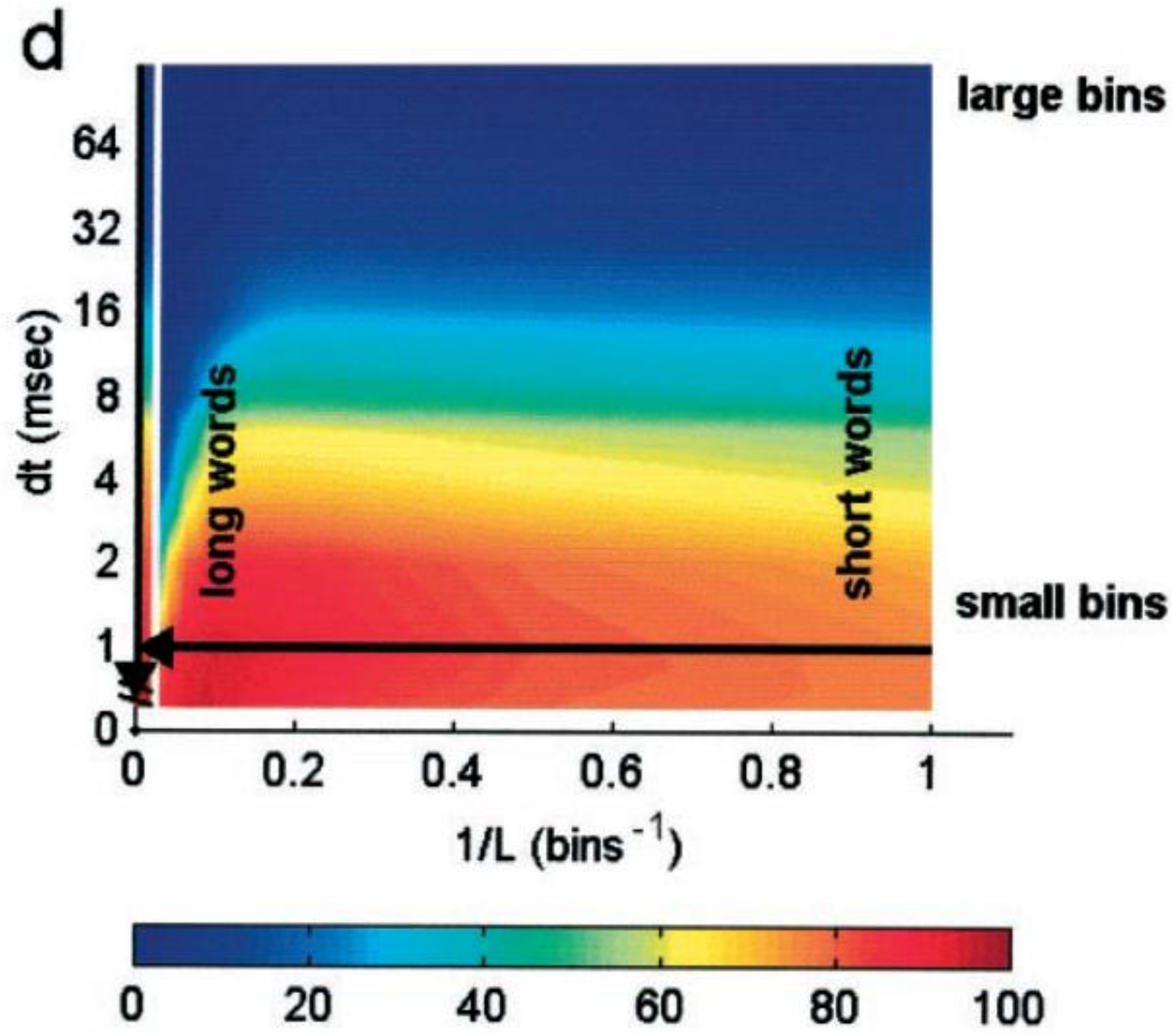


Apply the same procedure:
collect word distributions
for a random, then repeated stimulus.



Information in the LGN

Use this to quantify how
precise the code is,
and over what timescales
correlations are important.



Information in single spikes

How much information does a single spike convey about the stimulus?

Key idea: the information that a spike gives about the stimulus is the reduction in entropy between the distribution of spike times not knowing the stimulus, and the distribution of times knowing the stimulus.

The response to an (arbitrary) stimulus sequence ***s*** is r(t).

Without knowing that the stimulus was ***s***, the probability of observing a spike in a given bin is proportional to $\bar{r}$, the mean rate, and the size of the bin.

Consider a bin Δt small enough that it can only contain a single spike. Then in the bin at time t,

$$\begin{aligned} P(r=1) &= \bar{r}\Delta t, \\ P(r=0) &= 1-\bar{r}\Delta t, \\ P(r=1|s) &= r(t)\Delta t, \\ P(r=0|s) &= 1-r(t)\Delta t. \end{aligned}$$

Information in single spikes

Now compute the entropy difference: $p = \bar{r}\Delta t$, $p(t) = r(t)\Delta t$

$$I(r,s) = -p\log p - (1-p)\log(1-p) +$$ ← prior

$$+\frac{1}{T}\int_0^T dt\ [p(t)\log p(t) + (1-p(t))\log(1-p(t))] .$$ ← conditional

Note substitution of a time average for an average over the *r* ensemble.

