Lecture 4: The Bayesian brain

Samuel Gershman

Harvard University

Roadmap

- The sensory data received by the brain provides incomplete and noisy information about the environment state.

Roadmap

- The sensory data received by the brain provides incomplete and noisy information about the environment state.
- This lecture describes models of how the brain computes a probability distribution over (or point estimate of) hidden states.

Roadmap

- The sensory data received by the brain provides incomplete and noisy information about the environment state.
- This lecture describes models of how the brain computes a probability distribution over (or point estimate of) hidden states.
- Problem: behavior seems to deviate from Bayes-optimal inference.

Roadmap

- The sensory data received by the brain provides incomplete and noisy information about the environment state.
- This lecture describes models of how the brain computes a probability distribution over (or point estimate of) hidden states.
- Problem: behavior seems to deviate from Bayes-optimal inference.
- Can we understand these deviations through the lens of computational and representational constraints on inference?

Is human behavior Bayesian?

- Answering this question is trickier than it might seem, because we need to know what (if any) prior, likelihood, posterior, and utility function the brain uses.

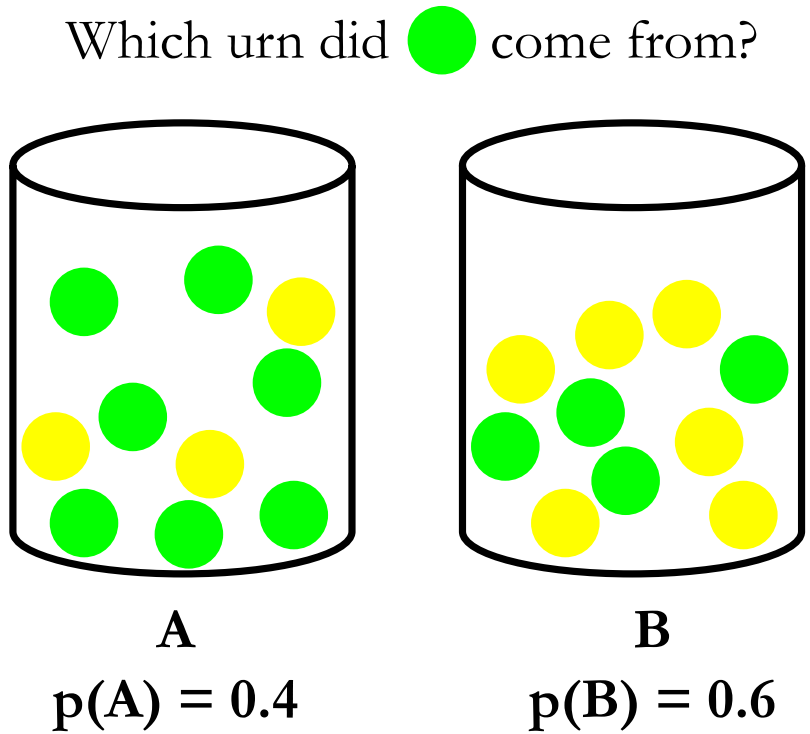
Is human behavior Bayesian?

- Answering this question is trickier than it might seem, because we need to know what (if any) prior, likelihood, posterior, and utility function the brain uses.
- One approach is to manufacture experimental tasks that tightly control all of these factors and impose them on human subjects.

Is human behavior Bayesian?

- Answering this question is trickier than it might seem, because we need to know what (if any) prior, likelihood, posterior, and utility function the brain uses.
- One approach is to manufacture experimental tasks that tightly control all of these factors and impose them on human subjects.
- This approach has the advantage of allowing us to precisely answer the question, but it has the disadvantage of being contrived.

The urn task



Posterior log odds

We can reduce the binary inference problem to a one-dimensional log odds:

$$\log \frac{p(A|\bullet)}{p(B|\bullet)} = \log \frac{p(\bullet|A)}{p(\bullet|B)} + \log \frac{p(A)}{p(B)},$$

where the first term is the likelihood log odds and the second term is the prior log odds.

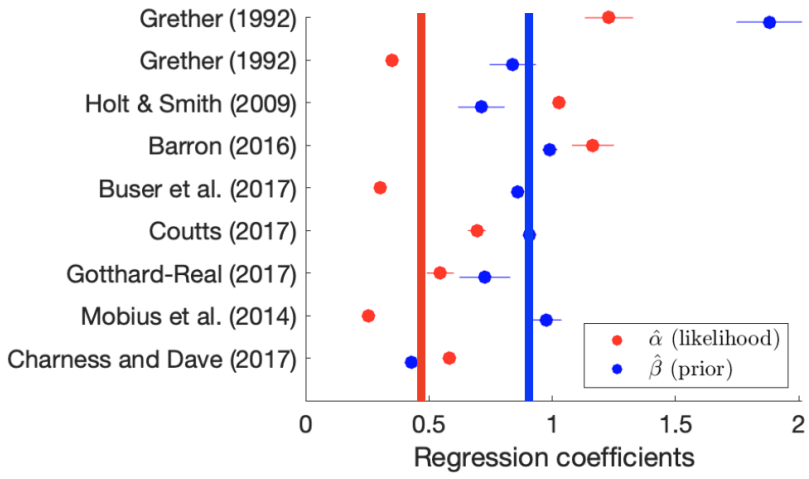
Generalized log odds

To quantitatively evaluate the Bayesian hypothesis, we can generalize this equation to a more flexible model with coefficients α and β:

$$y = \alpha \log \frac{p(\bullet|A)}{p(\bullet|B)} + \beta \log \frac{p(A)}{p(B)},$$

where y is the response generated by human subjects.

