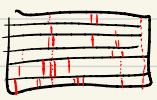
Information and Coding Principles



← image

Information Theory as a way to systematize how useful we can make the data we collect from spikes

Goals:
- Defining entropy and information
- Computing information for neural spike trains
- What can information tell us about coding?

Suppose $P(1) = p$
$P(0) = 1 - p$

Surprise $\text{Information}(1) = -\log_2(p)$

Each bit of information specifies location by an additional factor of 2!

$-\log(\frac{1}{2} \cdot \frac{1}{2}) = 2$ bits

1 bit: know which half of region something is located in

↳ where probabilities multiply, bits will add

Entropy = average information

$= -\sum_i \log_2(p_i)$

$= -\int dx\, p(x) \log_2 p(x)$ * Assume base of log in this context will be 2 hereon out

(Units are bits)

→ Count # of yes-no questions as it takes to specify a variable

- Where's my card?

↳ 3 divisions

Entropy $H = -\sum p_i \log p_i$

$p_i = 1/8$ (all cards have 1/8 prob of being mine)

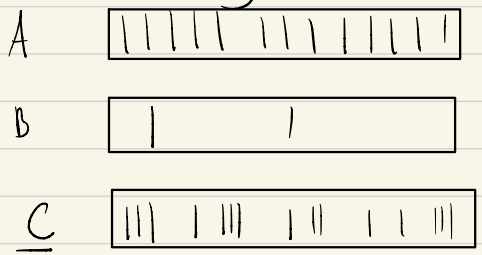
$H = -\sum (1/8 \log_2(1/8))$ $\quad 2^3 = 8$

$= -\sum_i (1/8(-3))$

Note that $i = \{1, 2, \ldots, 8\}$ as there are 8 possibilities

$= 1/8(3) \cdot 8 = 3$

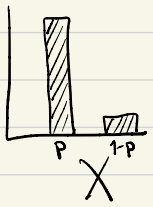
Quantifying Variability

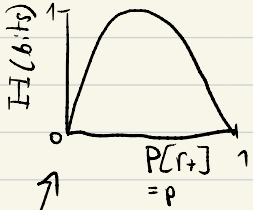


C has the most intrinsic capability for encoding, ie, ability to generate stimulus driven variation in output

Recall Entropy = avg information
$= -\sum_i^2 p_i \log_2(p_i)$

Maximizing Entropy $P(x)$





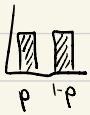
- More intrinsic variability, more capacity for representation

$$-\sum p_i \log(p_i)$$

$$= -p\log(p) - (1-p)\log(1-p)$$

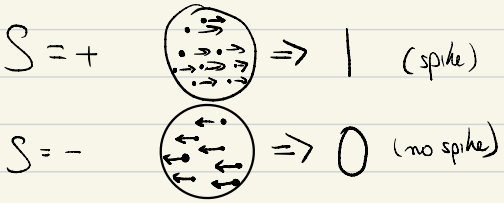
$P(x)$

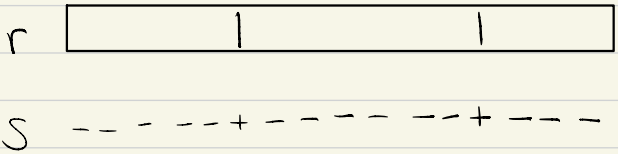
maximum @ $p = 1/2$:

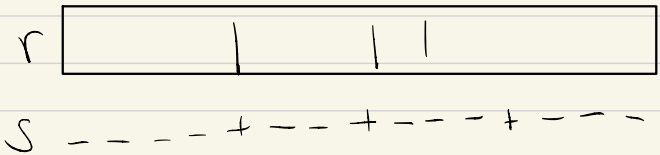


- Generally our $P(x)$ is not uniform, but convenient if it were!

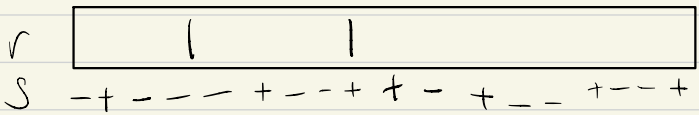
Relating to stimulus







• most likely related



?

Q: How much of the variability in r is encoding s?