Assuming $p \ll 1$, $\log(1-p) \sim p$ and using $\frac{1}{T}\int_0^T dt\, p(t) \to p$

$$I(r,s) = \frac{1}{T}\int_0^T dt\,\Delta t\, r(t)\log\frac{r(t)}{\bar{r}} + Var(p(t))/2ln2 + O(p^3).$$

In terms of information per spike (divide by $\bar{r}\Delta t$):

$$I(r,s) = \frac{1}{T}\int_0^T dt\,\frac{r(t)}{\bar{r}}\log\frac{r(t)}{\bar{r}}$$

Information in single spikes

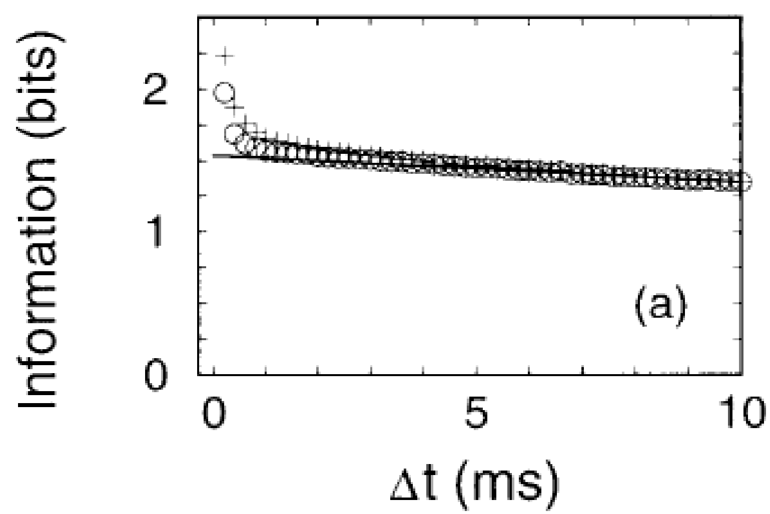
Given $$I(r,s) = \frac{1}{T}\int_0^T dt\, \frac{r(t)}{\bar{r}} \log \frac{r(t)}{\bar{r}}$$

note that:

- It doesn’t depend explicitly on the stimulus
- The rate *r* does not have to mean rate of spikes; rate of any event.
- Information is limited by spike precision, which blurs r(t), and the mean spike rate.

Compute as a function of Δt:

Undersampled for small bins



Adaptation and coding efficiency


SIEGFRI

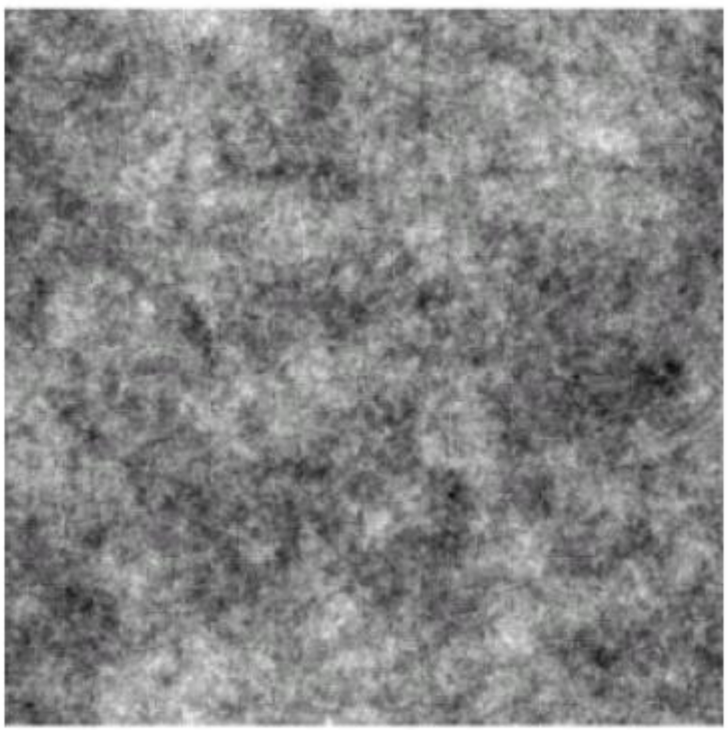
Natural stimuli

1. Huge dynamic range: variations over many orders of magnitude

Natural stimuli

1. Huge dynamic range: variations over many orders of magnitude

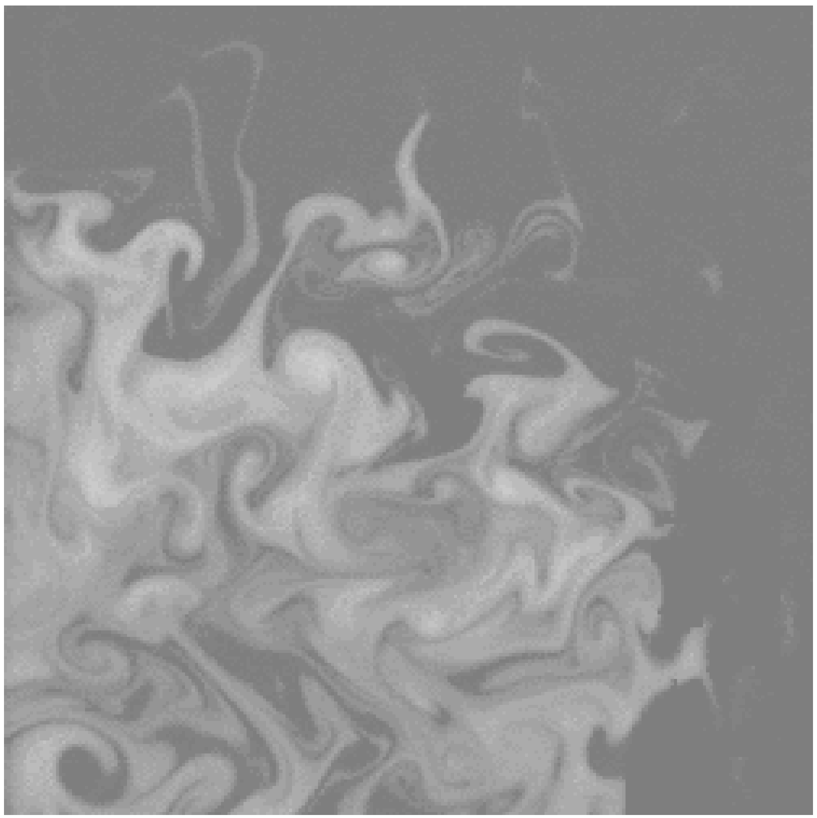
2. Power law scaling: highly nonGaussian