Fitted coefficients



[Zhu & Griffiths 2023]

Take-aways

- Both coefficients are systematically below 1 (under-reaction), though the prior coefficient (β) is pretty close to 1.

Take-aways

- Both coefficients are systematically below 1 (under-reaction), though the prior coefficient (β) is pretty close to 1.
- Thus, even in this idealized setting, people don't perfectly execute Bayes' rule: Although they update in the correct direction, they systematically under-react to the likelihood (i.e., the urn composition in this case).

Resource-rational analysis of costly inference

- Under-reaction suggests that updating from the prior to the (approximate) posterior is costly.

Resource-rational analysis of costly inference

- Under-reaction suggests that updating from the prior to the (approximate) posterior is costly.
- Let's replace the true posterior $p(s|x)$ with an approximate posterior $q(s|x)$ to make explicit that we are no longer assuming exact Bayesian inference.

Resource-rational analysis of costly inference

- Under-reaction suggests that updating from the prior to the (approximate) posterior is costly.
- Let's replace the true posterior $p(s|x)$ with an approximate posterior $q(s|x)$ to make explicit that we are no longer assuming exact Bayesian inference.
- Assume the action a output deterministically by policy π is the approximate posterior: $a = q$.

Resource-rational analysis of costly inference

Cost of updating after observing signal x using the Kullback-Leibler (KL) divergence:

$$\mathcal{D}[q(s|x)||p(s)] = \sum_s q(s|x) \log \frac{q(s|x)}{p(s)}.$$

Belief updates that move the approximate posterior $q(s|x)$ farther from the prior $p(s)$ are more costly.

Resource-rational analysis of costly inference

Expected cost $c(\pi)$ under policy π, which averages over signals:

$$c(\pi) = \sum_x p(x)\mathcal{D}[q(s|x)||p(s)].$$

Resource-rational analysis of costly inference

Utility should be higher when our beliefs are closer to the posterior. Suppose rewards are signals ($r = x$) and that the utility derived from these signals is the negative KL divergence between the approximate and true posterior:

$$u(r) = -\mathcal{D}[q(s|x)||p(s|x)]$$

Expected utility:

$$\bar{u}(\pi) = \mathbb{E}[u(r)|\pi] = -\sum_x p(x)\mathcal{D}[q(s|x)||p(s|x)]$$

Resource-rational analysis of costly inference

Optimal policy:

$$\pi^* = \underset{\pi:\, c(\pi) \leq \mathcal{C}}{\text{argmax}} \; \bar{u}(\pi)$$

where $\mathcal{C}$ is the capacity limit.

Resource-rational analysis of costly inference

Equivalent unbounded optimization problem (Lagrangian):

$$\pi^* = \underset{\pi}{\text{argmax}}\ \bar{u}(\pi) - \lambda c(\pi)$$

where the Lagrange multiplier is:

$$\lambda = \frac{\partial \bar{u}(\pi^*)}{\partial c(\pi^*)}$$

with $c(\pi^*) = \mathcal{C}$ (i.e., the optimal policy operates at the capacity limit).

Resource-rational analysis of costly inference

Closed-form optimal policy [Zhu & Griffiths 2023]:

$$q^*(s|x) \propto p(x|s)^{1/(1+\lambda)} p(s)$$

This is just Bayes' rule with a down-weighted likelihood. This implies under-reaction to the likelihood, as seen experimentally.

Neural implementation

- We now construct a neural model that approximates the posterior over $s \in \{A, B\}$.

Neural implementation

- We now construct a neural model that approximates the posterior over $s \in \{A, B\}$.
- Our basic primitive is the integrate-and-fire neuron (no leak) with membrane potential dynamics governed by:

$$C\dot{\mu} = I(t)$$

where C is the membrane capacitance, and $I(t) = \sum_d w_d z_d(t)$ is the input current, which linearly integrates presynaptic spikes.

Neural implementation

- Consider a population of presynaptic neurons, where neuron d has tuning function $f_d(s)$. The log-likelihood under Poisson spiking is given by:

$$\log p(x|s) = \sum_d x_d \log f_d(s) - f_d(s) - \log x_d!$$

where x_d is the spike count for neuron d over some time window.

Neural implementation

- Consider a population of presynaptic neurons, where neuron d has tuning function $f_d(s)$. The log-likelihood under Poisson spiking is given by:

$$\log p(x|s) = \sum_d x_d \log f_d(s) - f_d(s) - \log x_d!$$

  where x_d is the spike count for neuron d over some time window.

- The third term doesn't depend on s, so we can ignore it. We will also ignore the second term under the assumption that $\sum_d f_d(s)$ is a constant.