$s = +$  $\Rightarrow 1$, $\rightarrow 0$

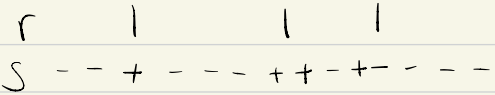
$P(r_+ | +) = 1 - q$

$P(r_- | +) = q$ (error probability q), call noise

$s = -$  $\Rightarrow 0$, $\rightarrow 1$

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

$P(r_+ | -) = q$



How much Entropy is accounted for from noise?

$$H = -\sum_{i=1}^{2} p_i \log_2(p_i)$$

Total Entropy: $H[R] = -P(r_+)\log P(r_+) - P(r_-)\log P(r_-)$

Noise Entropy: $H[R|+] = -q\log q - (1-q)\log(1-q)$

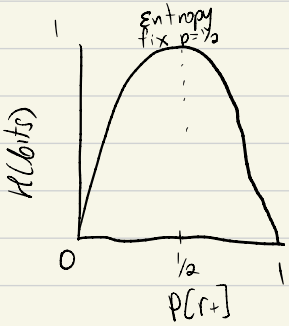
Goal: Determine how much entropy to assign to the noise

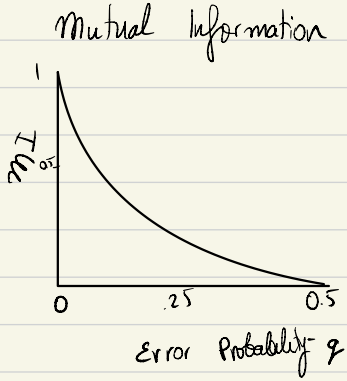
Mutual Information

- Amount of information used in coding stimulus

$MI(S,R)$ = Total entropy − average noise entropy

$$\Rightarrow MI = -\underbrace{\sum_r p(r)\log_2 p(r)}_{\text{Total}} - \underbrace{\sum_s p(s)\left[-\underbrace{\sum_r p(r|s)\log_2 p(r|s)}_{\text{conditional}}\right]}_{\text{averaged over } s}$$





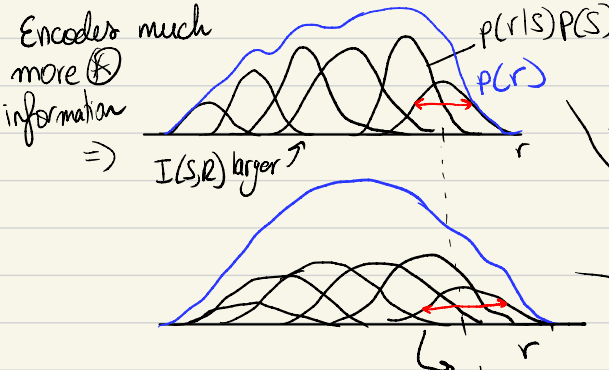
- if no noise entropy, information is maximal, 1 bit
- as error probability increases, spiking is less and less likely to accurately represent the stimulus S
- and Mutual Information Decreases
- When error probability reaches ½, that is, response occurs at total chance (50%), there is no mutual information between r and S.

What are limits of this

- If Response unrelated to stimulus, then independent &
  $P(r|s) = P(r)$
  $MI = 0$

- If Response Perfectly related to stimulus, then
  MI = Total entropy
  ↳ all of the response's coding capacity is used in encoding the stimulus

For continuous variables



Entropy of blue: Total Entropy H[R]
Entropy of conditionals: H[R|s]
avg of stimulus that drove them: $\sum_i P(s_i) H[R|s_i]$

Differ by amount of intrinsic noise that response has

↳ the stimulus affects different variance (in second, degree of noise stretches a much larger range of response distribution)

- much more of the variability of r accounted by variability in responses to specific stimuli

$$I(R,S) = H[R] - \sum_s p(s)\, H[R|s]$$

Information quantifies how independent R and S are

$$I(S;R) = D_{KL}\left[\underbrace{P(R,S)}_{\text{joint distribution}}, \underbrace{P(R)P(S)}_{\text{dist. if independent}}\right] \quad \text{(K-L divergence)}$$

(difference b/w)

