Mathematical Tools
for Neural and Cognitive Science

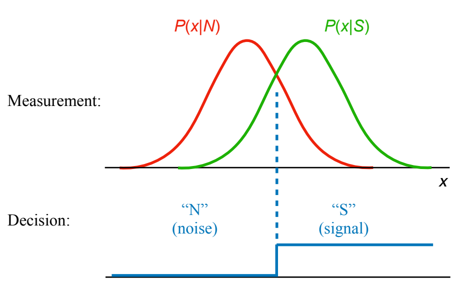
Fall semester, 2024

Section 5a:

Statistical Decision Theory
+
Signal Detection Theory

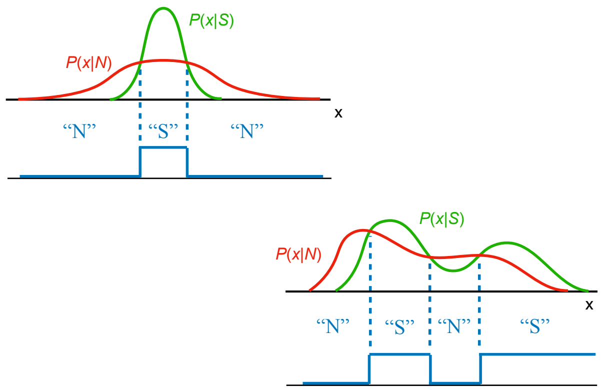
Tumor, or not?

Signal Detection Theory (binary estimation)

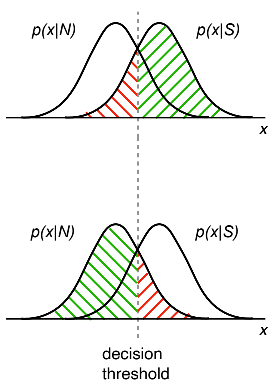


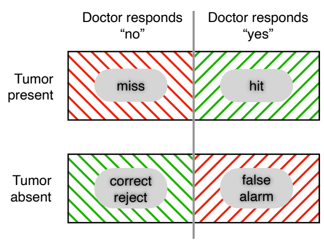
For equal-shape, unimodal, symmetric distributions,
the ML decision rule is a *threshold* function.

More generally, decision rule can have multiple thresholds…



Signal Detection Theory: Potential outcomes



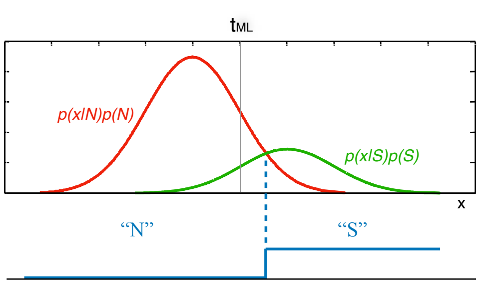


For threshold *t*, cumulatives *c()*

- $c(t\,|\,S)$ = p(miss)
- $1\text{-}c(t\,|\,S)$ = p(hit)
- $c(t\,|\,N)$ = p(correct reject)
- $1\text{-}c(t\,|\,N)$ = p(false alarm)

MAP decision rule?

MAP solution maximizes proportion of correct answers, *taking prior probability into account*.



Compared to ML threshold, the MAP threshold moves *away* from higher-probability option.

Bayes decision rule?

Incorporate values for the four possible outcomes:

Payoff Matrix

Stimulus \ Response	Yes	No
S+N	V_{S+N}^{Yes}	V_{S+N}^{No}
N	V_{N}^{Yes}	V_{N}^{No}

Bayes Optimal Criterion

Stimulus \ Response	Yes	No
S+N	V_{S+N}^{Yes}	V_{S+N}^{No}
N	V_{N}^{Yes}	V_{N}^{No}

$$\mathbb{E}(Yes\,|\,x) = V_{S+N}^{Yes}p(S+N\,|\,x) + V_{N}^{Yes}p(N\,|\,x)$$
$$\mathbb{E}(No\,|\,x) = V_{S+N}^{No}p(S+N\,|\,x) + V_{N}^{No}p(N\,|\,x)$$