Natural stimuli

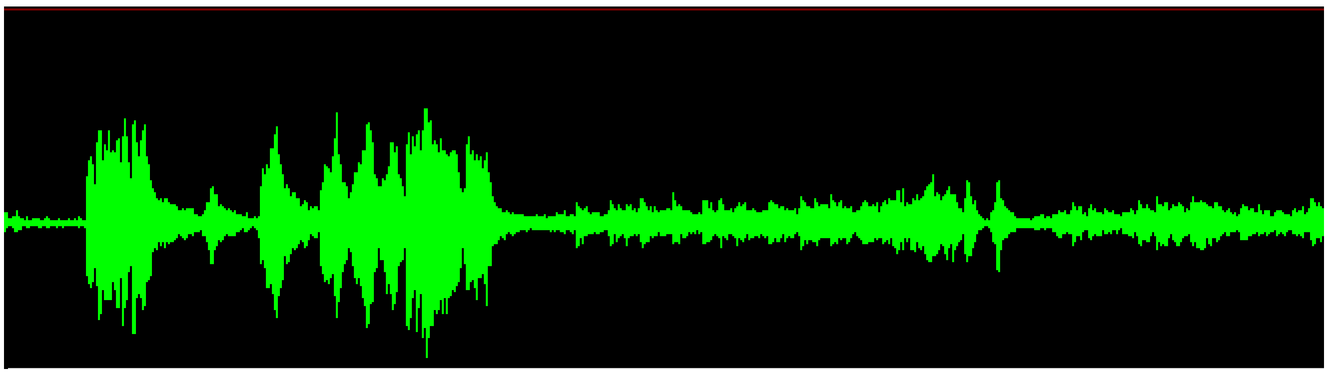
1. Huge dynamic range: variations over many orders of magnitude

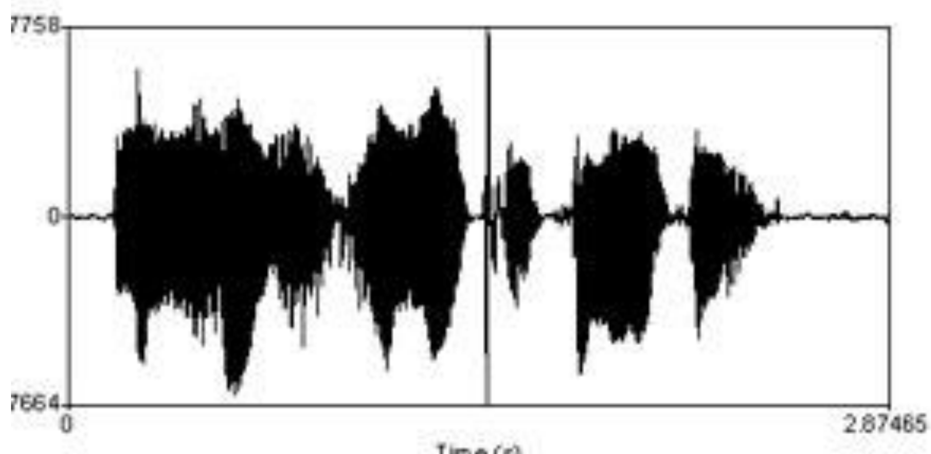
2. Power law scaling: highly nonGaussian



Natural stimuli

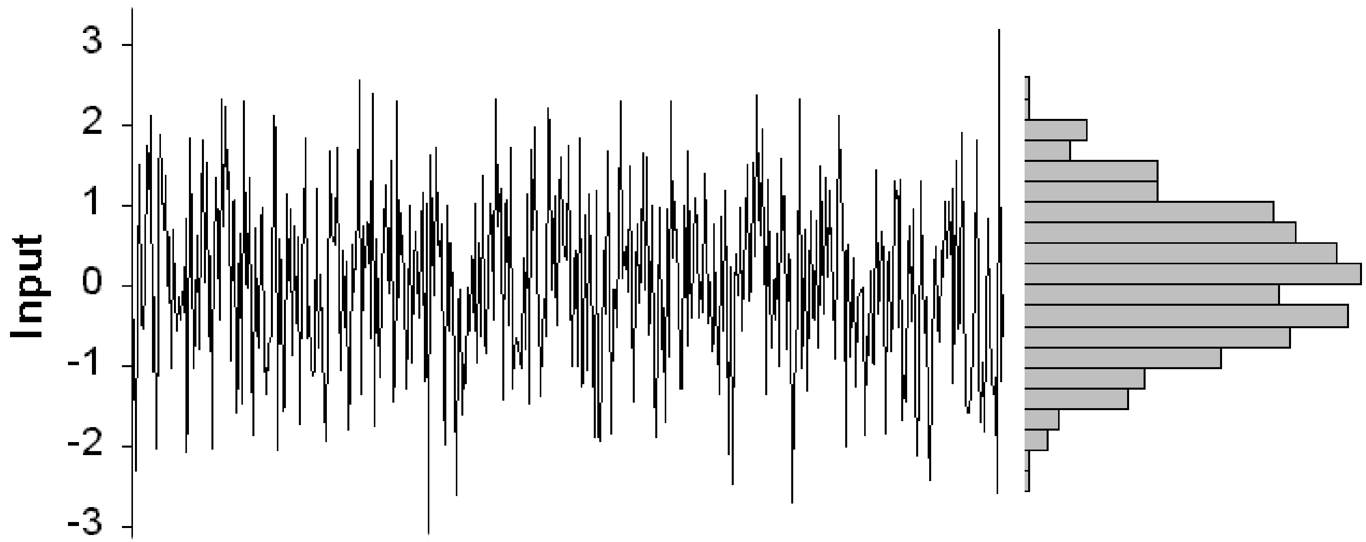
1. Huge dynamic range: variations over many orders of magnitude
2. Power law scaling: highly nonGaussian