Neural implementation

- Consider a population of presynaptic neurons, where neuron d has tuning function $f_d(s)$. The log-likelihood under Poisson spiking is given by:

$$\log p(x|s) = \sum_d x_d \log f_d(s) - f_d(s) - \log x_d!$$

  where x_d is the spike count for neuron d over some time window.

- The third term doesn't depend on s, so we can ignore it. We will also ignore the second term under the assumption that $\sum_d f_d(s)$ is a constant.
- After discarding these terms, the log-likelihood ratio becomes:

$$\log \frac{p(x|s=A)}{p(x|s=B)} = \sum_d x_d \log \frac{f_d(A)}{f_d(B)}$$

Neural implementation

If we set the synaptic strength of neuron d to be $w_d = \log \frac{f_d(A)}{f_d(B)}$, the postsynaptic neuron will accumulate weighted spike counts over time such that its membrane potential represents the posterior log-odds:

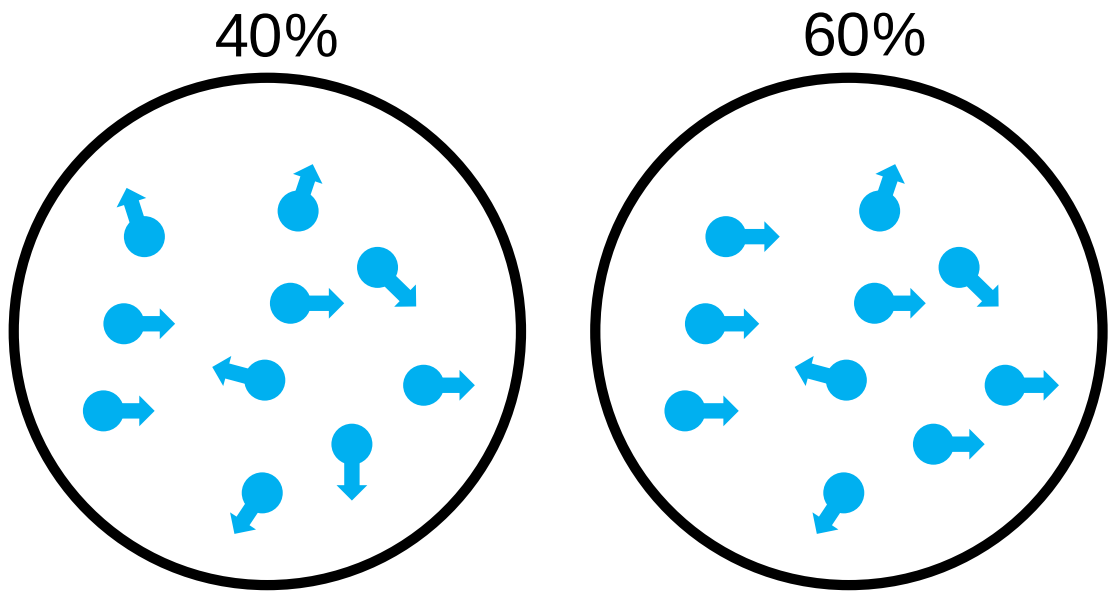
$$I(t) = \sum_d x_d w_d, \quad w_d = \log \frac{p(s = A|x)}{p(s = B|x)},$$

provided the resting potential is given by the prior log odds:

$$\mu(0) = \log \frac{p(s = A)}{p(s = B)}$$

Example: motion direction discrimination

Task is to determine the direction of coherently moving dots.



Example: motion direction discrimination

- Relevant presynaptic population is in extrastriate area MT, where most neurons are tuned to particular motion directions.

Example: motion direction discrimination

- Relevant presynaptic population is in extrastriate area MT, where most neurons are tuned to particular motion directions.
- Tuning functions can be modeled as a cosine function defined over the space of motion directions ($s \in [0, 360]$):

$$f_d(s) = \exp[\cos(s - s_d^*)/\nu]$$

where s_d^* is the preferred direction for neuron d and ν is the tuning width.

Example: motion direction discrimination

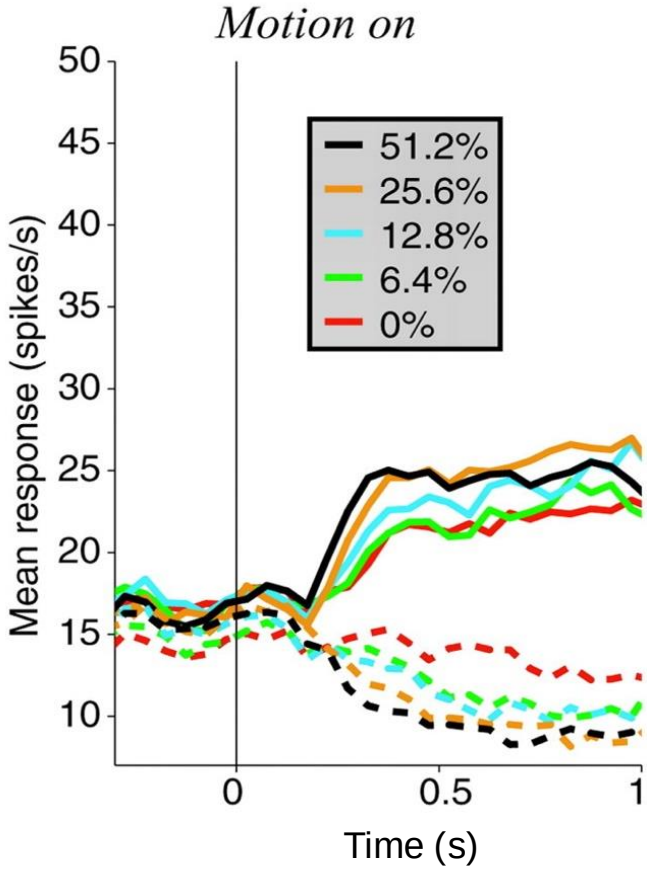
- Relevant presynaptic population is in extrastriate area MT, where most neurons are tuned to particular motion directions.
- Tuning functions can be modeled as a cosine function defined over the space of motion directions ($s \in [0, 360]$):

