Information Theory

Mark van Rossum

School of Informatics, University of Edinburgh

January 24, 2018

[0]Version: January 24, 2018

Why information theory

Understanding the neural code.

- Encoding and decoding. We imposed coding schemes, such as 2nd-order kernel, or NLP. We possibly lost information in doing so.
- Instead, use information:
  - Don't need to impose encoding or decoding scheme (non-parametric).
    In particular important for 1) spike timing codes, 2) higher areas.
  - Estimate how much information is coded in certain signal.

  Caveats:
  - No easy decoding scheme for organism (upper bound only)
  - Requires more data and biases are tricky

Overview

- Entropy, Mutual Information
- Entropy Maximization for a Single Neuron
- Maximizing Mutual Information
- Estimating information
- Reading: Dayan and Abbott ch 4, Rieke

Definition

The *entropy* of a quantity is defined as
$H(X) = -\sum_x P(x) \log_2 P(x)$
This is not 'derived', nor fully unique, but it fulfills these criteria:

- Continuous
- If $p_i = \frac{1}{n}$, it increases monotonically with n. $H = \log_2 n$.
- Parallel independent channels add.

"Unit": bits
Entropy can be thought of as physical entropy, "richness" of distribution
[Shannon and Weaver, 1949, Cover and Thomas, 1991, Rieke et al., 1996]

Entropy

Discrete variable

$$H(R) = -\sum_r p(r) \log_2 p(r)$$

Continuous variable at resolution Δr

$$H(R) = -\sum_r p(r)\Delta r \log_2(p(r)\Delta r) = -\sum_r p(r)\Delta r \log_2 p(r) - \log_2 \Delta r$$

letting $\Delta r \to 0$ we have

$$\lim_{\Delta r \to 0} [H + \log_2 \Delta r] = -\int p(r) \log_2 p(r) dr$$

(also called differential entropy)

Joint, Conditional entropy

Joint entropy:

$$H(S, R) = -\sum_{r,s} P(S, R) \log_2 P(S, R)$$

Conditional entropy:

$$\begin{aligned} H(S|R) &= \sum_r P(R = r) H(S|R = r) \\ &= -\sum_r P(r) \sum_s P(s|r) \log_2 P(s|r) \\ &= H(S, R) - H(R) \end{aligned}$$

If S, R are independent

$$H(S, R) = H(S) + H(R)$$

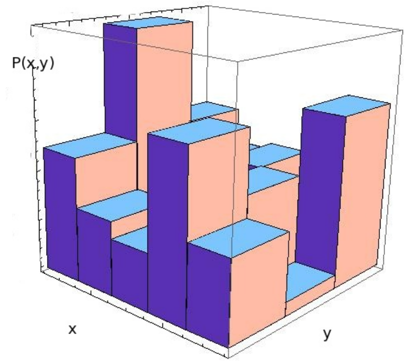
Mutual information

Mutual information:

$$\begin{aligned} I_m(R; S) &= \sum_{r,s} p(r, s) \log_2 \frac{p(r, s)}{p(r)p(s)} \\ &= H(R) - H(R|S) = H(S) - H(S|R) \end{aligned}$$

- Measures reduction in uncertainty of R by knowing S (or vice versa)
- $I_m(R; S) \geq 0$
- The continuous version is the *difference* of two entropies, the Δr divergence cancels

Mutual Information



The joint histogram determines mutual information.
Given $P(r, s) \Rightarrow I_m$.

