Decoding

The brain doesn't decode, but we do:
this bounds how much information the brain has.

The brain could have more info, or use less!
Look for a match between theory and data
(eg. behavioral performance on some task
vs
optimal linear decoding of some population)

$s \longrightarrow r$ What can we say about s given observations r?

Two approaches:
Estimation: pick one $\hat{s}$
Bayesian: find distribution $p(s|r)$ (posterior)

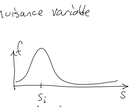
What is s?
Image? Feature? Observable? Rewarded task variable? Causal/latent variable? (requires internal model)

Example: orientation

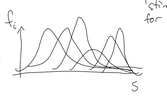
s = angle: 0 ... π ...

c = contrast: 0 ... 1 nuisance variable

Tuning curves: $\langle r|s\rangle = f(s)$



Population:



Neural responses are stochastic: $p(r|s)$

Maximum (likelihood) (ML) estimation: $\hat{s} = \arg\max_s p(r|s)$

Example: poisson: $r \sim \text{Poiss}(f(s))$

$$p(r|s) = \prod_i \frac{e^{-f_i} f_i^{r_i}}{r_i!} = \frac{e^{-\sum_i f_i(s) + \sum_i r_i \log f_i(s)}}{Z(r)}$$

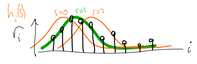
View as function of s, not r, so $Z(r)$ doesn't matter.

If $\sum_i f_i(s) \approx$ const, then ML is template matching:

$p(r|s) \propto e^{r \cdot \log f(s)} = e^{r \cdot h(s)}$

$\arg\max p = \arg\max \log p \propto \arg\max r \cdot h(s)$

Which $h(s)$ has greatest overlap with r?



More generally, ML can be viewed as template matching with a metric that depends on the noise distribution.

Including $\eta(s) = \sum_i f_i(s)$ gives a weighting to different possible s.
This has the same effect as a prior $p(s)$

If we include info about the prior, and base our estimate on $p(s|r)$ then we can find

MAP = Maximum A Posteriori

In general, we have
$p(r,s) = e^{A(s) + B(s,r) + C(r)}$
A(s): indep of r; B(s,r): couples s to r; C(r): indep of s

Max decoding is not the only estimator...

Different cost functions favor different estimators

- min absolute error $\langle|\hat{s}-s|\rangle$: median
- min MSE $\langle(\hat{s}-s)^2\rangle$: mean
- min binary loss $\langle\delta(\hat{s}-s)\rangle$: mode (eg. MAP)

Properties of estimators:

bias: $\langle\hat{s}|s\rangle$ vs s (if = then $\hat{s}$ is unbiased)

Consistency: $p(\hat{s}|\text{data}) \xrightarrow{N\to\infty} \delta(\hat{s}-s)$ (asymptotically unbiased)

efficiency: best possible estimator given data & loss fcn

robustness: low sensitivity to outliers (the exception makes the rule)

Let's evaluate the performance of an estimator.

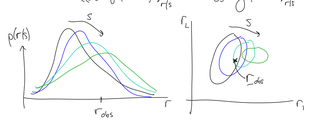
$\langle(\hat{s}-s)^2\rangle$ = MSE

If $\langle\hat{s}\rangle = s$ (unbiased) then $\langle(\hat{s}-s)^2\rangle = \langle(\hat{s}-\langle\hat{s}\rangle)^2\rangle = \sigma_{\hat{s}}^2$

and $\sigma_{\hat{s}}^2 \geq \frac{1}{J(s)}$ where $J(s)$ = Fisher Info

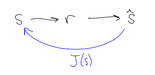
Fisher Information

$$J(s) = \left\langle \left(\frac{\partial}{\partial s} \log p(r|s)\right)^2 \right\rangle_{r|s} = -\left\langle \frac{\partial^2}{\partial s^2} \log p(r|s) \right\rangle_{r|s}$$

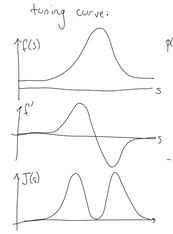


$J(s)$ is a local measure of how much variability in r affects estimation of s. It depends on s.

You can think of $J(s)$ as the variance of an optimal estimator (although an efficient estimator may not exist)



$J(s)$ for one neuron is largest on flanks of tuning curve:



$p(r|s) = \frac{e^{-f(s)} f(s)^r}{r!}$

$\log p(r|s) = -f + r\log f - \log r!$

$\partial_s \log p = -f' + \frac{r}{f} f'$

$\partial_s^2 \log p = -f'' - \frac{r}{f^2}(f')^2 + \frac{r}{f} f''$

$-\langle \partial_s^2 \log p \rangle = f'' + \frac{\langle r\rangle}{f^2}(f')^2 - \frac{\langle r\rangle}{f} f''$

$J(s) = \frac{(f')^2}{f}$

For multiple independent neurons,

$p(r|s) = \prod_i p(r_i|s)$

$J(s) = \sum_i^N \frac{(f_i'(s))^2}{f_i(s)}$ ← large slope is better; small variance is better; more neurons is better

Linear decoding

$r = f(s) + \eta$ but now with N neurons

$\eta \sim N(0, \Sigma)$ Σ: covariance

$\Sigma_{ij} = \langle r_i r_j\rangle - \langle r_i\rangle\langle r_j\rangle$

so $p(r|s) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} \exp\left[-\frac{1}{2}(r-f(s))^T \Sigma^{-1} (r-f(s))\right]$

Consider a decoded signal estimate

$\hat{s} = w^T r + c$ w: vector of decoding weights

What is the bias?

