Information Theory & the Efficient Coding Hypothesis

Jonathan Pillow

Mathematical Tools for Neuroscience (NEU 314)
Spring, 2016

lecture 19

Information Theory

A mathematical theory of communication,
Claude Shannon 1948

- Entropy
- Conditional Entropy
- Mutual Information
- Data Processing Inequality
- Efficient Coding Hypothesis (Barlow 1961)

Entropy

$$H(x) = -\underbrace{\sum_x p(x)}_{\text{averaged over } p(x)} \underbrace{\log p(x)}_{\text{"surprise" of } x}$$

averaged over p(x)

"surprise" of x

- average "surprise" of viewing a sample from p(x)
- number of "yes/no" questions needed to identify x (on average)

for distribution on K bins,

- maximum entropy = log K (achieved by uniform dist)
- minimum entropy = 0 (achieved by all probability in 1 bin)

Entropy

$$H(x) = -\sum_x p(x) \log p(x)$$

$$= -\mathbb{E}[\log p(x)]$$

aside: log-likelihood and entropy

model: $P(x|\theta)$

entropy H: $-\mathbb{E}[\log P(x|\theta)]$

How would we compute a Monte Carlo estimate of this?

draw samples: $x_i \sim P(x|\theta)$ for i = 1,…,N

compute average: $\hat{H} = -\frac{1}{N} \underbrace{\sum_{i=1}^{N} \log P(x_i|\theta)}_{\text{log-likelihood}}$

- Neg Log likelihood = Monte Carlo estimate for entropy!
- maximizing likelihood ⇒ minimizing entropy of P(xl θ)

Conditional Entropy

$$H(x|y) \quad = \quad -\underbrace{\boxed{\sum_y p(y)}}_{\text{averaged over p(y)}}\underbrace{\boxed{\sum_x p(x|y)\log p(x|y)}}_{\text{entropy of x given some fixed value of y}}$$

Conditional Entropy

$$H(x|y) \quad = \quad -\boxed{\sum_y p(y)}\boxed{\sum_x p(x|y)\log p(x|y)}$$

averaged over p(y)

entropy of x given some fixed value of y

$$= \quad -\sum_{x,y} p(x,y)\log p(x|y)$$

$$= H(x) \quad \text{if} \quad P(x,y) = P(x)P(y)$$

"On average, how uncertain are you about *x* if you know *y*?"