$$f_d(s) = \exp[\cos(s - s_d^*)/\nu]$$

  where s_d^* is the preferred direction for neuron d and ν is the tuning width.
- One synapse downstream, neurons in parietal area LIP integrate the spiking of MT neurons, enabling them to compute the posterior log odds.

Example: motion direction discrimination

LIP neurons ramp up over time during viewing of the random dot motion stimulus.



[Shadlen & Gold 2001]

Another example: discrete evidence accumulation

- Drawback of the random dot motion stimulus: difficult to precisely quantify the information value of the stimulus at any given time.

Another example: discrete evidence accumulation

- Drawback of the random dot motion stimulus: difficult to precisely quantify the information value of the stimulus at any given time.
- Yang & Shadlen [2007] addressed this issue, recording LIP neurons while monkeys viewed a sequence of abstract shapes.

Another example: discrete evidence accumulation

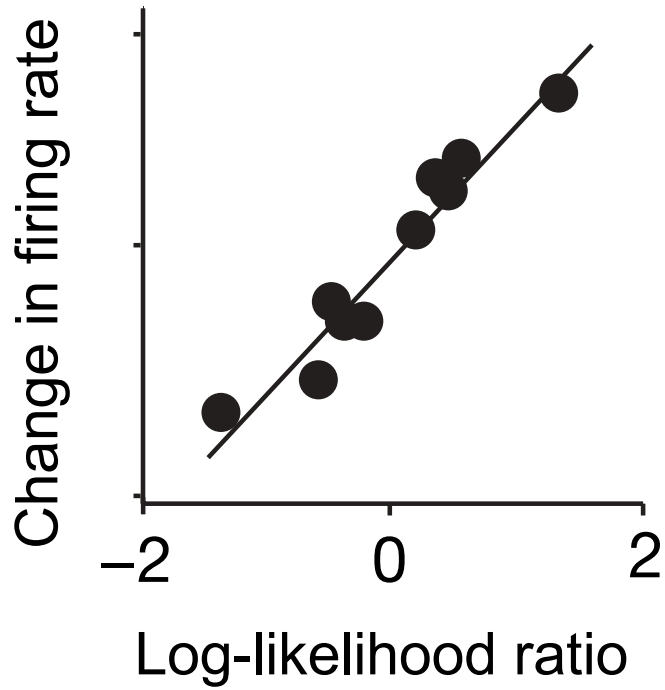
- Drawback of the random dot motion stimulus: difficult to precisely quantify the information value of the stimulus at any given time.
- Yang & Shadlen [2007] addressed this issue, recording LIP neurons while monkeys viewed a sequence of abstract shapes.
- At the end of the sequence, the monkey needed to choose one of two visual targets.

Another example: discrete evidence accumulation

- Drawback of the random dot motion stimulus: difficult to precisely quantify the information value of the stimulus at any given time.
- Yang & Shadlen [2007] addressed this issue, recording LIP neurons while monkeys viewed a sequence of abstract shapes.
- At the end of the sequence, the monkey needed to choose one of two visual targets.
- The correct target was determined by the shape sequence: each shape was associated with a particular log-likelihood ratio, such that the total log-likelihood ratio could be obtained by summing up the contributions of the shapes in the sequence.

Another example: discrete evidence accumulation

Changes in the firing rate of LIP neurons are linearly related to the log-likelihood ratio.



[Yang & Shadlen 2007]

Incorporating costly inference

- Cost parameter λ enters through a multiplier of the log-likelihood ratio.

Incorporating costly inference

- Cost parameter λ enters through a multiplier of the log-likelihood ratio.
- This can be interpreted as a global modulation:

$$\log \frac{p(x|s=A)^{1/(1+\lambda)}}{p(x|s=B)^{1/(1+\lambda)}} = \frac{1}{1+\lambda} \sum_d x_d \log \frac{f_d(A)}{f_d(B)}.$$

Incorporating costly inference

- Cost parameter λ enters through a multiplier of the log-likelihood ratio.
- This can be interpreted as a global modulation:

$$\log \frac{p(x|s=A)^{1/(1+\lambda)}}{p(x|s=B)^{1/(1+\lambda)}} = \frac{1}{1+\lambda} \sum_d x_d \log \frac{f_d(A)}{f_d(B)}.$$

- As λ increases (lower capacity C), the log-likelihood is suppressed.

