Module 3: Reinforcement Learning

Andrew Howes

Overview

- Reinforcement learning (RL) problems versus RL algorithms
- POMDPs
- Bayesian estimation
- An example RL model of gaze-based interaction

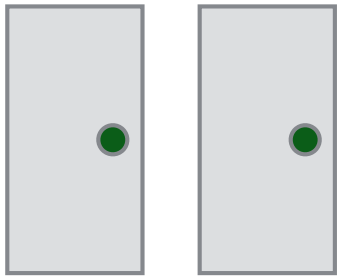
Problems v Algorithms

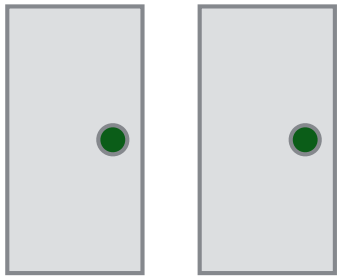
- RL is the problem of learning from the environment.
- Some prefer to say that it is the problem of learning from experience.
- An RL **algorithm** learns a **policy** that maps observations of the state of the world into actions.
- There are many RL algorithms that learn incrementally, sometimes by tuning a policy, sometimes by searching a space of policies, for example.
- Some of these algorithms use ANNs.

Problems v Algorithms

- RL learns policies for specific **RL problems**.
- For example, chess is one problem, balancing a beam is another.
- For many (though not all) cognitive modelling problems we can focus on the **RL problem specification**.
- … and then use off-the-shelf RL algorithms to solve this problem. This is the approach taken in this course.

The tiger problem



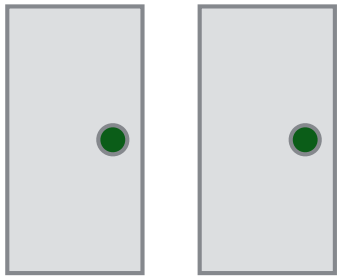


The partially observable setting. Consider the following example: A person must choose to open either the left door or the right door.

There is a Tiger behind one of the doors.

The person can listen and try and determine which door hides the tiger.

But listening is not 100% accurate. In the example, the frequency with which listening gives the right answer is 85 in 100.

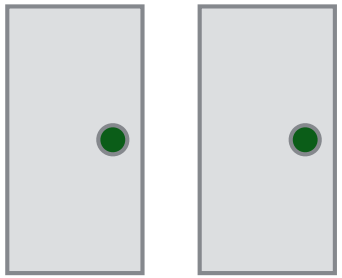


State estimation. If the person hears that the Tiger is behind the left door then they can believe with probability p = 0.85 that the Tiger is behind that door.

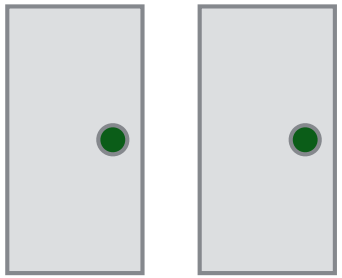
The probability represents a degree of certainty.

The accuracy of the belief can be improved by making multiple observations.

p2 = 0.85 * 0.85 / (0.85 * 0.85 + 0.15 * 0.15)



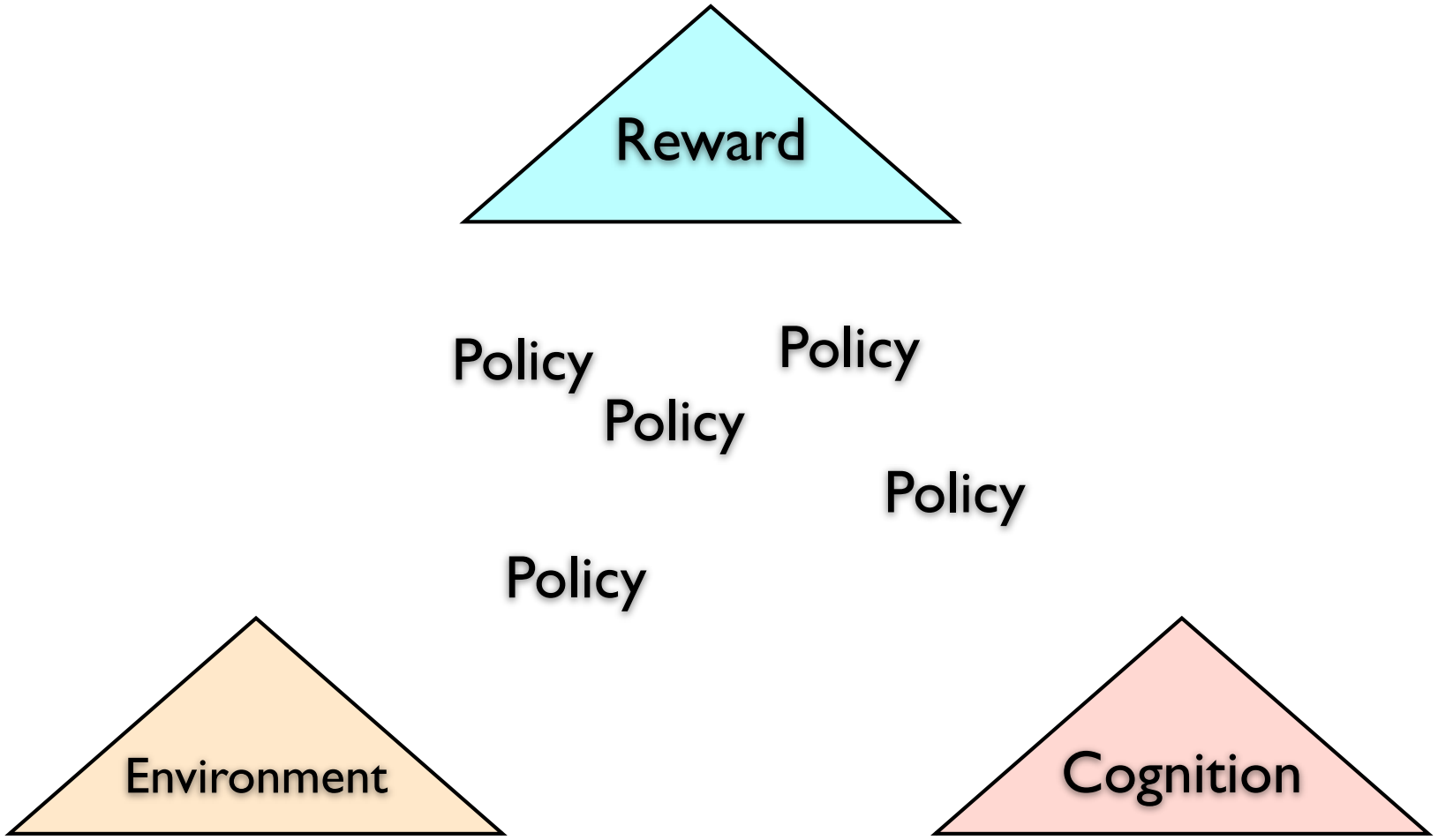
Reward. Opening a door with a tiger behind it has a severe cost and a much lower cost (or perhaps a benefit) if there is a domestic cat.