Mutual Information

$$I(x, y) = H(x) - H(x|y)$$

total entropy in X minus
conditional entropy of X given Y

$$= H(y) - H(y|x)$$

total entropy in Y minus
conditional entropy of Y given X

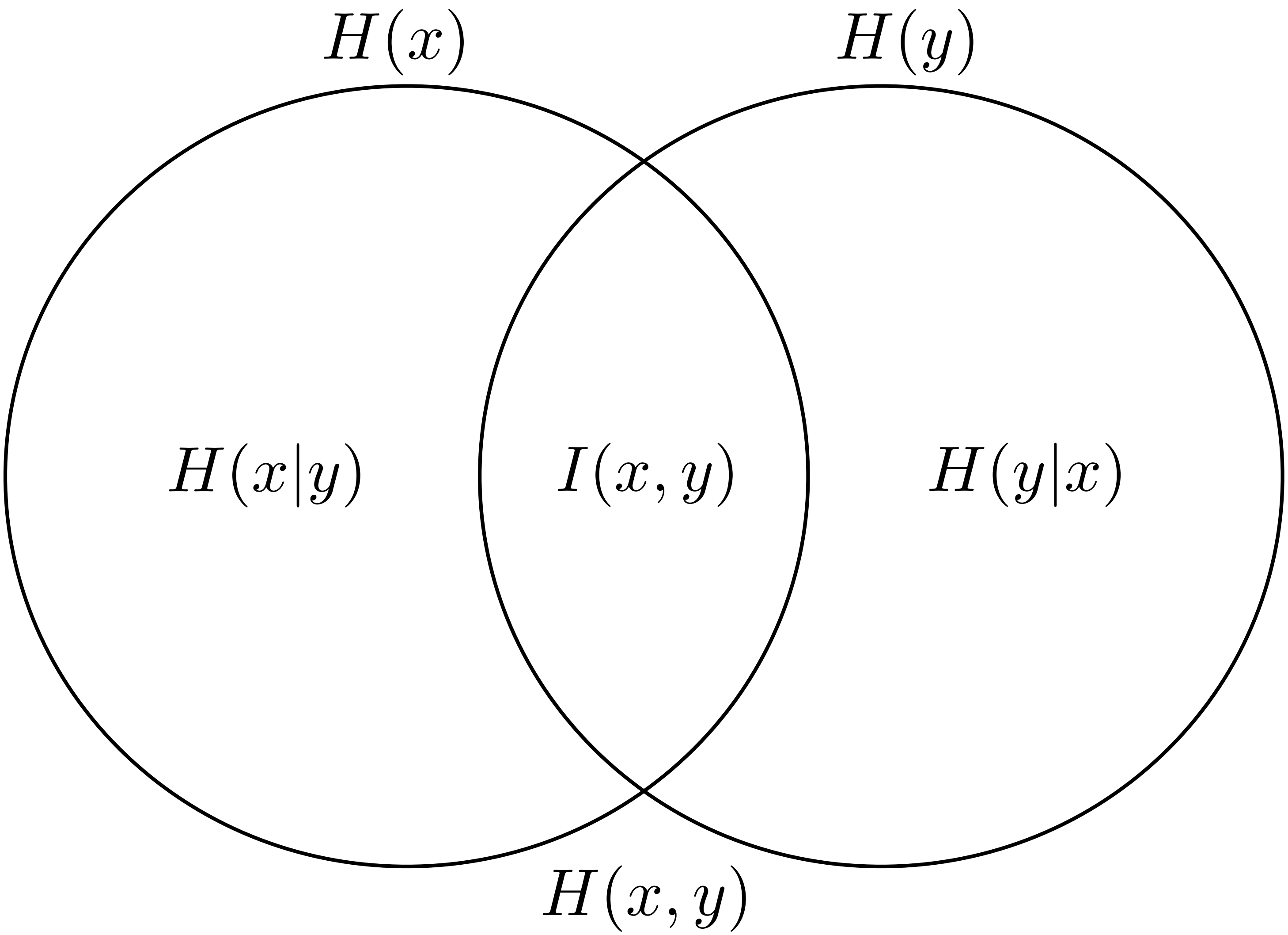
$$= H(x) + H(y) - H(x, y)$$

sum of entropies
minus joint entropy

"How much does X tell me about Y (or vice versa)?"

"How much is your uncertainty about X reduced from knowing Y?"

Venn diagram of entropy and information



Data Processing Inequality

Suppose $S \rightarrow R_1 \rightarrow R_2$ form a Markov chain, that is

$$P(R_1, R_2|S) = P(R_2|R_1)P(R_1|S)$$

Then necessarily: $I(S, R_2) \leq I(S, R_1)$

- in other words, we can only lose information during processing

Barlow 1961
Atick & Redlich 1990

Efficient Coding Hypothesis:

- goal of nervous system: maximize information about environment
  (one of the core "big ideas" in theoretical neuroscience)

redundancy: $R = 1 - \frac{I}{C}$

mutual information (I)

channel capacity (C)

Barlow 1961
Atick & Redlich 1990

Efficient Coding Hypothesis:

- goal of nervous system: maximize information about environment
  (one of the core "big ideas" in theoretical neuroscience)

redundancy: $$R = 1 - \frac{I}{C}$$

mutual information (I)

channel capacity (C)

mutual information:

$$I(x,y) = \underbrace{H(y)}_{\text{response entropy}} - \underbrace{H(y|x)}_{\text{"noise" entropy}}$$

- avg # yes/no questions you can answer about x given y ("bits")

channel capacity:

$$C = \sup_{P_x} I(x,y)$$

- upper bound on mutual information
- determined by physical properties of encoder