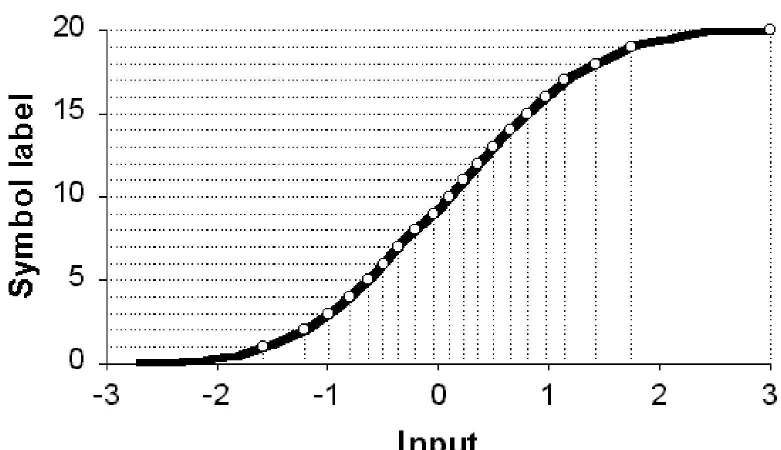




Efficient coding

In order to encode stimuli effectively, an encoder should match its outputs to the statistical distribution of the inputs





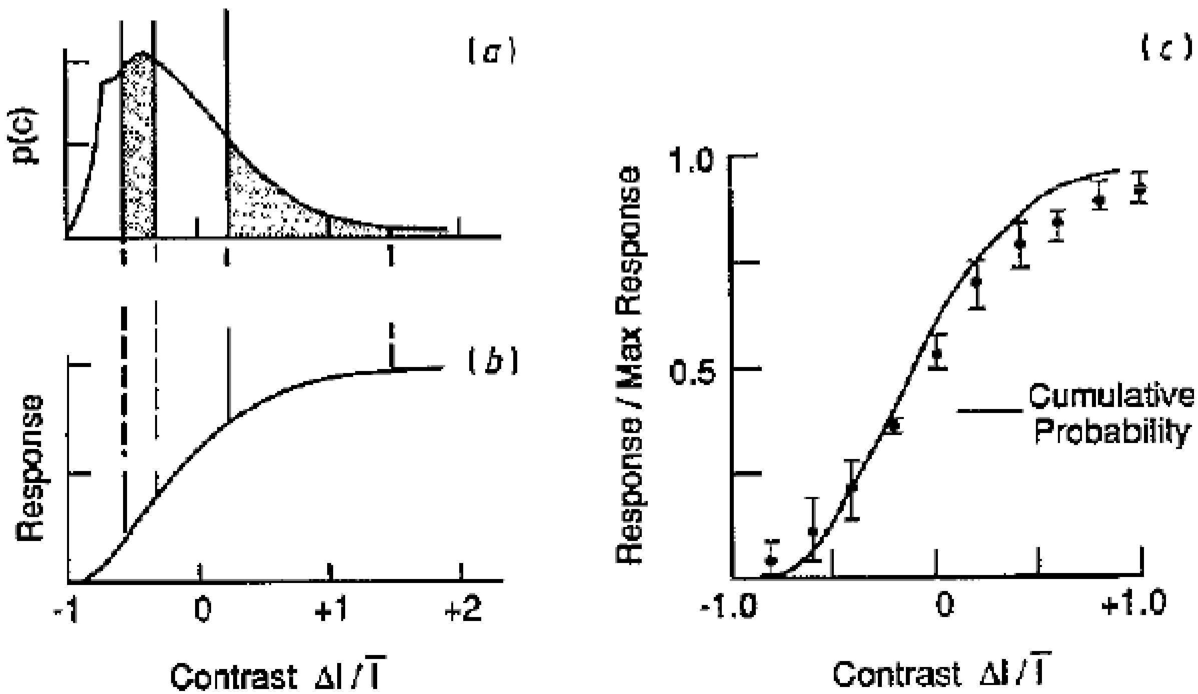
Shape of the I/O function should be determined by the distribution of natural inputs