Say yes if $\mathbb{E}(Yes\,|\,x) \geq \mathbb{E}(No\,|\,x)$

Optimal Criterion

$$\mathbb{E}(Yes\,|\,x) = V_{S+N}^{Yes}p(S+N\,|\,x) + V_{N}^{Yes}p(N\,|\,x)$$
$$\mathbb{E}(No\,|\,x) = V_{S+N}^{No}p(S+N\,|\,x) + V_{N}^{No}p(N\,|\,x)$$

Say yes if $\mathbb{E}(Yes\,|\,x) \geq \mathbb{E}(No\,|\,x)$

Say yes if $\underbrace{\dfrac{p(S+N\,|\,x)}{p(N\,|\,x)}}_{\text{Posterior odds}} \geq \dfrac{V_{N}^{No} - V_{N}^{Yes}}{V_{S+N}^{Yes} - V_{S+N}^{No}} = \dfrac{V(\text{Correct}\,|\,N)}{V(\text{Correct}\,|\,S+N)}$

Posterior odds

Apply Bayes' Rule

Posterior Likelihood Prior

$$p(S+N\,|\,x) = \frac{p(x\,|\,S+N)\,p(S+N)}{p(x)}$$

Nuisance normalizing term

$$p(N\,|\,x) = \frac{p(x\,|\,N)\,p(N)}{p(x)}, \text{ hence}$$

$$\frac{p(S+N\,|\,x)}{p(N\,|\,x)} = \left(\frac{p(x\,|\,S+N)}{p(x\,|\,N)}\right)\left(\frac{p(S+N)}{p(N)}\right)$$

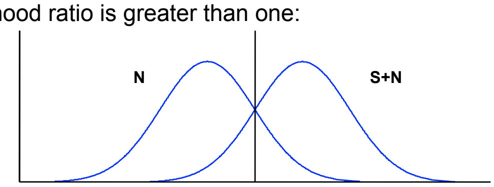
Posterior odds Likelihood ratio Prior odds

Optimal Criterion

$$\text{Say yes if } \frac{p(S+N\,|\,x)}{p(N\,|\,x)} \geq \frac{V(\text{Correct}\,|\,N)}{\text{V(Correct}\,|\,S+N)}$$

$$\text{i.e., if } \frac{p(x\,|\,S+N)}{p(x\,|\,N)} \geq \frac{p(N)}{p(S+N)}\,\frac{V(\text{Correct}\,|\,N)}{\text{V(Correct}\,|\,S+N)} = \beta$$

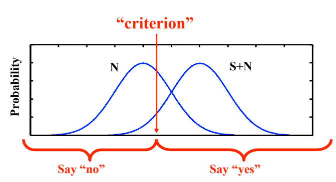
Example, if equal priors and equal payoffs, say yes if the likelihood ratio is greater than one:



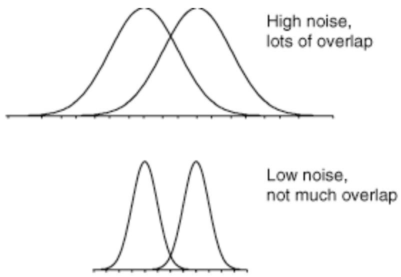
Example applications of SDT

- Vision
  - Detection (something vs. nothing)
  - Discrimination (lower vs greater level of: intensity, contrast, depth, slant, size, frequency, loudness, ...
- Memory (internal response = trace strength = familiarity)
- Neurometric function/discrimination by neurons (internal response = spike count)

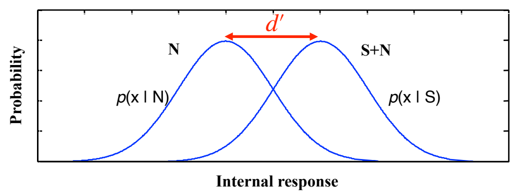
From experimental measurements, assuming human is optimal, can we determine the underlying distributions and criterion?