Barlow 1961
Atick & Redlich 1990

Barlow's original version:

redundancy: $R = 1 - \frac{I}{C}$ mutual information

mutual information:

$$I(x,y) = \underbrace{H(y)}_{\text{response entropy}} - \underbrace{\cancel{H(y|x)}}_{\text{"noise" entropy}}$$

if responses are noiseless

Barlow 1961
Atick & Redlich 1990

Barlow's original version:

response entropy

redundancy: $R = 1 - \frac{H(Y)}{C}$

mutual information:

$$I(x,y) = \underbrace{H(y)}_{\text{response entropy}} - \underbrace{\cancel{H(y|x)}}_{\text{“noise” entropy}}$$

noiseless system

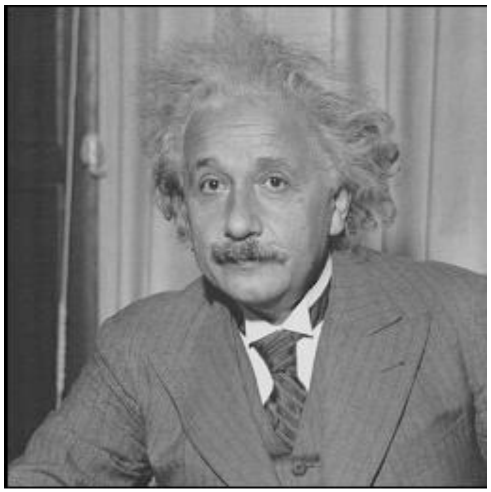
$\Longrightarrow$ brain should maximize response entropy

- use full dynamic range
- decorrelate (“reduce redundancy”)

• mega impact: huge number of theory and experimental papers focused on decorrelation / information-maximizing codes in the brain

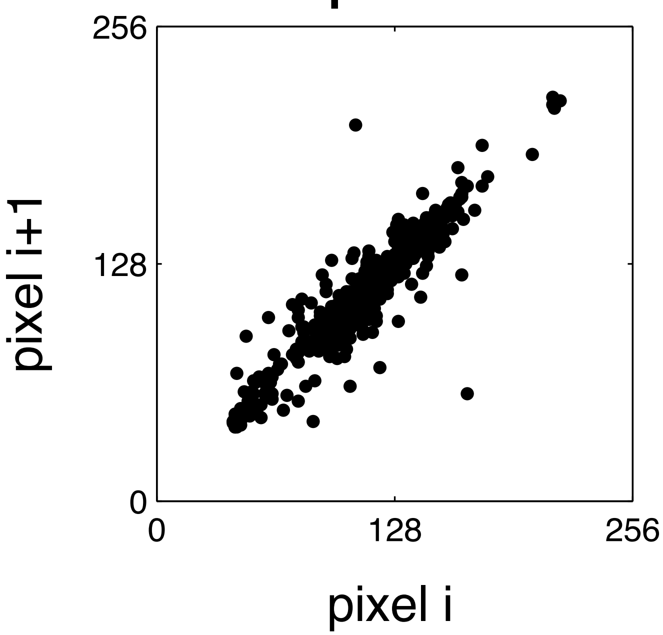
basic intuition

natural image



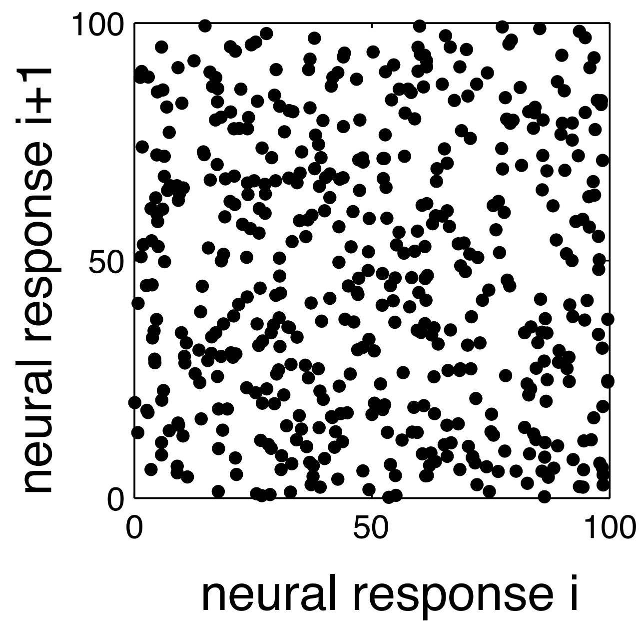
nearby pixels exhibit strong dependencies

