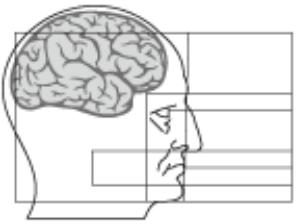


MAX PLANCK UCL CENTRE
for Computational Psychiatry and Ageing Research

UCL

HGF Workshop

Christoph Mathys

Wellcome Trust Centre for Neuroimaging at UCL,
Max Planck UCL Centre for Computational Psychiatry and Ageing Research

Monash University, March 3, 2015

Systems theory places constraints on the mind

- It is widely taken for granted that the mind functions within the confines of neurobiology.
- It is less appreciated that there are also systems theoretic constraints on how the mind has to operate.
- In systems theory, the mind (and its substrate, the body including the brain) is seen as a regulator of its environment.
- In order to survive, the mind has to be a good regulator of its environment.
- That is, the mind has to regulate its environment in way that ensures its – the mind's – further existence.

"Every good regulator of a system must be a model of that system"

- This is the title of a paper by Conant & Ashby (1970) where they give a proof of this statement (the "good regulator theorem").

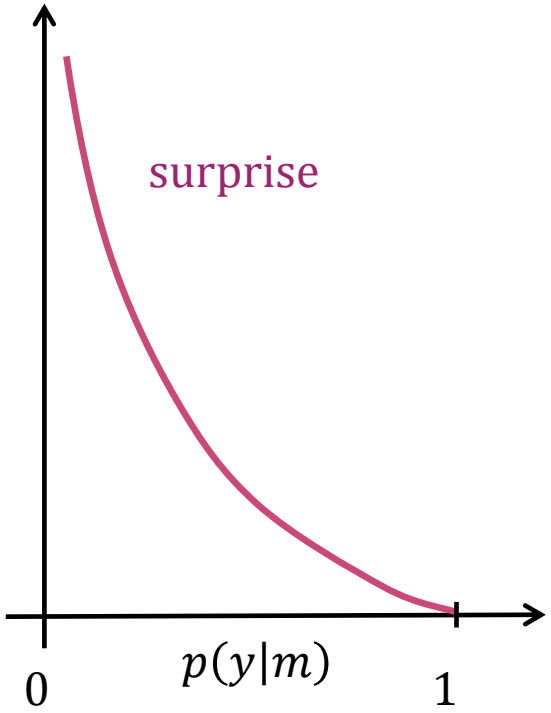
- In addition to a systems theorist, Ashby was a psychiatrist and as such immediately understood the consequences of his theorem for the brain:

"The theorem has the interesting corollary that the living brain, so far as it is to be successful and efficient as a regulator for survival, *must* proceed, in learning, by the formation of a model (or models) of its environment."

There's more: in order to be a good regulator, the brain needs to minimize surprise

- The formal definition of surprise in words: *the surprise associated with an event is the negative logarithm of that event's probability.*

- As a graph:



Minimizing the time-average of surprise is equivalent to minimizing entropy

- Under ergodic* assumptions, the sum (or, more precisely, the integral) of surprise over time is entropy.

  *Ergodic systems are such where time spent in a given state is proportional to the probability of that state.

- This gives us an additional perspective on what it means to stay alive: we have to keep the entropy of our sensations (ie, of the states we visit) low.

- Here we have the link between information entropy and physical entropy: an organism that wants stay alive has to resist the second law of thermodynamics (an increase in its own physical entropy would mean death), and the way it achieves this is by minimizing information entropy (ie, by sampling its environment such that external and internal states are predictable).

But there's a problem: surprise is intractable

- In all but the simplest cases, the equation for surprise has no closed-form solutions.

- One way to deal with this is to introduce approximations. Since the minds we know are certainly not optimal, it's a safe assumption that they are not minimizing surprise, but an approximation to it.

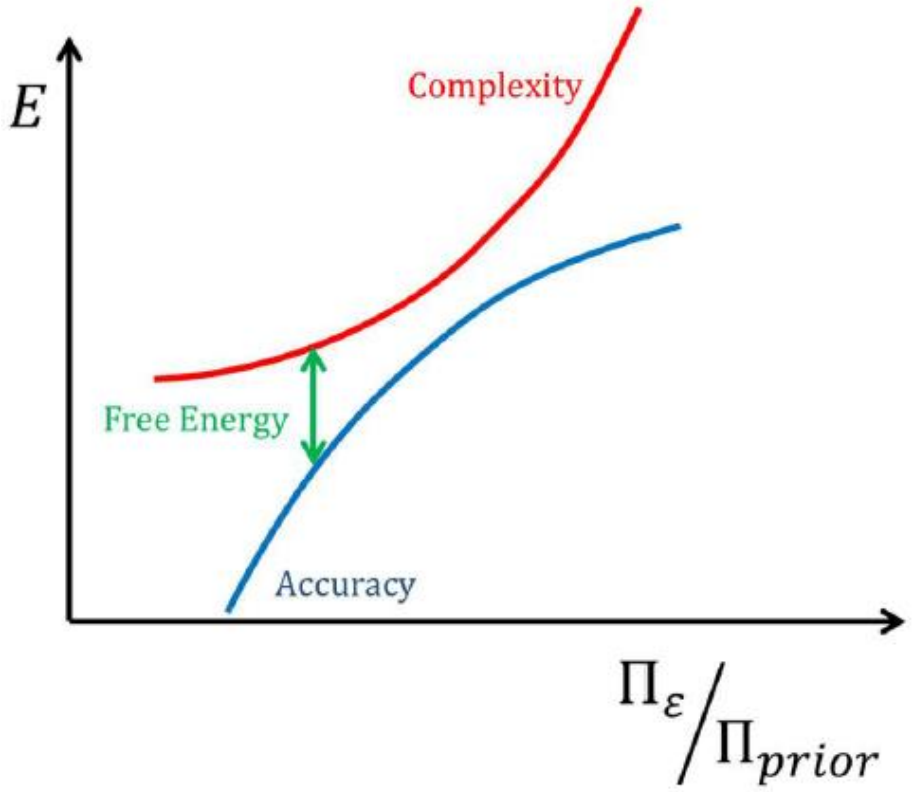
- One possible and plausible approximation to surprise is **variational free energy** (cf. Friston, 2009; Feynman, 1972).

What does it mean to have a model?

- It means ascribing hidden states to the environment which are related to each other by parameters.
- Hidden states change in time, parameters do not.
- Hidden states are hidden in the sense that they are not directly accessibly to the sensorium but have to be inferred on the basis of sensory evidence.
- The probability of a certain sensation given hidden states and parameters is called the **likelihood**.
- The likelihood alone is not a complete description of the model. We still need the probability of the hidden states and parameters. These are called the **priors**.
- The product of likelihood and priors is the **joint** probability (of sensations, hidden states, and parameters) and constitutes a **generative model**.

Optimal inference depends on optimal precision

- *A* can be decomposed into complexity minus accuracy.



- *A* is minimized when the precision of the likelihood is optimal relative to the precision of the prior.

Varieties of free energy

At least three kinds of free energy have to be kept apart:

- Thermodynamic free energy
- Informational free energy
- Variational free energy

First however, we need to know the reason why thermodynamic quantities show up (at least in name) in information theory.

Thermodynamic free energy

- Two kinds: Gibbs and Helmholtz
- Helmholtz free energy:

$$A := U - TS$$

- U: internal energy; T: temperature; S: entropy
- [Gibbs free energy: $G := U + pV - TS$]

Informational free energy

Here is where it gets interesting, but first we need some new concepts.

- A **generative model** m of an observation y has two components.

- First, the **likelihood:**

$$p(y|\vartheta, m)$$

- This is the probability of the observation, given the model and a particular set of parameter values ϑ.

- Second, the **prior:**

$$p(\vartheta|m)$$

- This is the probability that the particular set of parameter values ϑ had to begin with (therefore: "prior").

Informational free energy

- Multiplied together, the likelihood and the prior give the **joint** probability of observations and parameter settings:

$$p(y, \vartheta|m) = p(y|\vartheta, m)p(\vartheta|m)$$

- This equality holds because of the product rule of probability theory

- Such a joint probability consisting of a likelihood and a prior is what we mean when we speak of a generative model.

Informational free energy

- The next important concept is the **posterior** probability $p(\vartheta|y,m)$. This is the probability of a particular set of parameter values given the observation and the model.

- Like the joint probability, it can be calculated using the product rule:

$$p(\vartheta|y,m) = \frac{p(y|\vartheta,m)p(\vartheta|m)}{p(y|m)}$$

- This particular application of the product rule is called **Bayes' theorem**.

Informational free energy

- Bayes' theorem now contains another new concept: the **model evidence** or **marginal likelihood** $p(y|m)$. This is the overall probability of making observation y given model m, regardless of parameter values (i.e., after taking account of all possible parameter values according to their probability):

$$p(y|m) = \int p(y|\vartheta, m)p(\vartheta|m)\mathrm{d}\vartheta$$

- It makes intuitive sense to take the negative logarithm of $p(y|m)$ as a measure of **surprise**: if $p(y|m) = 1$, the outcome was certain and there was no surprise at all ($-\log\big(p(y|m)\big) = 0$); if, however, $p(y|m) = 0$, the outcome was impossible and surprise is infinite ($-\log\big(p(y|m)\big) = \infty$). In between, surprise is greater than zero and increases for less probable observations.

Informational free energy

- Surprise is essential as a measure of how good a model is. When we compare models, we calculate the **Bayes factor**:

$$BF = \frac{p(y|m_1)}{p(y|m_0)}$$

- This is a measure of whether model m_1 is more surprised by the outcome y than model m_0.

Entropy

- The more ignorant we are about a quantity, the greater is the surprise we may expect when observing it.
- Expected surprise is called the **entropy** S of a probability distribution p:

$$S[p] \coloneqq -\int p(\vartheta) \log p(\vartheta) \, \mathrm{d}\vartheta$$

- Entropy is a **measure of ignorance.**
- Its name is due to an analogous quantity in thermodynamics.

Entropy example

- As a simple example, let's look at a coin toss.
- There are two possible outcomes: $\vartheta \in \{\text{heads}, \text{tails}\}$
- Since outcomes are discrete and binary, we use a sum instead of an integral and the binary logarithm to define the entropy:

$$S[p] := -\sum_{\vartheta} p(\vartheta) \log_2 p(\vartheta)$$

- For a fair coin (i.e., $p(\text{heads}) = p(\text{tails}) = \frac{1}{2}$), $S[p] = 1$
- However, for $p(\text{heads}) = \frac{9}{10}$, $p(\text{tails}) = \frac{1}{10}$, we get $S[p] \approx 0.47$ because expected surprise is much lower.

Informational free energy

- We can now begin to understand the connection with free energy. First, we perform a series of algebraic operations on the negative logarithm of surprise $p(y|m)$:

$$A \coloneqq -\log p(y|m) = -\int p(\vartheta|y,m)\log p(y|m)\,\mathrm{d}\vartheta$$

$$= -\int p(\vartheta|y,m)\log\frac{p(y,\vartheta|m)}{p(\vartheta|y,m)}\,\mathrm{d}\vartheta$$

$$= \underbrace{-\int p(\vartheta|y,m)\log p(y,\vartheta|m)\,\mathrm{d}\vartheta}_{\coloneqq U} - \underbrace{\left(-\int p(\vartheta|y,m)\log p(\vartheta|y,m)\,\mathrm{d}\vartheta\right)}_{\coloneqq S}$$

- This gives us an **information theoretic analogon** to the definition of Helmholtz free energy in thermodynamics

Variational free energy

- The problem with informational free energy is that we cannot calculate it except in trivial cases. Whenever models are complicated enough to be interesting, the integrals involved are intractable.

$$A \coloneqq \underbrace{-\int p(\vartheta|y,m)\log p(y,\vartheta|m)\,\mathrm{d}\vartheta}_{\coloneqq U} - \underbrace{\left(-\int p(\vartheta|y,m)\log p(\vartheta|y,m)\,\mathrm{d}\vartheta\right)}_{\coloneqq S}$$

- The solution to this is variational free energy, where we replace the true posterior $p(\vartheta|y,m)$ by an approximation $q(\vartheta)$:

$$A_v \coloneqq \underbrace{-\int q(\vartheta)\log p(y,\vartheta|m)\,\mathrm{d}\vartheta}_{\coloneqq U_v} - \underbrace{\left(-\int q(\vartheta)\log q(\vartheta)\,\mathrm{d}\vartheta\right)}_{\coloneqq S_v}$$

Variational free energy

- What makes variational free energy A_v such an extremely useful concept is the following theorem:

$$A_v \geq A \text{ for all } q(\vartheta)$$

- This means that **whatever $q(\vartheta)$ we plug into A_v, we get an A_v that is greater than A**. So without having to know anything about A, we can vary $q(\vartheta)$ such that it minimizes A_v.

$$A_v := -\int q(\vartheta) \log p(y, \vartheta|m) \, \mathrm{d}\vartheta + \int q(\vartheta) \log q(\vartheta) \, \mathrm{d}\vartheta$$

- The branch of mathematics that describes how to carry out the minimization of A_v with respect to $q(\vartheta)$ is called **variational calculus**, hence "variational" free energy.

- Minimizing A_v with respect to $q(\vartheta)$ leads to an approximation of $p(\vartheta|y, m)$ by $q(\vartheta)$ because of the theorem above and because $A_v = A$ for $q(\vartheta) = p(\vartheta|y, m)$.

- The remarkable thing here is that we can use variational calculus to find a $q(\vartheta)$ that approximates $p(\vartheta|y, m)$ **without ever having to know $p(\vartheta|y, m)$ itself**.

- This is how the brain can build, update, and compare models of the world without ever "seeing behind the scenes" of its sensory input.

Variational free energy

Proof that $A_v \geq A$ for all $q(\vartheta)$:

$$
\begin{aligned}
A &:= -\log p(y|m) \\
&= -\log \int p(y,\vartheta|m)\mathrm{d}\vartheta \\
&= -\log \int q(\vartheta)\frac{p(y,\vartheta|m)}{q(\vartheta)}\mathrm{d}\vartheta \\
&\leq -\int q(\vartheta)\log\frac{p(y,\vartheta|m)}{q(\vartheta)}\mathrm{d}\vartheta \\
&= -\int q(\vartheta)\log p(y,\vartheta|m)\,\mathrm{d}\vartheta + \int q(\vartheta)\log q(\vartheta)\,\mathrm{d}\vartheta \\
&=: A_v
\end{aligned}
$$

Jensen's inequality

□

Three ways to decompose A_v

$$A_v := -\int q(\vartheta)\log\frac{p(y,\vartheta|m)}{q(\vartheta)}\,\mathrm{d}\vartheta$$

$$= \underbrace{-\int q(\vartheta)\log p(y,\vartheta|m)\,\mathrm{d}\vartheta}_{\text{Expected energy } U_v} \underbrace{-\left(-\int q(\vartheta)\log q(\vartheta)\,\mathrm{d}\vartheta\right)}_{\text{Entropy } S_v}$$

$$= -\int q(\vartheta)\log\frac{p(\vartheta|y,m)p(y|m)}{q(\vartheta)}\,\mathrm{d}\vartheta = KL[q(\vartheta),p(\vartheta|y,m)]\underbrace{-\log p(y|m)}_{=A}$$

$$= -\int q(\vartheta)\log\frac{p(y|\vartheta,m)p(\vartheta|m)}{q(\vartheta)}\,\mathrm{d}\vartheta = \underbrace{KL[q(\vartheta),p(\vartheta|m)]}_{\text{Complexity}} \underbrace{-\int q(\vartheta)\log p(y|\vartheta,m)\,\mathrm{d}\vartheta}_{\text{Accuracy}}$$

The first decomposition of A_v

$$A_v = -\int q(\vartheta) \log p(y, \vartheta|m)\, \mathrm{d}\vartheta - \left(-\int q(\vartheta) \log q(\vartheta)\, \mathrm{d}\vartheta \right)$$

$$= U_v - S_v$$

$$= \text{Expected energy} - \text{Entropy}$$

- This first decomposition illustrates the mathematical analogy to statistical mechanics.
- More importantly, it only contains quantities known to the model-builder: the joint density $p(y, \vartheta|m)$, consisting of likelihood and prior, and the arbitrary density $q(\vartheta)$.
- Because it only contains known quantities, this decomposition shows that A_v is, in principle, computable up to an arbitrarily small error.

The second decomposition of A_v

$$A_v = KL[q(\vartheta), p(\vartheta|y,m)] \underbrace{-\log p(y|m)}_{=A}$$

$$= \text{Divergence between approixmate and true posterior} + \text{log–model evidence}$$

- The **Kullback-Leibler divergence** between two distributions is defined as

$$KL[p_1, p_2] := \int p_1(\vartheta) \log \frac{p_1(\vartheta)}{p_2(\vartheta)} \mathrm{d}\vartheta$$

- It is zero if and only if $p_1 = p_2$, otherwise positive. It is not symmetric (i.e., $KL[p_1, p_2] \neq KL[p_2, p_1]$ in general).