Mutual Information: Examples

		Y_1: smoker	Y_1: non smoker
Y_2	lung cancer	1/3	0
	no lung cancer	0	2/3

		Y_1: smoker	Y_1: non smoker
Y_2	lung cancer	1/9	2/9
	no lung cancer	2/9	4/9

Mutual Information: Examples

		Y_1: smoker	Y_1: non smoker
Y_2	lung cancer	1/3	0
	no lung cancer	0	2/3

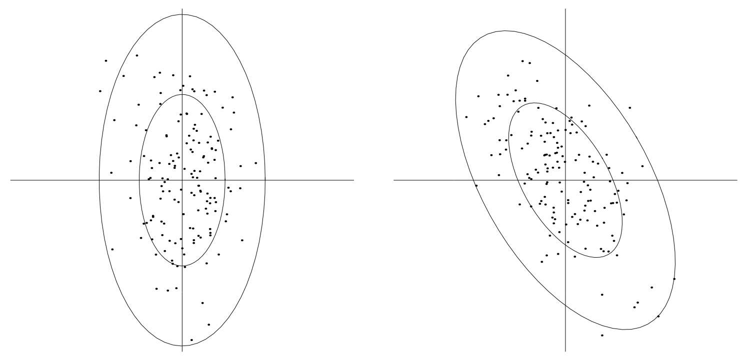
		Y_1: smoker	Y_1: non smoker
Y_2	lung cancer	1/9	2/9
	no lung cancer	2/9	4/9

Only for the left joint probability $I_m > 0$ (how much?). On the right, knowledge about Y_1 does not inform us about Y_2.

Kullback-Leibler divergence

- KL-divergence measures distance between two probability distributions
- $D_{KL}(P||Q) = \int P(x) \log_2 \frac{P(x)}{Q(x)} dx$, or $D_{KL}(P||Q) \equiv \sum_i P_i \log_2 \frac{P_i}{Q_i}$
- Not symmetric, but can be symmetrized
- $I_m(R;S) = D_{KL}(p(r,s)||p(r)p(s))$.
- Often used as probabilistic cost function: $D_{KL}(data||model)$.
- Other probability distances exist (e.g. earth-movers distance)

Mutual info between jointly Gaussian variables



$$I(Y_1;Y_2) = \int \int P(y_1, y_2) \log_2 \frac{P(y_1, y_2)}{P(y_1)P(y_2)} \, dy_1 \, dy_2 = -\frac{1}{2} \log_2(1-\rho^2)$$

ρ is (Pearson-r) correlation coefficient.

Populations of Neurons

Given

$$H(\mathbf{R}) = -\int p(\mathbf{r}) \log_2 p(\mathbf{r}) d\mathbf{r} - N \log_2 \Delta r$$

and

$$H(R_i) = -\int p(r_i) \log_2 p(r_i) d\mathbf{r} - \log_2 \Delta r$$

We have

$$H(\mathbf{R}) \leq \sum_i H(R_i)$$

(proof, consider KL divergence)

Mutual information in populations of Neurons

Reduncancy can be defined as (compare to above)

$$R = \sum_{i=1}^{n_r} I(r_i; \mathbf{s}) - I(\mathbf{r}; \mathbf{s}).$$

Some codes have $R > 0$ (redundant code), others $R < 0$ (synergistic)
Example of synergistic code: $P(r_1, r_2, s)$ with
$P(0,0,1) = P(0,1,0) = P(1,0,0) = P(1,1,1) = \frac{1}{4}$

Entropy Maximization for a Single Neuron

$$I_m(R; S) = H(R) - H(R|S)$$

- If noise entropy $H(R|S)$ is independent of the transformation $S \to R$, we can maximize mutual information by maximizing $H(R)$ under given constraints
- Possible constraint: response r is $0 < r < r_{\max}$. Maximal $H(R)$ if $\Rightarrow p(r) \sim U(0, r_{\max})$ (U is uniform dist)
- If average firing rate is limited, and $0 < r < \infty$: exponential distribution is optimal $p(x) = 1/\bar{x} exp(-x/\bar{x})$. $H = \log_2 e\bar{x}$
- If variance is fixed and $-\infty < r < \infty$: Gaussian distribution. $H = \frac{1}{2}\log_2(2\pi e\sigma^2)$ (note funny units)