Signal Detection Theory: discriminability (*d′*)



Internal response: probability of occurrence curves

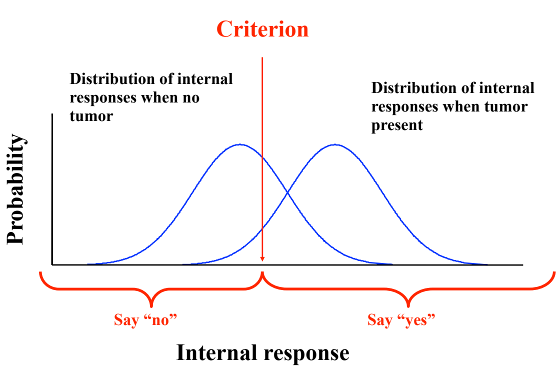


$$d' = \frac{\text{“separation”}}{\text{“width”}}$$

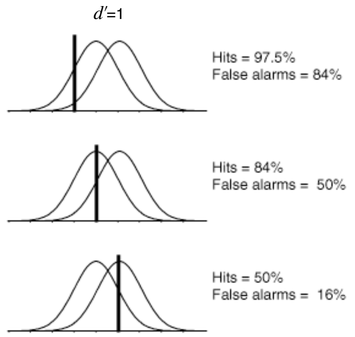
Discriminability (“d-prime”) is the normalized separation between the two distributions

Error rate is a function of *d′*

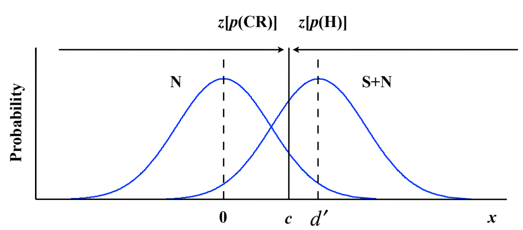
Criterion



Signal Detection Theory: Criterion



SDT: Gaussian case

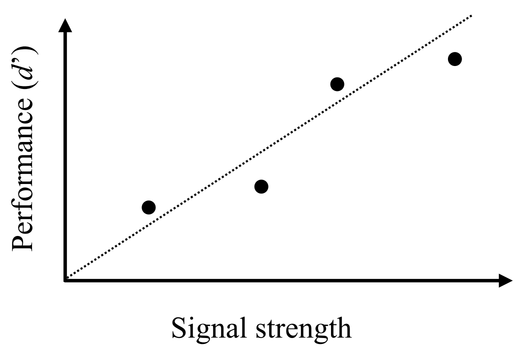


$$d' = z[p(H)] + z[p(CR)] = z[p(H)] - z[p(FA)]$$

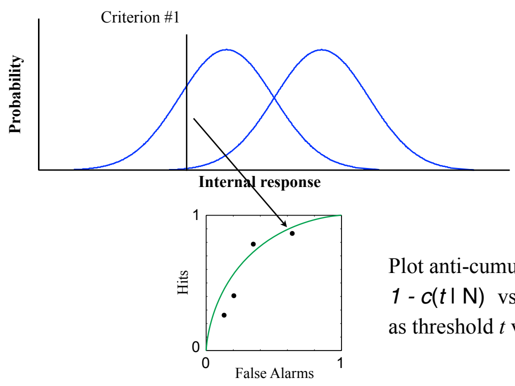
$$c = z[p(CR)] \qquad G(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

$$\beta = \frac{p(x=c \mid S+N)}{p(x=c \mid N)} = \frac{e^{-(c-d')^2/2}}{e^{-c^2/2}} \qquad \text{(Fix } \sigma = 1\text{)}$$

SDT: Psychometric function

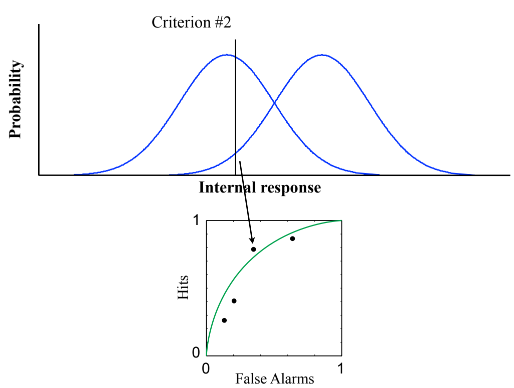


ROC (Receiver Operating Characteristic)