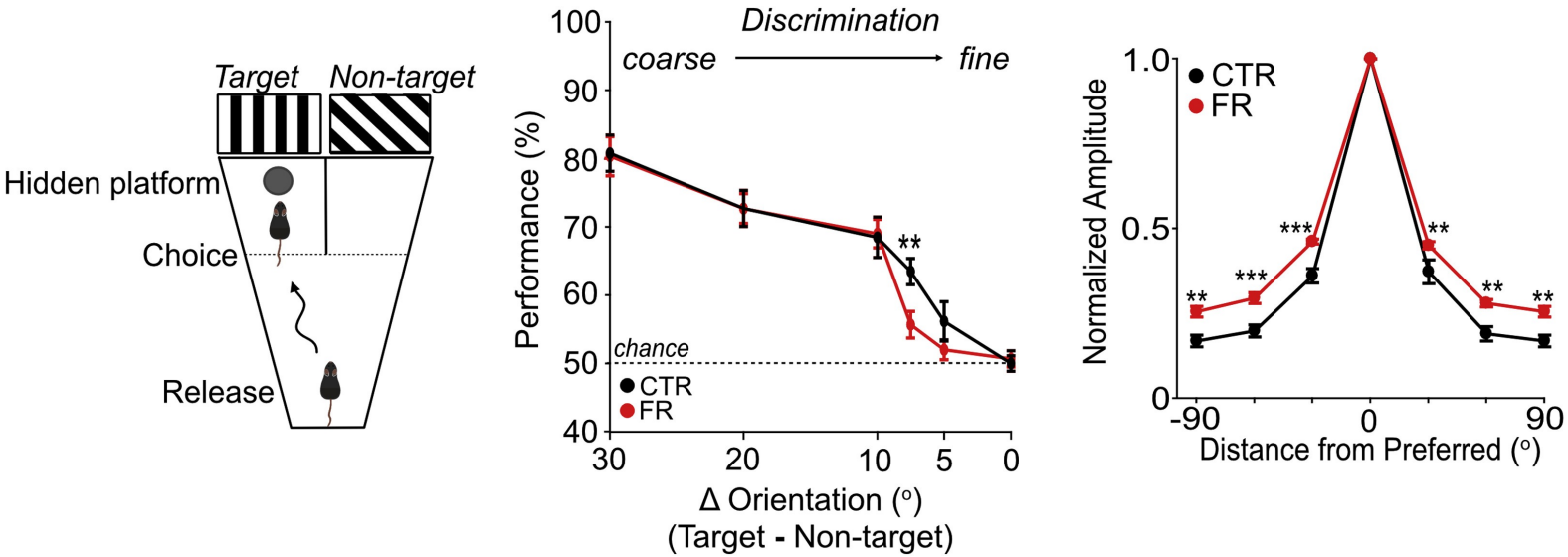
Incorporating costly inference

- Cost parameter λ enters through a multiplier of the log-likelihood ratio.
- This can be interpreted as a global modulation:

$$\log \frac{p(x|s=A)^{1/(1+\lambda)}}{p(x|s=B)^{1/(1+\lambda)}} = \frac{1}{1+\lambda} \sum_d x_d \log \frac{f_d(A)}{f_d(B)}.$$

- As λ increases (lower capacity C), the log-likelihood is suppressed.
- Possible mechanistic interpretations: suppression of firing, suppression of synaptic strengths, or suppression of the postsynaptic membrane potential.

Case study: effects of food restriction on orientation discrimination



[Padamsey et al 2022]

Case study: effects of food restriction on orientation discrimination

- The mechanism underlying the tuning change was a reduction in AMPA receptor conductance, which was compensated for by increased input resistance and depolarization of the membrane potential.

