Population Coding

Maneesh Sahani
`maneesh@gatsby.ucl.ac.uk`

Gatsby Computational Neuroscience Unit
University College London

Term 1, Autumn 2010

Coding so far ...

- Time-series for both spikes and stimuli
- Empirical — estimate function(s) given measured stimuli or movements and spikes

Population codes

- High dimensionality (cells × stimulus × time).
  - usually limited to simple rate codes.
  - even prosthetic work assumes instantaneous (lagged) coding

- Limited empirical data
  - can record 10s - 100s of neurons.
  - population size more like 10^4 - 10^6.
  - *theoretical* inferences, based on single-cell and aggregate (fMRI, LFP, optical) measurements .

Common approach

The most common sort of questions asked of population codes:

- given assumed encoding functions, how well can we (or downstream areas) decode the encoded stimulus value?
- what encoding schemes would be optimal, in the sense of allowing decoders to estimate stimulus values as well as possible.

Before considering populations, we need to formulate some ideas about rate coding in the context of single cells.

Rate coding

In the rate coding context, we imagine that the firing rate of a cell r represents a single (possibly multidimensional) stimulus value s at any one time:

$$r = f(s).$$

Even if s and r are embedded in time-series we assume:

1. that coding is instantaneous (with a fixed lag),
2. that r (and therefore s) is constant over a short time Δ.

The actual number of spikes n produced in Δ is then taken to be distributed around $r\Delta$, often according to a Poisson distribution.

Tuning curves

The function $f(s)$ is known as a tuning curve.

Commonly assumed forms:

- Gaussian $\quad r_0 + r_{\text{max}} \exp\left[-\frac{1}{2\sigma^2}(x - x_{\text{pref}})^2\right]$
- Cosine $\quad r_0 + r_{\text{max}} \cos(\theta - \theta_{\text{pref}})$
- Wrapped Gaussian $\quad r_0 + r_{\text{max}} \sum_n \exp\left[-\frac{1}{2\sigma^2}(\theta - \theta_{\text{pref}} - 2\pi n)^2\right]$
- von Mises (“circular Gaussian”) $\quad r_0 + r_{\text{max}} \exp\left[\kappa \cos(\theta - \theta_{\text{pref}})\right]$

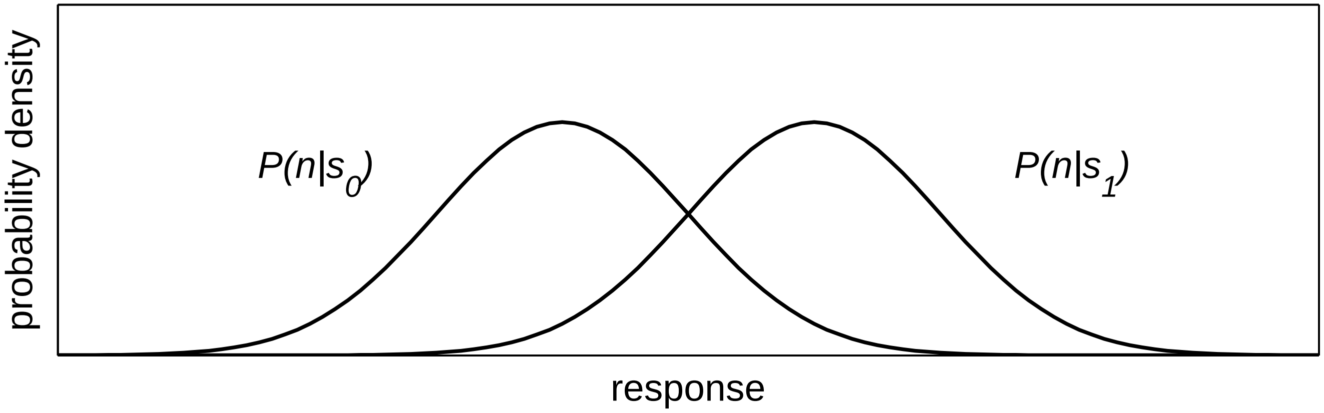
Measuring the performance of rate codes: Discrete choice

Suppose we want to make a binary choice based on firing rate:

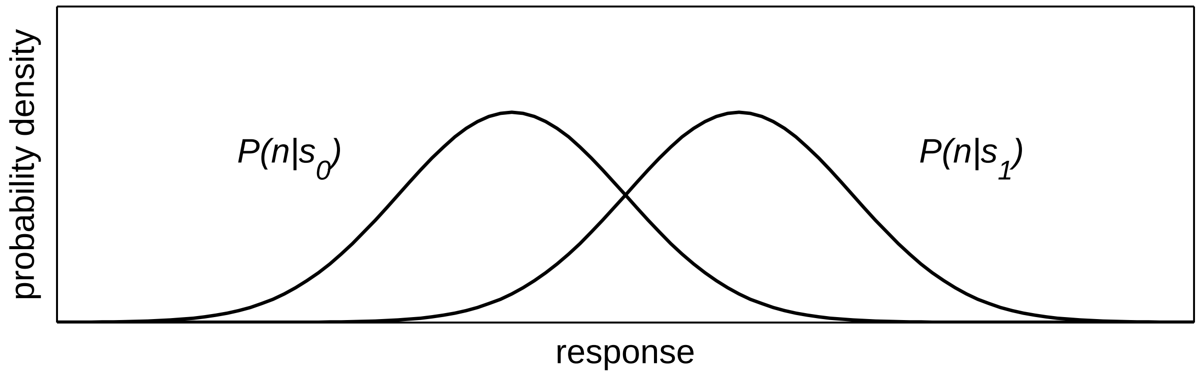
- present / absent (signal detection)
- up / down
- horizontal / vertical