- This second decomposition again shows that $A_v \geq A$ for all $q(\vartheta)$ (because the divergence is non-negative).

- Crucially, it again shows that **minimizing A_v with respect to $q(\vartheta)$ leads to an approximation of $p(\vartheta|y,m)$ by $q(\vartheta)$**.

The third decomposition of A_v

$$A_v = KL[q(\vartheta), p(\vartheta|m)] - \int q(\vartheta) \log p(y|\vartheta, m)\, \mathrm{d}\vartheta$$

$$= \text{Complexity} - \text{Accuracy}$$

- The expected log-likelihood $\log p(y|\vartheta, m)$ under the approximate posterior $q(\vartheta)$ is a measure of the accuracy we may expect under the current model.
- The divergence between the approximate posterior $q(\vartheta)$ and the prior $p(\vartheta|m)$ is a measure for how much the data y have forced the model to adapt. As such, it is a measure of **model complexity.**
- It is important to note that complexity cannot be assessed in the absence of data. Different data will lead to different complexity. One way to remind oneself of this is to think of model complexity as the **complexity of the data under the current model.**

The third decomposition of A_v

$$A_v = KL[q(\vartheta), p(\vartheta|m)] - \int q(\vartheta) \log p(y|\vartheta, m)\, \mathrm{d}\vartheta$$

$$= \text{Complexity} - \text{Accuracy}$$

- This decomposition illustrates why A_v is a good measure of model quality: a good model is one that makes good predictions.
- This means that inferences based on currently available data have to generalize to new data.
- There are two dangers to this: seeing patterns where there are none (i.e., too much complexity) and missing patterns (i.e., too little accuracy).
- A_v is a measure that balances these two opposing demands because it rewards accuracy while penalizing complexity.

The third decomposition of A_v

$$A_v = KL[q(\vartheta), p(\vartheta|m)] - \int q(\vartheta) \log p(y|\vartheta, m)\, \mathrm{d}\vartheta$$

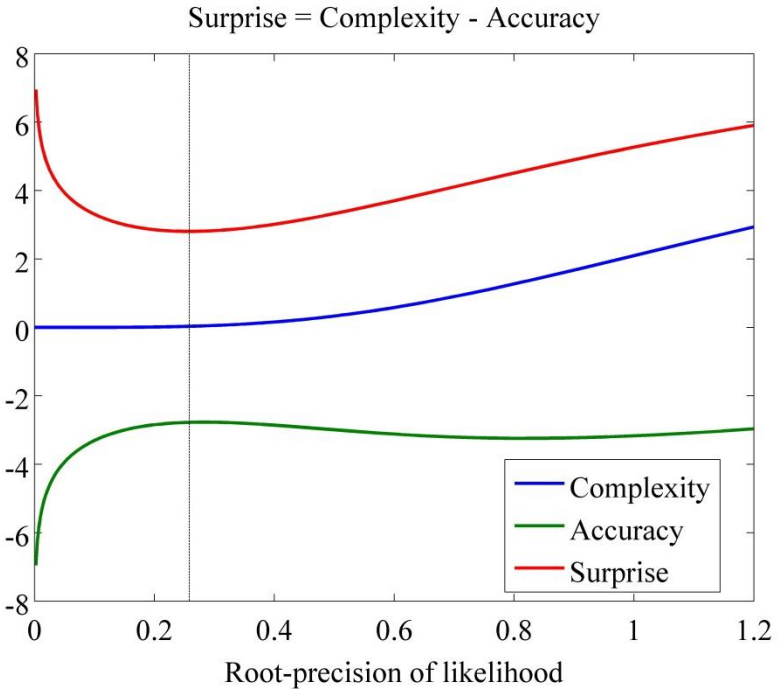
$$= \text{Complexity} - \text{Accuracy}$$

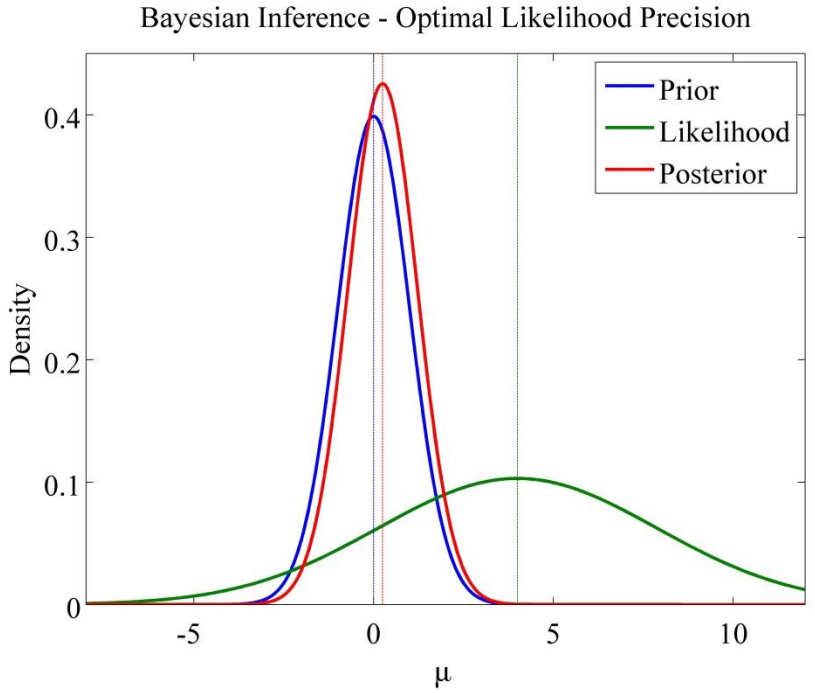
- The principled reason why A_v is a good measure of model quality is that the difference in A_v is an approximation to the log-Bayes factor.
- AIC (the Akaike Information Criterion) and BIC (the Bayesian Information Criterion) are approximations to A_v where the complexity term is replaced by a function of the number of parameters.

The third decomposition of A_v

$$A_v = KL[q(\vartheta), p(\vartheta|m)] - \int q(\vartheta) \log p(y|\vartheta, m)\, \mathrm{d}\vartheta$$

$$= \text{Complexity} - \text{Accuracy}$$





Bayesian inference

Movie!

Bayesian inference

Since variational free energy is a tool for Bayesian inference, it will be worth our while to look at Bayesian inference more deeply and to explore its connections with logic and with classical statistics.

«Bayesian» = logical and logical = probabilistic

«The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.»

— James Clerk Maxwell, 1850

«Bayesian» = logical and logical = probabilistic

But in what sense is probabilistic reasoning (i.e., reasoning about uncertain quantities according to the rules of probability theory) «logical»?

R. T. Cox showed in 1946 that the rules of probability theory can be derived from three basic desiderata:

1. Representation of degrees of plausibility by real numbers
2. Qualitative correspondence with common sense (in a well-defined sense)
3. Consistency

The rules of probability

By mathematical proof (i.e., by deductive reasoning) the three desiderata as set out by Cox imply the rules of probability (i.e., the rules of inductive reasoning).

This means that anyone who accepts the desiderata must accept the following rules:

1. $\sum_a p(a) = 1$ (Normalization)
2. $p(b) = \sum_a p(a, b)$ (Marginalization – also called the **sum rule**)
3. $p(a, b) = p(a|b)p(b) = p(b|a)p(a)$ (Conditioning – also called the **product rule**)

«Probability theory is nothing but common sense reduced to calculation.»

— Pierre-Simon Laplace, 1819

Conditional probabilities

- The probability of ***a* given *b*** is denoted by
- $p(a|b)$.
- In general, this is different from the probability of a alone (the *marginal* probability of a), as we can see by applying the sum and product rules:
- $p(a) = \sum_b p(a, b) = \sum_b p(a|b)p(b)$
- Because of the product rule, we also have the following rule (**Bayes' theorem**) for going from $p(a|b)$ to $p(b|a)$:
- $p(b|a) = \frac{p(a|b)p(b)}{p(a)} = \frac{p(a|b)p(b)}{\sum_{b\prime} p(a|b\prime)p(b\prime)}$

A simple example of Bayesian inference

(adapted from Jaynes (1976))

Two manufacturers, A and B, deliver the same kind of components that turn out to have the following lifetimes (in hours):

A:

59.5814
37.3953
47.5956
40.5607
48.6468
36.2789
31.5110
31.3606
45.6517

B:

48.8506
48.7296
59.1971
51.8895

Assuming prices are comparable, from which manufacturer would you buy?

A simple example of Bayesian inference

How do we compare such samples?

- By comparing their arithmetic means

Why do we take means?

- If we take the mean as our estimate, the error in our estimate is the mean of the errors in the individual measurements
- Taking the mean as maximum-likelihood estimate implies a **Gaussian error distribution**
- A Gaussian error distribution appropriately reflects our **prior** knowledge about the errors whenever we know nothing about them except perhaps their variance

A simple example of Bayesian inference

What next?

- Let's do a t-test (but first, let's compare variances with an F-test):

```
>> [fh,fp,fci,fstats] = vartest2(xa,xb)

fh =          fp =          fci =          fstats =

     0             0.3297       0.2415          fstat: 3.5114
                              19.0173            df1: 8
                                                 df2: 3
```

Variances not significantly different!

```
>> [h, p, ci, stats]= ttest2(xa,xb)

h =           p =           ci =           stats =

     0             0.0665     -21.0191          tstat: -2.0367
                               0.8151             df: 11
                                                  sd: 8.2541
```

Means not significantly different!

Is this satisfactory? No, so what can we learn by turning to probability theory (i.e., Bayesian inference)?

A simple example of Bayesian inference

The procedure in brief:

- Determine your question of interest («What is the probability that...?»)
- Specify your model (likelihood and prior)
- Calculate the full posterior using Bayes' theorem
- [Pass to the uninformative limit in the parameters of your prior]
- Integrate out any nuisance parameters
- Ask your question of interest of the posterior

All you need is the rules of probability theory.

(Ok, sometimes you'll encounter a nasty integral – but that's a technical difficulty, not a conceptual one).

A simple example of Bayesian inference

The question:

- What is the probability that the components from manufacturer B have a longer lifetime than those from manufacturer A?
- More specifically: given how much more expensive they are, how much longer do I require the components from B to live.
- Example of a decision rule: if the components from B live 3 hours longer than those from A with a probability of at least 80%, I will choose those from B.

A simple example of Bayesian inference

The model (bear with me, this **will** turn out to be simple):

- likelihood (Gaussian):

$$p(\{x_i\}|\mu,\lambda) = \prod_{i=1}^{n}\left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}}\exp\left(-\frac{\lambda}{2}(x_i-\mu)^2\right)$$

- prior (Gaussian-gamma):

$$p(\mu,\lambda|\mu_0,\kappa_0 a_0,b_0) = \mathcal{N}(\mu|\mu_0,(\kappa_0\lambda)^{-1})\mathrm{Gam}(\lambda|a_0,b_0)$$

A simple example of Bayesian inference

The posterior (Gaussian-gamma):

$$p(\mu, \lambda | \{x_i\}) = \mathcal{N}(\mu | \mu_n, (\kappa_n \lambda)^{-1}) \mathrm{Gam}(\lambda | a_n, b_n)$$

Parameter updates:

$$\mu_n = \mu_0 + \frac{n}{\kappa_0 + n}(\bar{x} - \mu_0), \qquad \kappa_n = \kappa_0 + n, \qquad a_n = a_0 + \frac{n}{2}$$

$$b_n = b_0 + \frac{n}{2}\left(s^2 + \frac{\kappa_0}{\kappa_0 + n}(\bar{x} - \mu_0)^2\right)$$

with

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i\,, \qquad s^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

A simple example of Bayesian inference

The limit for which the prior becomes uninformative:

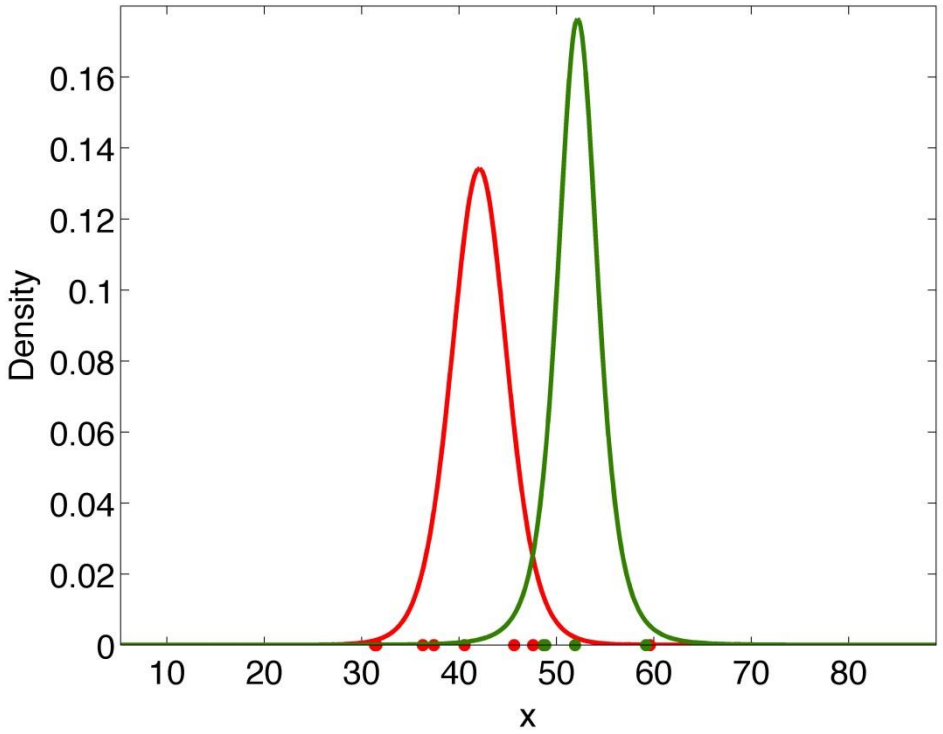
- For $\kappa_0 = 0, a_0 = 0, b_0 = 0$, the updates reduce to:

$$\mu_n = \bar{x} \qquad \kappa_n = n \qquad a_n = \frac{n}{2} \qquad b_n = \frac{n}{2} s^2$$

- As promised, this is really simple: **all you need is n, the number of datapoints; $\bar{x}$, their mean; and s^2, their variance.**
- This means that only the data influence the posterior and all influence from the parameters of the prior has been eliminated.
- The uninformative limit should only ever be taken **after** the calculation of the posterior using a proper prior.

A simple example of Bayesian inference

Integrating out the nuisance parameter λ gives rise to a t-distribution:



A simple example of Bayesian inference

The joint posterior $p(\mu_A, \mu_B | \{x_i\}_A, \{x_k\}_B)$ is simply the product of our two independent posteriors $p(\mu_A | \{x_i\}_A)$ and $p(\mu_B | \{x_k\}_B)$. It will now give us the answer to our question:

$$p(\mu_B - \mu_A > 3) = \int_{-\infty}^{\infty} \mathrm{d}\mu_A \, p(\mu_A | \{x_i\}_A) \int_{\mu_A+3}^{\infty} \mathrm{d}\mu_B \, p(\mu_B | \{x_k\}_B) = 0.9501$$

Note that the t-test told us that there was «no significant difference» even though there is a >95% probability that the parts from B will last at least 3 hours longer than those from A.

Bayesian inference

The procedure in brief:

- Determine your question of interest («What is the probability that...?»)
- Specify your model (likelihood and prior)
- Calculate the full posterior using Bayes' theorem
- [Pass to the uninformative limit in the parameters of your prior]
- Integrate out any nuisance parameters
- Ask your question of interest of the posterior

All you need is the rules of probability theory.

Variational Laplace

- **Variational Laplace** is a powerful implementation of Bayesian inference based on variational free energy.

- "Variational Laplace" is shorthand for "variational Bayes under the mean field approximation and the Laplace assumption".

- The **mean field approximation** is the assumption that the true posterior $p(\vartheta|y,m)$ can be approximated by an approximate posterior $q(\vartheta)$ that factorizes across subsets of ϑ:

$$p(\vartheta|y,m) \approx q(\vartheta) = q_1(\vartheta_1) \cdot q_2(\vartheta_2) \cdot \ \ldots \cdot q_n(\vartheta_n)$$

- The **Laplace assumption** is that the posterior is Gaussian. In particular, $q(\vartheta)$ will be Gaussian if each of the $q_i(\vartheta_i)$ is Gaussian.

Variational Laplace

- Reminder: in order to approximate the true posterior $p(\vartheta|y,m)$ and to minimize surprise, we need to find the $q(\vartheta)$ that minimizes variational free energy A_v.