pixels



desired encoding

neural representation



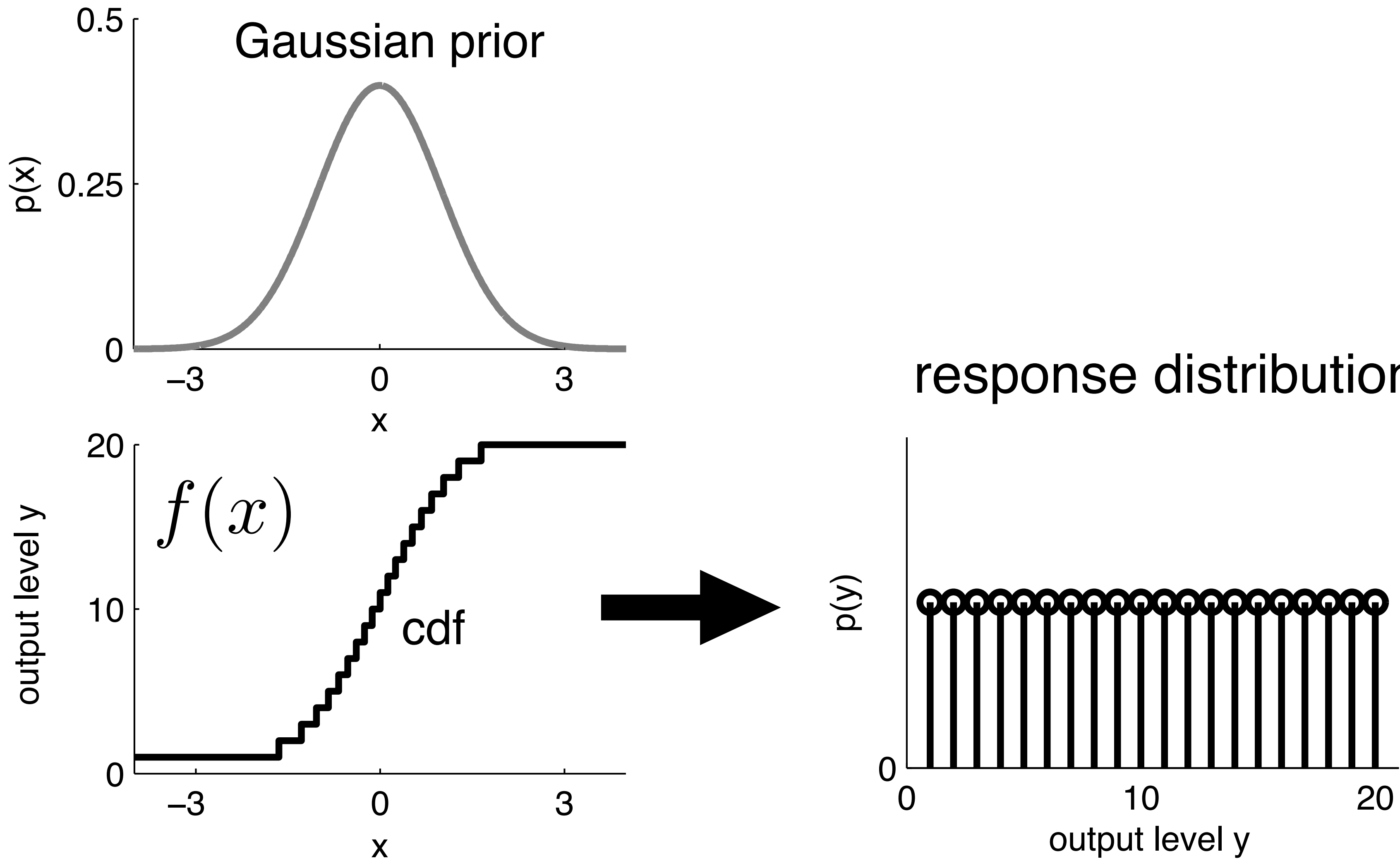
Example: single neuron encoding stimuli from a distribution P(x)

stimulus prior	$x \sim P(x)$	
noiseless, discrete encoding	$y = f(x)$	(with constraint on range of y values)

Application Example: single neuron encoding stimuli from a distribution P(x)