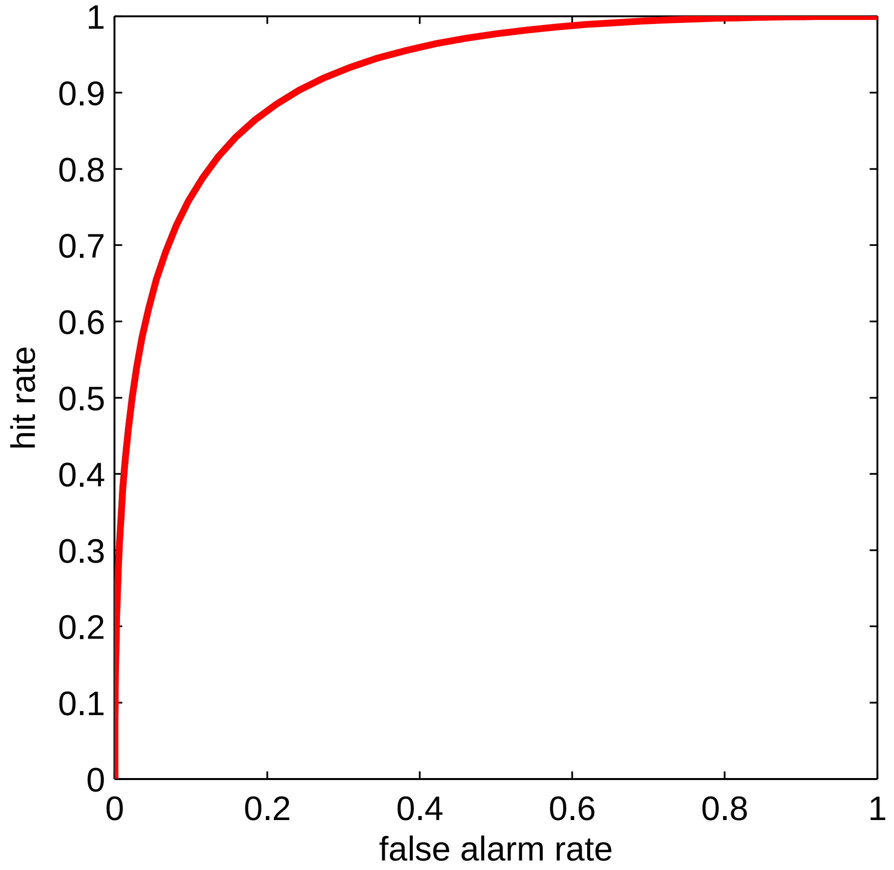
Call one potential stimulus s_0, the other s_1.

$P(n|s)$:

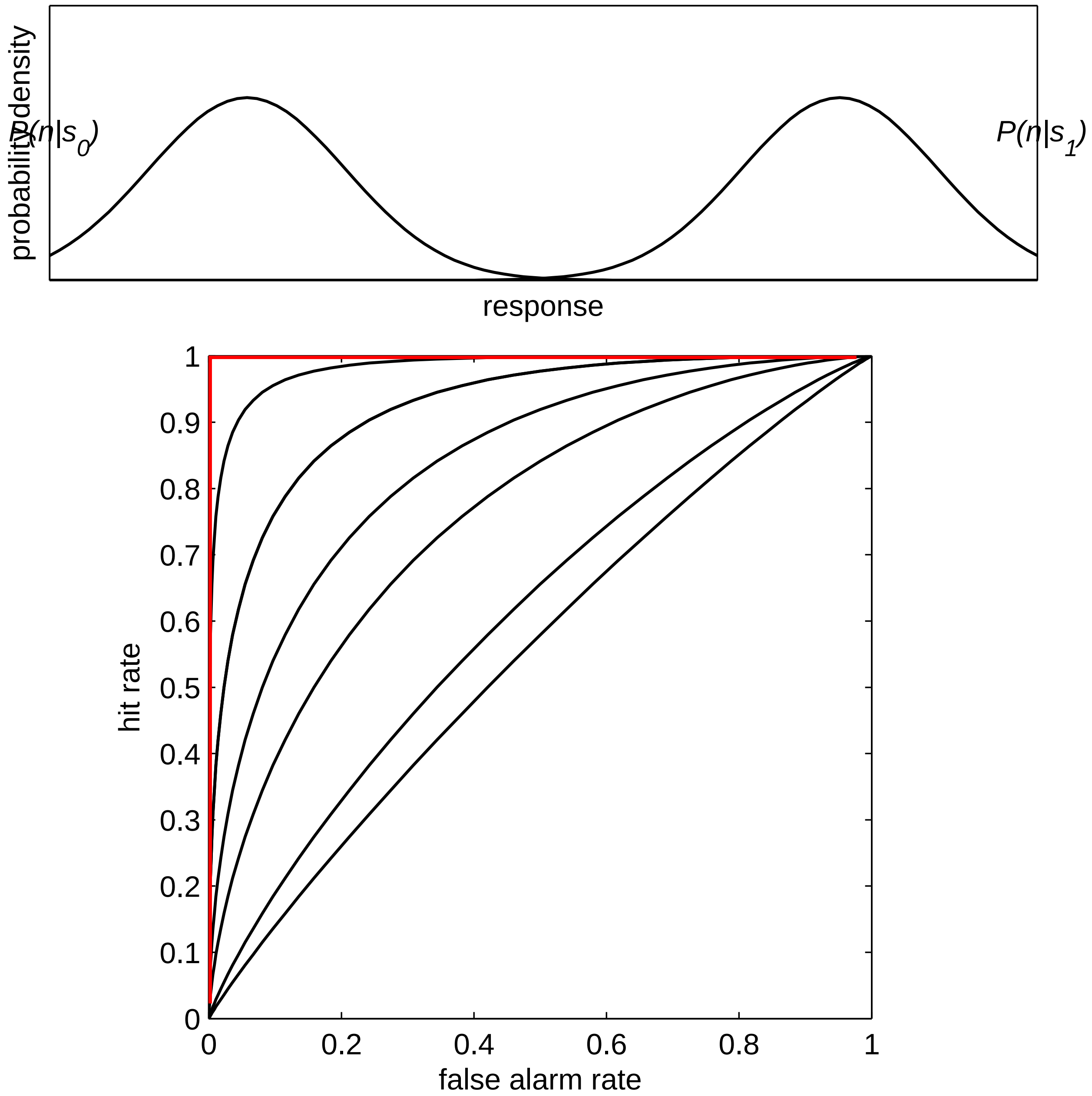


ROC curves


probability density
P(n|s0)
P(n|s1)
response


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
hit rate
0
0.2
0.4
0.6
0.8
1
false alarm rate

ROC curves


probability density
P(n|s0)
P(n|s1)
response
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
hit rate
0
0.2
0.4
0.6
0.8
1
false alarm rate

Summary measures

- area under the ROC curve
  - given $n_1 \sim P(n|s_1)$ and $n_0 \sim P(n|s_0)$, this equals $P(n_1 > n_0)$

- discriminability d'
  - for equal variance Gaussians $d' = \frac{\mu_1 - \mu_0}{\sigma}$.
  - for any threshold $d' = \Phi^{-1}(1 - \text{FA}) - \Phi^{-1}(1 - \text{HR})$ where Φ is a standard normal cdf.
  - definition unclear for non-Gaussian distributions.

Continuous estimation

Now consider a (one dimensional) stimulus that takes any real value (or angle).

- contrast
- orientation
- motion direction
- movement speed

Consider a neuron that fires n spikes in response to a stimulus s, according

$$P(n|f(s)\Delta)$$

Given n we attempt to estimate s. How well can we do?

Continuous estimation

Useful to consider a limit given $N \to \infty$ measurements n_i all generated by the same stimulus s^*. Then the posterior over s is

$$\log P(s|\{n_i\}) = \sum_i \log P(n_i|s) + \log P(s) - \log Z(\{n_i\})$$

and so taking $N \to \infty$

$$\frac{1}{N} \log P(s|\{n_i\}) \to \left\langle \log P(n|s) \right\rangle_{n|s^*} + 0 - \log Z(s^*)$$

and so

$$\begin{aligned} P(s|\{n_i\}) &\to e^{N\left\langle \log P(n|s) \right\rangle_{n|s^*}} / Z \\ &= e^{-N\mathbf{KL}\left[P(n|s^*)\middle\|P(n|s)\right]} / Z \end{aligned}$$

Continuous estimation

Now, Taylor expand the KL divergence in s around s^*:

$$
\begin{aligned}
&\textbf{KL}\left[P(n|s^*)\|P(n|s)\right] \\
&= -\left\langle \log P(n|s)\right\rangle_{n|s^*} + \left\langle \log P(n|s^*)\right\rangle_{n|s^*} \\
&= -\left\langle \log P(n|s^*)\right\rangle_{n|s^*} - (s-s^*)\left\langle \frac{d\log P(n|s)}{ds}\Big|_{s^*}\right\rangle_{s^*} - \frac{1}{2}(s-s^*)^2\left\langle \frac{d^2\log P(n|s)}{ds^2}\Big|_{s^*}\right\rangle_{s^*} + \ldots \\
&\quad + \left\langle \log P(n|s^*)\right\rangle_{n|s^*} \\
&= -\frac{1}{2}(s-s^*)^2\left\langle \frac{d^2\log P(n|s)}{ds^2}\Big|_{s^*}\right\rangle_{s^*} + \ldots \\
&= \frac{1}{2}(s-s^*)^2 J(s^*) + \ldots
\end{aligned}
$$

So in asymptopia, the posterior $\to \mathcal{N}\left(s^*, 1/J(s^*)\right)$.

$J(s^*)$ is called the *Fisher Information*.

$$
J(s^*) = -\left\langle \frac{d^2\log P(n|s)}{ds^2}\Big|_{s^*}\right\rangle_{s^*} = \left\langle \left(\frac{d\log P(n|s)}{ds}\Big|_{s^*}\right)^2\right\rangle_{s^*}
$$

(You will show that these are identical in the homework.)