- To find this optimal $q^*(\vartheta)$, we make use of **variational calculus**, a branch of mathematics that tells us how to take the derivative of a function of functions (usually, we deal with functions of variables that are numbers, not functions). At the minimum of A_v with respect to $q_i(\vartheta_i)$, we need this derivative to vanish:

$$\frac{\delta A_v}{\delta q_i}[q_i^*] = 0$$

- Solving this equation for q_i^*, we find

$$q_i^*(\vartheta_i) \propto \exp\big(I(\vartheta_i)\big)$$

$$I(\vartheta_i) := \int q_{\setminus i}^*\big(\vartheta_{\setminus i}\big) \ln p(y,\vartheta|m)\, \mathrm{d}\vartheta_{\setminus i}$$

Variational Laplace

- $I(\vartheta_i) \coloneqq \int q^*_{\setminus i}(\vartheta_{\setminus i}) \ln p(y, \vartheta|m)\, \mathrm{d}\vartheta_{\setminus i}$ is the **variational energy**.

- The notation $\setminus i$ means "not i" (e.g., $q^*_{\setminus i}(\vartheta_{\setminus i}) = \prod_{j \neq i} q^*_j(\vartheta_j)$).

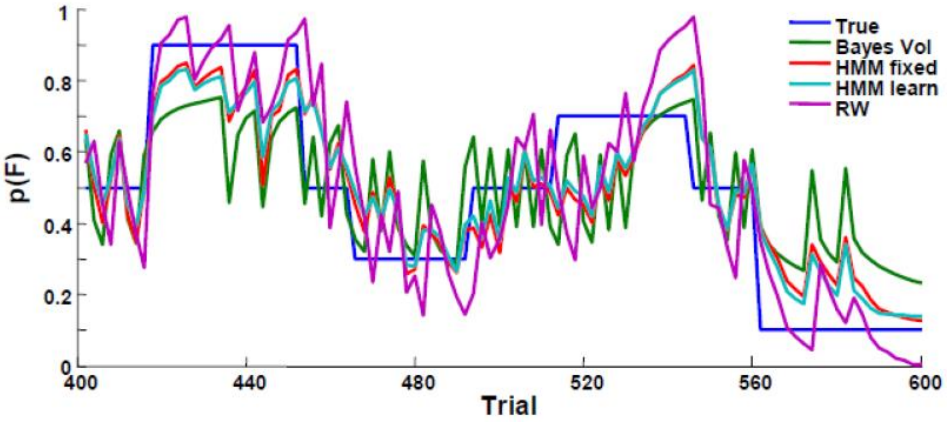
- Since $q^*_i(\vartheta_i) \propto \exp\big(I(\vartheta_i)\big)$ depends on all the other q^*_j with $j \neq i$ (which we don't know at the outset), we have to start with a reasonable guess for each of the q_i and keep updating them iteratively until we converge on q^*_i. This procedure is called **variational Bayes**.

- If we additionally constrain the q_i to be a Gaussian with its mean at the maximum of $I(\vartheta_i)$ and the negative Hessian of $I(\vartheta_i)$ as its precision, we have **variational Laplace**.

- This makes inference tractable even with complicated dynamic models and relevant prior information.

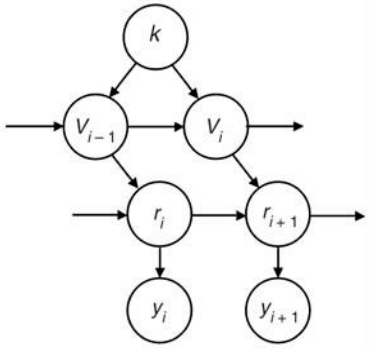
HGF: Context

- Hierarchical Bayesian models are natural candidates for explaining learning



den Ouden et al. (2010)

- However, their normative nature and computational cost pose problems

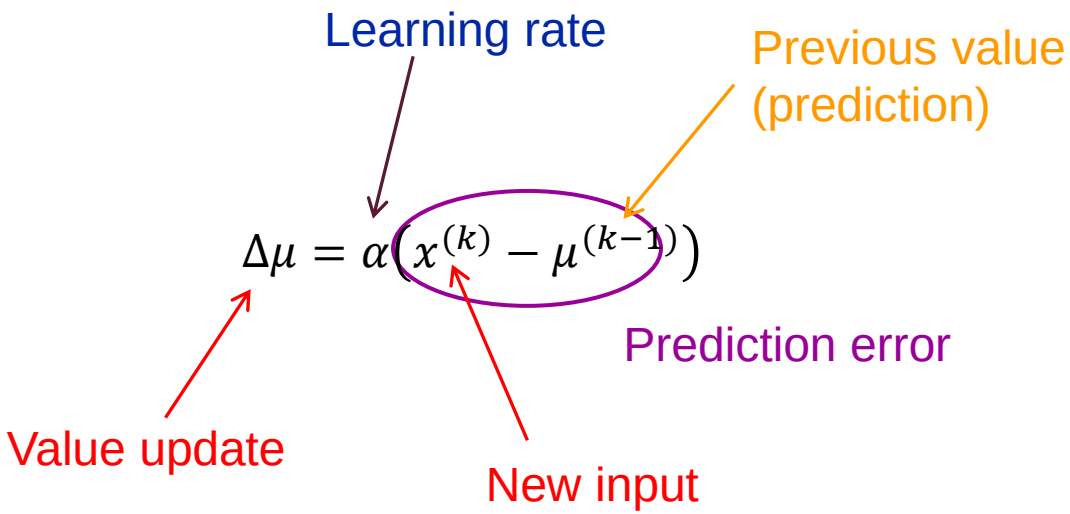


$$p(r_i, v_i, k \mid y_{\leq i}) \propto p(k) \int \cdots \int p(r_1) p(v_1) \prod_{j=1}^{i} [p(y_j \mid r_j) p(r_j \mid r_{j-1}, v_j) p(v_j \mid v_{j-1}, k)] dr_{\leq i-1} dv_{\leq i-1}$$

Behrens et al. (2007)

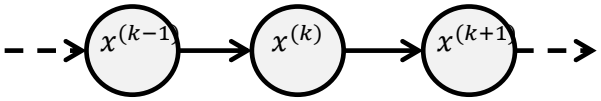
HGF: Context

Rescorla-Wagner learning:

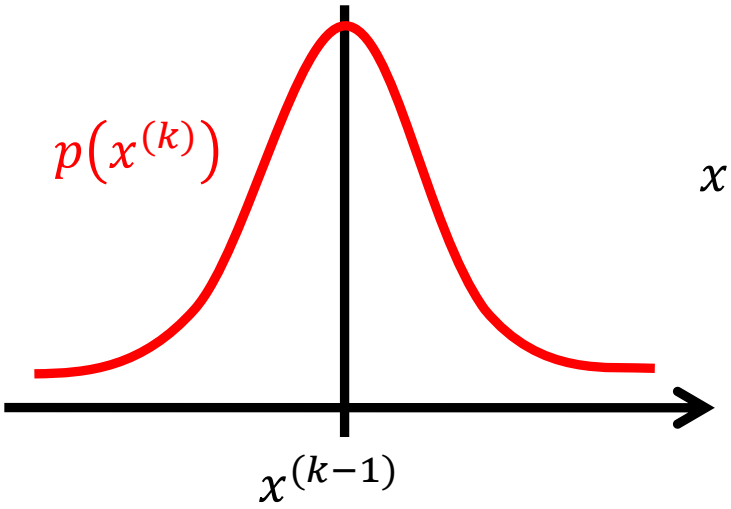


A generalized approach to learning

- A very general goal: to learn about a continuous quantity that changes



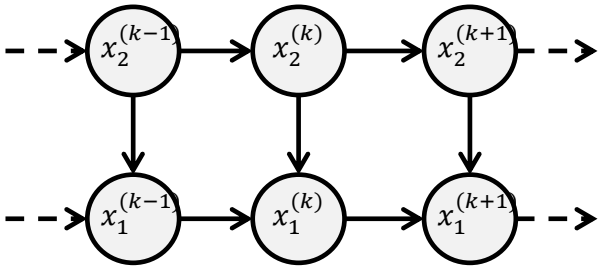
- Assumption: it performs a Gaussian random walk



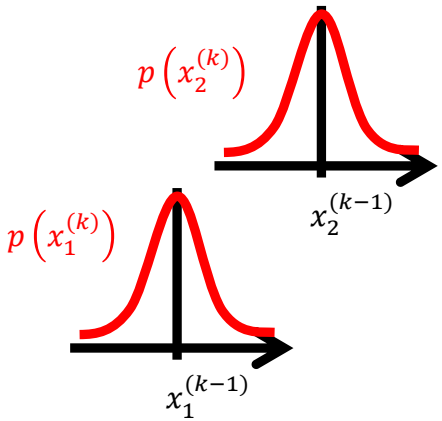
$$x^{(k)} \sim \mathcal{N}(x^{(k-1)}, \vartheta)$$

A generalized approach to learning

- To allow for changes in volatility, we take the variance of the random walk to be a positive function f of another state, x_2.



- We may then assume the volatility to perform its own Gaussian random walk.

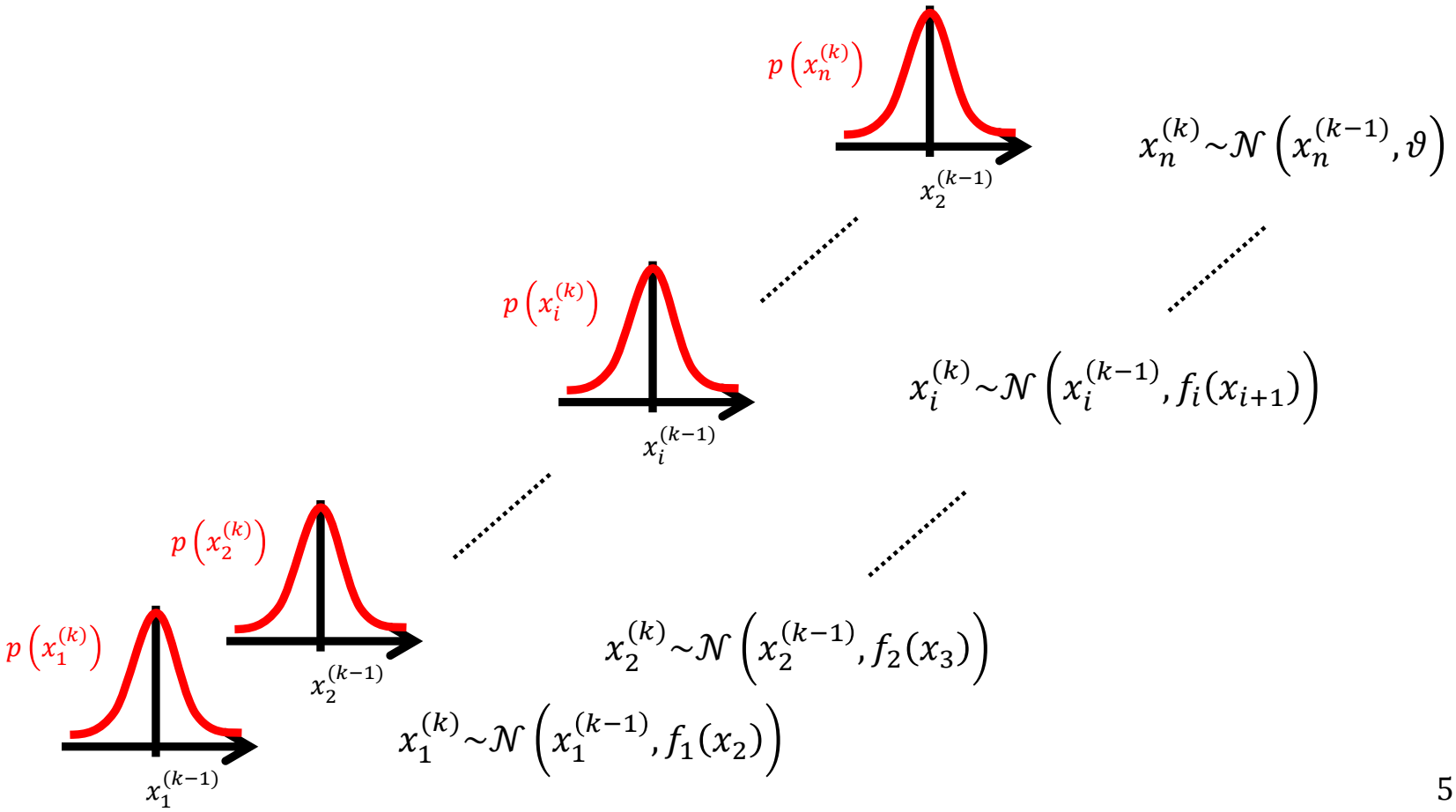


$$x_2^{(k)} \sim \mathcal{N}\left(x_2^{(k-1)}, \vartheta\right)$$

$$x_1^{(k)} \sim \mathcal{N}\left(x_1^{(k-1)}, f(x_2)\right)$$

A generalized approach to learning

This can be continued *ad infinitum*. In practice, we stop at some level n, where we assume the volatility to be constant.



Coupling between levels

Since f has to be everywhere positive, we cannot approximate it by expanding in powers. Instead, we expand its logarithm.

$$f(x) > 0 \; \forall \, x \implies \exists \, g \colon f(x) = \exp\big(g(x)\big) \; \forall \, x$$

$$g(x) = g(a) + g'(a) \cdot (x - a) + O(2) = \log f(x) =$$

$$= \log f(a) + \frac{f'(a)}{f(a)} \cdot (x - a) + O(2) =$$

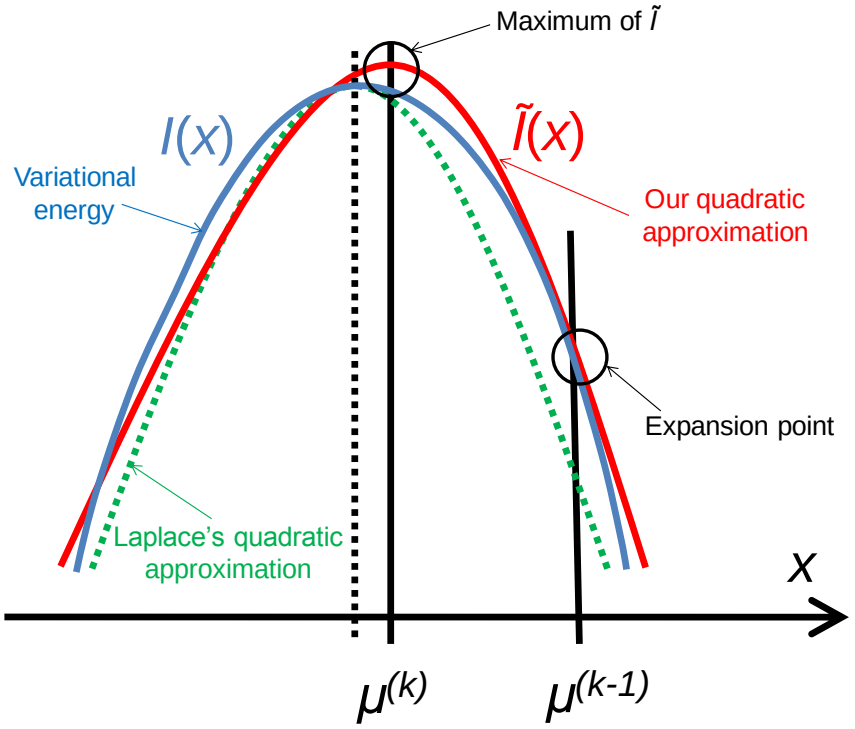
$$= \underbrace{\frac{f'(a)}{f(a)}}_{\stackrel{\text{def}}{=} \kappa} \cdot x + \underbrace{\log f(a) - a \cdot \frac{f'(a)}{f(a)}}_{\stackrel{\text{def}}{=} \omega} + O(2) =$$

$$= \kappa x + \omega + O(2)$$

$$\implies f(x) \approx \exp(\kappa x + \omega)$$

Variational inversion

- A quadratic approximation is found by expanding to second order about the expectation $\mu^{(k-1)}$.
- The update in the sufficient statistics of the approximate posterior is then performed by analytically finding the maximum of the quadratic approximation.



$$\sigma_i^{(k)} = -\frac{1}{\partial^2 I\left(\mu_i^{(k-1)}\right)}$$

$$\mu_i^{(k)} = \mu_i^{(k-1)} - \frac{\partial I\left(\mu_i^{(k-1)}\right)}{\partial^2 I\left(\mu_i^{(k-1)}\right)} = \mu_i^{(k-1)} + \sigma_i^{(k)} \partial I\left(\mu_i^{(k-1)}\right)$$

Variational inversion and update equations

- Inversion proceeds by introducing a mean field approximation and fitting quadratic approximations to the resulting variational energies (Mathys et al., 2011).
- This leads to **simple one-step update equations** for the sufficient statistics (mean and precision) of the approximate Gaussian posteriors of the states x_i.
- The updates of the means have the same structure as value updates in Rescorla-Wagner learning:

$$\Delta\mu_i \propto \frac{\hat{\pi}_{i-1}}{\pi_i}\delta_{i-1}$$

Prediction error

Precisions determine learning rate

- Furthermore, the updates are **precision-weighted prediction errors**.