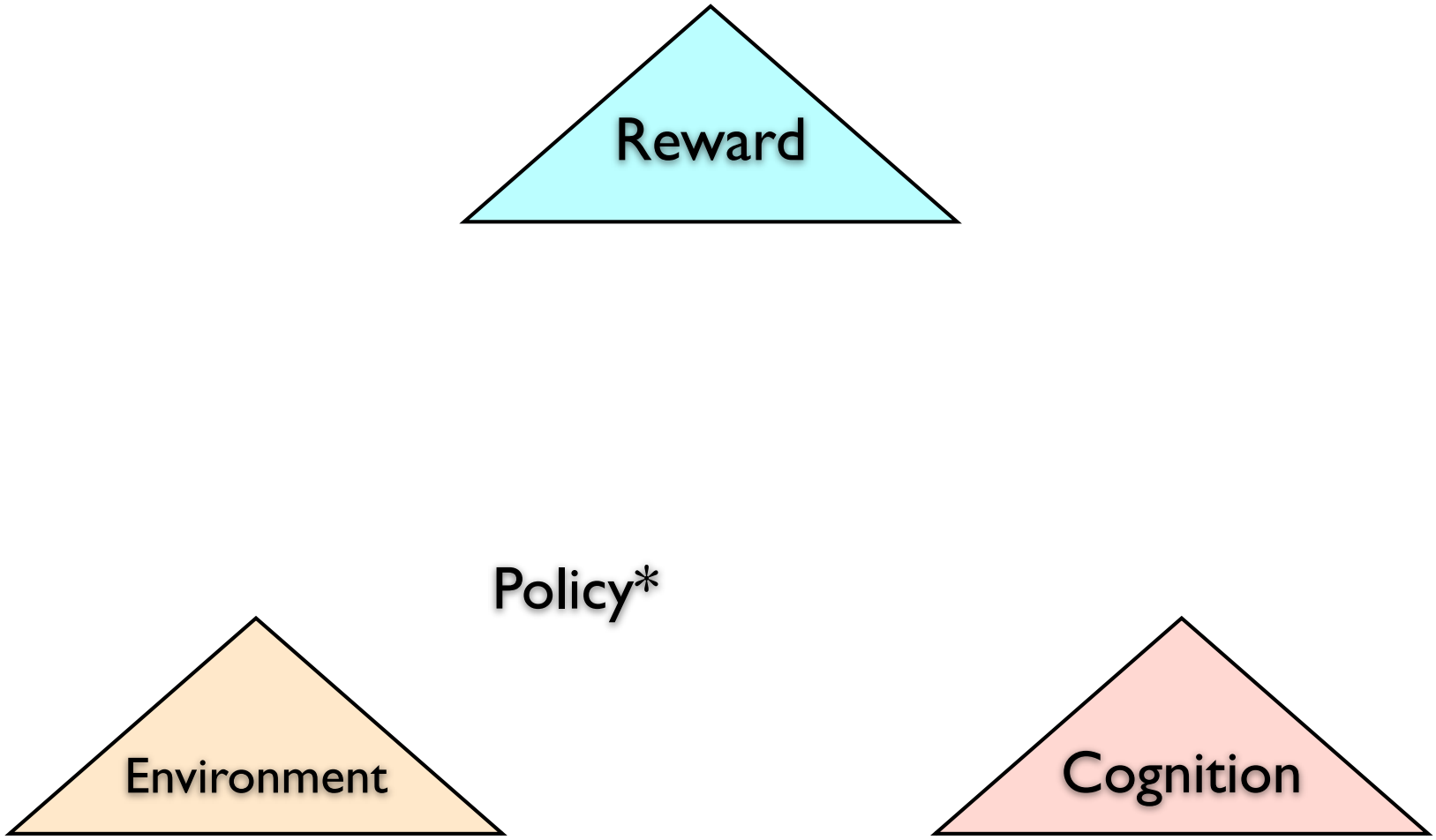


Policy. Determining what to do depends on the costs and benefits of action. Opening the wrong door will have a severe cost if there is a tiger behind it.