$\langle\hat{s}|s\rangle = \langle w^T r + c|s\rangle = w^T\langle f(s)+\eta|s\rangle + c$

$= w^T f(s) + c$

For unbiased-ness, want $\langle\hat{s}|s\rangle - s = 0$

$w^T f + c - s = 0$

$c = s - w^T f(s)$ ← depends on s, which we don't know!

Local discrimination: $f(s) \approx f(0) + s f'(0)$ (around s=0)

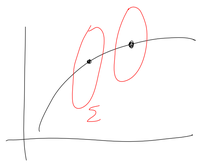
Then $c = s - w^T f(0) - w^T s f'(0)$

If $w^T f' = 1$ and $c = -w^T f(0)$

Then we are unbiased!

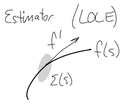
What is estimator variance? (for unbiased estimator)

$$\begin{aligned}
\sigma_{\hat{s}}^2 &= \langle(\hat{s}-\langle\hat{s}\rangle)^2\rangle = \langle(\hat{s}-s)^2\rangle \\
&= \left\langle \left(w^T r - w^T f(0) - s\right)^2 \right\rangle \\
&= \left\langle \left(w^T(\cancel{f(0)} + s f'(0) + \eta) - w^T f(0) - s\right)^2 \right\rangle \\
&= \left\langle \left(w^T f' s - s + w^T\eta\right)^2 \right\rangle \\
&= \left\langle (w^T\eta)^2 \right\rangle \\
&= \left\langle w^T \eta\eta^T w \right\rangle \\
&= w^T \langle\eta\eta^T\rangle w \\
&= w^T \textstyle\sum w
\end{aligned}$$



Locally Optimal Linear Estimator (LOLE)

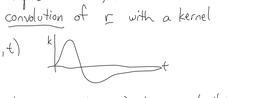
$w \propto \Sigma^{-1} f'$



Globally Optimal Linear Estimator (OLE)

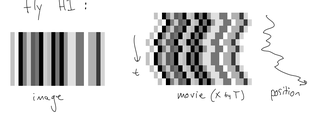
$w \propto \langle rr^T\rangle^{-1}\langle rs\rangle$

For time-dependent cases, this estimator is the convolution of r with a kernel

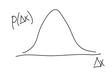


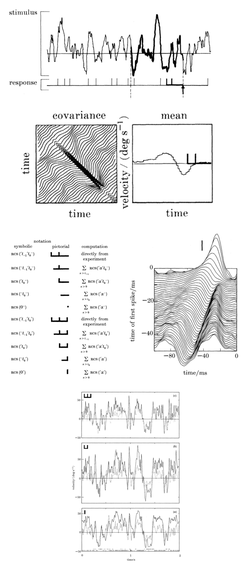
The estimator can be used to reconstruct the entire stimulus or some aspect of it.

Classic example is horizontal motion detection in fly H1:



Δposition = speed is drawn i.i.d. from a gaussian distribution. p(Δx)



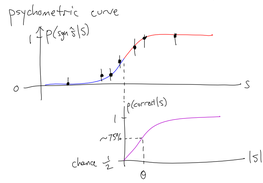


Later work found that linear filtering performed as well as this more complicated scheme!

J(s) can be related to discrimination performance, which is how many neuroscience experiments are quantified.

Experiment: 2 Alternative Forced-Choice (2AFC)

Use behavioral choices where available, to plot psychometric curve

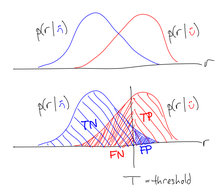


$\theta = \sigma_{\hat{s}} = \frac{1}{\sqrt{J(s)}}$ High θ ⇔ Low info J; Low θ ⇔ High info J

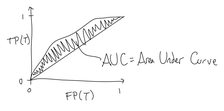
What if we don't have binary choices? e.g. for neurons?

Use signal detection theory ("ROC analysis") see Green & Swets

to generate "neurometric" curves.



Plot TP vs FP:



$AUC + \frac{1}{2} = p(r_+ < r_-)$ = prob. that random draw from $p(r|s=+)$ is less than random draw from $p(r|s=-)$.

To see this, note that FP(T) gives info about the whole distribution $p(r|s=-)$:

$FP(T) = \int_T^\infty dr\, p(r|s=-)$

Now $p(r_+ < r_-) = \int_{-\infty}^{\infty} dr_+ \int_{-\infty}^{\infty} dr_-\, p(r_+|s=+) p(r_-|s=-) \mathbb{1}(r_+ < r_-)$

$= \int_{-\infty}^{\infty} dr_+ \int_{r_+}^{\infty} dr_-\, p(r_+|s=+) p(r_-|s=-)$

$= \int_{-\infty}^{\infty} dr_+ \underbrace{p(r_+|s=+)}_{dTP} FP(r_+)$

$= \int_0^1 dTP \cdot FP(TP)$

$= AUC + \frac{1}{2}$

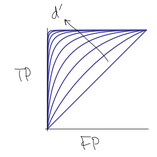
When AUC = 0, performance is at chance

AUC = 1, perfect discrimination

For gaussian r, AUC depends only on

$(d')^2 = \frac{\Delta f^2}{\sigma^2} = J(s)$ in limit of small Δs

d' is a generalization of Fisher Info for non-local discriminations



Experiments use 2AFC & ROC analysis all the time, because binary variables are easier to analyze