Precision-weighting of updates

- Updates are weighted by belief precisions.
- To see this, first consider a simple non-hierarchical model with one parameter ϑ.
- Likelihood and prior are Gaussian, therefore the posterior also:

$$p(\vartheta) = \mathcal{N}(\vartheta; \mu_\vartheta, \pi_\vartheta) \quad \textbf{Prior}$$

$$p(y|\vartheta) = \mathcal{N}(y; \vartheta, \pi_\varepsilon) \quad \textbf{Likelihood}$$

$$\Longrightarrow \quad p(\vartheta|y) = \mathcal{N}\left(\vartheta; \mu_{\vartheta|y}, \pi_{\vartheta|y}\right) \quad \textbf{Posterior}$$

- The exact Bayesian update then is the precision-weighted prediction error:

$$\pi_{\vartheta|y} = \pi_\vartheta + \pi_\varepsilon \qquad \mu_{\vartheta|y} = \mu_\vartheta + \frac{\pi_\varepsilon}{\pi_{\vartheta|y}}(y - \mu_\vartheta)$$

Precision-weighted prediction error

Precision-weighting of updates

Comparison to the simple non-hierarchical Bayesian update:

HGF:

$$\mu_i^{(k)} = \mu_i^{(k-1)} + \frac{1}{2}\kappa_{i-1}\, v_{i-1}^{(k)} \cdot \frac{\hat{\pi}_{i-1}^{(k)}}{\pi_i^{(k)}} \cdot \delta_{i-1}^{(k)}$$

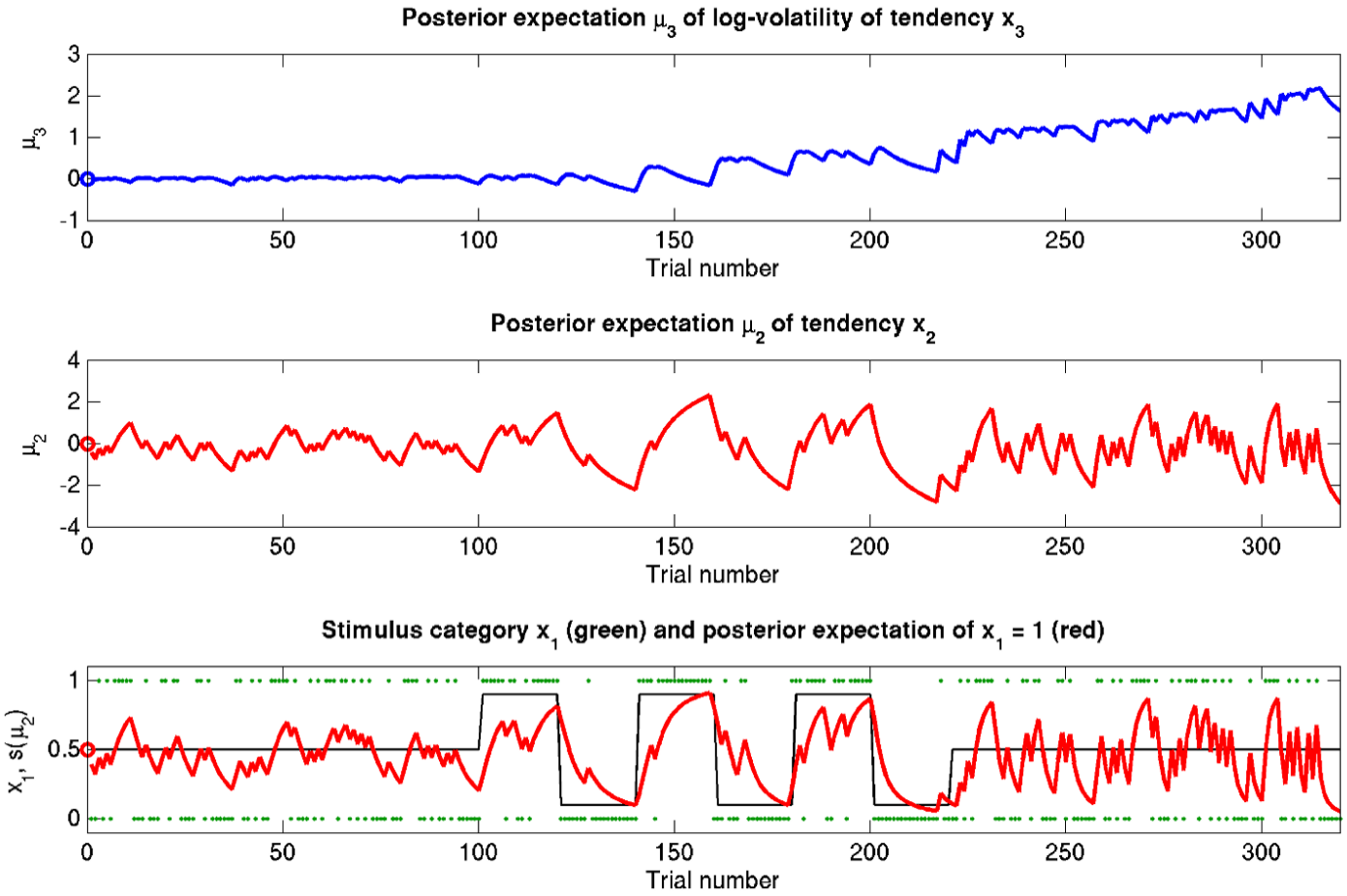
Precision-weighted prediction error

Simple Gaussian:

$$\mu_{\vartheta|y} = \mu_\vartheta + \frac{\pi_\varepsilon}{\pi_{\vartheta|y}}(y - \mu_\vartheta)$$

Context effects on the learning rate

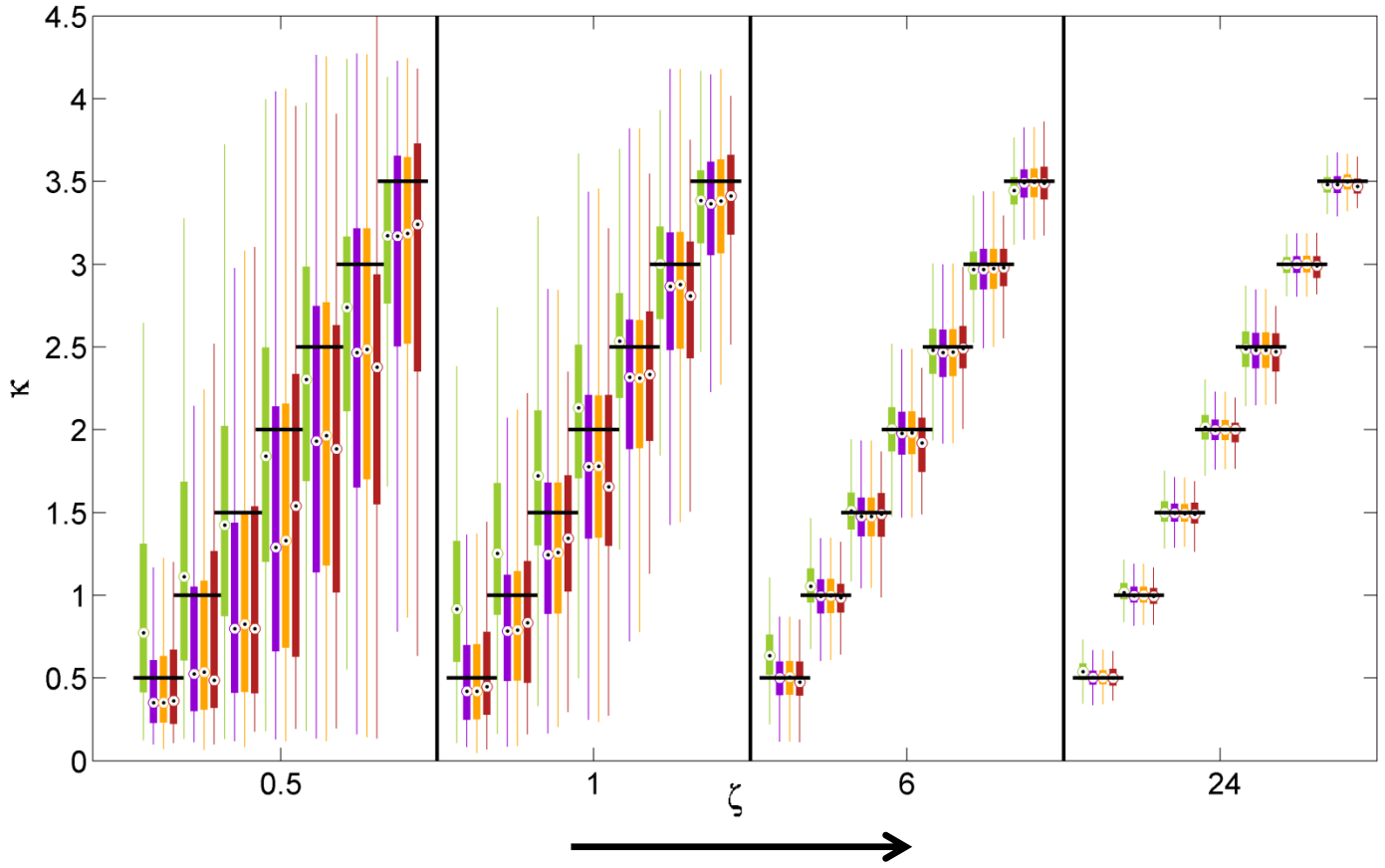
Simulation: $\vartheta = 0.5,\ \omega = -2.2,\ \kappa = 1.4$



Parameter estimation

4 estimation methods:

NMSA
GPGO
VB
MCMC

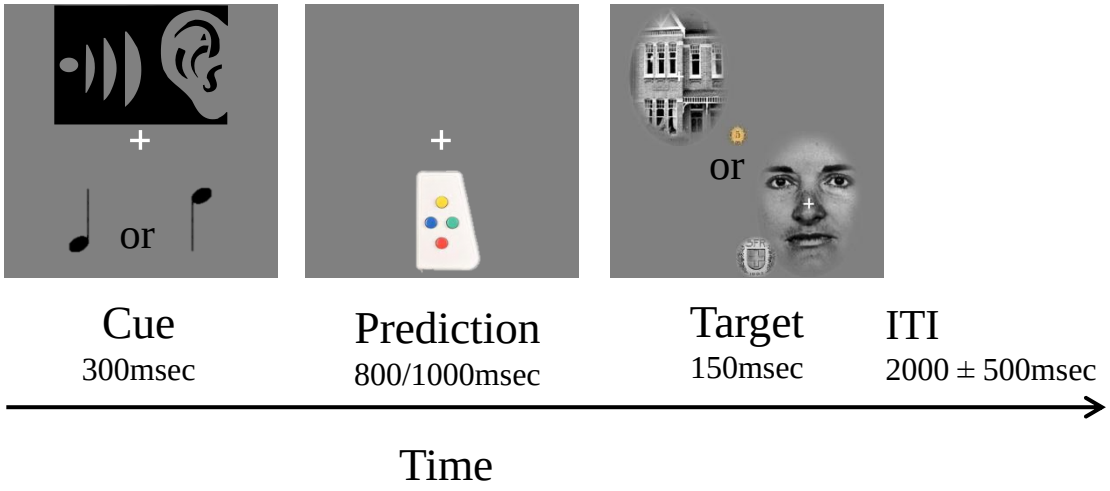


Decreasing noise

Practical uses

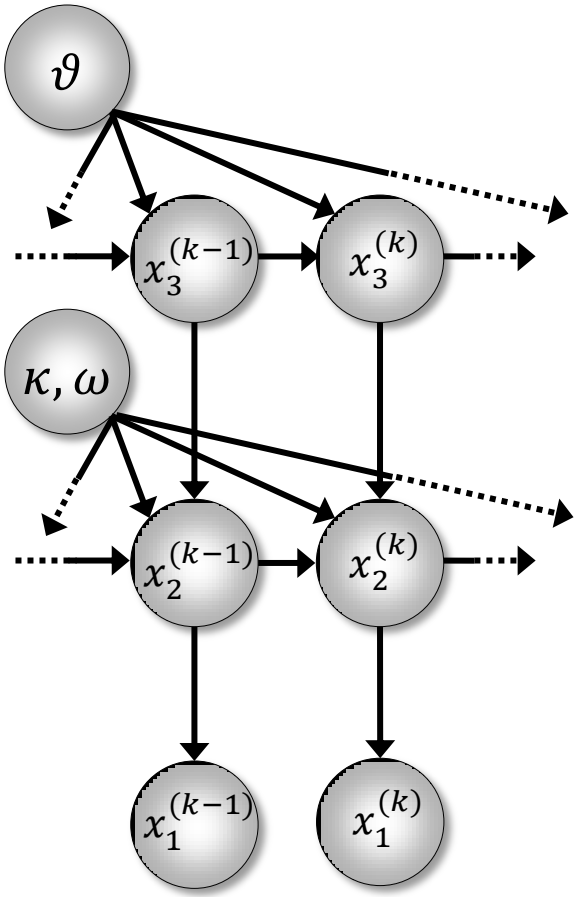
- Updates have a **general** and **interpretable** structure.
- They are **computationally** extremely **efficient.**
- They contain **parameters** that can differ from subject to subject and can be **individually estimated** from experimental data.
- This enables the **comparison of parameter estimates** between subjects and of **evolving beliefs on states** within subjects.
- Furthermore, it provides a basis for **model selection** on the basis of log-model evidence (e.g., comparison of **learning models** with different hierarchical depths, comparison of **decision models**).

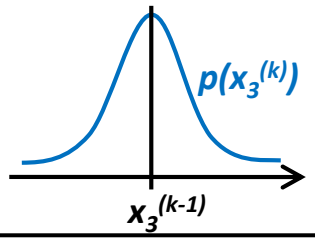
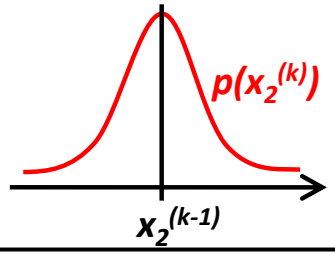
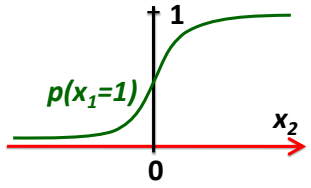
Associative learning task (Iglesias et al., 2013)



- fMRI
- 10 blocks of changing association strength:
  0.1 / 0.3 / 0.5 / 0.7 / 0.9
- 320 trials + 64 null events

Application to binary data



State of the world	Model	
Log-volatility x_3 of tendency	Gaussian random walk with constant step size ϑ	$p(x_3^{(k)}) \sim N(x_3^{(k-1)}, \vartheta)$ 
Tendency x_2 towards category "1"	Gaussian random walk with step size $\exp(\kappa x_3+\omega)$	$p(x_2^{(k)}) \sim N(x_2^{(k-1)}, \exp(\kappa x_3+\omega))$ 
Stimulus category x_1 ("0" or "1")	Sigmoid trans-formation of x_2	$p(x_1=1) = s(x_2)$ $p(x_1=0) = 1-s(x_2)$ 

Update equation for binary observations

- $x_1 \in \{0,1\}$ is observed by the agent. Each observation leads to an update in the belief on x_2, x_3, ..., and so on up the hierarchy.
- The updates for x_2 can be derived in the same manner as above.

$$I\left(x_2^{(k)}\right) = \ln s\left(x_2^{(k)}\right) + x_2^{(k)}\left(x_1^{(k)} - 1\right) - \frac{1}{2}\hat{\pi}_2^{(k)}\left(x_2^{(k)} - \mu_2^{(k-1)}\right)^2$$

$$\mu_2^{(k)} = \mu_2^{(k-1)} + \sigma_2^{(k)}\delta_1^{(k)}$$

- At first, this simply looks like an uncertainty-weighted update. However, when we unpack σ_2 and do a Taylor expansion in powers of $\hat{\pi}_1$, we see that it is again proportional to the precision of the prediction on the level below:

$$\sigma_2^{(k)} = \frac{\hat{\pi}_1^{(k)}}{\hat{\pi}_2^{(k)}\hat{\pi}_1^{(k)} + 1} = \hat{\pi}_1^{(k)} - \hat{\pi}_2^{(k)}\left(\hat{\pi}_1^{(k)}\right)^2 + \left(\hat{\pi}_2^{(k)}\right)^2\left(\hat{\pi}_1^{(k)}\right)^3 + O(4)$$

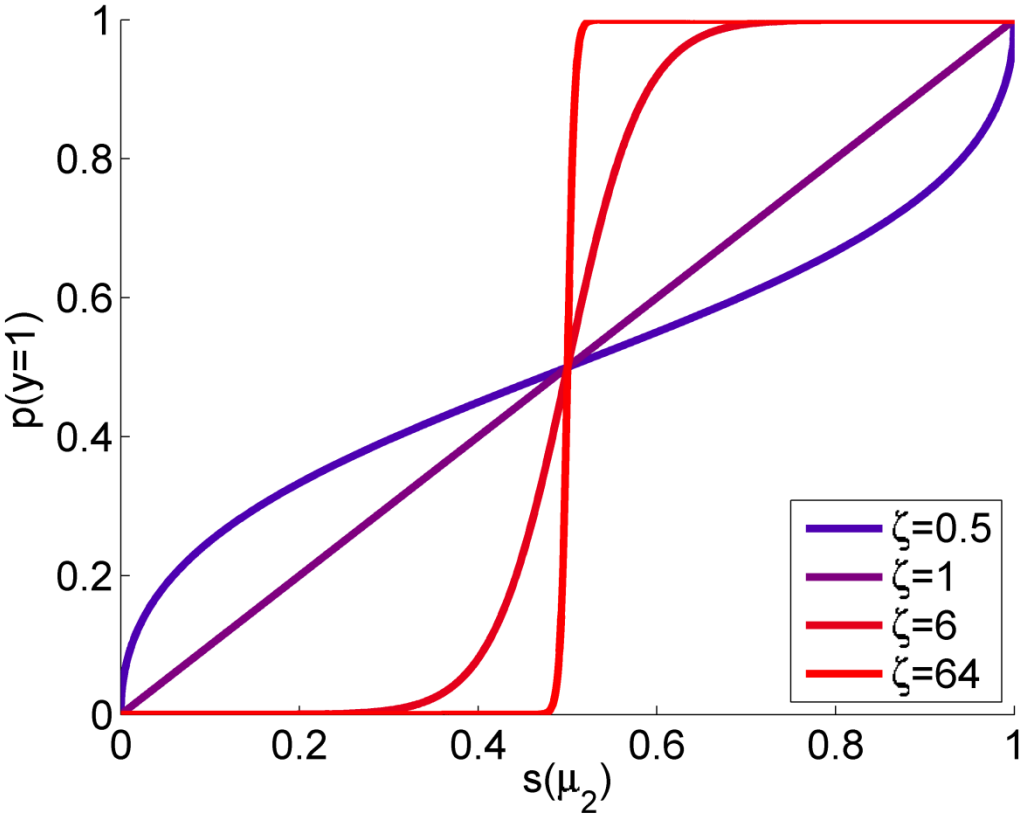
- At all higher levels, the updates are as previously derived.