Recall $D_{KL}(P,Q) = \int dx\, P(x) \log \frac{P(x)}{Q(x)}$

$$\Rightarrow \int ds\, dr\, P(s,r) \log \frac{P(r,s)}{P(r)P(s)} = \int ds\, dr\, P(s,r) \log \frac{P(r|s)\,P(s)}{P(r)P(s)}$$

$$= \int ds\, dr\, P(s,r)\left[\log P(r|s) - \log P(r)\right]$$

$$= -\int ds\, dr\, P(s,r) \log(P_r) \Leftarrow H[P(r)]$$

$$+ \int ds\, dr\, P(s)P(r|s) \log P(r|s)$$

$\int ds P(s) H[P(r|s)]$

Hence $I(S,R) = H[R] - \sum_s P(s)\, H[R|s]$

and $I(S,R) = H[S] - \sum_r P(r)\, H[S|R]$

*Info symmetric b/w two variables

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

Algorithm for Mutual Information

- Take one stimulus s and repeat many times to obtain P(R|s)
- Compute variability due to noise: noise entropy H(R|s)
- Repeat for all s and average $\sum_s P(s)\, H[R|s]$
- Compute $P(R) = \sum P(s)P(R|s)$ and total energy H(R)

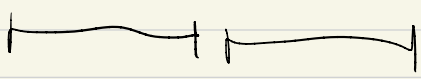
Next using this for calculating information in spike trains

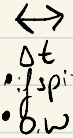
2 methods:
- Information in spike patterns
- Information in single spikes

- What information is being carried by patterns of spikes?

Idea: Take binary strings and group together

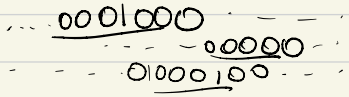
1 0 0 1 0 1 1 0 1 0 0 0 1 0 0 1 0 0 1 1 1 0



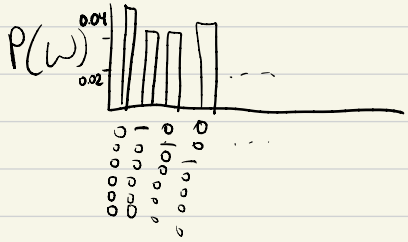


& record the generated words of length T

Ex) Binary words w w/ letter size Δt, length T



Entropy



Compute $p(w_i)$

$$H[w] = -\sum p(w_i) \log_2 p(w_i)$$

Information: difference b/w total variability driven by stimuli and that due to noise, averaged over stimuli

noise entropy:

*Repeated stimulus is still random, but the same segment of random noise is repeated over and over

Applying the algorithm for mutual information:

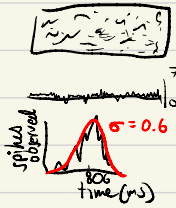
- Take a stimulus sequence and repeat many times
- Sample $P(s)$: Average over $s \rightarrow$ average over time (time standing in for avg over stimulus)
  - For each time in repeated stimulus, get a set of words $P(w|s(t))$

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

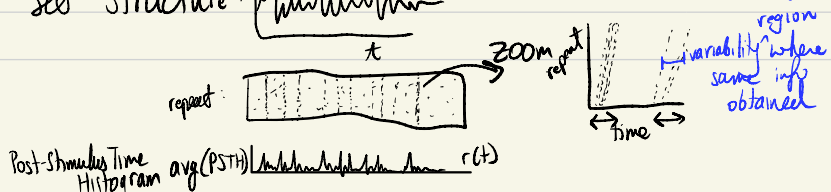
- Choose length of repeated sequence long enough to sample the noise entropy adequately

Ex: Reinagel & Reid (2000)

- Ran Random Stimulus over many trials



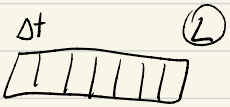
- Ran fixed Stimulus say, $s(t)$, frozen white noise w/ set structure



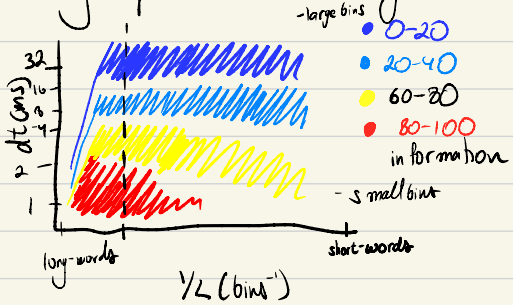
Time scale of above jitter

- Some jitter expected as a result of noise inherent in neural activity
- Information about a stimulus can still be reliably transmitted even if jitter exists
- The greater the width of this jitter, the less accurate we are that the cell is accurately representing the stimulus