POMDPs for cognition

- The tiger problem is an example of a Partially Observable Markov Decision Problem (POMDP).
- We can use POMDPs to model cognition.
- To do so we must specify theories of the bounds and of the reward function and then find the **computationally rational policy.**


Reward
Policy
Policy
Policy
Policy
Policy
Environment
Cognition



Learning selects the optimal policy.
This is the “computationally rational” or “resource rational” policy.
Lewis, Howes, Singh, 2014; Lieder and Griffiths, 2020

How to model human behaviour as a POMDP

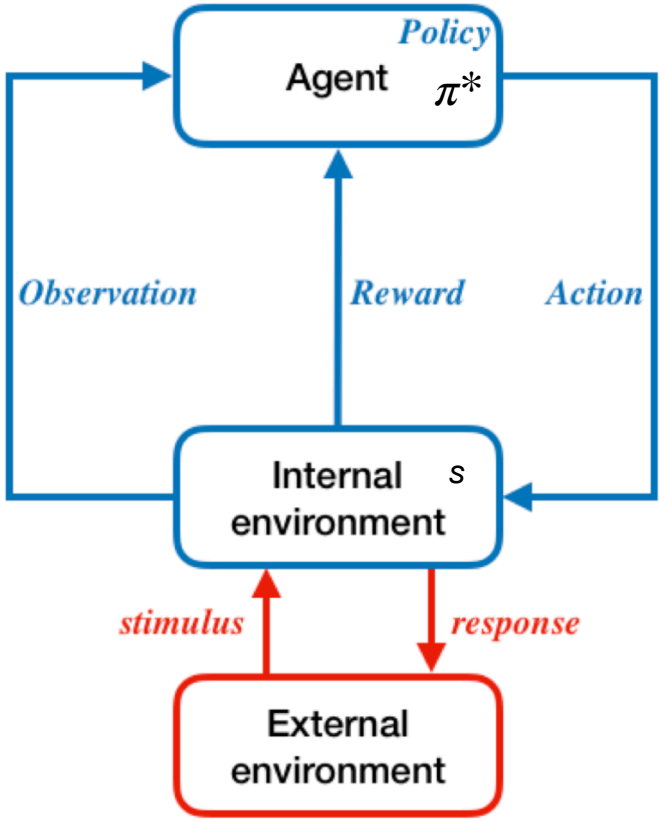
- A POMDP models an agent decision process where the agent cannot directly observe the underlying state.
- A policy is a mapping from the observations to the actions.
- The POMDP framework is general enough to model cognition in a wide range of HCI and HCI-related settings.

- An exact solution to a POMDP yields the optimal action for each possible belief over the world states.
- The optimal policy maximizes the expected reward of the agent.

A POMDP models the interaction between an agent and its environment.

Formally, a POMDP is a tuple (S,A,T,R,Ω,O,γ), where

* ***S*** **is a set of states (the internal states of the user)**
* ***A*** **is a set of actions**
* ***T*** **is a state transition function**
* ***R*****: is the reward function**
* **Ω is a set of observations**
* **O is an observation function**
* **γ is a discount factor.**



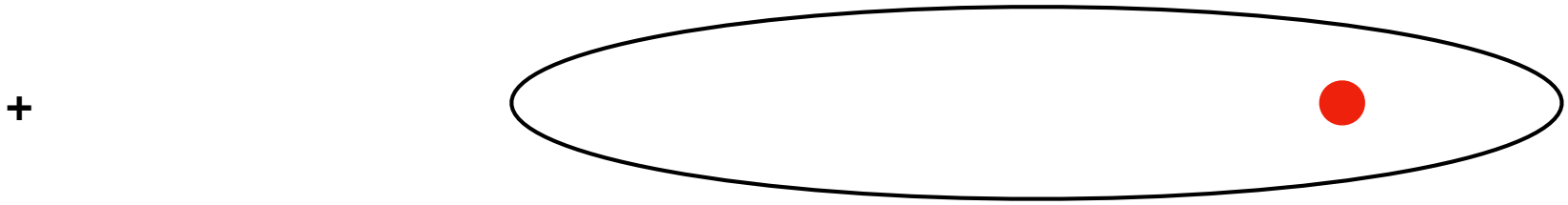
State estimation with Bayesian inference

Noise

- Observations are partial and they are noisy.
- Look at the + and then estimate the location of the red dot.
- Try physically pointing to the red dot while looking at the cross. How close were you?

Noise

- The open circle represents a prior probability of the location of the red circle.
- The size of the circle represents observation uncertainty with the centre of the circle representing the most likely location.