Decision model

- Softmax decision rule
- Curve shape is determined by the parameter ζ
- Translates beliefs into decision probabilities

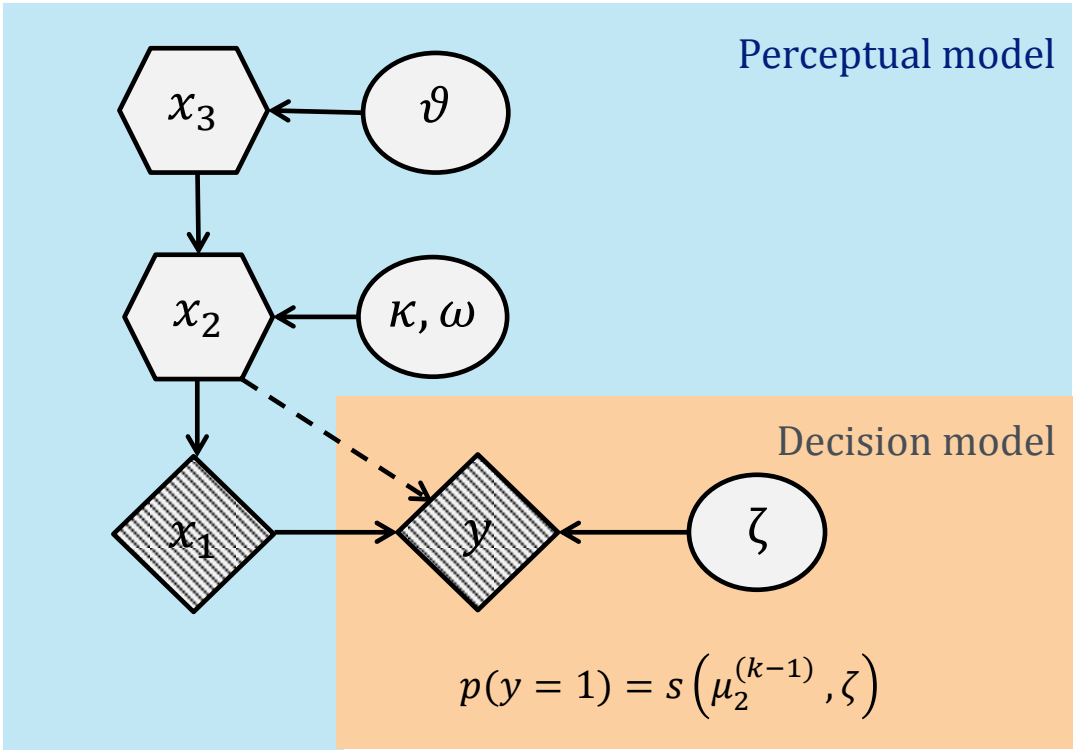
Probability of decision "1",

(i.e., of betting on "1")



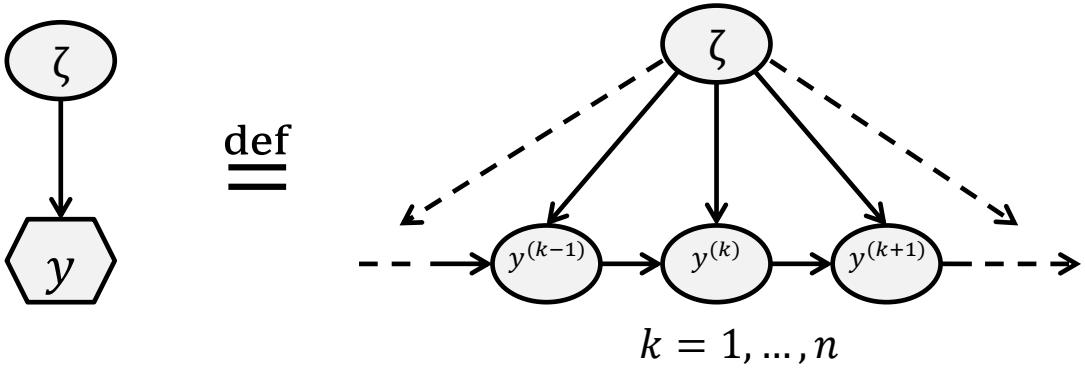
Prediction that next stimulus is "1"

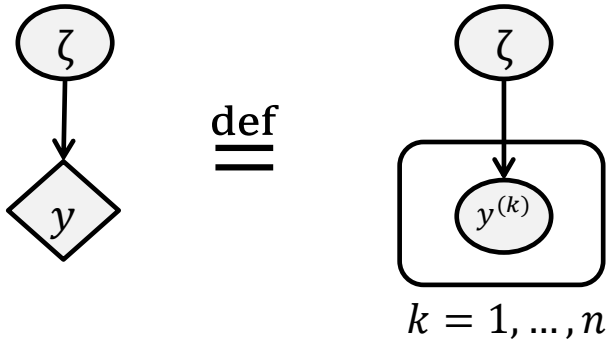
Taking it all together: perception and decision

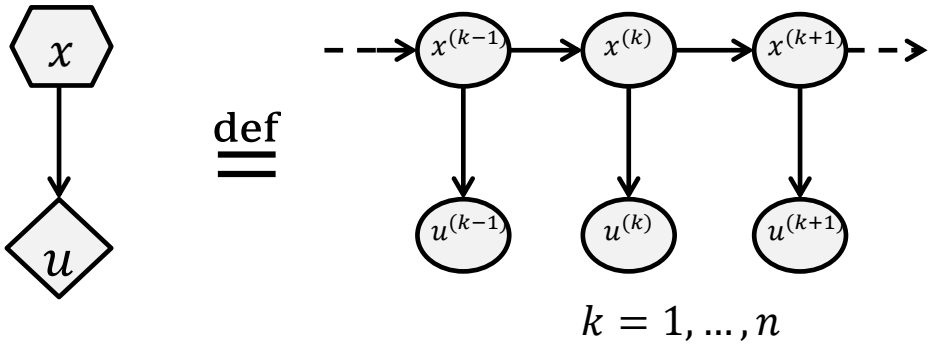


cf. Daunizeau et al. (2010a,b)