- Let $r = f(s)$ and $s \sim p(s)$. Which f (assumed monotonic) maximizes $H(R)$ using max firing rate constraint? Require: $P(r) = \frac{1}{r_{\max}}$

$$p(s) = p(r)\frac{dr}{ds} = \frac{1}{r_{\max}}\frac{df}{ds}$$

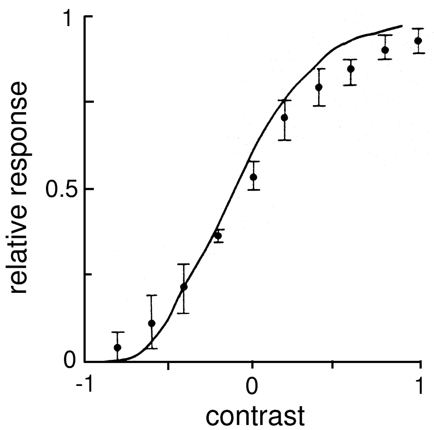
Thus $df/ds = r_{\max} p(s)$ and

$$f(s) = r_{\max} \int_{s_{\min}}^{s} p(s')ds'$$

- This strategy is known as *histogram equalization* in signal processing

Fly retina

Evidence that the large monopolar cell in the fly visual system carries out histogram equalization



Contrast response for fly large monopolar cell (points) matches environment statistics (line) [Laughlin, 1981] (but changes in high noise conditions)

V1 contrast responses

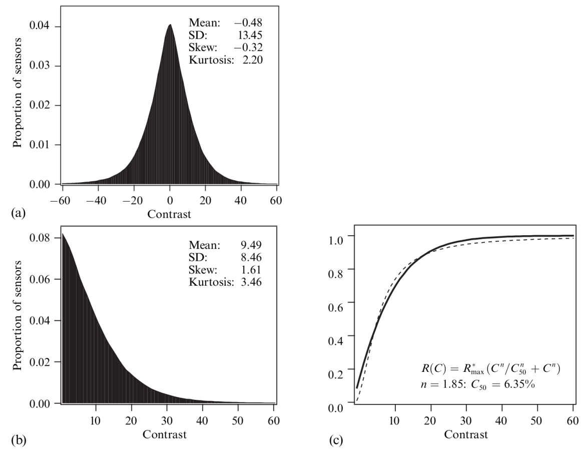


Figure 3. The distribution of image contrast in natural scenes: (a) both positive and negative, and (b) positive alone. In this study, sensor responses were pooled across 46 images, 5 spatial frequencies, and 4 orientations. The contrast bin width was 1%. (c) The integral of the positive-contrast histogram shown by the solid line defines the optimal contrast-response function. A hyperbolic function shown by the dotted line with $R_{max} = 1.0$, $C_{50} = 6.35\%$, and $n = 1.85$ provides a good fit to the data. SD = standard deviation.

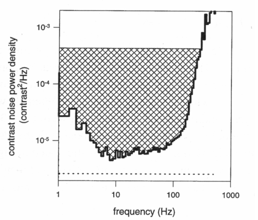
Similar in V1, but On and Off channels [Brady and Field, 2000]

Information of time varying signals

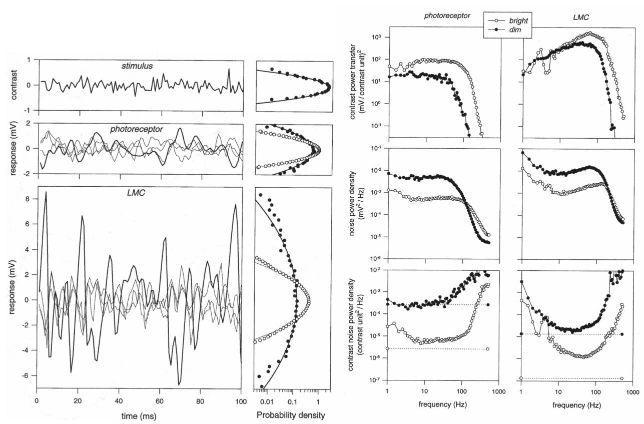
Single analog channel with Gaussian signal s and Gaussian noise η:
$r = s + \eta$

$$I = \frac{1}{2}\log_2\left(1 + \frac{\sigma_s^2}{\sigma_\eta^2}\right) = \frac{1}{2}\log_2(1 + SNR)$$