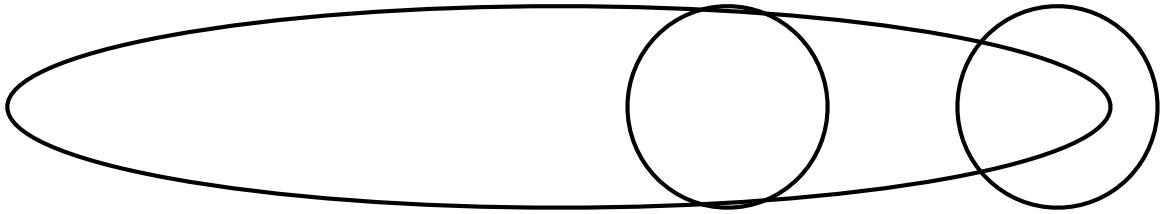
Noise

- An observation provides more information but, by itself, it is just as uncertain.

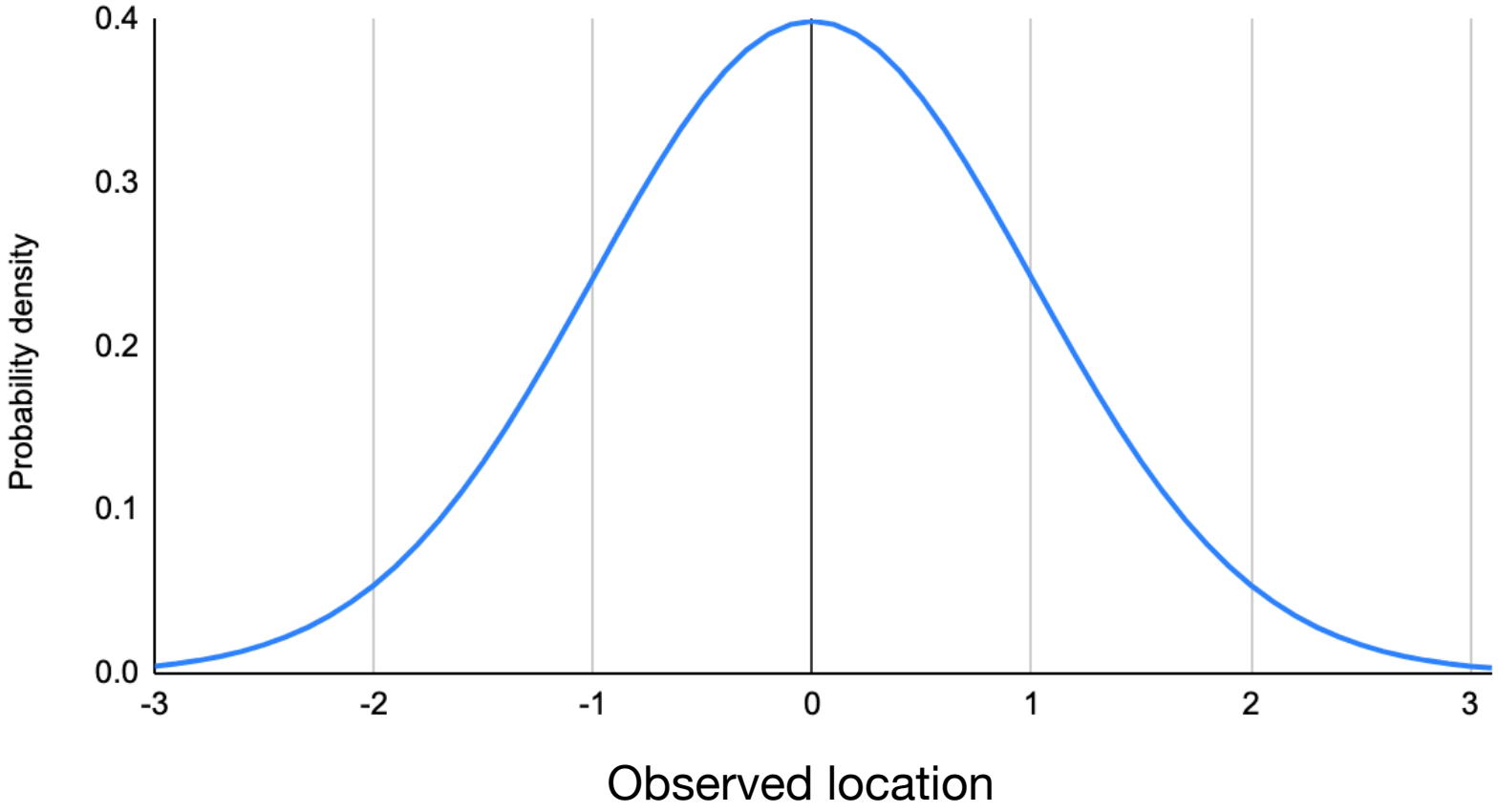
Noise

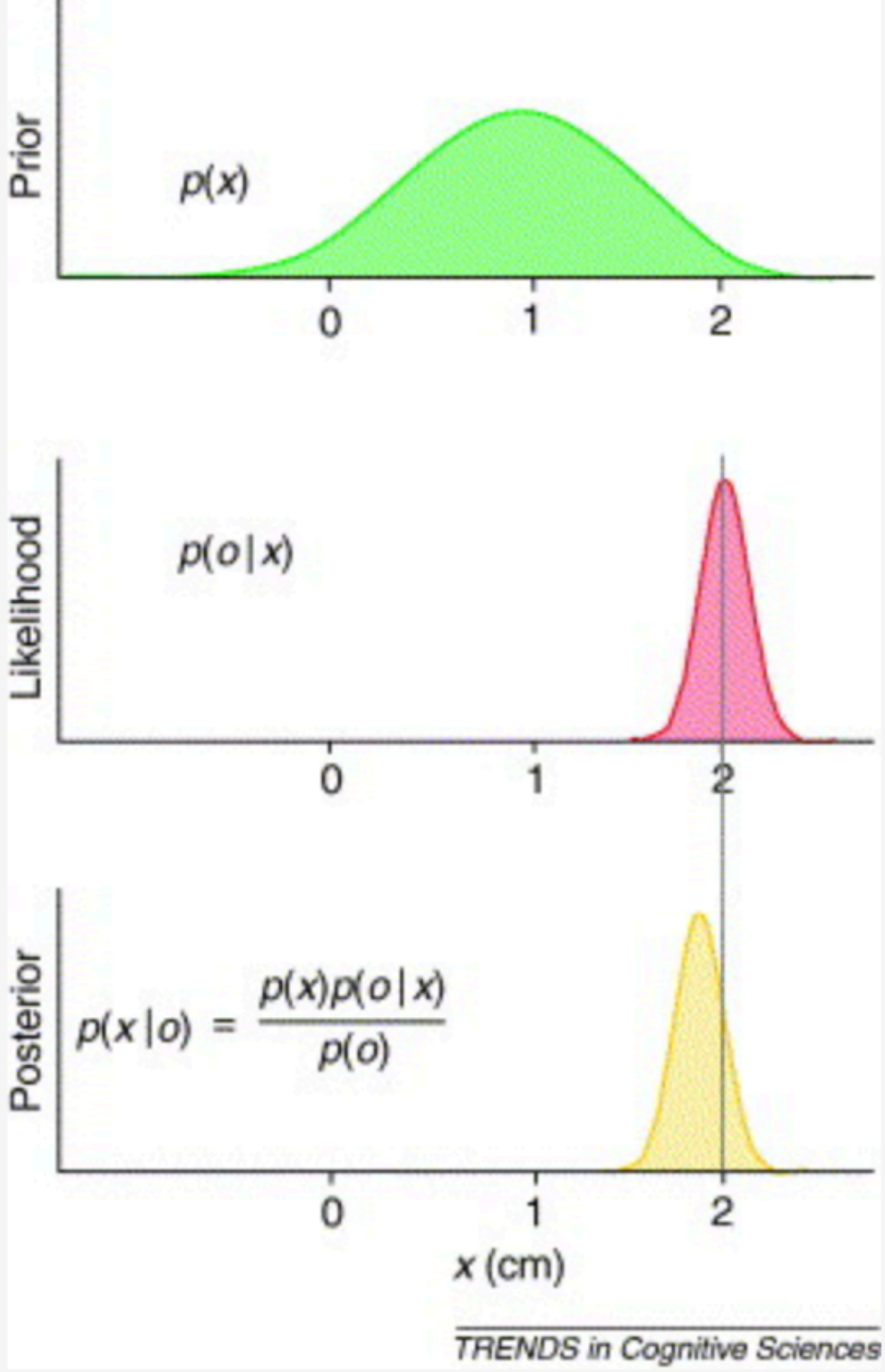
- A second observation provides more information but, by itself, it is just as uncertain.

How do people put together a prior and multiple observations into a single estimate of the state of the world?



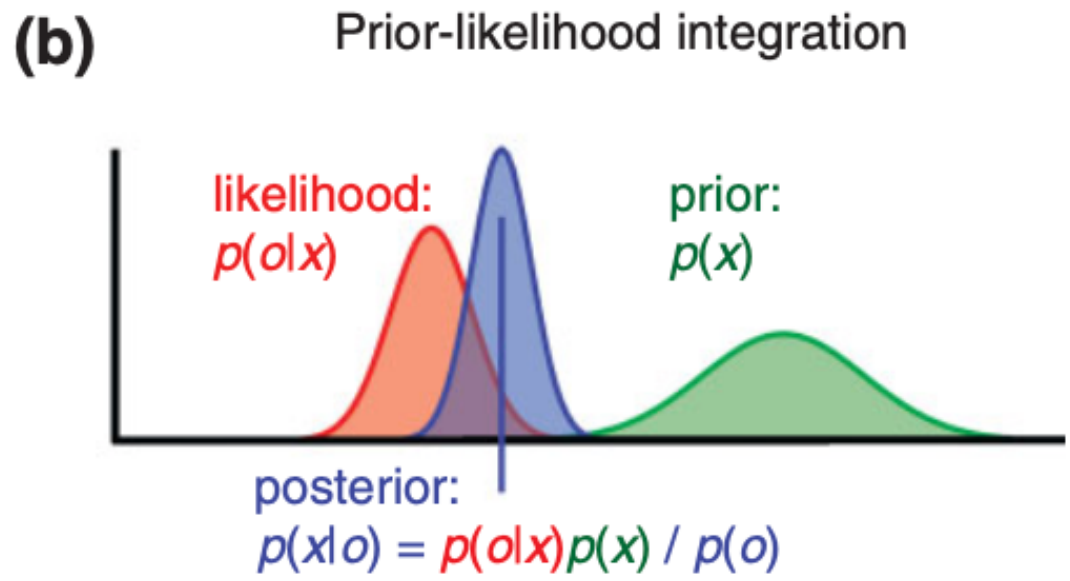
Assume that location observation noise can be modelled with the normal distribution




Prior
p(x)
0
1
2
Likelihood
p(o|x)
0
1
2
Posterior
p(x|o) = p(x)p(o|x) / p(o)
0
1
2
x (cm)
TRENDS in Cognitive Sciences

Bayesian estimation as a model of human estimation

- For a location x, people have a prior expectation as to its location. This can be represented as a Gaussian distribution with mean and standard deviation (green distribution).
- Prior and Observation can be integrated into a posterior estimate.