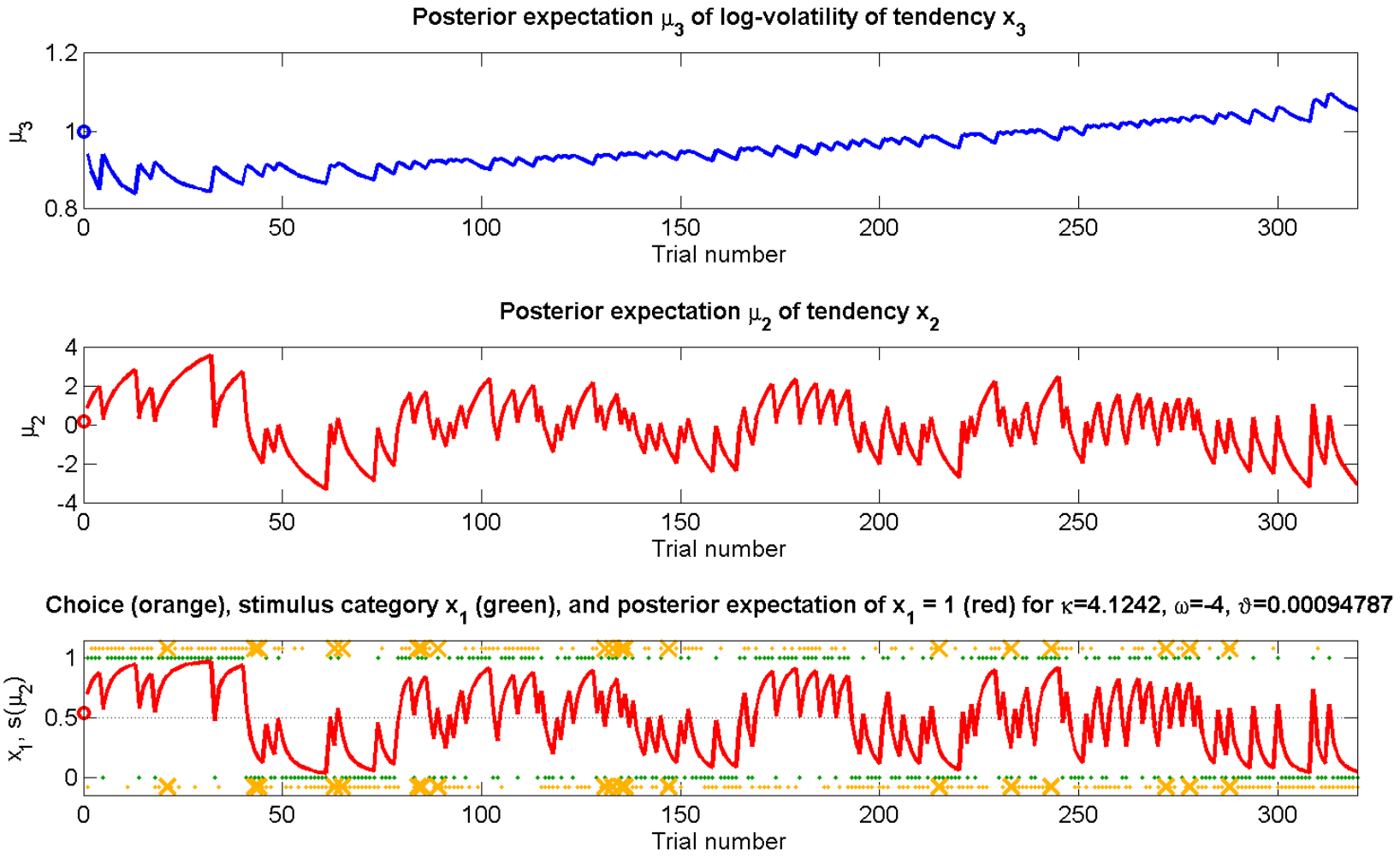
Taking it all together: notation



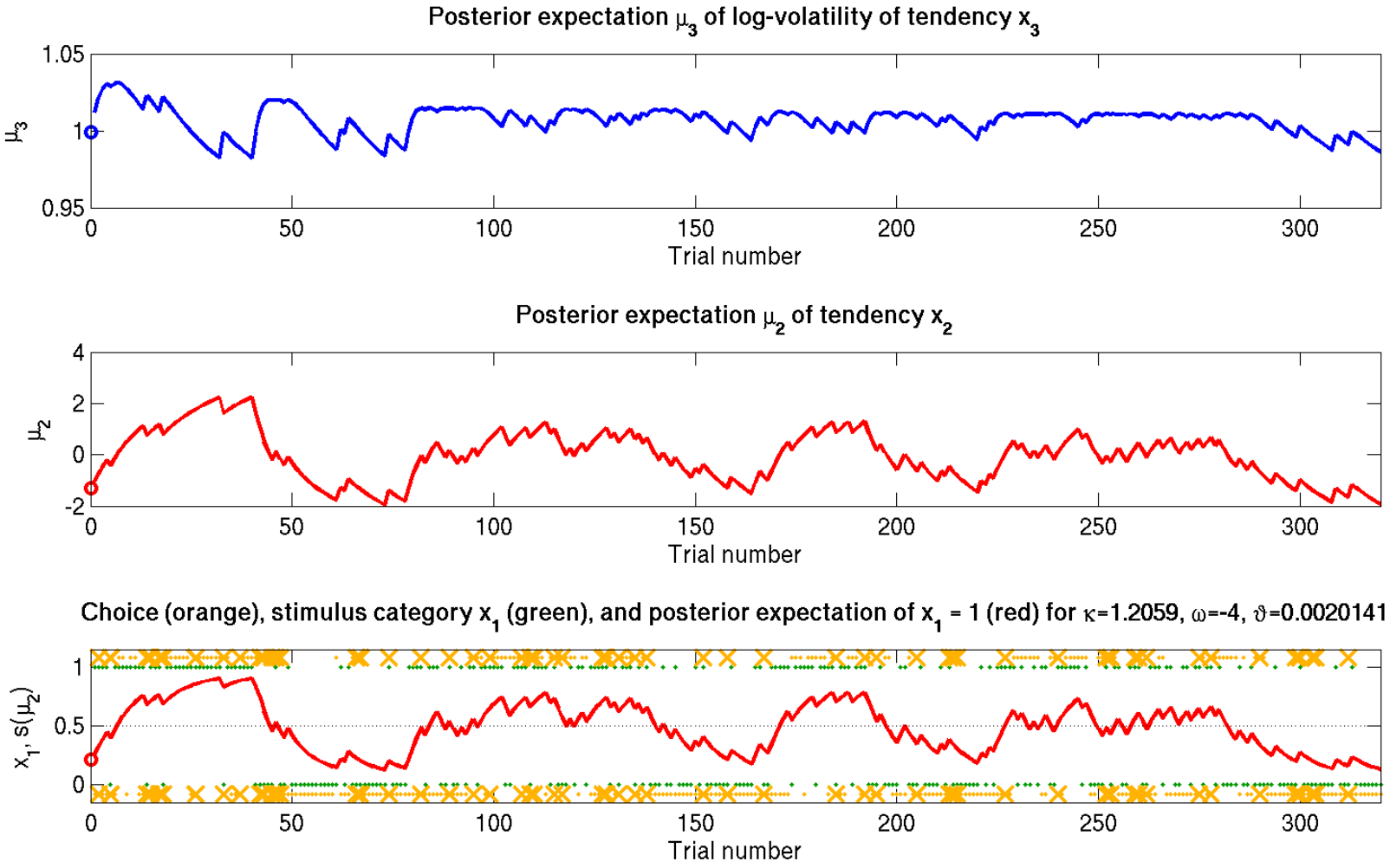




Individual belief trajectories

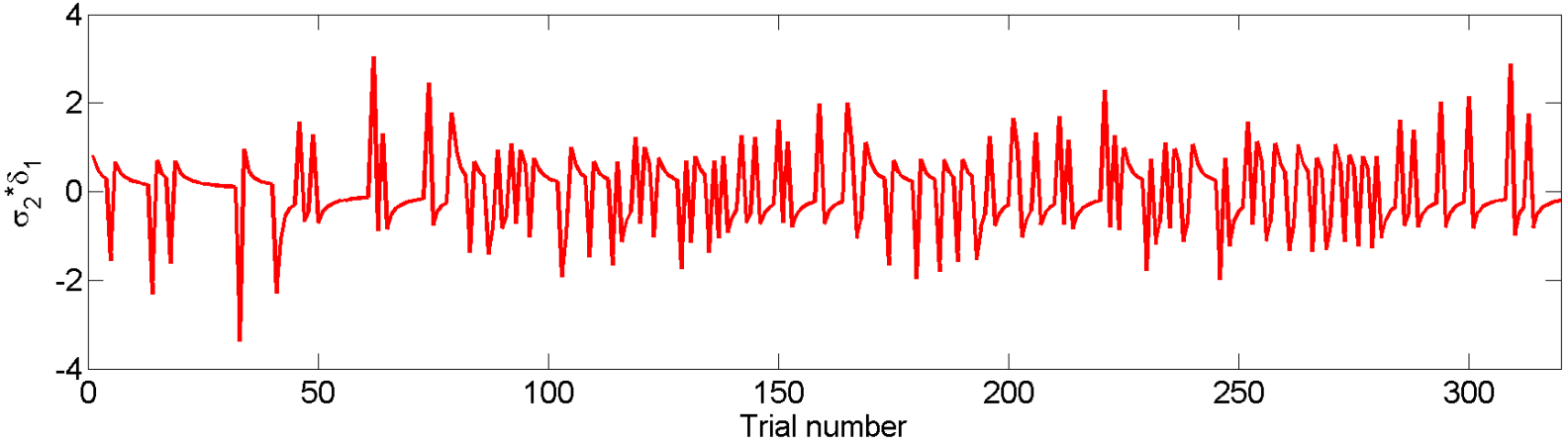


Individual belief trajectories



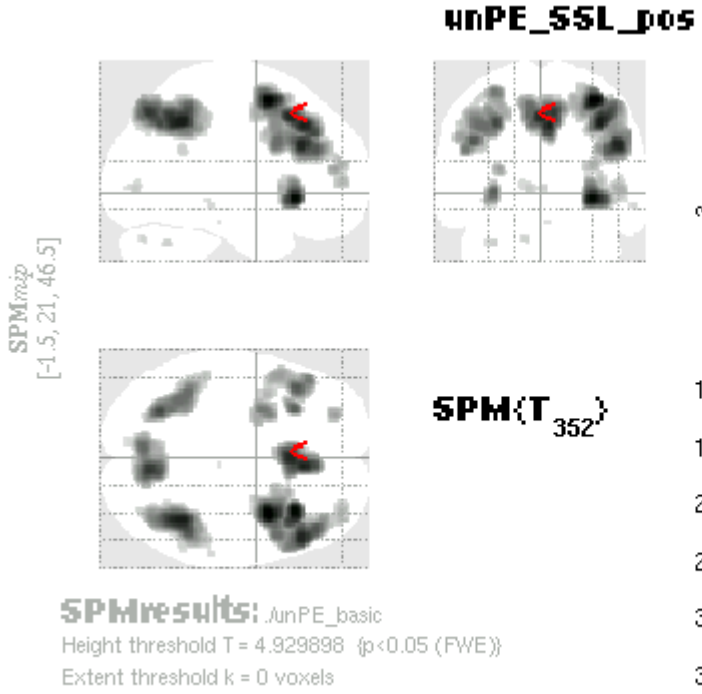
Individual regressors

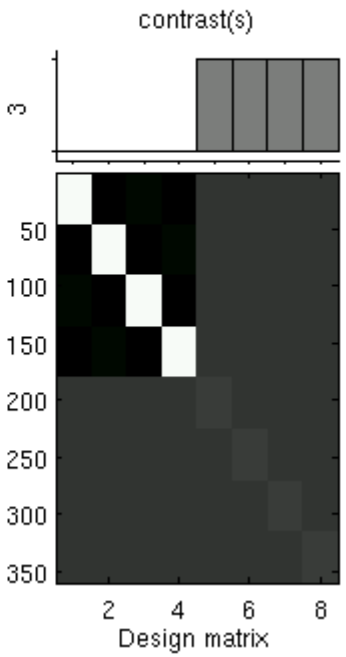
Uncertainty-weighted prediction error $\sigma_2 \cdot \delta_1$

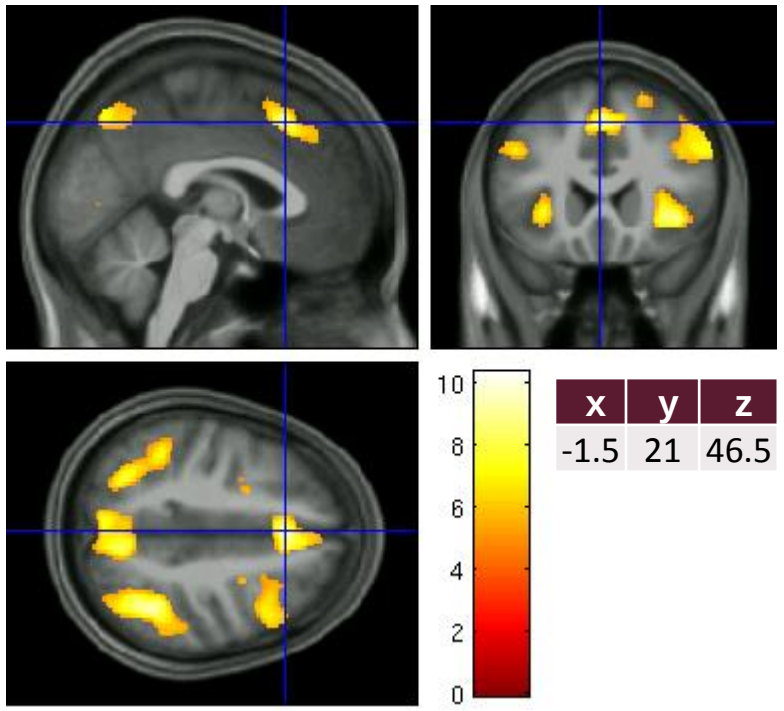


$$\varepsilon_2 = \sigma_2^{(k)} \delta_1^{(k)}$$

positive correlation





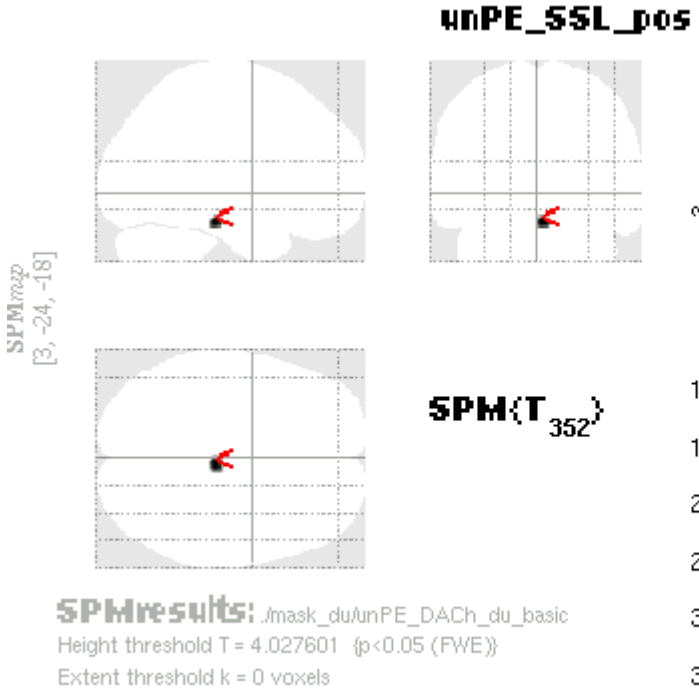


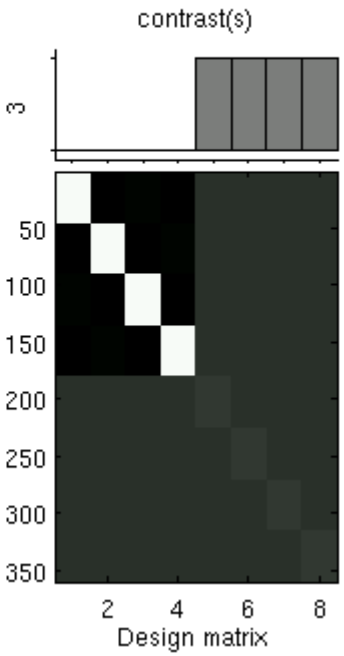
x	y	z
-1.5	21	46.5

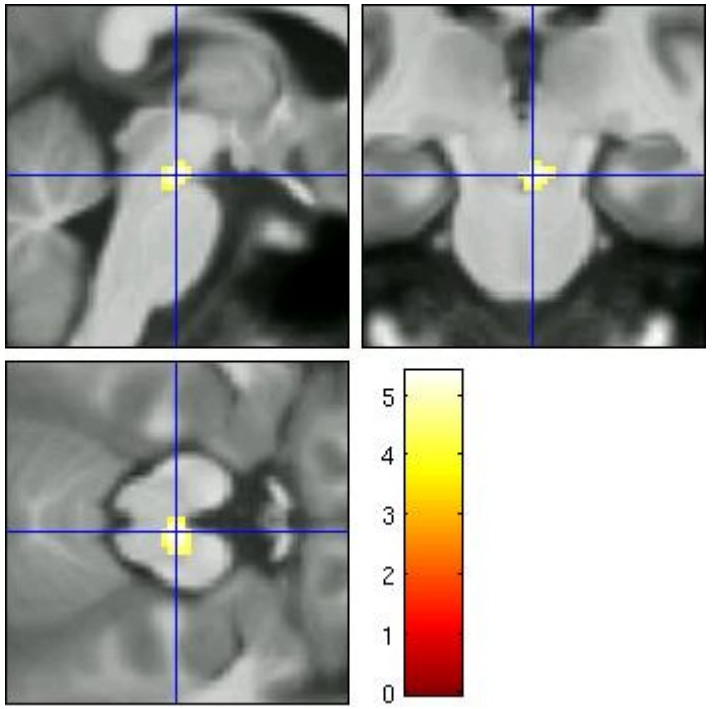
p < 0.05 FWE whole-brain corrected

$$\varepsilon_2 = \sigma_2^{(k)} \delta_1^{(k)}$$

positive correlation

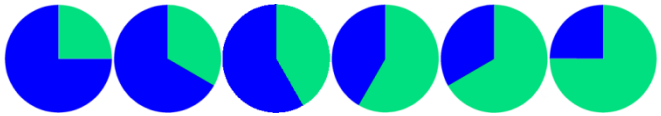






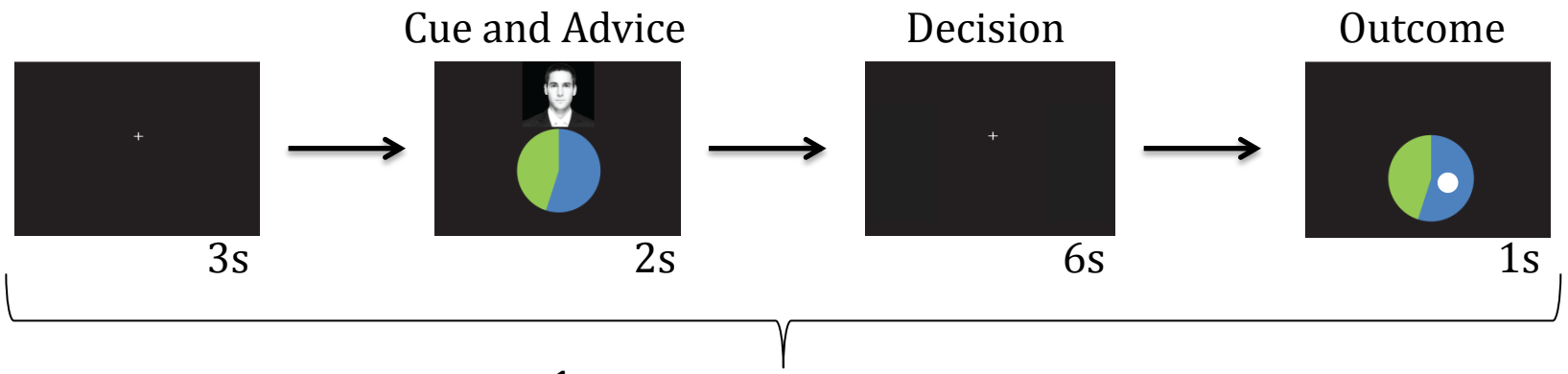
p < 0.05 FWE mask

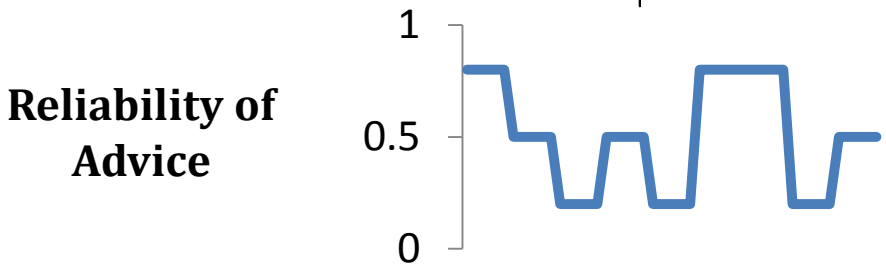
Social learning (Diaconescu et al., in prep.)

Cue: 

Advice: Video of adviser holding up blue or green card

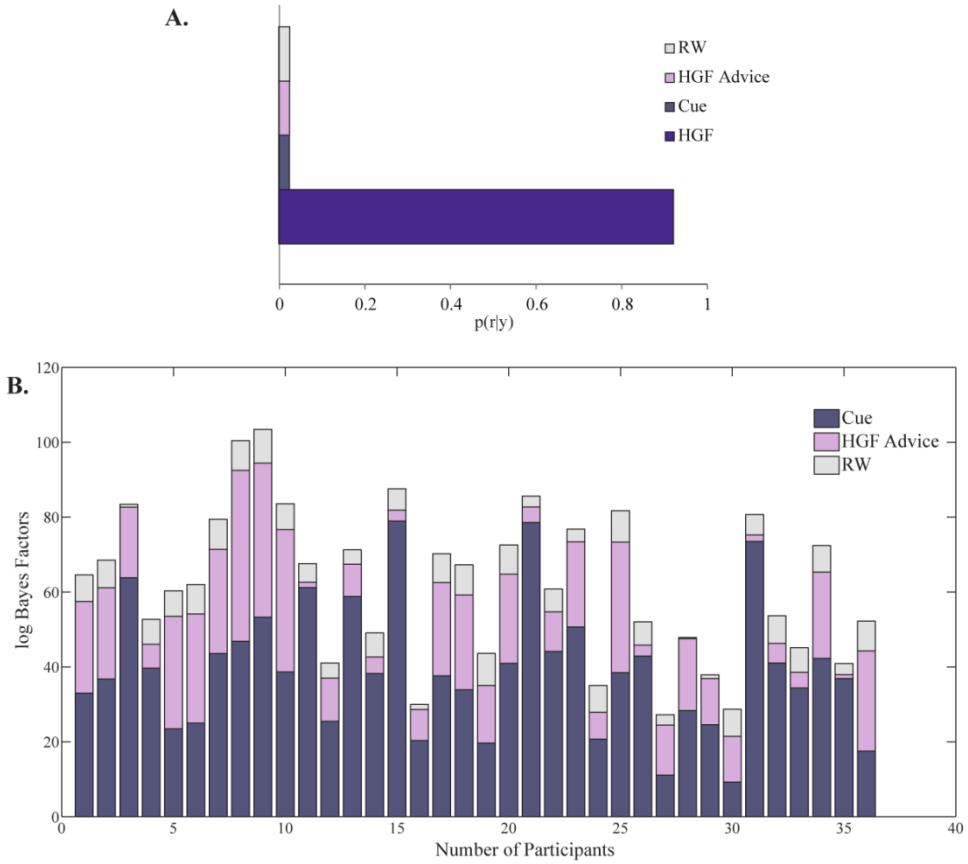
Subjects must compute: $p(\text{blue}|\ \text{cue} = 65\%\ \text{blue}, \text{advice} = \text{green}, \text{history of advice})$





Social learning (Diaconescu et al., in prep.)