Case study: effects of food restriction on orientation discrimination

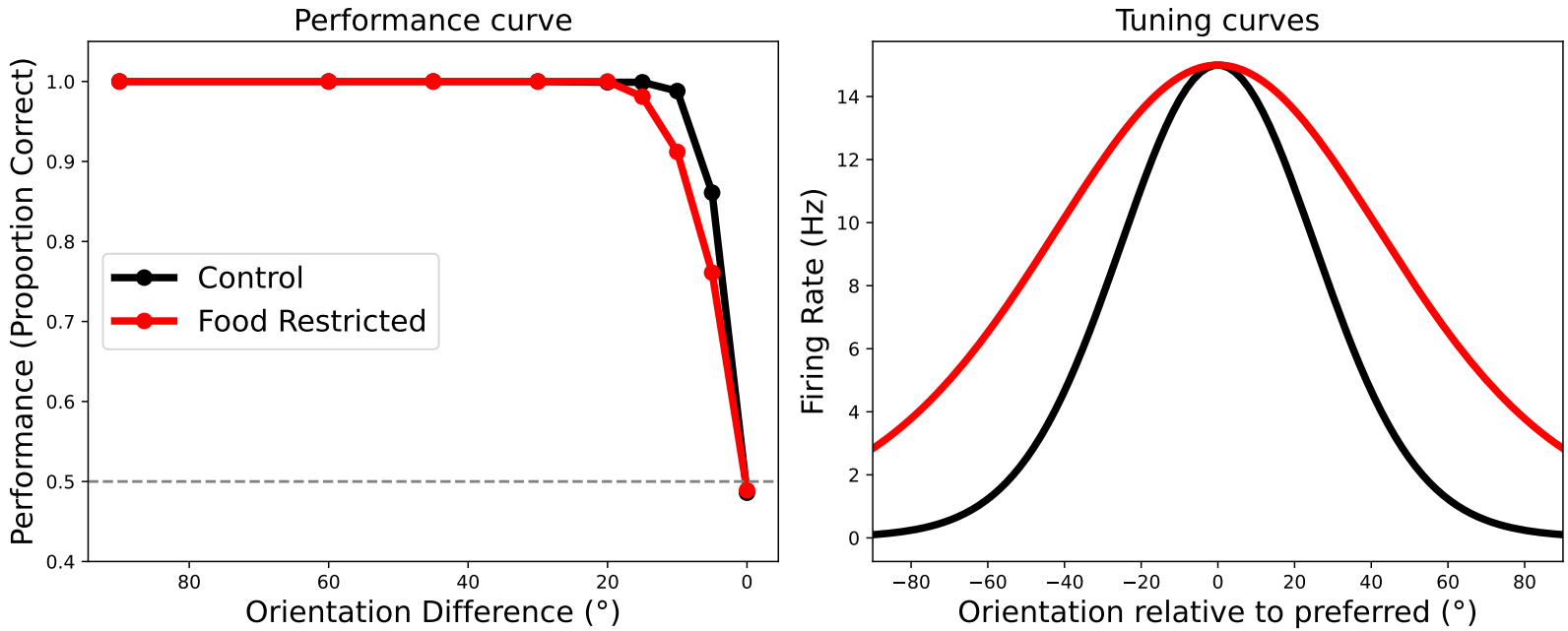
- The mechanism underlying the tuning change was a reduction in AMPA receptor conductance, which was compensated for by increased input resistance and depolarization of the membrane potential.
- This had the effect of maintaining roughly the same firing rates but making firing more variable.

Case study: effects of food restriction on orientation discrimination

- The mechanism underlying the tuning change was a reduction in AMPA receptor conductance, which was compensated for by increased input resistance and depolarization of the membrane potential.
- This had the effect of maintaining roughly the same firing rates but making firing more variable.
- The broader orientation tuning essentially reflects this higher variability (i.e., a higher probability of randomly responding to stimuli farther away from a neuron's preferred stimulus).

Case study: effects of food restriction on orientation discrimination

Simulation of the costly inference model:



Study question

How would you modify the random dot motion discrimination task to directly test predictions of the resource-rational inference model?

Magnitude estimation

Some examples:

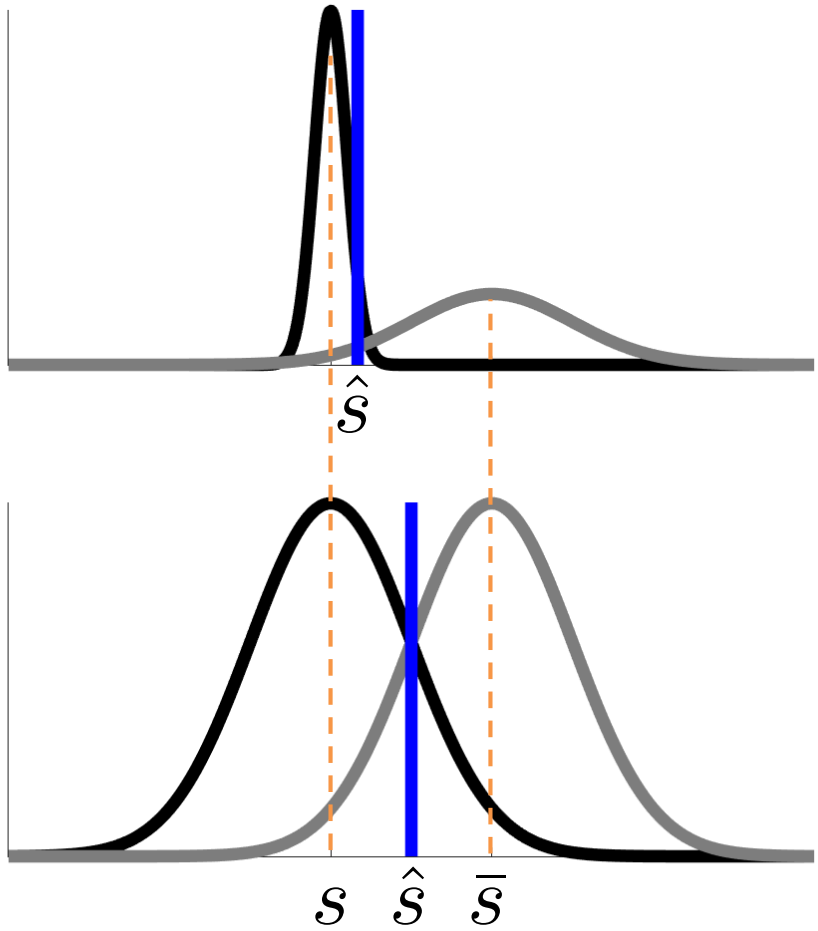
- How far?
- How big?
- How long?
- How many?