Cramér-Rao bound

The Fisher Information is important even outside the large data limit due to a deeper result that is due to Cramér and Rao.

This states that for any N, any *unbiased* estimator $\hat{s}(\{n_i\})$ of s will have the property that

$$\left\langle (\hat{s}(\{n_i\}) - s^*)^2 \right\rangle_{n_i|s^*} \geq \frac{1}{J(s^*)}.$$

Thus, Fisher Information gives a lower bound on the variance of any unbiased estimator. This is called the Cramér-Rao bound. (There is also a version for biased estimators).

The Fisher Information will be our primary tool to quantify the performance of a population code.

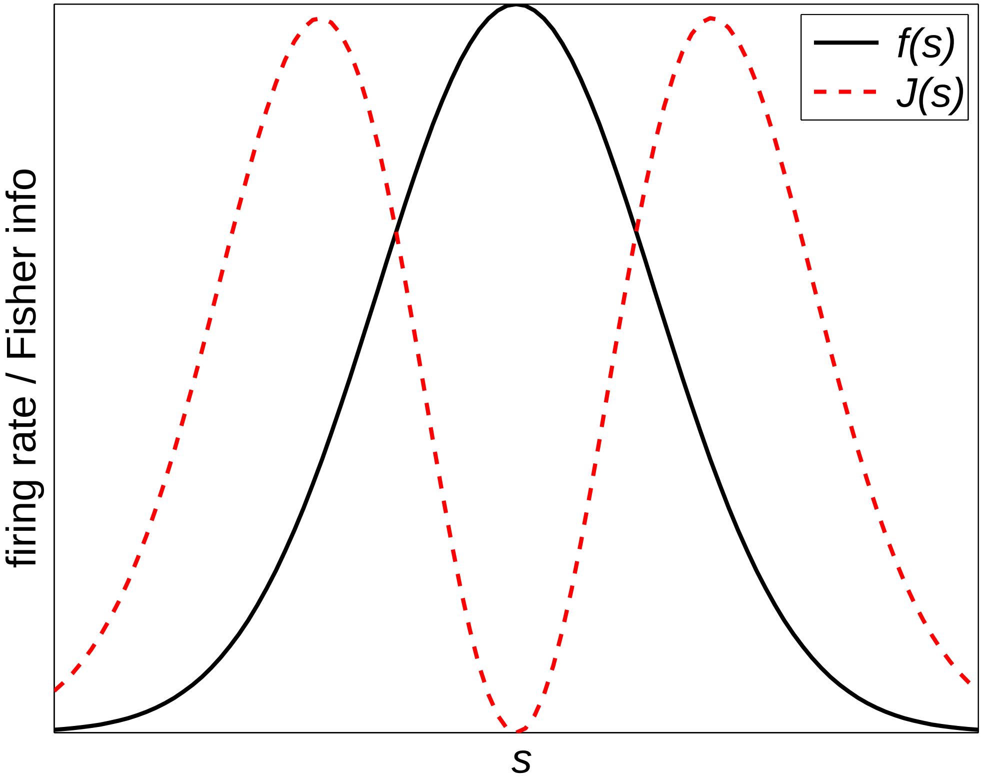
Fisher Info and tuning curves

$$n = r\Delta + \text{noise}; \quad r = f(s) \Rightarrow$$

$$J(s^*) = \left\langle \left(\frac{d}{ds}\bigg|_{s^*} \log P(n|s) \right)^2 \right\rangle_{s^*}$$

$$= \left\langle \left(\frac{d}{dr\Delta}\bigg|_{f(s^*)} \log P(n|r\Delta)\Delta f'(s^*) \right)^2 \right\rangle_{s^*}$$

$$= J_{\text{noise}}(r\Delta)\Delta^2 f'(s^*)^2$$



Fisher info for Poisson neurons

For Poisson neurons

$$P(n|f(s)) = \frac{e^{-f(s)} f(s)^n}{n!}$$

so

$$\begin{aligned} J_{\text{noise}}[f(s^*)] &= \left\langle \left(\frac{d}{df}\bigg|_{f(s^*)} \log P(n|f(s)) \right)^2 \right\rangle_{s^*} \\ &= \left\langle \left(\frac{d}{df}\bigg|_{f(s^*)} - f(s) + n \log f(s) - \log n! \right)^2 \right\rangle_{s^*} \\ &= \left\langle \left(-1 + n/f(s^*) \right)^2 \right\rangle_{s^*} \\ &= \left\langle \frac{(n - f(s^*))^2}{f(s^*)^2} \right\rangle_{s^*} \\ &= \frac{f(s^*)}{f(s^*)^2} = \frac{1}{f(s^*)} \end{aligned}$$