For time dependent signals $I = \frac{1}{2} T \int \frac{d\omega}{2\pi} \log_2\left(1 + \frac{s(\omega)}{n(\omega)}\right)$
To maximize information, *when* variance of the signal is constrained, use all frequency bands such that signal+noise = constant.
Whitening. Water filling analog:



Information of graded synapses



Light - (photon noise) - photoreceptor - (synaptic noise) - LMC
At low light levels photon noise dominates, synaptic noise is negligible.
Information rate: 1500 bits/s
[de Ruyter van Steveninck and Laughlin, 1996].

Spiking neurons: maximal information

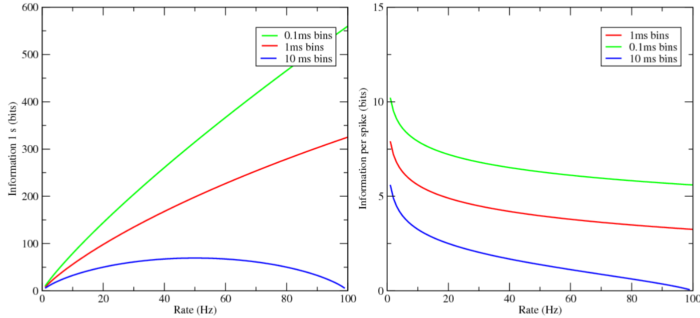
Spike train with $N = T/\delta t$ bins [Mackay and McCullogh, 1952] δt *"time-resolution"*.
$pN = N_1$ events, #words $= \frac{N!}{N_1!(N-N_1)!}$
Maximal entropy if all words are equally likely.
$H = \sum p_i \log_2 p_i = \log_2 N! - \log_2 N_1! - \log_2 (N - N_1)!$
Use for large x that $\log x! \approx x(\log x - 1)$

$$H = \frac{-T}{\delta t}[p \log_2 p + (1-p)\log_2(1-p)]$$

For low rates $p \ll 1$, setting $\lambda = (\delta t)p$:

$$H = T\lambda \log_2\left(\frac{e}{\lambda \delta t}\right)$$

Spiking neurons



Calculation incorrect when multiple spikes per bin. Instead, for large bins maximal information for exponential distribution:
$P(n) = \frac{1}{Z}\exp[-n \log(1 + \frac{1}{\langle n \rangle})]$
$H = \log_2(1 + \langle n \rangle) + \langle n \rangle \log_2(1 + \frac{1}{\langle n \rangle}) \approx \log_2(1 + \langle n \rangle) + 1$

Spiking neurons: rate code

[Stein, 1967]

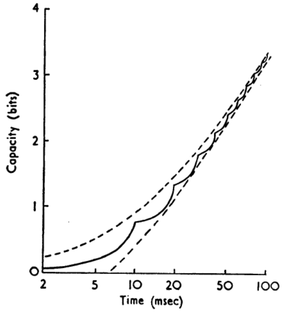


FIGURE 2 Information capacity of a completely regular neuron (solid line) as a function of the duration of a maintained stimulus. The dashed lines are upper and lower limits which converge rapidly as time (on a logarithmic scale) increases. The values were calculated for the example described in the text. The range of neuronal impulse frequencies was from 10 to 100 impulses/sec.