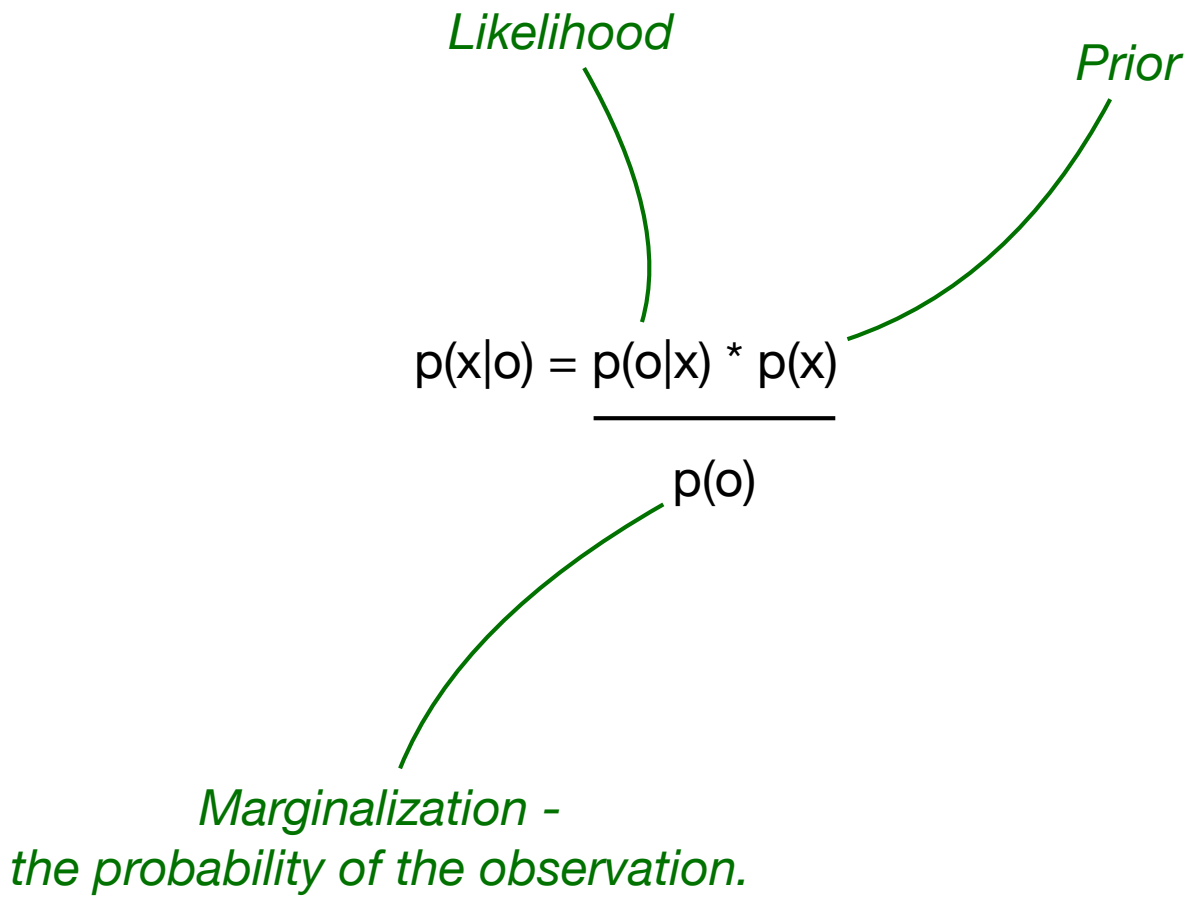


Berniker and Kording, 2011

Bayesian estimation

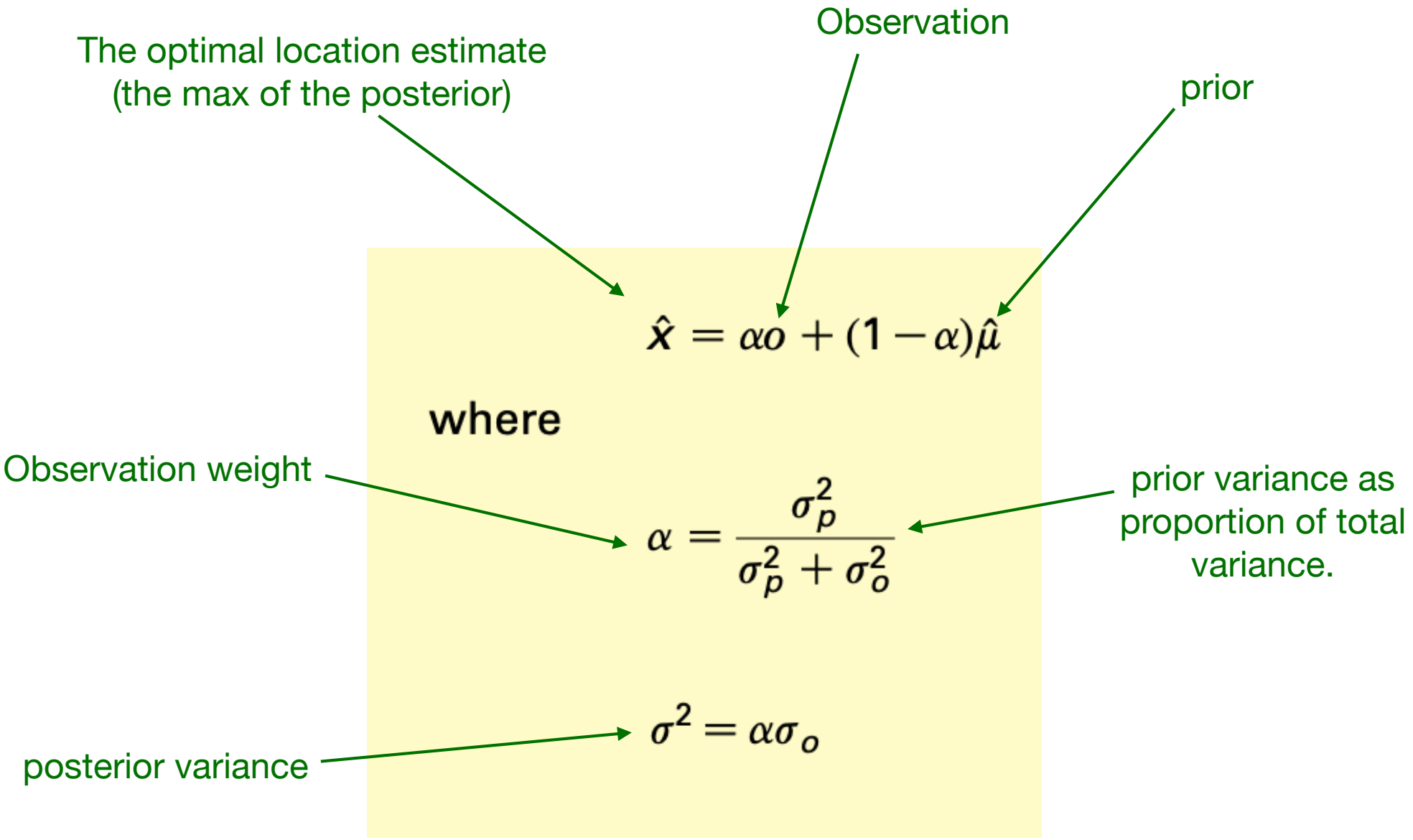
- Given an observation o,
- then for a possible location x
- We can calculate the probability p(x|o).
- p(x|o) is the probability of location x *given* observation o.
- This is known as the **posterior** probability.