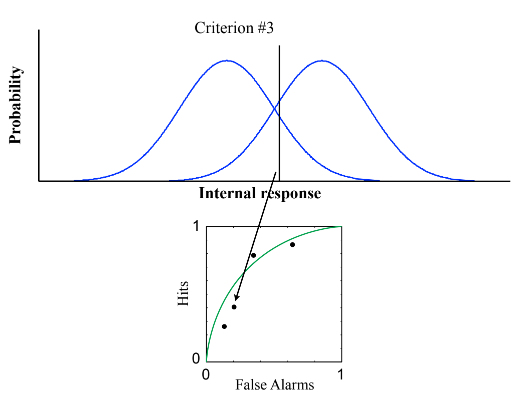


Plot anti-cumulatives:
$1 - c(t\,|\,N)$ vs. $1 - c(t\,|\,S)$
as threshold t varies

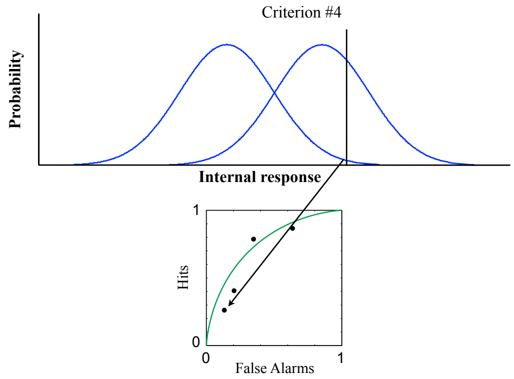
ROC (Receiver Operating Characteristic)



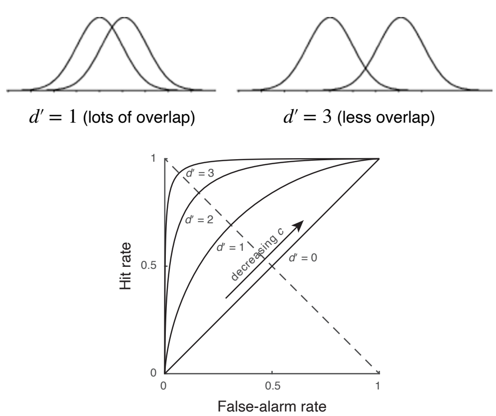
ROC (Receiver Operating Characteristic)



ROC (Receiver Operating Characteristic)

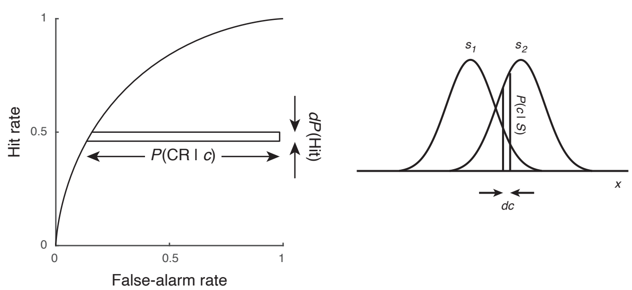


ROC (Receiver Operating Characteristic)



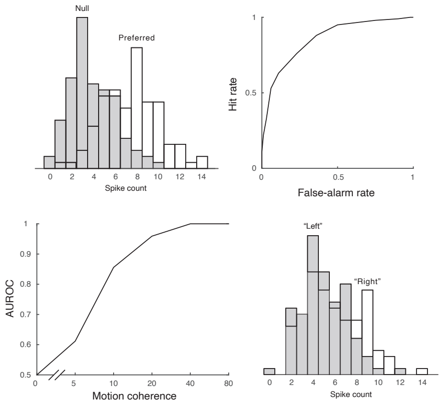
Area under the ROC

Area under curve = %correct in a 2AFC task]



Slope of the ROC = likelihood ratio or posterior ratio if a prior is used

Area under the ROC - Poisson case or with data: Neurometric function and Choice probability

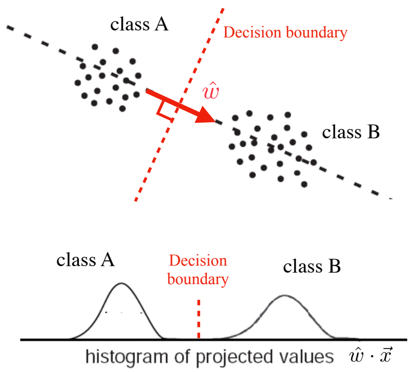


Decision/classification in multiple dimensions

- Data-driven linear classifiers:
  - Prototype Classifier - minimize distance to class mean
  - Fisher Linear Discriminant (FLD) - maximize d'
  - Support Vector Machine (SVM) - maximize margin
- Statistical:
  - ML/MAP/Bayes under a probabilistic model
  - e.g.: Gaussian, identity covariance (same as Prototype)
  - e.g.: Gaussian, equal covariance (same as FLD)
  - e.g.: Gaussian, general case (Quadratic Discriminator)
- Some Examples:
  - Visual gender classification
  - Neural population decoding