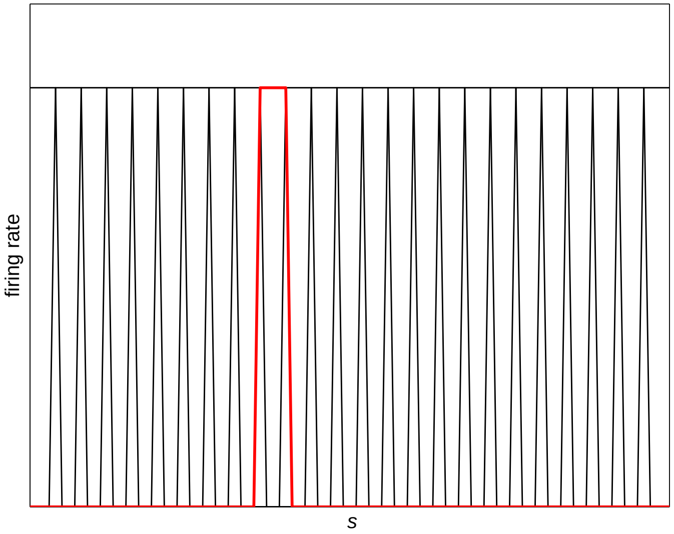
[not surprising! $\widehat{f(s)} = n$]