Magnitude estimation

Some examples:

- How far?
- How big?
- How long?
- How many?

Gaussian magnitude estimation



Gray curve: prior, $p(s|\bar{s})$. Black curve: signal distribution, $p(x|s)$. Blue line: posterior mean $\hat{s}$.

Gaussian magnitude estimation

Signal-generating process:

$$x \sim \mathcal{N}(s, \sigma_x^2), \quad s \sim \mathcal{N}(\bar{s}, \sigma_s^2).$$

The posterior mean (also the posterior mode) is a convex combination of the signal x and the prior mean $\bar{s}$:

$$\hat{s} = wx + (1 - w)\bar{s},$$

where $w = \frac{\sigma_s^2}{\sigma_x^2 + \sigma_s^2}$ is the signal sensitivity.

Central tendency effect

The bias for Gaussian magnitude estimation is given by:

$$\mathbb{E}[\hat{s} - s|s] = (1 - w)(\bar{s} - s).$$

The prior mean attracts the posterior mean (bias always pushes $\hat{s}$ towards $\bar{s}$), and the strength of this attraction is inversely proportional to the signal sensitivity w. Many studies show such a central tendency effect.

Moderators of central tendency

- Central tendency should be stronger when the signal variance is large relative to the prior variance.

Moderators of central tendency

- Central tendency should be stronger when the signal variance is large relative to the prior variance.
- Signal variance tends to increase with magnitude, possibly due to a nonlinear transformation from objective to subjective magnitude (more on this later).

Moderators of central tendency

- Central tendency should be stronger when the signal variance is large relative to the prior variance.
- Signal variance tends to increase with magnitude, possibly due to a nonlinear transformation from objective to subjective magnitude (more on this later).
- Consistent with this prediction, central tendency is stronger for larger magnitudes, and shorter stimulus durations [Xiang et al 2021].

Repulsion

- Sometimes human judgments are *repulsed* from the prior mean—an apparently "anti-Bayesian" bias.

Repulsion

- Sometimes human judgments are *repulsed* from the prior mean—an apparently "anti-Bayesian" bias.
- For example, people judge a smaller object to be heavier than a larger object with the same mass, the *size-weight illusion*.

Repulsion

- Sometimes human judgments are *repulsed* from the prior mean—an apparently "anti-Bayesian" bias.
- For example, people judge a smaller object to be heavier than a larger object with the same mass, the *size-weight illusion*.
- This seems to defy the prior that larger objects tend to be more massive.

Formalizing attraction/repulsion

- Bias is attractive when pointing towards the prior mode, repulsive when pointing away from the prior mode.

Formalizing attraction/repulsion

- Bias is attractive when pointing towards the prior mode, repulsive when pointing away from the prior mode.
- Let $p'(s)$ denote the derivative of the prior $p(s)$. Then repulsion when $\mathbb{E}[\hat{s} - s|s]p'(s) < 0$.

Formalizing attraction/repulsion

- The direction bias can be approximated [Hahn & Wei 2024]:

$$\mathbb{E}[\hat{s} - s|s]p'(s) \approx \frac{1}{J(s)}\left[\frac{p'(s)^2}{p(s)} - \frac{J'(s)p'(s)}{J(s)}\right]$$

where $J(s)$ is the Fisher information (measure of coding precision).

Formalizing attraction/repulsion

- The direction bias can be approximated [Hahn & Wei 2024]:

$$\mathbb{E}[\hat{s} - s|s]p'(s) \approx \frac{1}{J(s)}\left[\frac{p'(s)^2}{p(s)} - \frac{J'(s)p'(s)}{J(s)}\right]$$

  where $J(s)$ is the Fisher information (measure of coding precision).
- Repulsion will occur when $J'(s)$ and $p'(s)$ have the same sign and their product is large enough to outweigh the first term.

Formalizing attraction/repulsion

- The direction bias can be approximated [Hahn & Wei 2024]:

$$\mathbb{E}[\hat{s} - s|s]p'(s) \approx \frac{1}{J(s)}\left[\frac{p'(s)^2}{p(s)} - \frac{J'(s)p'(s)}{J(s)}\right]$$

  where $J(s)$ is the Fisher information (measure of coding precision).
- Repulsion will occur when $J'(s)$ and $p'(s)$ have the same sign and their product is large enough to outweigh the first term.
- Consistent with Bayes' rule!

Diminishing sensitivity

- Many psychophysical studies show diminishing sensitivity for larger magnitudes. Neurally this corresponds to decreasing coding precision, $J'(s) < 0$.

Diminishing sensitivity

- Many psychophysical studies show diminishing sensitivity for larger magnitudes. Neurally this corresponds to decreasing coding precision, $J'(s) < 0$.
- This implies that repulsion should tend to occur when $p'(s) < 0$.