Optimizes information between output and input

Fly visual system

$$P(r)dr = P(s)ds$$

$$r = g(s) = \frac{1}{\alpha}\int_{-1}^{s} ds'\, P(s').$$

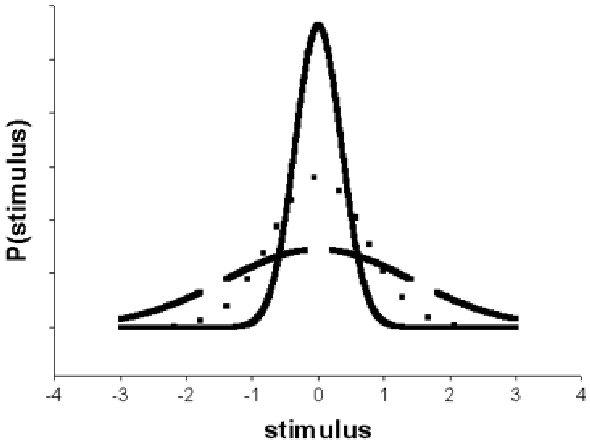


Laughlin, ‘81

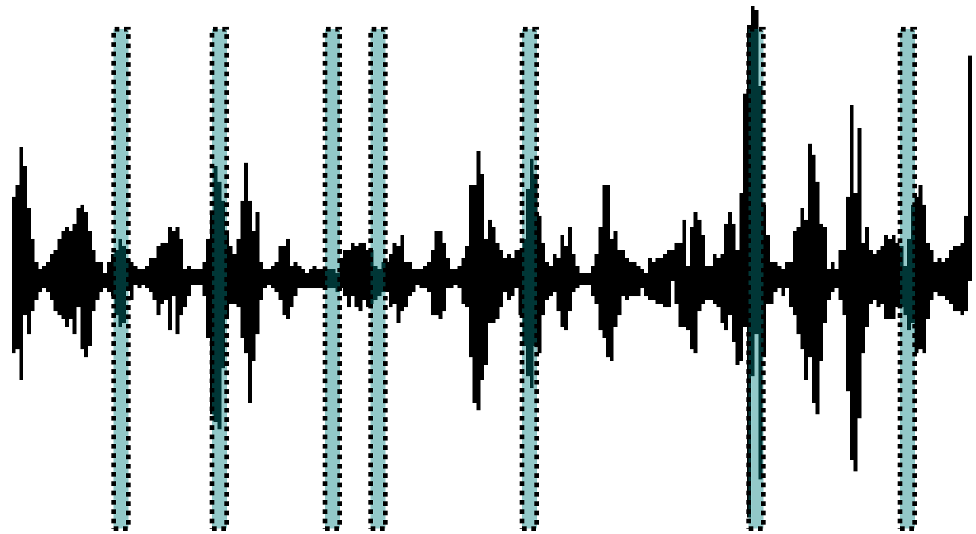
Variation in time

Contrast varies hugely in time.

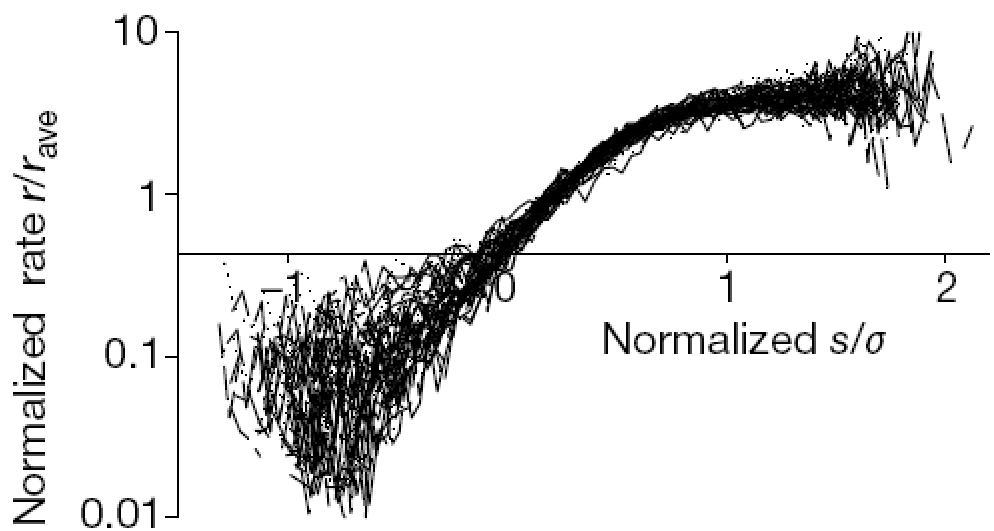
Should a neural system optimize over evolutionary time or locally?