Linear Classifier

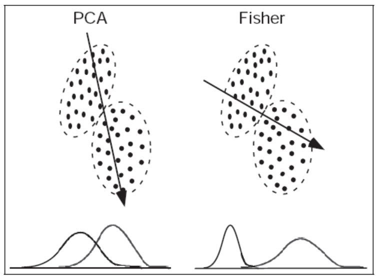
Find unit vector $\hat{w}$ ("discriminant") that best separates the distributions



Simplest linear discriminant: the Prototype Classifier

$$\hat{w} = \frac{\vec{\mu}_A - \vec{\mu}_B}{\|\vec{\mu}_A - \vec{\mu}_B\|}$$

Fisher Linear Discriminant



$$\max_{\hat{w}} \frac{\left[\hat{w}^T(\vec{u}_A - \vec{u}_B)\right]^2}{\left[\hat{w}^T C_A \hat{w} + \hat{w}^T C_B \hat{w}\right]}$$

(note: this is d' squared!)

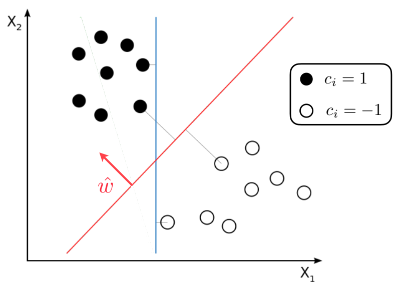
optimum: $\hat{w} = C^{-1}(\vec{u}_A - \vec{u}_B)$, where $C = \frac{1}{2}(C_A + C_B)$

Support Vector Machine (SVM)

(widely used in machine learning, but no closed form solution)

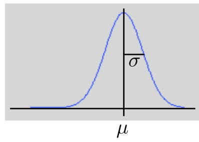
Maximize the "margin" (gap between data sets):

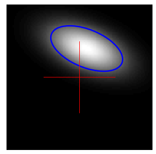
find largest m, and $\{\hat{w}, b\}$ s.t. $c_i(\hat{w}^T \vec{x}_i - b) \geq m, \quad \forall\ i$



Reminder: Multi-D Gaussian densities

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



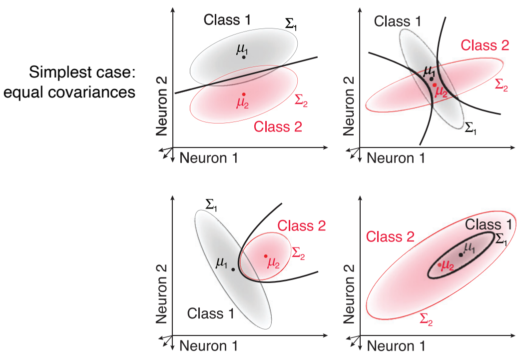


$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^N |C|}}\, e^{-(\vec{x}-\vec{\mu})^T C^{-1} (\vec{x}-\vec{\mu})/2}$$

mean: $[0.2, 0.8]$
cov: [1.0 -0.3;
-0.3 0.4]

ML (or MAP) classifier for two Gaussians

Decision boundary is *quadratic*, with four possible geometries:



[on board]

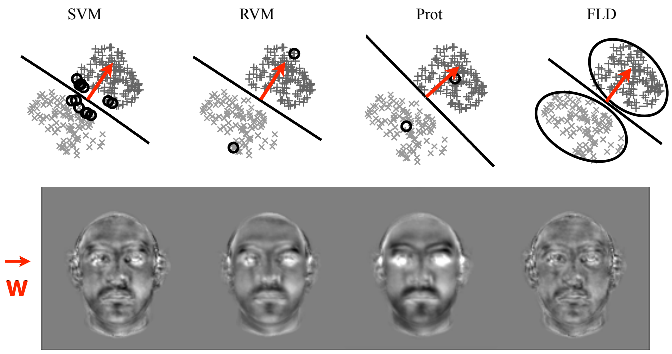
[figure: Pagan et al. 2016]