Method for determining width of cells



- Vary parameters of word δt (δt) and L



Here good width ~ 1 ms

Sampling + Bias

- Never enough data!
- Correction for finite same size
- Panzeri, Nemenman, etc

Info in Single Spikes

- By how much does knowing that a particular stimulus occurred reduce the entropy of the response

Probability a single spike occured, $P(r=1) = \bar{r}\Delta t$, avg firing rate * bin size

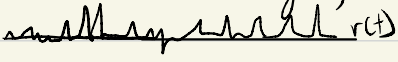
prob no spike, $P(r=0) = 1-\bar{r}\Delta t$

Probability of spike during presentation of stimulus $P(r=1|s) = r(t)\Delta t$

$r(t)$: time varying rate caused by changing stimulus

$P(r=0|s) = 1-r(t)\Delta t$

*$r(t)$ = PSTH in case before,



Computing Entropy difference: $p=\bar{r}\Delta t$, $p(t) = r(t)\Delta t$

- Recall Entropy of Response = Total Entropy - Noise Entropy

$$I(r,s) = -p\log p - (1-p)\log(1-p) \quad \text{Total}$$

$$+ \frac{1}{T}\int_0^T dt\,[p(t)\log p(t) + (1-p(t))\log(1-p(t))] \quad \text{Noise (function of } t\text{)}$$

- Every time t stands in for a sample of s
- A time average is equivalent to averaging over the ensemble
  - ↳ Ergodicity

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

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

$\Rightarrow$ To get information per spike, divide by $\bar{r}\Delta t$:

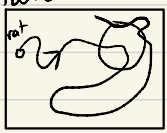
$$I(r,s) = \frac{1}{T}\int_0^T dt\, \frac{r(t)}{\bar{r}} \log \frac{r(t)}{\bar{r}}$$

- no explicit stimulus dependence (no coding/decoding model)
- rate r does not just define rate of spikes, rate of any event

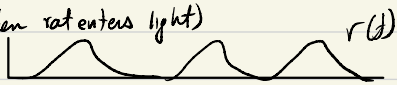
limits of info: - spike precision, blurring $r(t)$
- mean spike rate

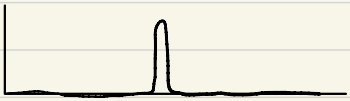
place-field cartoon

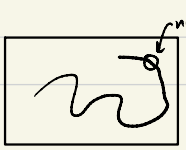
ex:



firing rate:





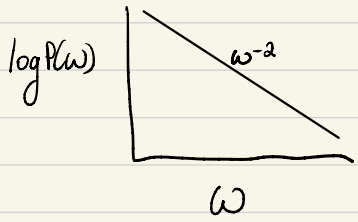


Next: Information & Coding efficiency

- What do information-theoretic concepts suggest that neural systems should do?
- What principles seem to work in shaping neural code?

Natural stimuli: ① Huge dynamic range - variations over many orders of magnitude

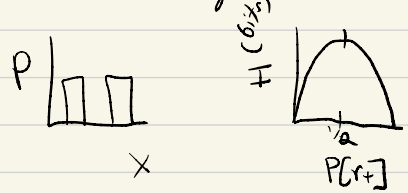
② Power Law Scaling



- Structure at many scales

Goal ∴ encode information as well as possible

What makes a good code?

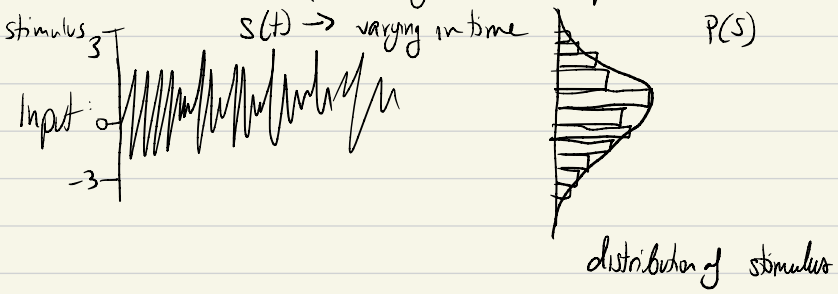


- We found entropy maximized @ ½, equal prob.