- Measure rate in window T, during which stimulus is constant.
- Periodic neuron can maximally encode $[1 + (f_{max} - f_{min})T]$ stimuli
- $H \approx log_2[1 + (f_{max} - f_{min})T]$. Note, only $\propto \log(T)$

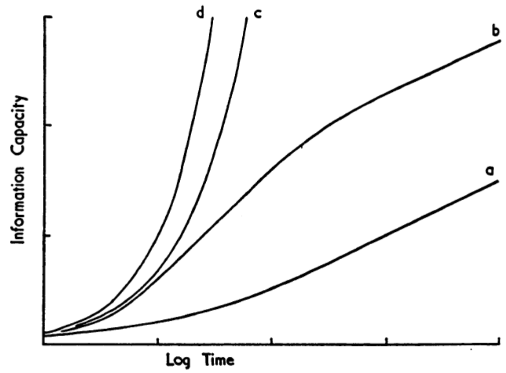


FIGURE 7 Schematic representation of the information capacity as a function of stimulus duration for a neuron, (a) discharging randomly and using a frequency code, (b) discharging fairly regularly and using a frequency code, (c) using a binary pulse code, and (d) using an interval code. Explanation in text.

[Stein, 1967]
Similar behaviour for Poisson : $H \propto \log(T)$

Spiking neurons: dynamic stimuli

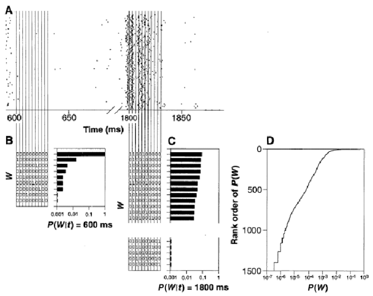
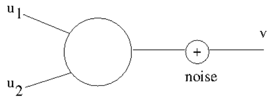


Fig. 3. Word frequency distributions and information transfer. **(A)** Two segments from 100 response traces of H1, starting at about 600 and 1800 ms, respectively, after onset of the repeated stimulus of Fig. 2. **(B)** Construction of local word frequencies. We start with a set of spike trains in response to a repeated random velocity sequence. Beginning at 600 ms these spike trains are divided in 10 contiguous 3-ms bins, as indicated by the array of vertical lines. For each trial, the spikes in each of the 10 bins are counted, and this set of 10 numbers forms a word, W. Here almost all words are binary strings, as two spikes occur only very rarely within 3 ms. This procedure gives us as many words as there are trials (here 900). From this set we compute the probability for each word, and the resulting distribution is depicted in the histogram, $P(W|t) = 600$ ms, where the words are ordered according to their probability. **(C)** As in (B), but now starting at 1800 ms. **(D)** Distribution, $P(W)$, of all words throughout the experiment. Words are defined in the same way as in (B) and (C). However, here they are taken from the long (900 times 10 s) nonrepeated part of the stimulus sequence in order to obtain a large number of independent stimulus samples. Thus, stepping in 3-ms bins, $\sim 3 \times 10^6$ words are sampled, and the distribution shown here describes their ranked frequencies. In these windows, by far the most likely word is 0000000000, and roughly 1500 different words are observed.

[de Ruyter van Steveninck et al., 1997], but see [Warzecha and Egelhaaf, 1999].

Maximizing Information Transmission: single output



Single linear neuron with post-synaptic noise

$$v = \mathbf{w} \cdot \mathbf{u} + \eta$$

where η is an independent noise variable

$$I_m(\mathbf{u}; v) = H(v) - H(v|\mathbf{u})$$

- Second term depends only on $p(\eta)$
- To maximize I_m need to maximize $H(v)$; sensible constraint is that $\|\mathbf{w}\|^2 = 1$
- If $\mathbf{u} \sim N(\mathbf{0}, Q)$ and $\eta \sim N(0, \sigma_\eta^2)$ then $v \sim N(0, \mathbf{w}^T Q \mathbf{w} + \sigma_\eta^2)$

- For a Gaussian RV with variance σ^2 we have $H = \frac{1}{2} \log 2\pi e \sigma^2$. To maximize $H(v)$ we need to maximize $\mathbf{w}^T Q \mathbf{w}$ subject to the constraint $\|\mathbf{w}\|^2 = 1$
- Thus $\mathbf{w} \propto \mathbf{e}_1$ so we obtain PCA
- If v is non-Gaussian then this calculation gives an *upper bound* on $H(v)$ (as the Gaussian distribution is the maximum entropy distribution for a given mean and covariance)