Time-varying stimulus representation

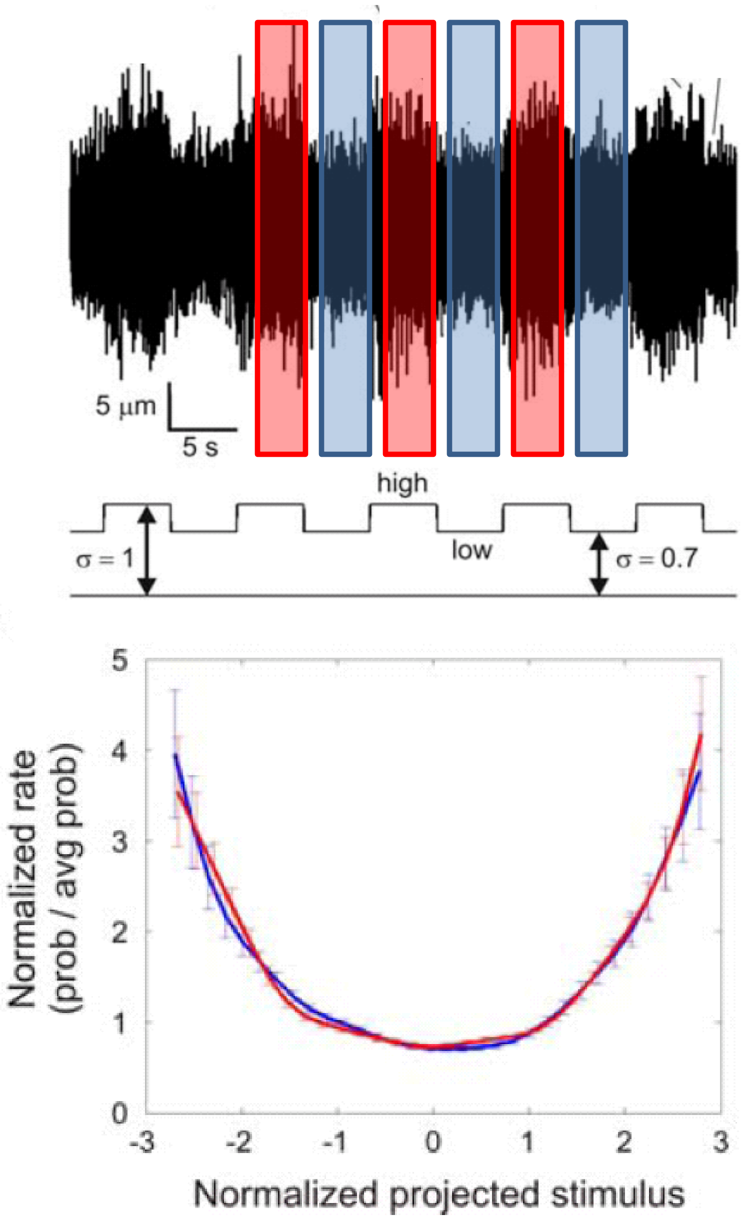


For fly neuron H1, determine the input/output relations throughout the stimulus presentation



A. Fairhall, G. Lewen, R. R. de Ruyter and W. Bialek (2001)

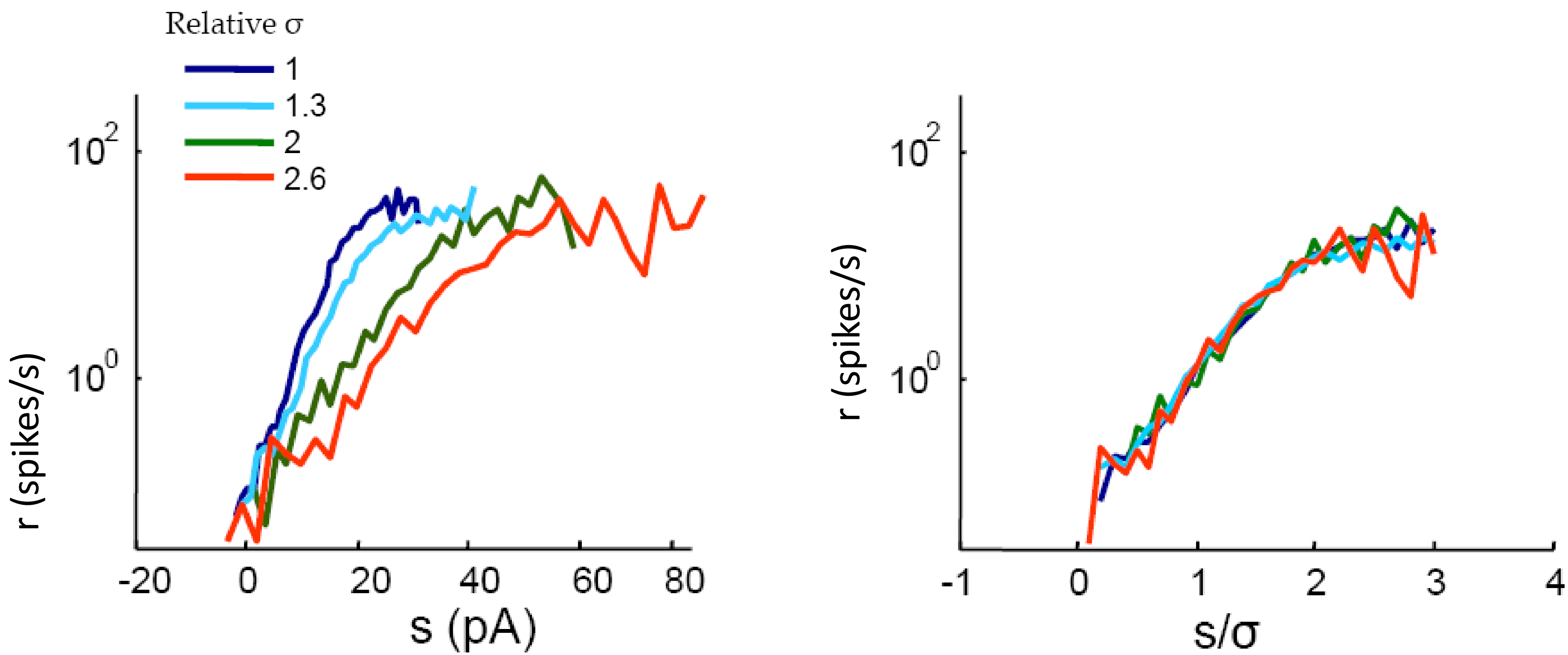
Barrel cortex



Extracellular *in vivo* recordings of responses to whisker motion in rat S1 barrel cortex in the anesthetized rat