Now goal ∴ maximize Maximum Information

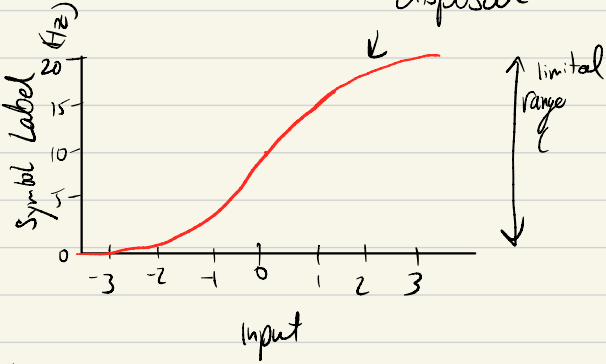
$$I(R,S) = H[R] - \langle H[R|S] \rangle_S$$

*In order to maximize entropy output, a good encoder should match its outputs to the distribution of its inputs



Our job as encoder: map stimulus onto symbols at our disposal

Say, we are constrained to max. firing rate.



- The input/output function should be determined by the distribution of natural inputs
- Optimizes mutual info. b/w input & output.

(!) Get most info by maximizing output entropy, that is, using all our symbols about equally as often.

In our example, 20 symbols @ disposal,
- set first 1Hz as the P(S)ds
- Called Histogram equalization

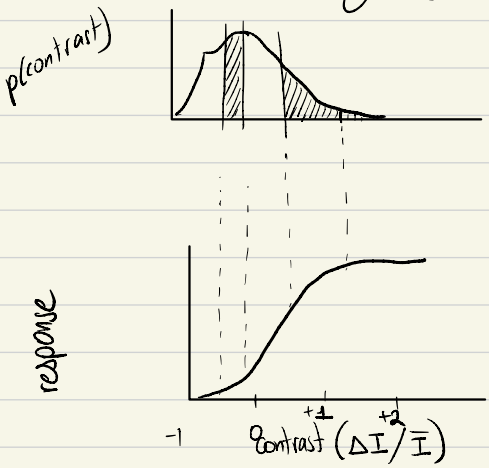
Study of ~~this~~ system: Laughlin, 1981

- Fly visual system

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

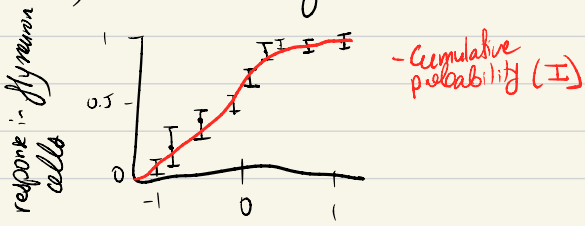
$$r = g(s) = \frac{1}{2}\int_{-1}^{s} d\sigma\, P(\sigma).$$

Experiment: go out into world w/ its natural stimuli and record contrast, (deviations in light-level)/mean light level, experienced by, say, a fly.

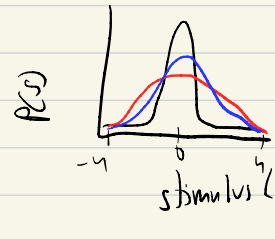


- if the response does indeed follow the distribution of natural inputs, then response curve shown should look like cumulative probability determined by integrating P(σ)

* Indeed, it was a good match to what was observed



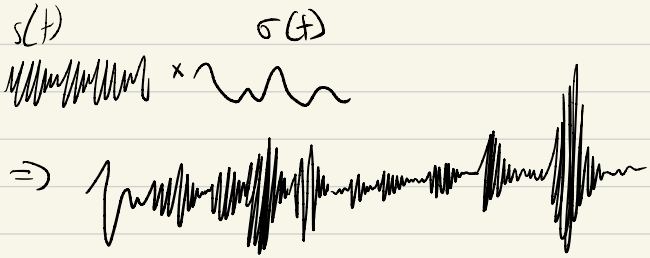
patches of visual info can vary widely



*Our code should take the widths of these distributions into account in setting up our local input-output curve accomodating current statistics of input.