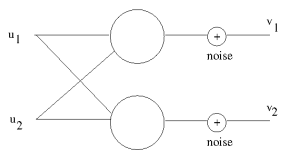
Infomax

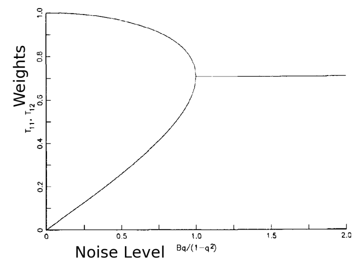
Infomax: maximize information in *multiple* outputs wrt weights [Linsker, 1988]

$$\mathbf{v} = W\mathbf{u} + \boldsymbol{\eta}$$

$$H(\mathbf{v}) = \frac{1}{2} \log \det(\langle \mathbf{v}\mathbf{v}^T \rangle)$$

Example: 2 inputs and 2 outputs. Input is correlated. $w_{k1}^2 + w_{k2}^2 = 1$.

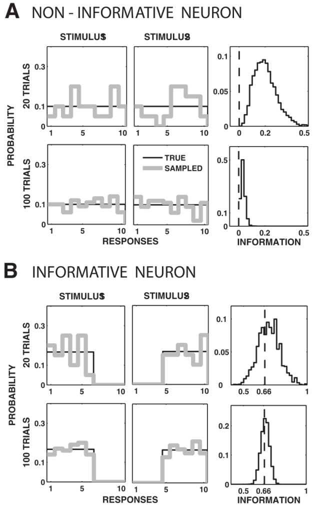




At low noise independent coding, at high noise joint coding.

Estimating information

Information estimation requires a lot of data.
Most statistical quantities are unbiased (mean, var,...).
But both entropy and noise entropy have bias.



[Panzeri et al., 2007]

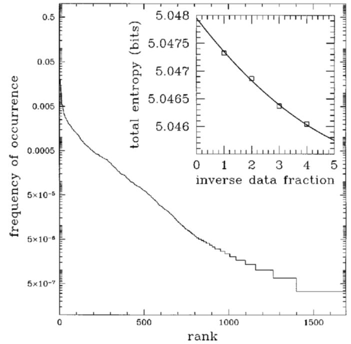


FIG. 2. The frequency of occurrence for different words in the spike train, with $\Delta\tau = 3$ ms and $T = 30$ ms. Words are placed in order so that the histogram is monotonically decreasing; at this value of T the most likely word corresponds to no spikes. Inset shows the dependence of the entropy, computed from this histogram according to Eq. (1), on the fraction of data included in the analysis. Also plotted is a least squares fit to the form $S = S_0 + S_1/\text{size} + S_2/\text{size}^2$. The intercept S_0 is our extrapolation to the true value of the entropy with infinite data [11].

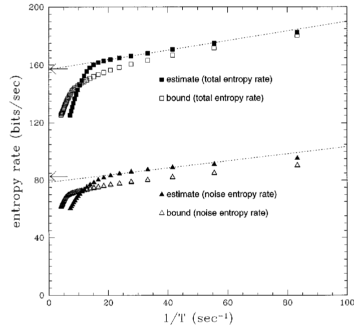


FIG. 3. The total and noise entropies per unit time are plotted versus the reciprocal of the window size, with the time resolution held fixed at $\Delta\tau = 3$ ms. Results are given both for the direct estimate and for the bounding procedure described in the text, and for each data point we apply the extrapolation procedures of Fig. 2 (inset). Dashed lines indicate extrapolations to infinite word length, as discussed in the text, and arrows indicate upper bounds obtained by differentiating $S(T)$ [7].