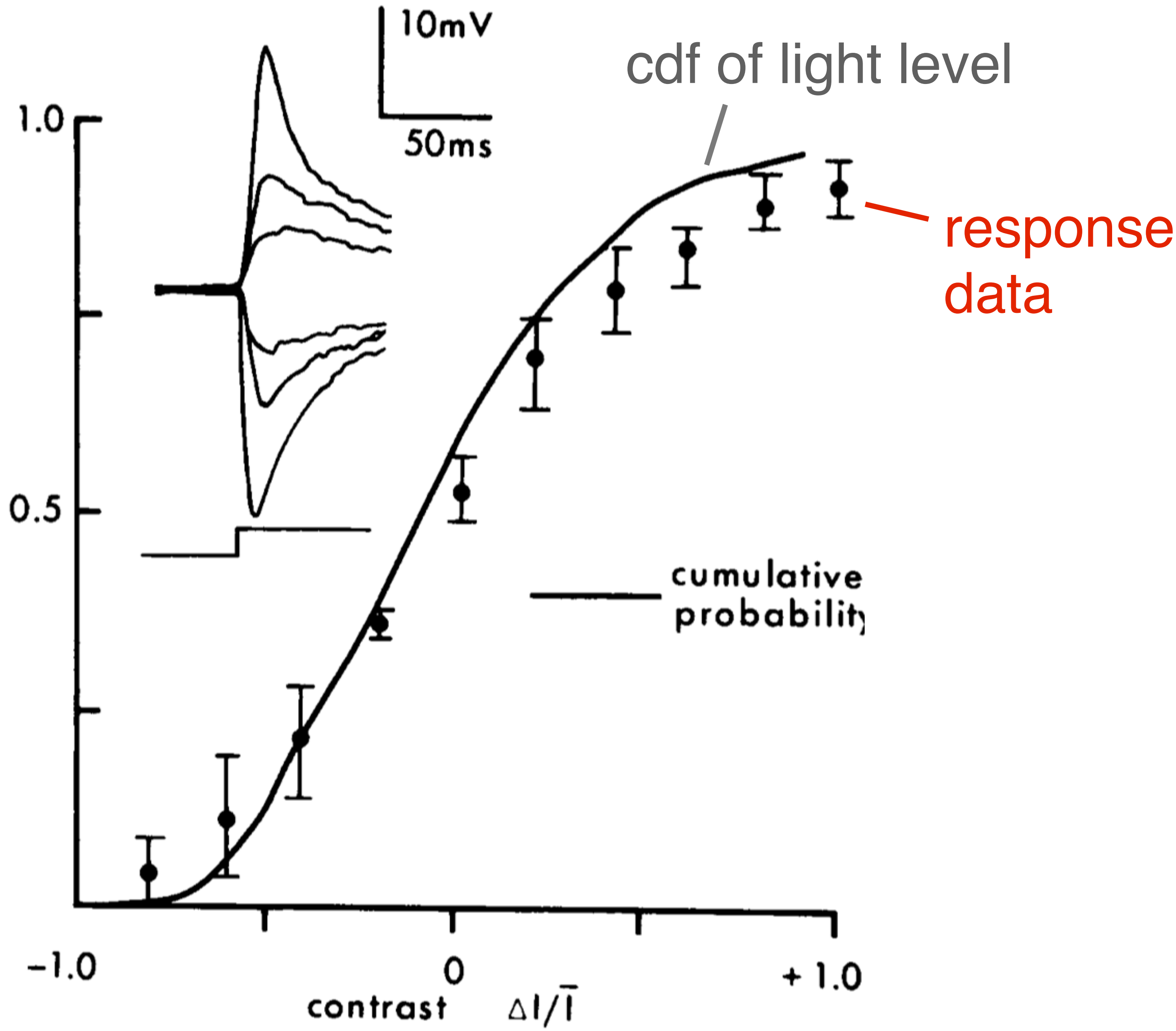
stimulus prior $x \sim P(x)$

noiseless, discrete encoding $y = f(x)$ (with constraint on range of y values)



Laughlin 1981: blowfly light response

- first major validation of Barlow's theory



summary

- entropy
- negative log-likelihood / N
- conditional entropy
- mutual information
- data processing inequality
- efficient coding hypothesis (Barlow)
  - neurons should "maximize their dynamic range"
  - multiple neurons: marginally independent responses
- direct method for estimating mutual information from data