Diminishing sensitivity

- Many psychophysical studies show diminishing sensitivity for larger magnitudes. Neurally this corresponds to decreasing coding precision, $J'(s) < 0$.
- This implies that repulsion should tend to occur when $p'(s) < 0$.
- Many natural magnitudes have this property. For example, the distribution of spatial frequencies in natural images falls off according to a power law: $p(s) \propto s^{-\alpha}$ with α between 1 and 2.

Diminishing sensitivity

- Many psychophysical studies show diminishing sensitivity for larger magnitudes. Neurally this corresponds to decreasing coding precision, $J'(s) < 0$.

Diminishing sensitivity

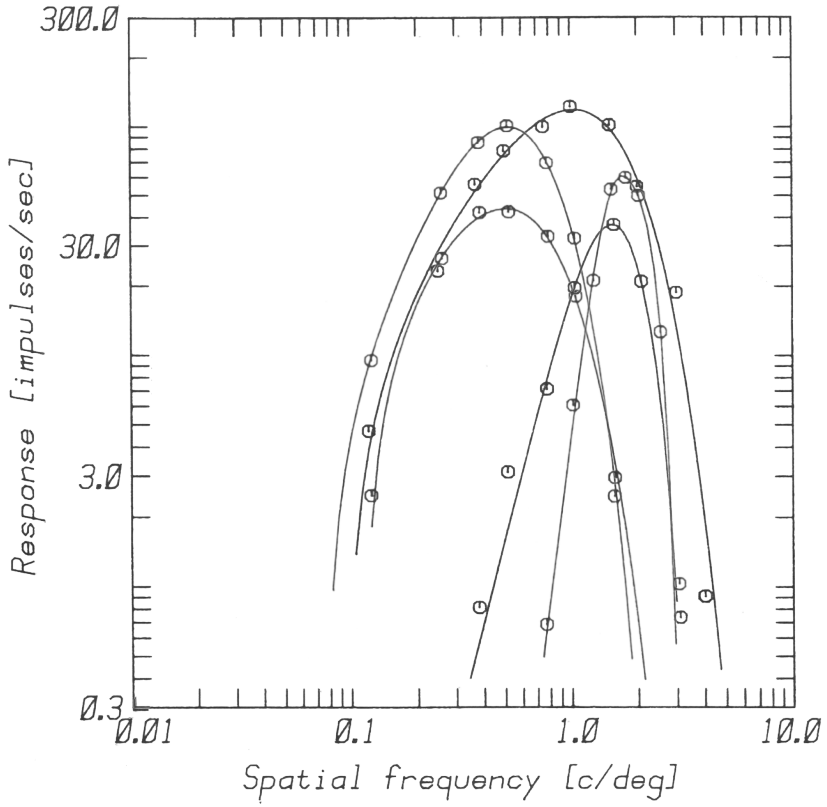
- Many psychophysical studies show diminishing sensitivity for larger magnitudes. Neurally this corresponds to decreasing coding precision, $J'(s) < 0$.
- This implies that repulsion should tend to occur when $p'(s) < 0$.

Diminishing sensitivity

- Many psychophysical studies show diminishing sensitivity for larger magnitudes. Neurally this corresponds to decreasing coding precision, $J'(s) < 0$.
- This implies that repulsion should tend to occur when $p'(s) < 0$.
- Many natural magnitudes have this property. For example, the distribution of spatial frequencies in natural images falls off according to a power law: $p(s) \propto s^{-\alpha}$ with α between 1 and 2.

Neural correlate of diminishing sensitivity

Spatial frequency tuning in V1 can be modeled using Gaussian tuning functions defined over log-transformed frequency.



[Movshon, unpublished]

Neural correlate of diminishing sensitivity

- If we assume Poisson spiking and a uniform density of preferred stimuli in log space, we get (in the limit of a large population):

$$J(s) \propto \frac{1}{s^2}$$

Neural correlate of diminishing sensitivity

- If we assume Poisson spiking and a uniform density of preferred stimuli in log space, we get (in the limit of a large population):

$$J(s) \propto \frac{1}{s^2}$$

- Thus, diminishing sensitivity, $J'(s) < 0$, can be derived from an approximation of the empirical tuning functions.

Neural correlate of diminishing sensitivity

- If we assume Poisson spiking and a uniform density of preferred stimuli in log space, we get (in the limit of a large population):

$$J(s) \propto \frac{1}{s^2}$$

- Thus, diminishing sensitivity, $J'(s) < 0$, can be derived from an approximation of the empirical tuning functions.
- This satisfies the assumptions underlying our analysis of repulsion.

Summary

- We started with the normative ideal of Bayesian inference, and then tried to explain both the successes and failures of this ideal as a model of inference in the brain.

Summary

- We started with the normative ideal of Bayesian inference, and then tried to explain both the successes and failures of this ideal as a model of inference in the brain.
- The key idea is that computational and representational constraints shape inference in ways that comport with empirical observations.

Summary

- We started with the normative ideal of Bayesian inference, and then tried to explain both the successes and failures of this ideal as a model of inference in the brain.
- The key idea is that computational and representational constraints shape inference in ways that comport with empirical observations.
- We also saw how these constraints can be realized in simple neural networks.

Study question

In what ways might resource-rational inference vary systematically across individuals (e.g., children, older adults, clinical populations)? How would you test this empirically?