Try to fit $1/N$ correction [Strong et al., 1998]

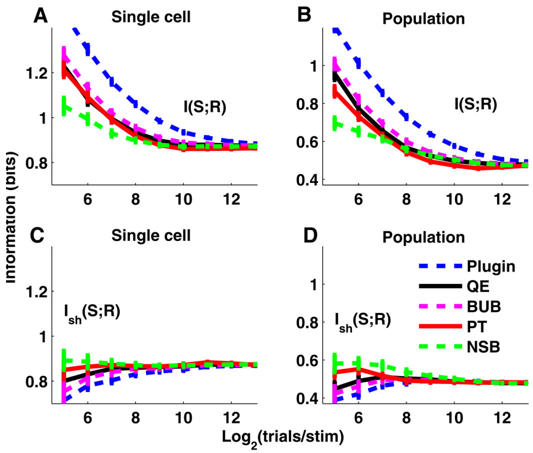


FIG. 3. Comparison of the performance of different bias correction methods. The information estimates $\mathbf{I}(\mathbf{S;R})$ and $\mathbf{I}_{sh}(\mathbf{S;R})$ are plotted as a function of the available number of trials per stimulus. *A* and *B*: mean ± SD (over 50 simulations) of $\mathbf{I}(\mathbf{S;R})$. *C* and *D*: mean ± SD (over 50 simulations) of $\mathbf{I}_{sh}(\mathbf{S;R})$. Various methods were used to correct for the bias: plug-in estimation (i.e., no bias correction), PT, QE, BUB, and NSB (see text). *A* and *C* and *B* and *D* report results using realistically simulated single-cell and population cortical spike trains, respectively (see main text).

Common technique for I_m: shuffle correction [Panzeri et al., 2007]
See also: [Paninski, 2003, Nemenman et al., 2002]

Summary

- Information theory provides non parametric framework for coding
- Optimal coding schemes depend strongly on noise assumptions and optimization constraints
- In data analysis biases can be substantial

References I

- Brady, N. and Field, D. J. (2000).
Local contrast in natural images: normalisation and coding efficiency.
Perception, 29(9):1041–1055.
- Cover, T. M. and Thomas, J. A. (1991).
Elements of information theory.
Wiley, New York.
- de Ruyter van Steveninck, R. R. and Laughlin, S. B. (1996).
The rate of information transfer at graded-potential synapses.
Nature, 379:642–645.
- de Ruyter van Steveninck, R. R., Lewen, G. D., Strong, S. P., Koberle, R., and Bialek, W. (1997).
Reproducibility and variability in neural spike trains.
Science, 275:1805–1809.
- Laughlin, S. B. (1981).
A simple coding procedure enhances a neuron's information capacity.
Zeitschrift für Naturforschung, 36:910–912.
- Linsker, R. (1988).
Self-organization in a perceptual network.
Computer, 21(3):105–117.

References II

- Mackay, D. and McCullogh, W. S. (1952).
The limiting information capacity of neuronal link.
Bull Math Biophys, 14:127–135.
- Nemenman, I., Shafee, F., and Bialek, W. (2002).
Entropy and Inference, Revisited.
nips, 14.
- Paninski, L. (2003).
Estimation of Entropy and Mutual Information.
Neural Comp., 15:1191–1253.
- Panzeri, S., Senatore, R., Montemurro, M. A., and Petersen, R. S. (2007).
Correcting for the sampling bias problem in spike train information measures.
J Neurophysiol, 98(3):1064–1072.
- Rieke, F., Warland, D., de Ruyter van Steveninck, R., and Bialek, W. (1996).
Spikes: Exploring the neural code.
MIT Press, Cambridge.
- Shannon, C. E. and Weaver, W. (1949).
The mathematical theory of communication.
Univeristy of Illinois Press, Illinois.

References III

- Stein, R. B. (1967).
The information capacity of nerve cells using a frequency code.
Biophys J, 7:797–826.
- Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R., and Bialek, W. (1998).
Entropy and Information in Neural Spike Trains.
Phys Rev Lett, 80:197–200.
- Warzecha, A. K. and Egelhaaf, M. (1999).
Variability in spike trains during constant and dynamic stimulation.
Science, 283(5409):1927–1930.