- The posterior probability can be calculated if we know the **prior** probability p(x) and the **likelihood**, p(o|x).

Bayes' rule


Likelihood
Prior
p(x|o) = p(o|x) * p(x)
p(o)
Marginalization -
the probability of the observation.

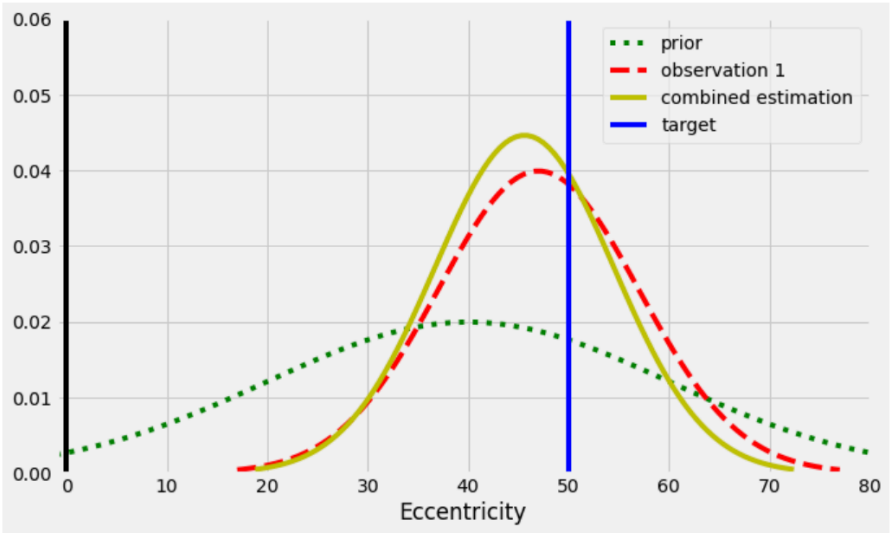
Bayes

- 1701-1761
- Many problems concerning the probability of certain events, can be solved.
- Revolutionised statistics in the 20th century and psychology in the 1990s and 2000s.