and

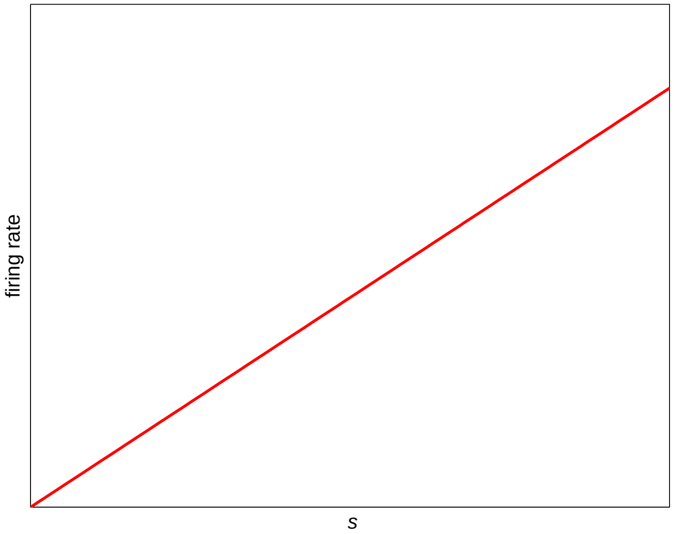
$$J[s^*] = f'(s^*)^2 / f(s^*)$$

Cooperative coding

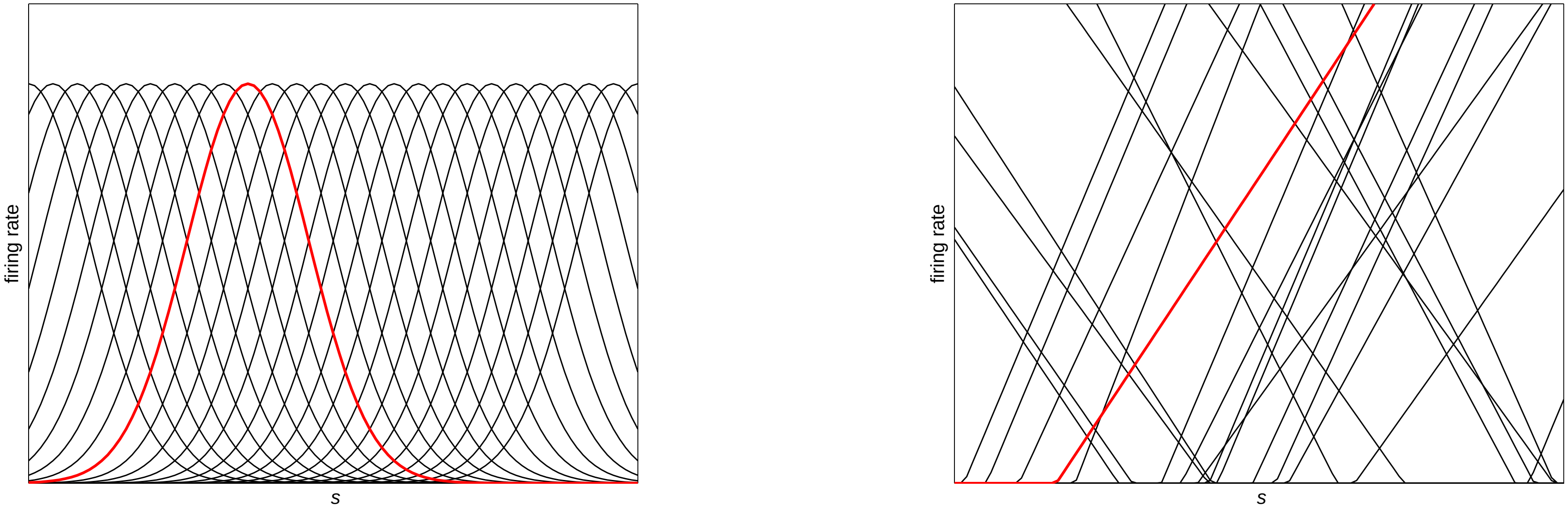
Labelled Line



Scalar coding



Distributed encoding



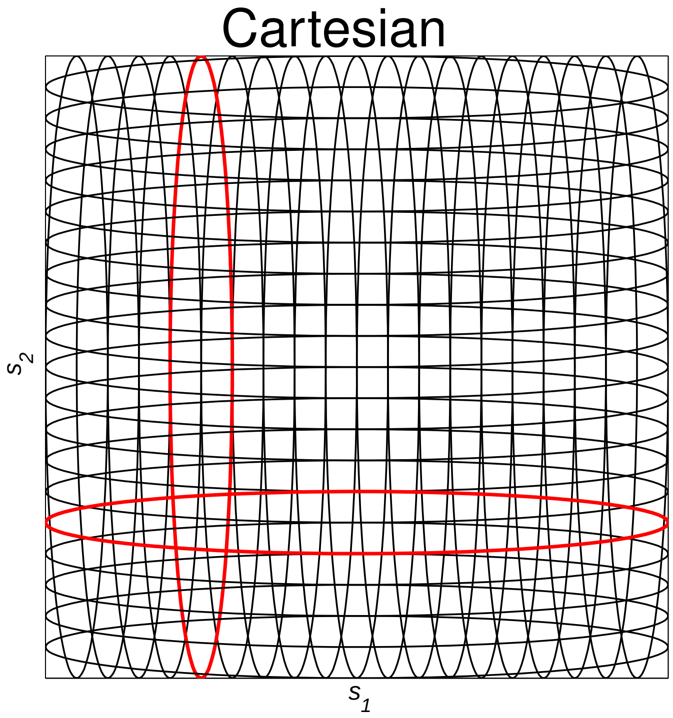
Cooperative coding

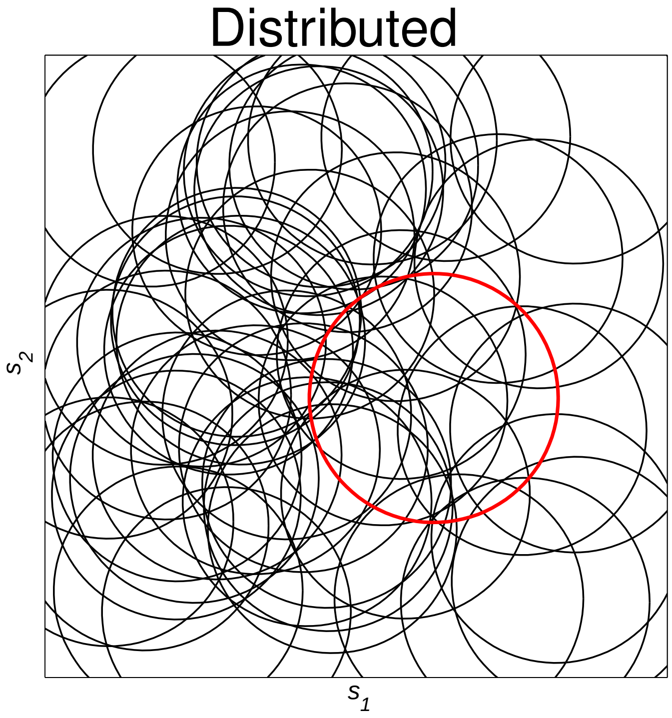
All of these are found in biological systems.

Issues:

1. redundancy and robustness (not scalar)
2. efficiency (not labelled line)
3. local computation (not scalar or distributed)
4. multiple values (not scalar)

Coding in multiple dimensions

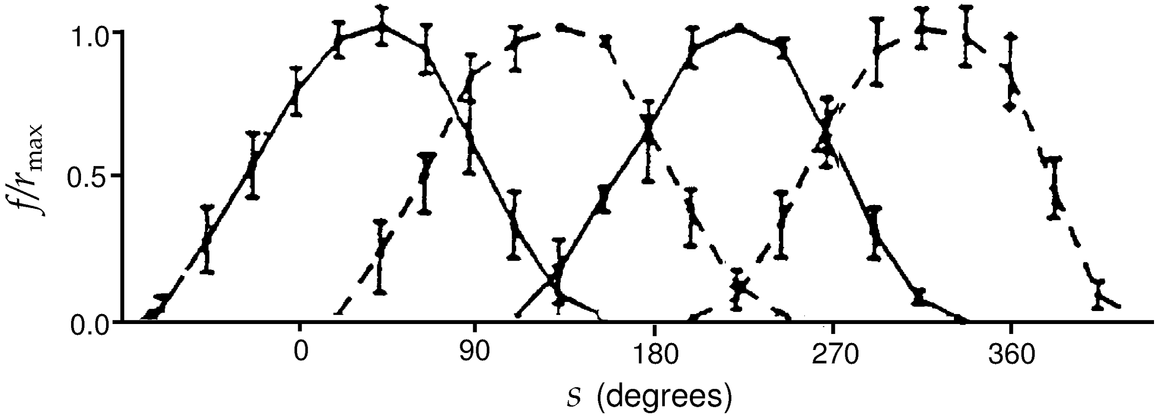




- efficient
- problems with multiple values

- represent multiple values
- may require more neurons

Cricket cercal system



$$r_a(s) = r_a^{\max}[\cos(\theta - \theta_a)]_+ = r_a^{\max}[\mathbf{c}_a^{\mathsf{T}}\mathbf{v}]_+$$

$$\mathbf{c}_1^{\mathsf{T}}\mathbf{c}_2 = 0$$
$$\mathbf{c}_3 = -\mathbf{c}_1$$
$$\mathbf{c}_4 = -\mathbf{c}_2$$