A perceptual example: Sex identification

- 200 face images (100 male, 100 female)
- Adjusted for position, size, intensity/contrast
- Labeled by 27 human subjects

[Graf & Wichmann, NIPS*03]

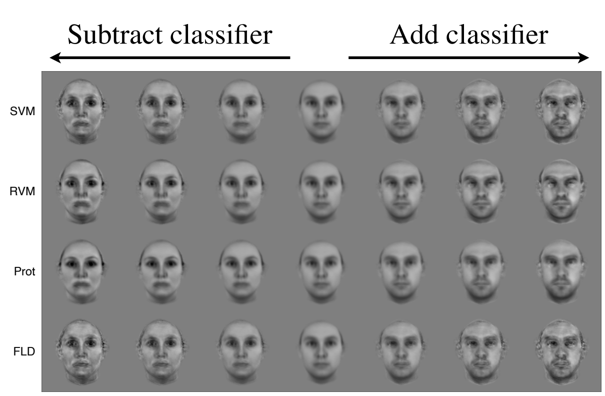
Linear classifiers



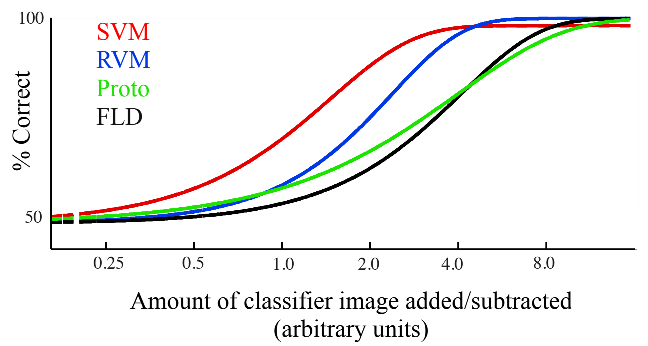
Four different linear classifiers, trained on human data

Model validation/testing

- Cross-validation: Subject responses [% correct, reaction time, confidence] are explained:
  - very well by SVM
  - moderately well by RVM / FLD
  - not so well by Prot
- Do these decision "models"make testable predictions? Synthesize optimally discriminable faces...

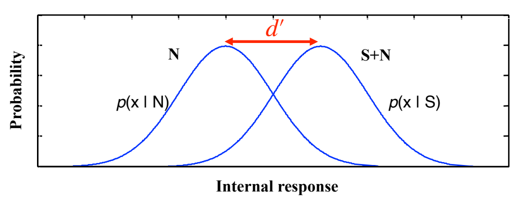


[Wichmann, et. al; NIPS*04]



[Wichmann, et. al; NIPS*04]

Internal response: probability of occurrence curves



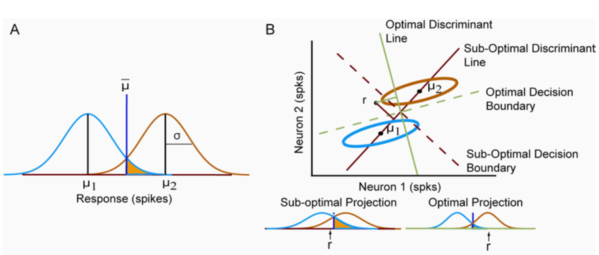
$$d' = \frac{\text{"separation"}}{\text{"width"}}$$

Discriminability ("d-prime") is the normalized separation between the two distributions

Error rate is a function of d'

Journal of Neurophysiology

Averbeck and Lee: Information and correlations



Summarizing error of ML estimators

Bias: the MLE is *asymptotically unbiased* and *Gaussian*, but can only rely on these if:

- the likelihood model is correct
- the likelihood can be maximized
- you have lots of data

Variance: (error bars)

- S.E.M. (relevant for sample averages only)
- second deriv of NLL (multi-D: "Hessian")
- simulation (resample from $p(x|\hat{\theta})$)
- bootstrapping (resample from *the data,* with replacement)