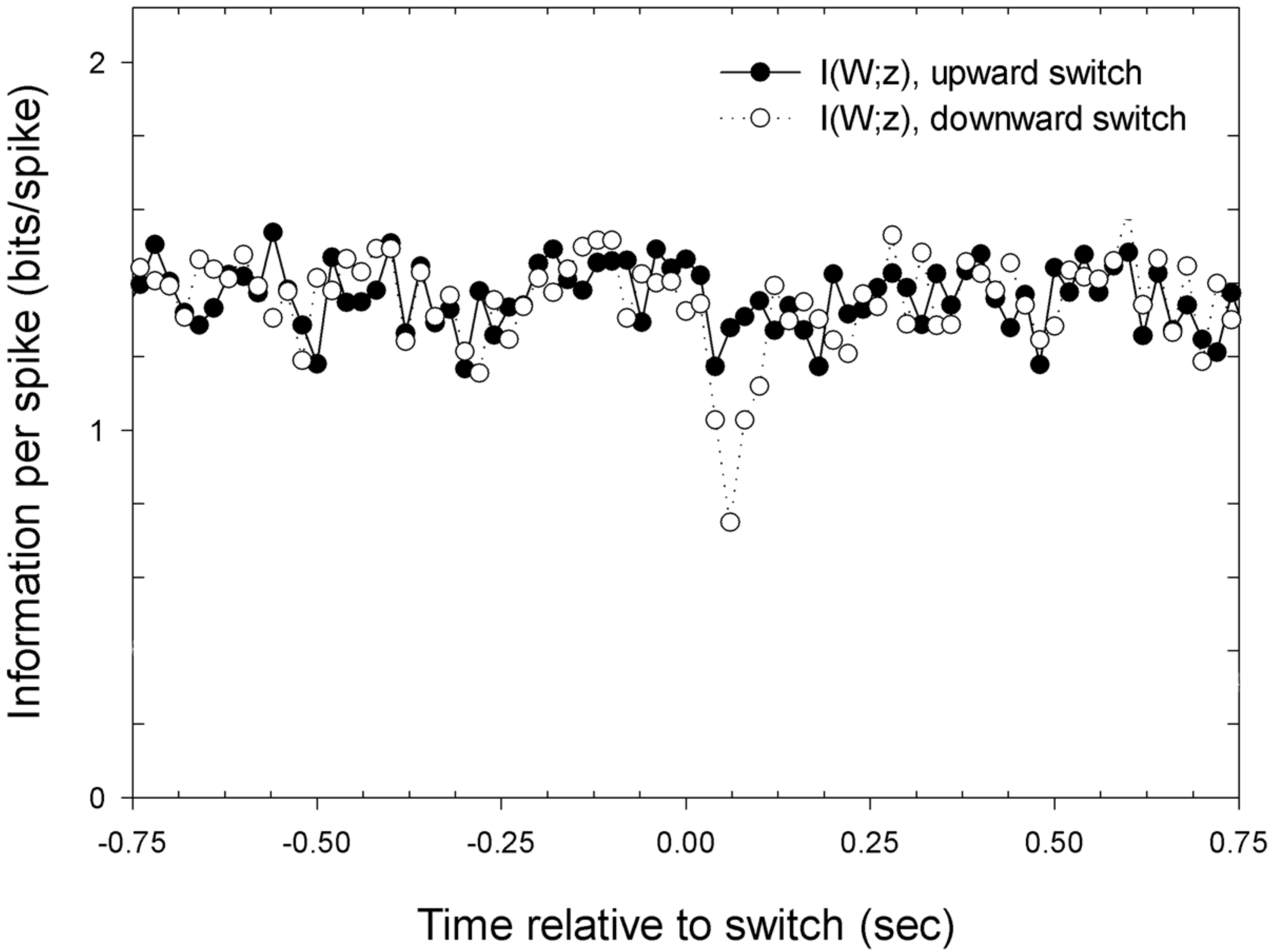
M. Maravall et al., (2007)