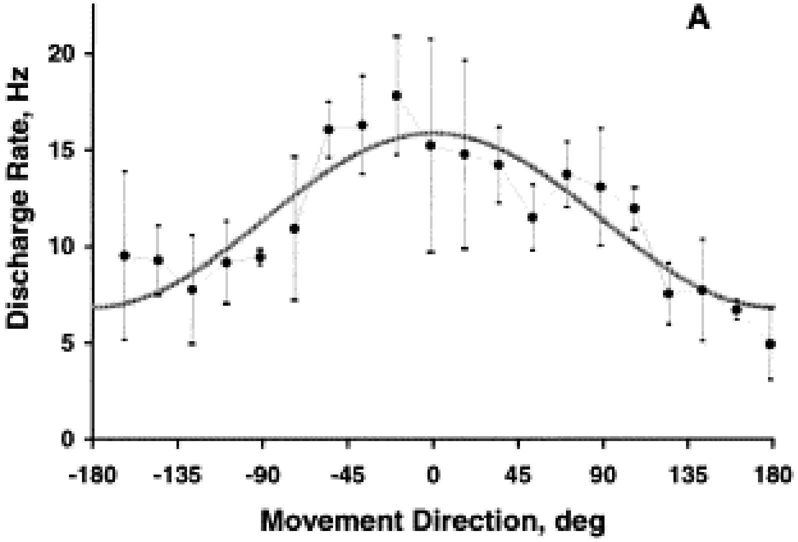
So, writing $\tilde{r}_a = r_a / r_a^{\max}$:

$$\begin{pmatrix} \tilde{r}_1 - \tilde{r}_3 \\ \tilde{r}_2 - \tilde{r}_4 \end{pmatrix} = \begin{pmatrix} \mathbf{c}_1^{\mathsf{T}} \\ \mathbf{c}_2^{\mathsf{T}} \end{pmatrix} \mathbf{v}$$

$$\mathbf{v} = (\mathbf{c}_1 \mathbf{c}_2) \begin{pmatrix} \tilde{r}_1 - \tilde{r}_3 \\ \tilde{r}_2 - \tilde{r}_4 \end{pmatrix} = \tilde{r}_1\mathbf{c}_1 - \tilde{r}_3\mathbf{c}_3 + \tilde{r}_2\mathbf{c}_2 - \tilde{r}_4\mathbf{c}_4 = \sum_a \tilde{r}_a\mathbf{c}_a$$

This is called population vector decoding.

Motor cortex (simplified)



Cosine tuning, randomly distributed preferred directions.

In general, population vector decoding works for

- cosine tuning
- cartesian or dense (*tight*) directions

But:

- is it optimal?
- does it generalise? (Gaussian tuning curves)
- how accurate is it?

Bayesian decoding

Take $n_a \sim$ Poisson$[f_a(s)\Delta]$, independently for different cells.

Then

$$P(\mathbf{n}|s) = \prod_a \frac{e^{-f_a(s)\Delta}(f_a(s)\Delta)^{n_a}}{n_a!}$$

and

$$\log P(s|\mathbf{n}) = -\sum_a f_a(s)\Delta + n_a log(f_a(s)\Delta) - \log n_a! + \log P(s)$$

Assume $\sum_a f_a(s)$ is independent of s for a homogeneous population, and prior is flat.

$$\begin{aligned}\frac{d}{ds}\log P(s|\mathbf{n}) &= \frac{d}{ds}\sum_a n_a log(f_a(s)\Delta) \\ &= \sum_a \frac{n_a}{f_a(s)\Delta} f_a'(s)\Delta\end{aligned}$$

Bayesian decoding

Now, consider $f_a(s) = e^{-(s-s_a)^2/2\sigma^2}$, so $f_a'(s) = -(s-s_a)/\sigma^2 e^{-(s-s_a)^2/2\sigma^2}$ and set the derivative to 0:

$$\sum_a n_a(s - s_a)/\sigma^2 = 0$$

$$\hat{s}_{\text{MAP}} = \frac{\sum_a n_a s_a}{\sum_a n_a}$$

So the MAP estimate is a population average of preferred directions. Not exactly a population vector.

Population Fisher Info

Fisher Informations for independent random variates add:

$$\begin{aligned} J_{\mathbf{n}}(s) &= \left\langle -\frac{d^2}{ds^2} \log P(\mathbf{n}|s) \right\rangle \\ &= \left\langle -\frac{d^2}{ds^2} \sum_a \log P(n_a|s) \right\rangle \\ &= \sum_a \left\langle -\frac{d^2}{ds^2} \log P(n_a|s) \right\rangle = \sum_a J_{n_a}(s). \\ &= \Delta \sum_a \frac{f_a'(s)^2}{f_a(s)} \end{aligned}$$

Optimal tuning properties

A considerable amount of work has been done in recent years on finding optimal properties of tuning curves for rate-based population codes. Here, we reproduce one such argument (from Zhang and Sejnowski, 1999).