Fisher Information

- Second-order expansion of the (expected) negative log likelihood:

$$I(s) = -\mathbb{E}\left[\frac{\partial^2 \log p(r|s)}{\partial s^2}\right]$$

- Provides a bound on "precision" of unbiased estimators: (the "Cramér-Rao bound") $\sigma^2(s) \geq \frac{1}{I(s)}$

- Perceptually, provides a bound on **discriminability**: (Series et. al. 2009) $D(s) \leq \sqrt{I(s)}$

- Examples: with mean stimulus response $\mu(s)$

Gaussian case: $p(r|s) \sim \mathcal{N}(\mu(s), \sigma^2)$ $\quad I(s) = [\mu'(s)]^2/\sigma^2$

Poisson case: $p(r|s) \sim \text{Poiss}(\mu(s))$ $\quad I(s) = [\mu'(s)]^2/\mu(s)$

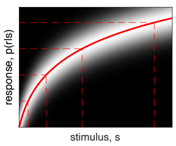
Example: Weber's law

[Weber, 1834]

$D(s) \propto \frac{1}{s}$ (discrimination thresholds proportional to stimulus strength)

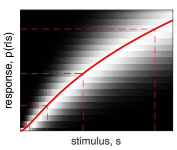
Assuming $I(s) \propto \frac{1}{s^2}$ what internal representation can explain this? Many!

additive Gaussian noise, with mean $\mu(s) = \log(s) + c$



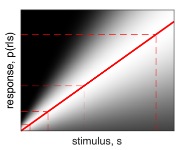
entirely due to response mean [Fechner, 1860]

Poisson noise, with mean $\mu(s) = [\log(s) + c]^2$



discrete representation, depends on both mean and variance

multiplicative Gaussian noise, with mean $\mu(s) = \alpha s$



entirely due to response variance

S.S. Stevens. “To Honor Fechner and Repeal His Law: A power function, not a log function, describes the operating characteristic of a sensory system” (1961)

Assuming Stevens is measuring internal mean, can we combine with Weber’s Law?

sensitivity

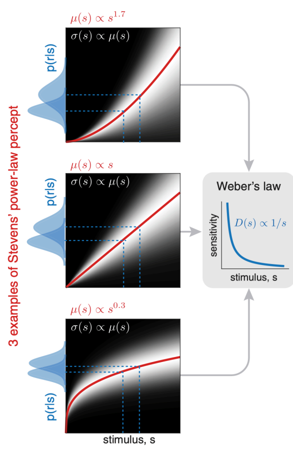
$\delta(s) \propto 1/s$

intensity

$\mu(s) \propto s^{\alpha}$

Loudness	0.67	Sound pressure of 3000 Hz tone
Vibration	0.95	Amplitude of 60 Hz on finger
Vibration	0.6	Amplitude of 250 Hz on finger
Brightness	0.33	5° target in dark
Brightness	0.5	Point source
Brightness	0.5	Brief flash
Brightness	1	Point source briefly flashed
Lightness	1.2	Reflectance of gray papers
Visual length	1	Projected line
Visual area	0.7	Projected square
Redness (saturation)	1.7	Red–gray mixture
Taste	1.3	Sucrose
Taste	1.4	Salt
Taste	0.8	Saccharin
Smell	0.6	Heptane
Cold	1	Metal contact on arm
Warmth	1.6	Metal contact on arm
Warmth	1.3	Irradiation of skin, small area
Warmth	0.7	Irradiation of skin, large area
Discomfort, cold	1.7	Whole-body irradiation
Discomfort, warm	0.7	Whole-body irradiation
Thermal pain	1	Radiant heat on skin
Tactual roughness	1.5	Rubbing emery cloths
Tactual hardness	0.8	Squeezing rubber
Finger span	1.3	Thickness of blocks
Pressure on palm	1.1	Static force on skin
Muscle force	1.7	Static contractions
Heaviness	1.45	Lifted weights
Viscosity	0.42	Stirring silicone fluids
Electric shock	3.5	Current through fingers
Vocal effort	1.1	Vocal sound pressure
Angular acceleration	1.4	5 s rotation
Duration	1.1	White-noise stimuli

S.S. Stevens. “To Honor Fechner and Repeal His Law: A power function, not a log function, describes the operating characteristic of a sensory system” (1961)



Three examples with different power-law mean response, each consistent with Weber’s law discriminability.

[Zhou, Duong & EPS, 2022]