Single cortical neurons

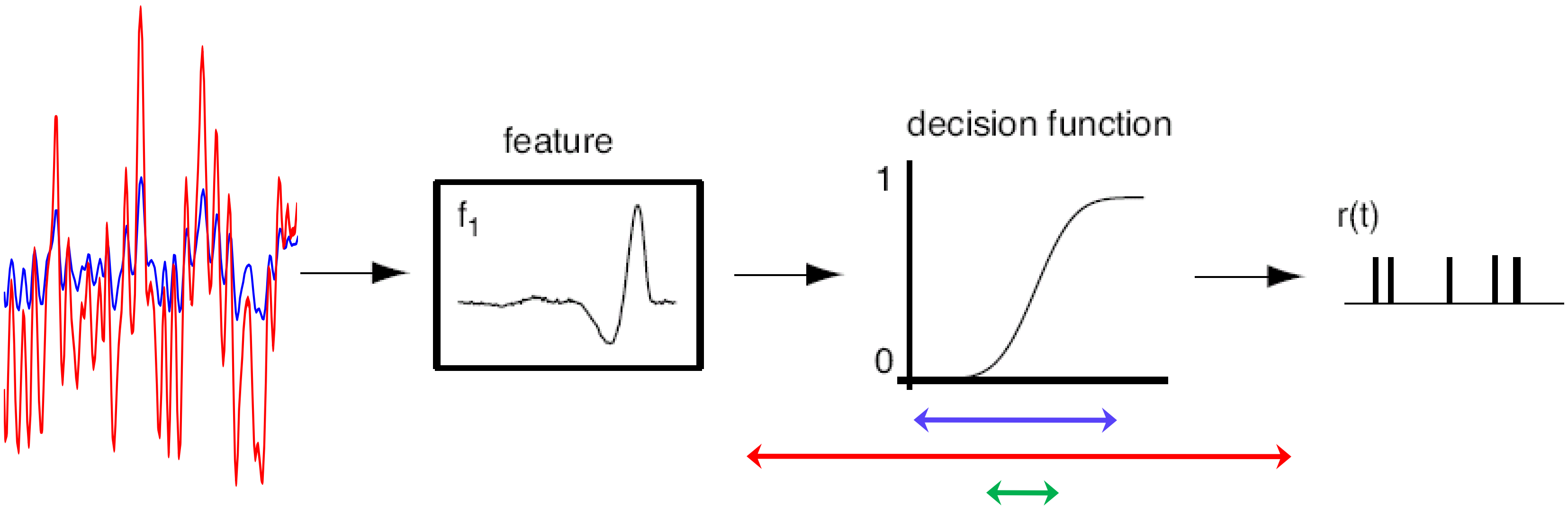


R. Mease, A. Fairhall and W. Moody, J. Neurosci.

Using information to evaluate coding



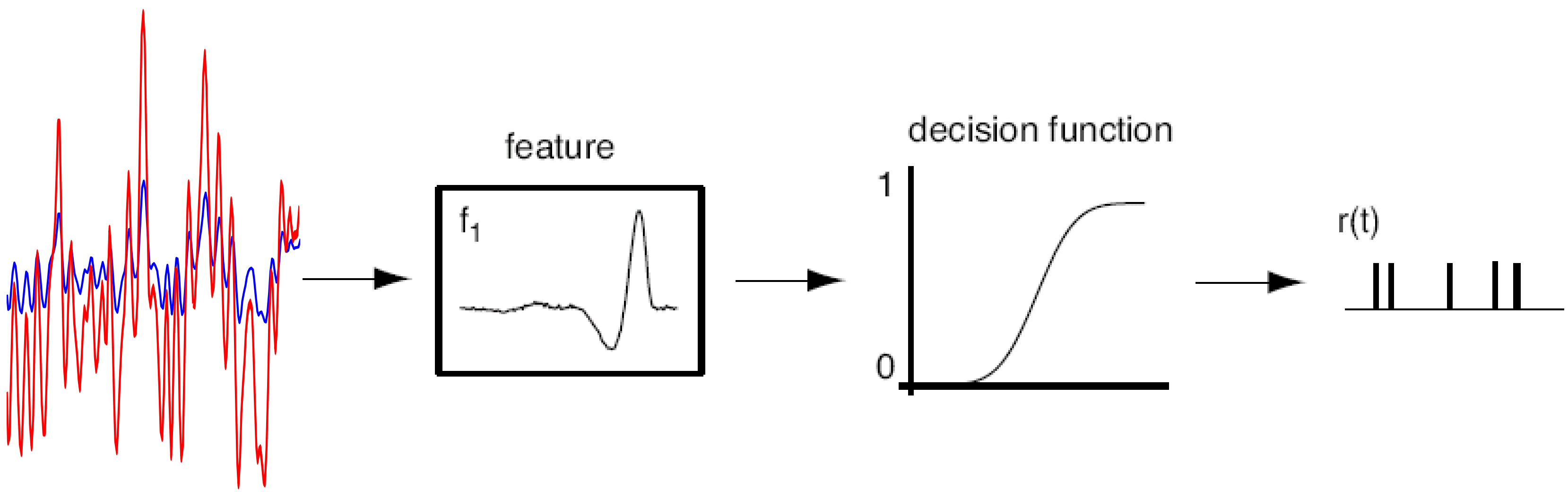
Adaptive representation of information



As one changes the characteristics of s(t), changes can occur both in the *feature* and in the *decision function*

Barlow ’50s, Laughlin ‘81, Shapley et al, ‘70s, Atick ‘91, Brenner ‘00

Feature adaptation



Barlow ’50s, Laughlin ‘81, Shapley et al, ‘70s, Atick ‘91, Brenner ‘00

Synergy and redundancy

The information in any given event can be computed as:

$$I(E;s) = \left\langle \left(\frac{r_E(t)}{\bar{r}_E} \right) \log_2 \left(\frac{r_E(t)}{\bar{r}_E} \right) \right\rangle_s ,$$

Define the *synergy,* the information gained from the joint symbol:

$$\mathrm{Syn}[E_1, E_2; s] = I[E_1, E_2; s] - (I[E_1; s] + I[E_2; s]).$$

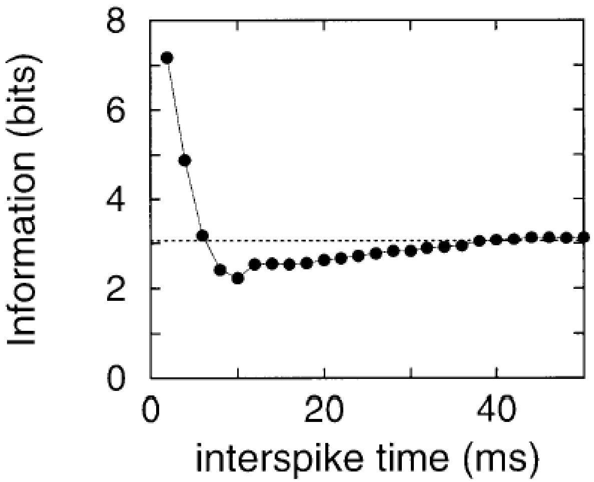
or equivalently,

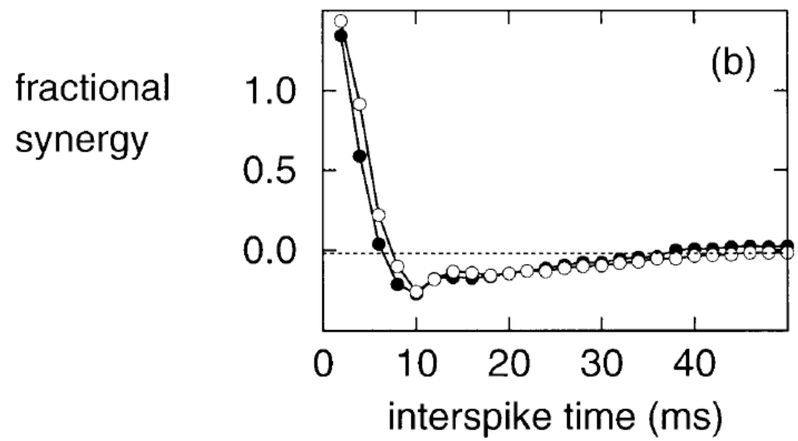
$$\mathrm{Syn}[E_1, E_2; s] = I[E_1; E_2|s] - I[E_1; E_2].$$

Negative synergy is called *redundancy*.

Multi-spike patterns

In the identified neuron H1, compute information in a spike pair, separated by an interval dt:





$$\mathrm{Syn}[E_1, E_2; s] = I[E_1, E_2; s] - (I[E_1; s] + I[E_2; s]).$$

Brenner et al., '00.