Bayesian model selection:

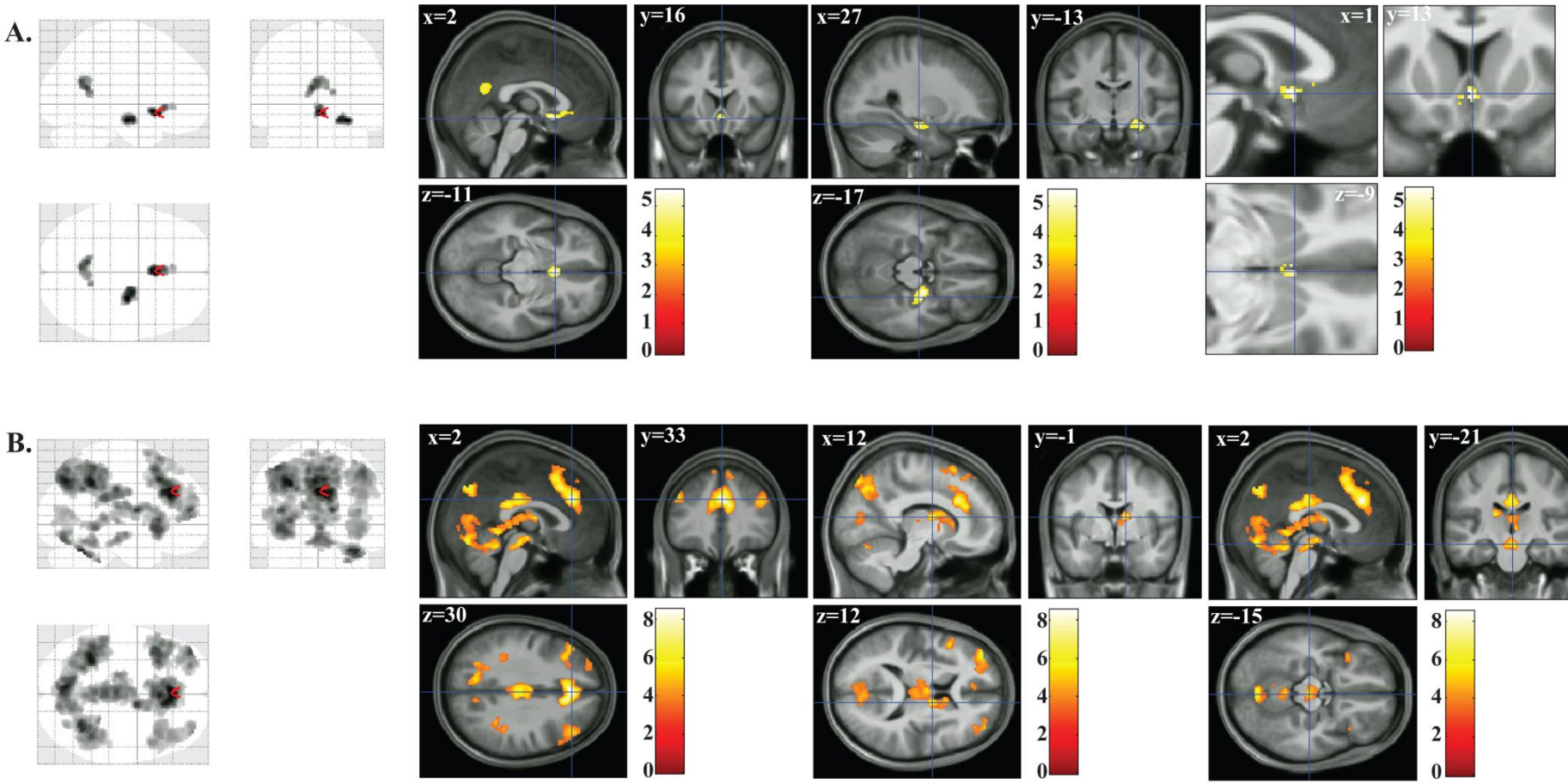


⟹ Use of volatility estimates in adjusting learning

⟹ Combination of both social and non-social cues in decision-making

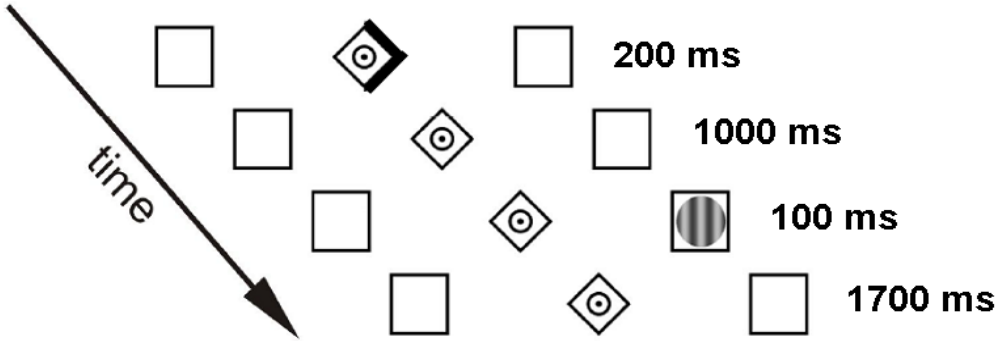
Integrated belief: positive and negative correlations

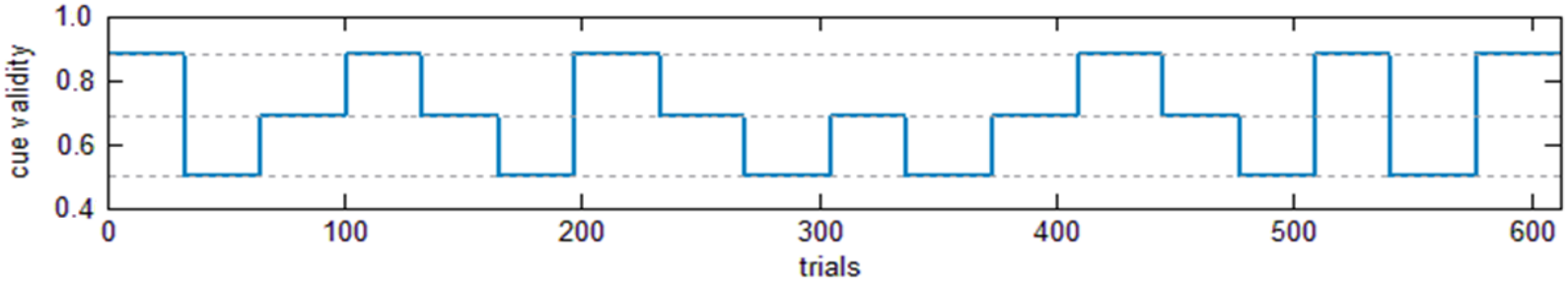
$$b = \zeta_1 \hat{\mu}_1 + (1 - \zeta_1)\tilde{c}$$



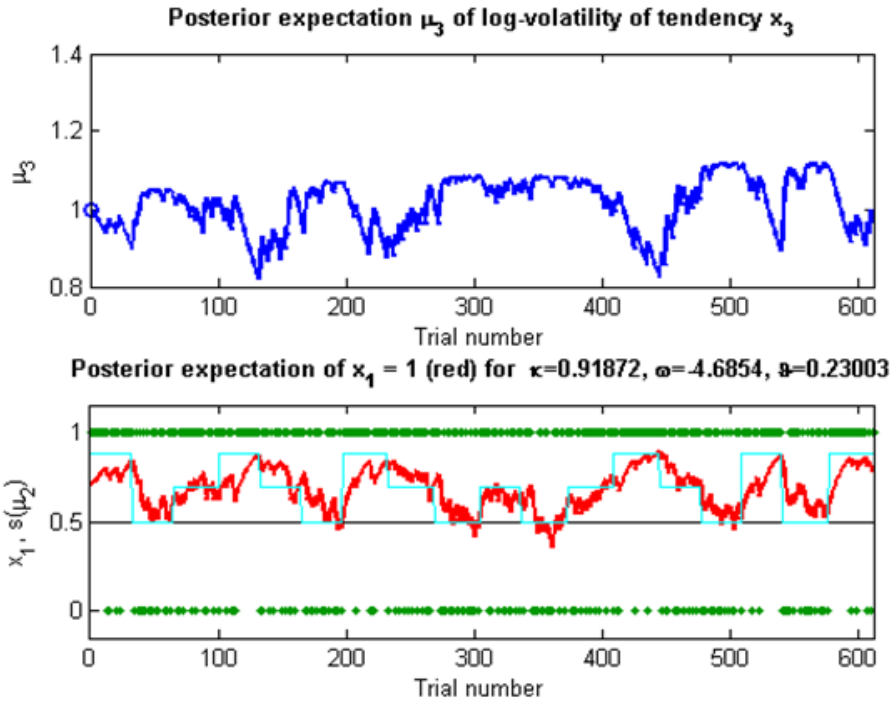
Posner task (Vossel & Mathys et al., 2014)

Measurement of saccadic reaction times in volatile Posner task

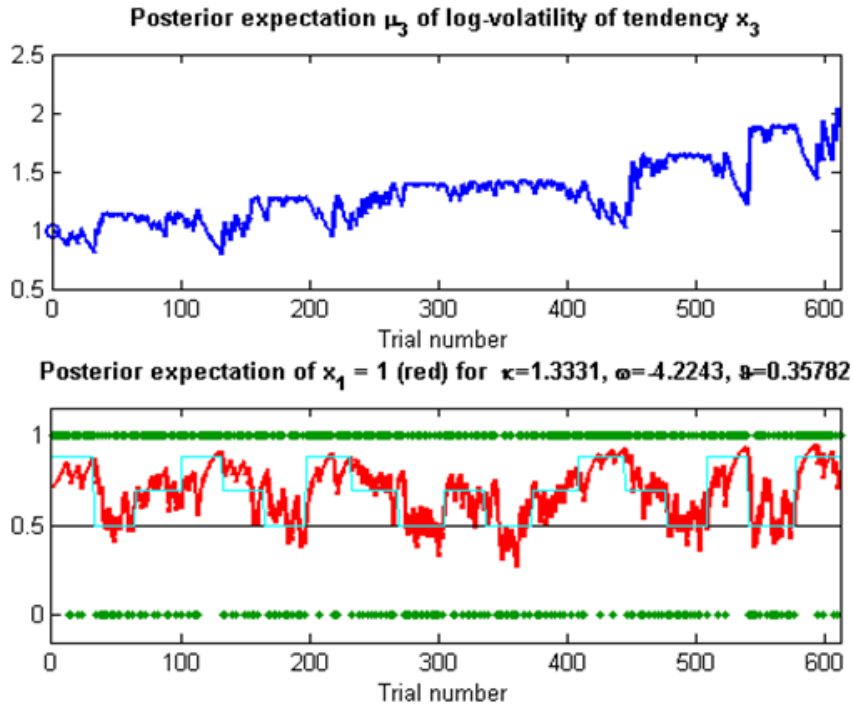




Posner task (Vossel & Mathys et al., 2014)



ω=-5.84; ϑ=0.02



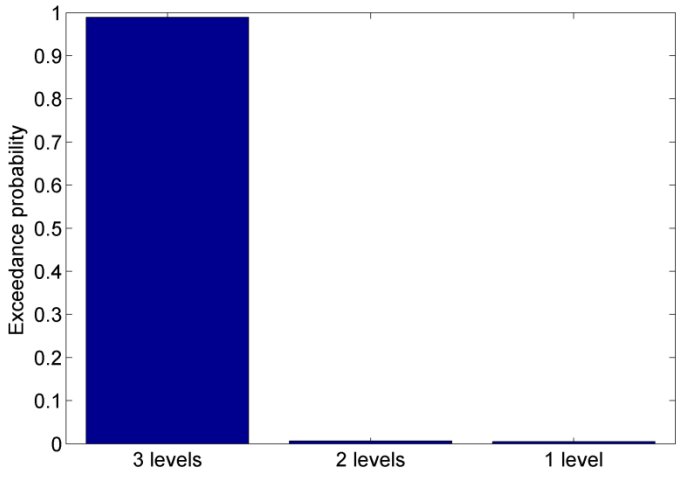
ω=-4.5; ϑ=0.98

Posner task (Vossel & Mathys et al., 2014)

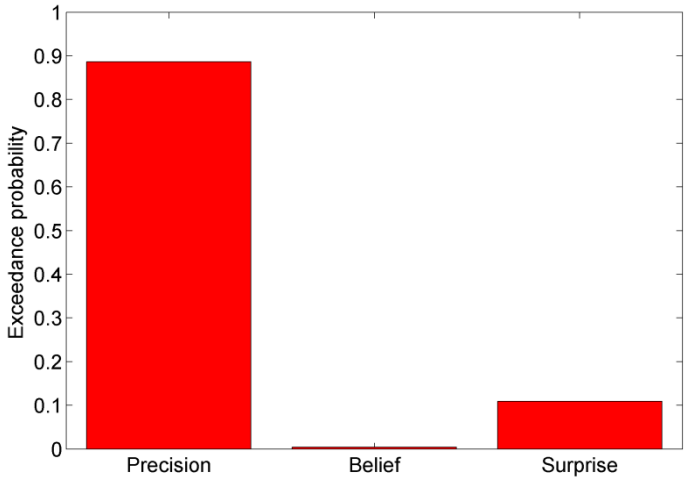
Decision

Learning	Precision	Belief	Surprise
1 level			
2 levels			
3 levels			

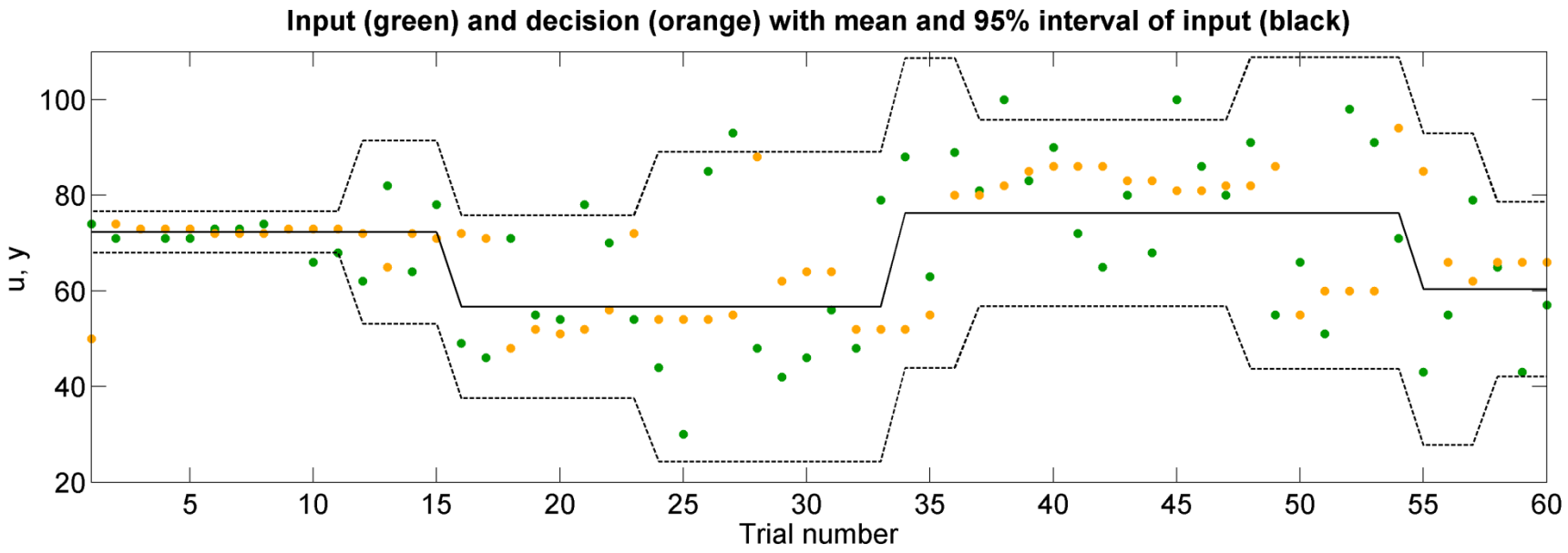
Learning models



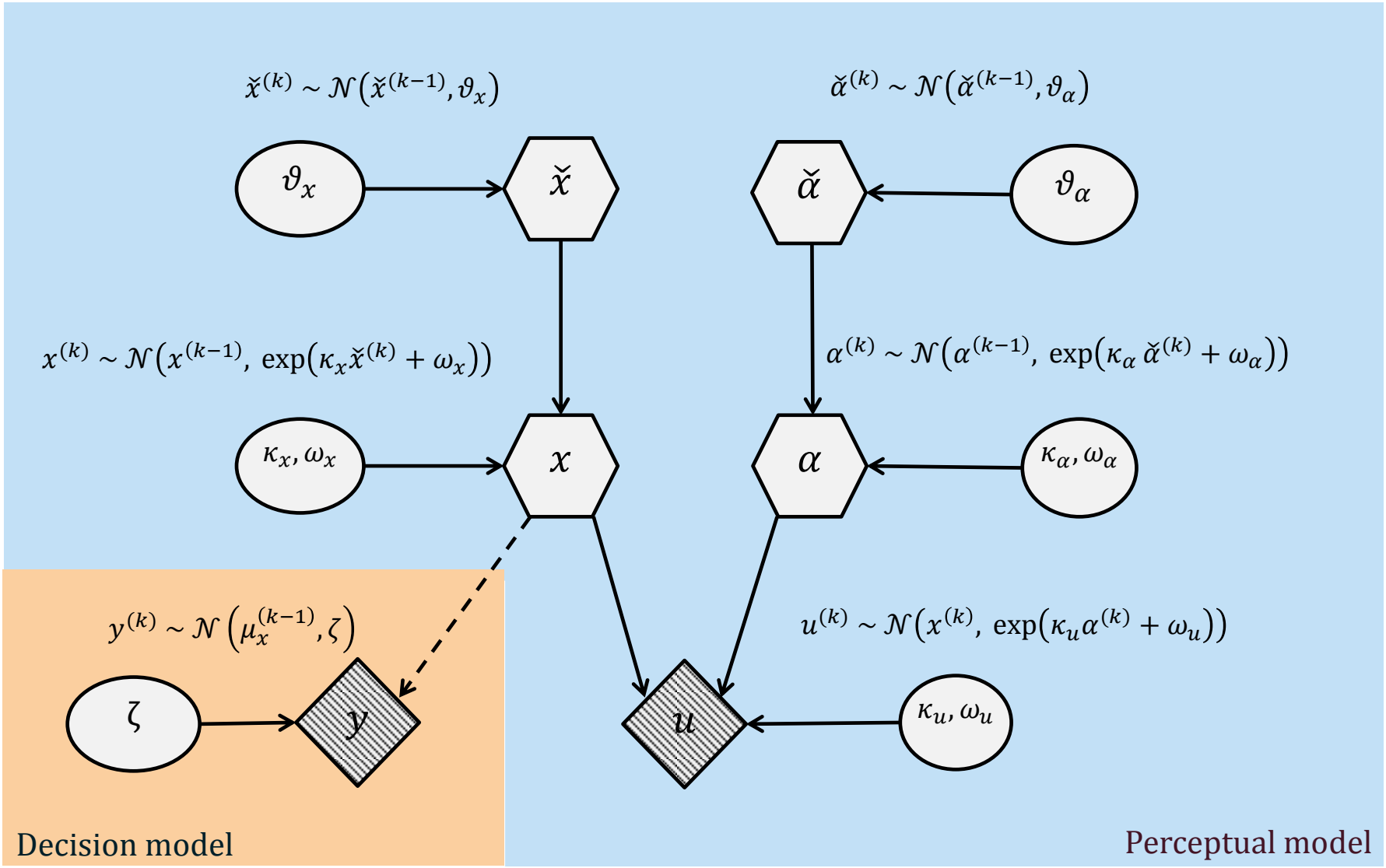
Decision models



Extensions (Guo et al., in prep.)