Time-varying stimulus representation

· take white noise input, s(t)



-repeat σ(t), change s(t) in every trial

* allowed us to pick at time-spikes that occured in each presentation

-Analyzing those spikes and plotting input-output functions,

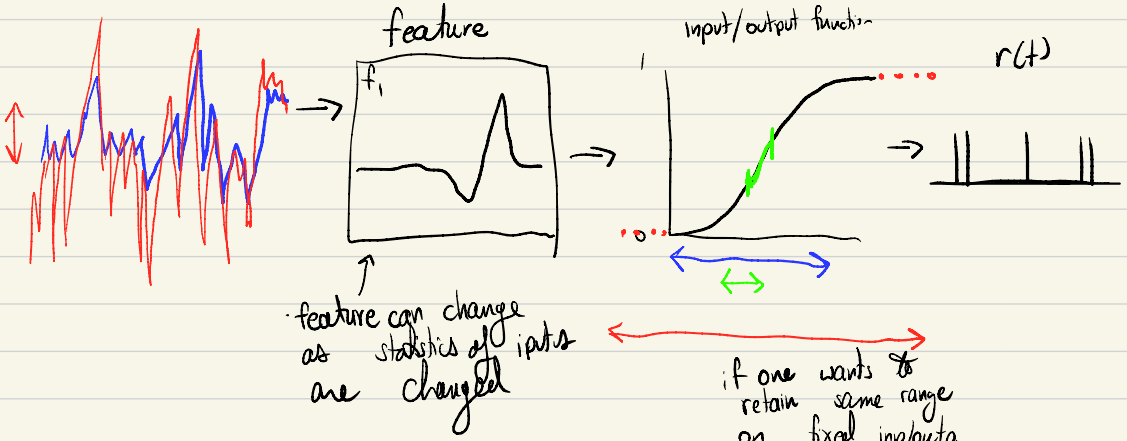
-varies based on behavior of spike, broad spike -broad curve

Conceptual ex:

- If fluctuations in some stimulus start out large and suddenly decrease in amplitude, the optimal input-output curve for a sensory neuron encoding the stimulus will contract, (ie, become more steep)

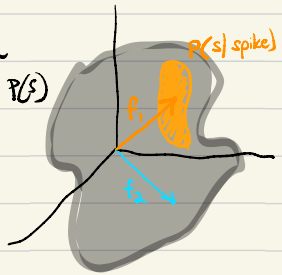
↳ If amplitude of stimulus fluctuations decreases, the prob. distribution of stimulus values become skinnier

* If divided by standard deviation, curves match.

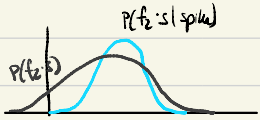


Information Theory ⇒ find what it is about stimulus that drives a neuron to fire

- Feature Adaption



return of maximally informative dim.



* Choose filter maximizing D_{KL} b/w spike-conditional and prior distributions

⇐⇒ equivalent to.

maximizing information the the spike provides about the stimulus

- Redundancy Reduction

hypothesis: max. info. when indep. (maximizes entropy)

- Population code (joint dist): $P(R_1, R_2) \sim P(R_1)P(R_2)$

$$H[R_1,R_2] \leq H[R_1] + H[R_2]$$

- However, realized some correlation can be good
  - error correction + robust coding
  - correlations can help discrimination
  - Indeed, neurons in retina appear semi-redundant

New hypothesis: as sparse as possible: redundancy redux
- as few neurons as possible are firing at any time

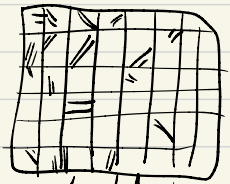
- Representing natural scenes sparsely

$$I(\bar{x}) \approx \sum_i a_i \phi_i(\bar{x}) + \varepsilon(\bar{x})$$

$$E = \sum_{\bar{x}} \left[I(\bar{x}) - \sum_i a_i \phi(\bar{x}) \right]^2 + \lambda \sum_i C(a_i) \qquad C(a_i) = |a_i|$$

(square difference of image & reconstruction; C: cost of coefficients needed)

⇒



↑ weighted constraint

goal: penalize too many basis functions

as few as possible & represent an image

Fourier Basis: Might code image really well, but definitely won't be sparse

Coding Principles
- coding efficiency
- adaptation & stimulus statistics
- Sparseness

Methods:

- Models for predicting how stimuli are coded in spikes
- Models for decoding stimuli from neural responses
- Information Theory and how it is used to evaluate coding schemes
- Quick glance at how coding strategies might be shaped by the statistics of natural inputs

Missed: Cycle of behavior