The optimal location estimate
(the max of the posterior)
Observation
prior
$\hat{x} = \alpha o + (1-\alpha)\hat{\mu}$
where
Observation weight
$\alpha = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_o^2}$
prior variance as
proportion of total
variance.
$\sigma^2 = \alpha\sigma_o$
posterior variance

Example

- prior: $\hat{\mu} = 40,\ \sigma_p = 30$
- observation: $o = 47,\ \sigma_o = 10$
- posterior $\hat{x} = 45.6,\ \sigma = 8.9$

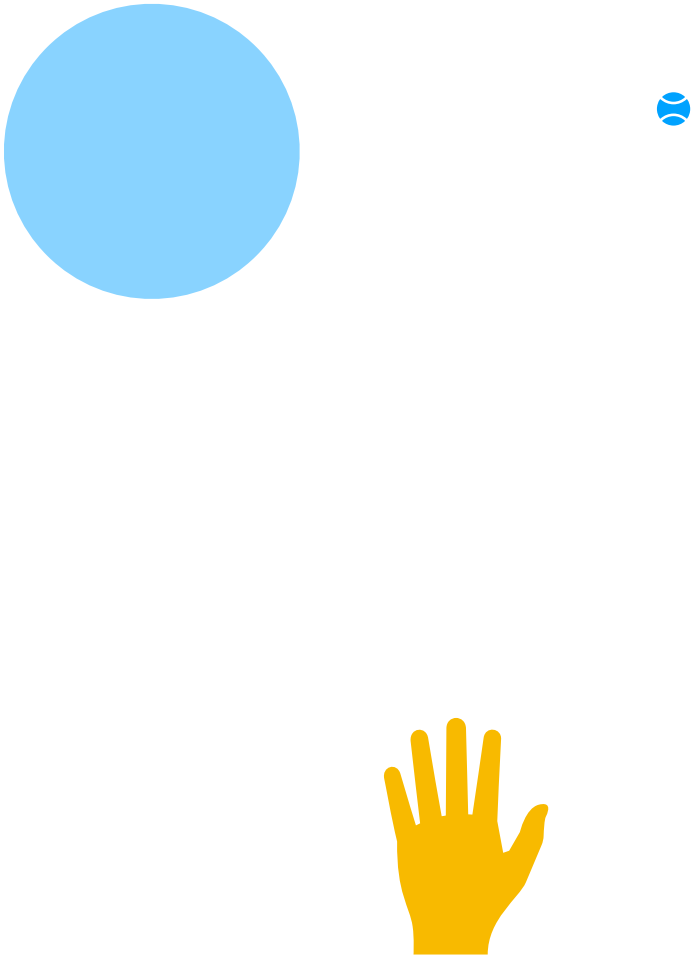


Why is this important to modelling interaction?

- People use a strategy that is Bayes optimal for perceptual estimation.

Evidence

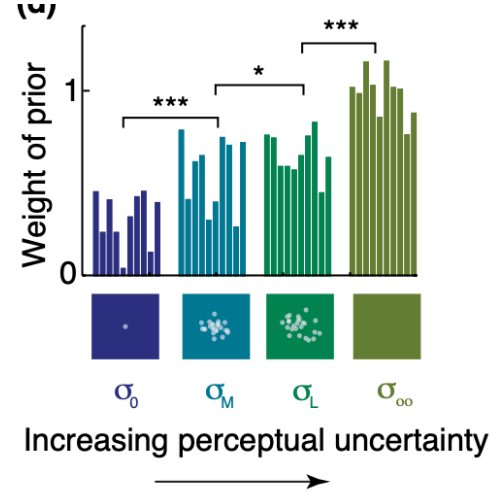
- Kording and Wolpert (2004) tested whether people use a Bayesian strategy.
- Subjects had to estimate the position of a cursor relative to their hand.
- Subjects could use two sources of information:
  - The distribution of displacements over the course of many trials (prior),
  - as well as what they see during the current trial (giving a likelihood).
- The quality of the visual feedback was also varied, in some cases a ball was shown at the position of the cursor giving precise feedback whereas in other trials a large cloud was shown at the position of the cursor thereby increasing the variability (noise) in the sensory input.
- Kording, K.P. and Wolpert, D.M. (2004) Bayesian integration in 42 sensorimotor learning. Nature 427, 244–247

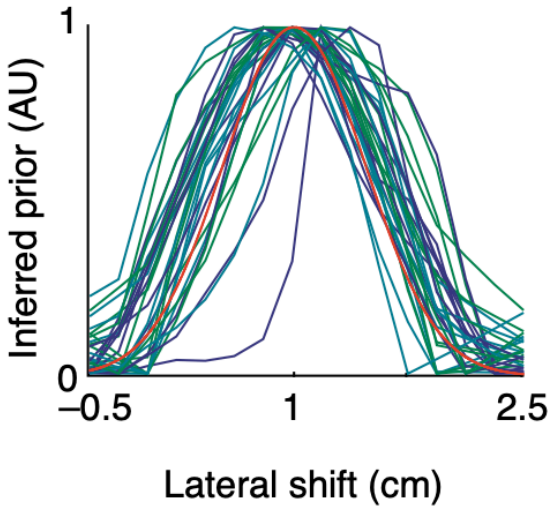


results

- The Bayesian estimation process predicts that with increasing noise in the sensory feedback subjects should increase the weight of the prior and decrease the weight of their sensory feedback in their final estimate of the location.

- People used a prior that was very close to optimal. (Optimal prior in red.)





Practical exercise 2: Bayesian inference