Consider a population of cells that codes the value of a D dimensional stimulus, $\mathbf{s}$. Let the ath cell emit r spikes in an interval τ with probability distribution that is conditionally independent of the other cells (given $\mathbf{s}$) and has the form

$$\mathsf{P}_a(r \mid \mathbf{s}, \tau) = S(r, f^a(\mathbf{s}), \tau).$$

The tuning curve of the ath cell, $f^a(\mathbf{s})$, has the form

$$f^a(\mathbf{s}) = F \cdot \phi\left((\xi^a)^2\right); \qquad (\xi^a)^2 = \sum_i^D (\xi_i^a)^2; \qquad \xi_i^a = \frac{s_i - c_i^a}{\sigma},$$

where F is a maximal rate and the function ϕ is monotically decreasing. The parameters $\mathbf{c}^a$ and σ give the centre of the ath tuning curve and the (common) width.

Optimal tuning properties

Now, the (ij)th term in the FI matrix for the ath cell is (by definition)

$$J^a_{ij}(\mathbf{s}) = \mathsf{E}\left[\frac{\partial}{\partial s_i}\log \mathsf{P}^a(r \mid \mathbf{s},\tau)\frac{\partial}{\partial s_j}\log \mathsf{P}^a(r \mid \mathbf{s},\tau)\right]$$

Applying the chain rule repeatedly, we find that

$$\begin{aligned}\frac{\partial}{\partial s_i}\log \mathsf{P}^a(r \mid \mathbf{s},\tau) &= \frac{1}{S(r, f^a(\mathbf{s}),\tau)}\frac{\partial}{\partial s_i}S(r, f^a(\mathbf{s}),\tau)\\ &= \frac{S^{(2)}(r, f^a(\mathbf{s}),\tau)}{S(r, f^a(\mathbf{s}),\tau)}\frac{\partial}{\partial s_i}f^a(\mathbf{s})\end{aligned}$$

(where $S^{(2)}$ indicates differentiation with respect to the second argument)

$$\begin{aligned}&= \frac{S^{(2)}(r, f^a(\mathbf{s}),\tau)}{S(r, f^a(\mathbf{s}),\tau)}F\phi'\left((\xi^a)^2\right)\frac{\partial}{\partial s_i}\sum_i^D(\xi^a_i)^2\\ &= \frac{S^{(2)}(r, f^a(\mathbf{s}),\tau)}{S(r, f^a(\mathbf{s}),\tau)}F\phi'\left((\xi^a)^2\right)\frac{2(s_i - c^a_i)}{(\sigma^a_i)^2}\end{aligned}$$

Optimal tuning properties

So,

$$\begin{aligned} J^a_{ij}(\mathbf{s}) &= \mathsf{E}\left[\left(\frac{S^{(2)}(r, f^a(\mathbf{s}), \tau)}{S(r, f^a(\mathbf{s}), \tau)}\right)^2\right] 4F^2 \left(\phi'\left((\xi^a)^2\right)\right)^2 \frac{(s_i - c^a_i)(s_j - c^a_j)}{\sigma^4} \\ &= A_\phi\left((\xi^a)^2, F, \tau\right) \frac{(s_i - c^a_i)(s_j - c^a_j)}{\sigma^4} \end{aligned}$$

where the function A_ϕ does not depend explicitly on σ.

Optimal tuning properties

We assumed neurons were independent ⇒ Fisher information adds. Approximate by integral over the tuning curve centres, assuming uniform density η of neurons.

$$\begin{aligned}
J_{ij}(\mathbf{s}) &= \sum_a J^a_{ij}(\mathbf{s}) \\
&\approx \int_{-\infty}^{+\infty} dc^a_1 \cdots \int_{-\infty}^{+\infty} dc^a_D \; \eta J^a_{ij}(\mathbf{s}) \\
&= \int_{-\infty}^{+\infty} dc^a_1 \cdots \int_{-\infty}^{+\infty} dc^a_D \; \eta A_\phi \left((\xi^a)^2, F, \tau\right) \frac{(s_i - c^a_i)(s_j - c^a_j)}{\sigma^4}
\end{aligned}$$

Change variables: $c^a_i \to \xi^a_i$

$$\begin{aligned}
&= \int_{-\infty}^{+\infty} \sigma d\xi^a_1 \cdots \int_{-\infty}^{+\infty} \sigma d\xi^a_D \; \eta A_\phi \left((\xi^a)^2, F, \tau\right) \frac{\xi^a_i \xi^a_j}{\sigma^2} \\
&= \frac{\sigma^D}{\sigma^2} \eta \int_{-\infty}^{+\infty} d\xi^a_1 \cdots \int_{-\infty}^{+\infty} d\xi^a_D \; A_\phi \left((\xi^a)^2, F, \tau\right) \xi^a_i \xi^a_j
\end{aligned}$$

Now, if $i \neq j$, integral is odd in both ξ^a_i and ξ^a_j, and thus vanishes. If $i = j$, then the integral has some value $D \cdot K_\phi(F, \tau, D)$, independent of σ. Thus,

$$J_{ii} = \sigma^{D-2} \eta D K_\phi(F, \tau, D)$$

and the total Fisher information is proportional to σ^{D-2}.

Optimal tuning properties

Thus optimal tuning width depends on the stimulus dimension.

- $D = 1$

  $\Rightarrow \sigma \to 0$ (although a lower limit is encountered when the tuning width falls below the inter-cell spacing)

- $D = 2$

  $\Rightarrow J$ independent of σ.

- $D > 2$

  $\Rightarrow \sigma \to \infty$ (actual limit set by valid stimuli).

More ...

- Correlated noise
- Extended s (feature maps etc.)
- Uncertainty