Extensions


$\breve{x}^{(k)} \sim \mathcal{N}\left(\breve{x}^{(k-1)}, \vartheta_x\right)$
$\breve{\alpha}^{(k)} \sim \mathcal{N}\left(\breve{\alpha}^{(k-1)}, \vartheta_\alpha\right)$
ϑ_x
$\breve{x}$
$\breve{\alpha}$
ϑ_α
$x^{(k)} \sim \mathcal{N}\left(x^{(k-1)}, \exp\left(\kappa_x \breve{x}^{(k)} + \omega_x\right)\right)$
$\alpha^{(k)} \sim \mathcal{N}\left(\alpha^{(k-1)}, \exp\left(\kappa_\alpha \breve{\alpha}^{(k)} + \omega_\alpha\right)\right)$
κ_x, ω_x
x
α
$\kappa_\alpha, \omega_\alpha$
$y^{(k)} \sim \mathcal{N}\left(\mu_x^{(k-1)}, \zeta\right)$
$u^{(k)} \sim \mathcal{N}\left(x^{(k)}, \exp\left(\kappa_u \alpha^{(k)} + \omega_u\right)\right)$
ζ
y
u
κ_u, ω_u
Decision model
Perceptual model

Extensions

$$\pi_{\breve{x}}^{(k)} = \hat{\pi}_{\breve{x}}^{(k)} + \frac{\kappa_x^2}{2} w_x^{(k)} \left(w_x^{(k)} + r_x^{(k)} \delta_x^{(k)} \right)$$

$$\mu_{\breve{x}}^{(k)} = \mu_{\breve{x}}^{(k-1)} + \frac{\kappa_x}{2} \frac{w_x^{(k)}}{\pi_{\breve{x}}^{(k)}} \delta_x^{(k)}$$

$$\pi_{\breve{\alpha}}^{(k)} = \hat{\pi}_{\breve{\alpha}}^{(k)} + \frac{\kappa_\alpha^2}{2} w_\alpha^{(k)} \left(w_\alpha^{(k)} + r_\alpha^{(k)} \delta_\alpha^{(k)} \right)$$

$$\mu_{\breve{\alpha}}^{(k)} = \mu_{\breve{\alpha}}^{(k-1)} + \frac{\kappa_\alpha}{2} \frac{w_\alpha^{(k)}}{\pi_{\breve{\alpha}}^{(k)}} \delta_\alpha^{(k)}$$

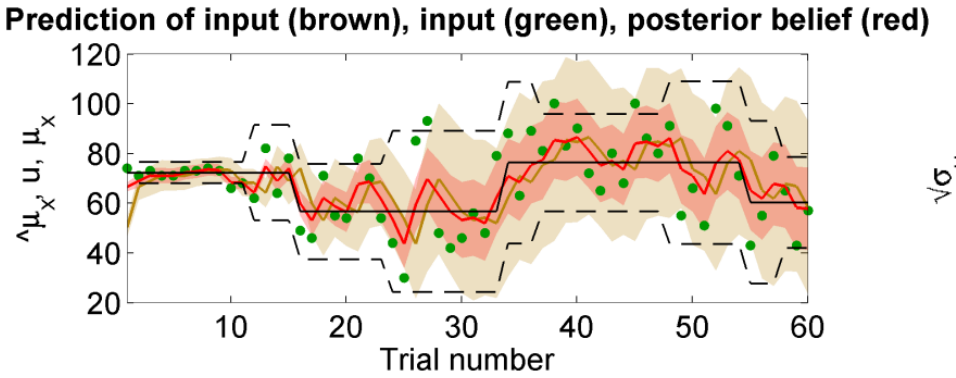
$$\pi_x^{(k)} = \hat{\pi}_x^{(k)} + \hat{\pi}_u^{(k)}$$

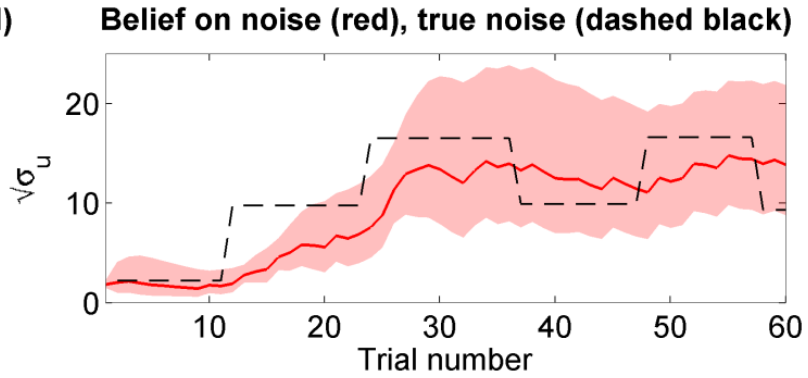
$$\mu_x^{(k)} = \mu_x^{(k-1)} + \frac{\hat{\pi}_u^{(k)}}{\pi_x^{(k)}} \delta_{ux}^{(k)}$$

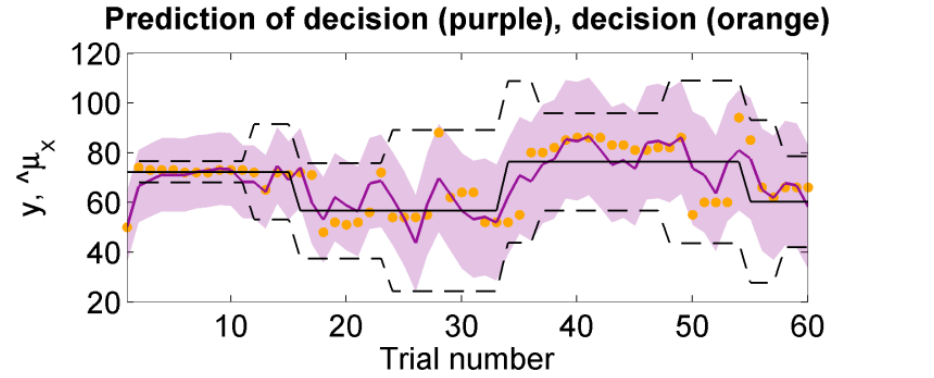
$$\pi_\alpha^{(k)} = \hat{\pi}_\alpha^{(k)} + \frac{\kappa_u^2}{2} \left(1 + \delta_{u\alpha}^{(k)} \right)$$

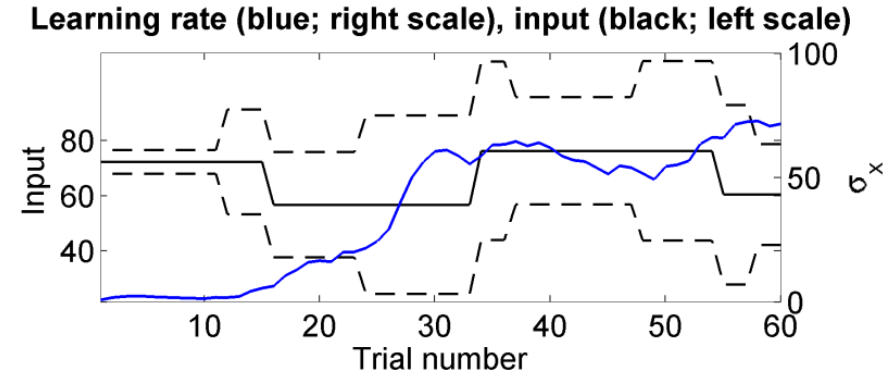
$$\mu_\alpha^{(k)} = \mu_\alpha^{(k-1)} + \frac{\kappa_u}{2} \frac{1}{\pi_\alpha^{(k)}} \delta_{u\alpha}^{(k)}$$

Extensions


Prediction of input (brown), input (green), posterior belief (red)
Trial number


Belief on noise (red), true noise (dashed black)
Trial number


Prediction of decision (purple), decision (orange)
Trial number


Learning rate (blue; right scale), input (black; left scale)
Input
Trial number

Closing remarks on technical issues

- A number of restrictions in the original formulation of the HGF can be lifted without destroying the simplicity of the update equations.

- Inputs can arrive at irregular intervals.

- The random walks may contain drift.

- This drift may itself be changing in time and modeled by its own HGF hierarchy.

- Instead of drift we may have first-order autoregressive (i.e., «AR(1)») processes.

Input at irregular intervals

$$x_i^{(k)} \sim \mathcal{N}\left(x_i^{(k-1)}, f_i(x_{i+1})\right), \quad i = 1, \ldots, n-1.$$

$$x_i^{(k)} \sim \mathcal{N}\left(x_i^{(k-1)}, t^{(k)} f_i(x_{i+1})\right), \quad i = 1, \ldots, n-1.$$

Input at irregular intervals: update equations

$$\mu_i^{(k)} = \hat{\mu}_i^{(k)} + \frac{1}{2}\kappa_{i-1}\, v_{i-1}^{(k)} \frac{\hat{\pi}_{i-1}^{(k)}}{\pi_i^{(k)}} \delta_{i-1}^{(k)}$$

$$\pi_i^{(k)} = \hat{\pi}_i^{(k)} + \frac{1}{2}\left(\kappa_{i-1}\, v_{i-1}^{(k)} \hat{\pi}_{i-1}^{(k)}\right)^2 \left(1 + \left(1 - \frac{1}{v_{i-1}^{(k)} \pi_{i-1}^{(k-1)}}\right) \delta_{i-1}^{(k)}\right)$$

with

$$\hat{\mu}_i^{(k)} \stackrel{\text{def}}{=} \mu_i^{(k-1)}$$

$$\hat{\pi}_i^{(k)} \stackrel{\text{def}}{=} \frac{1}{\sigma_i^{(k-1)} + t^{(k)} \exp\left(\kappa_i \mu_{i+1}^{(k-1)} + \omega_i\right)}$$

$$v_i^{(k)} \stackrel{\text{def}}{=} t^{(k)} \exp\left(\kappa_i \mu_{i+1}^{(k-1)} + \omega_i\right)$$

$$\delta_i^{(k)} \stackrel{\text{def}}{=} \frac{\sigma_i^{(k)} + \left(\mu_i^{(k)} - \hat{\mu}_i^{(k)}\right)^2}{\sigma_i^{(k-1)} + t^{(k)} \exp\left(\kappa_i \mu_{i+1}^{(k-1)} + \omega_i\right)} - 1$$

Constant drift

$$x_i^{(k)} \sim \mathcal{N}\left(x_i^{(k-1)}, t^{(k)} f_i(x_{i+1})\right), \quad i = 1, \ldots, n-1.$$

$$x_i^{(k)} \sim \mathcal{N}\left(x_i^{(k-1)} + t^{(k)}\varrho_i, t^{(k)} f_i(x_{i+1})\right), \quad i = 1, \ldots, n-1$$

leads to

$$\hat{\mu}_i^{(k)} \stackrel{\text{def}}{=} \mu_i^{(k-1)}$$

$$\hat{\mu}_i^{(k)} \stackrel{\text{def}}{=} \mu_i^{(k-1)} + t^{(k)}\varrho_i$$

AR(1) processes

$$x_i^{(k)} \sim \mathcal{N}\left(x_i^{(k-1)}, t^{(k)} f_i(x_{i+1})\right), \quad i = 1, \ldots, n-1.$$

$$x_i^{(k)} \sim \mathcal{N}\left(x_i^{(k-1)} + \varphi_i\left(m_i - x_i^{(k-1)}\right), f_i(x_{i+1})\right), \quad i = 1, \ldots, n-1,$$
$$0 < \varphi_i < 1$$

leads to

$$\hat{\mu}_i^{(k)} \stackrel{\text{def}}{=} \mu_i^{(k-1)}$$

$$\hat{\mu}_i^{(k)} \stackrel{\text{def}}{=} \mu_i^{(k-1)} + \varphi_i\left(m_i - \mu_i^{(k-1)}\right)$$

Variable drift

$$x_i^{(k)} \sim \mathcal{N}\left(x_i^{(k-1)} + t^{(k)} z_i^{(k)}, t^{(k)} f_i(x_{i+1})\right), \quad i = 1, \ldots, n-1$$

$$z_i^{(k)} \sim \mathcal{N}\left(z_i^{(k-1)}, t^{(k)} \vartheta_{z_i}\right), \quad i = 1, \ldots, n-1$$

leads to

$$\hat{\mu}_i^{(k)} \stackrel{\text{def}}{=} \mu_i^{(k-1)} + t^{(k)} \mu_{z_i}^{(k-1)}$$

$$\mu_{z_i}^{(k)} = \hat{\mu}_{z_i}^{(k)} + t^{(k)} \frac{\hat{\pi}_i^{(k)}}{\pi_{z_i}^{(k)}} \left(\mu_i^{(k)} - \hat{\mu}_i^{(k)}\right)$$

$$\pi_{z_i}^{(k)} = \hat{\pi}_{z_i}^{(k)} + \left(t^{(k)}\right)^2 \hat{\pi}_i^{(k)}$$

Variable drift: VAPEs and VOPEs

Note that the drift updates are driven by value prediction errors (VAPEs)

$$\mu_{z_i}^{(k)} = \hat{\mu}_{z_i}^{(k)} + t^{(k)} \frac{\hat{\pi}_i^{(k)}}{\pi_{z_i}^{(k)}} \left(\mu_i^{(k)} - \hat{\mu}_i^{(k)}\right), \text{ VAPE}$$

while the x_i-updates are driven by volatility prediction errors (VOPEs)

$$\mu_i^{(k)} = \hat{\mu}_i^{(k)} + \frac{1}{2} \kappa_{i-1} v_{i-1}^{(k)} \frac{\hat{\pi}_{i-1}^{(k)}}{\pi_i^{(k)}} \delta_{i-1}^{(k)} \quad \text{VOPE}$$

$$\delta_i^{(k)} \stackrel{\text{def}}{=} \frac{\sigma_i^{(k)} + \left(\mu_i^{(k)} - \hat{\mu}_i^{(k)}\right)^2}{\sigma_i^{(k-1)} + t^{(k)} \exp\left(\kappa_i \mu_{i+1}^{(k-1)} + \omega_i\right)} - 1$$

The HGF Toolbox

- Implements many of the models shown (and some not shown)
- Can be downloaded at

http://www.translationalneuromodeling.org/tapas/

Summary

- The HGF is a **general Bayesian model** for the learning of any changing quantity on the basis of a hierarchy of Gaussian random walks.

- We can derive **one-step updates** that are interpretable, have the structure of **precision-weighted prediction errors**, and can be understood in terms of Rescorla-Wagner learning and Bayesian belief updating.

- The resulting model is **modular** and **scalable,** can accommodate **drift** and **autoregressive processes**, and it can be combined with **many different decision models**.

- The parameters of the learning model **can reliably be estimated** by at least four methods.

Thanks

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

University of Zurich[UZH]

UCL

- Rick Adams
- Kay Brodersen
- Jean Daunizeau
- Andreea Diaconescu
- Chaohui Guo
- Karl Friston
- Sandra Iglesias
- Lars Kasper
- Ekaterina Lomakina
- Klaas Enno Stephan
- Simone Vossel
- Lilian Weber