# Modelling the Predictive Brain I

## Introduction / The predictive brain / Bayesian inference / Free energy

**Christoph Mathys**

Workshop at the University of Rijeka, Croatia

12-13 July 2018

# SISSA

- Founded 1978
- Graduate School
- 3 areas: Neuroscience, Mathematics, Physics
- 12 PhD programmes
- About 70 PhD students admitted per year
- Cognitive Neuroscience PhD programme
- Entrance exams in April and September

# Modelling (of the whole-system state) in neuroscience and psychiatry: overview

- The predictive brain
  According to Helmholtz, the brain actively predicts the input it will receive and reacts to violations of its predictions.

- The Bayesian brain
  Violations of predictions, i.e. prediction errors, can be used to update predictions. The optimal way to do that is Bayesian inference.

- (Bayesian) inference and learning
  When using input to **infer** on the states of the environment that are causing sensory inputs, the brain must use a model. This model contains parameters that are taken to be constant in the short term. However, in the long term, parameters have to be adjusted to reflect the structure of the environment. This is **learning**.

- Predictive coding
  How is all of this implemented in the living brain? This is covered by theories of predictive coding.

- What if inference and learning break down?
  Computational and neurobiological models of psychosis. Also, dreaming and the effects of psychoactive drugs.

# Modelling (of the whole-system state) in neuroscience : overview

- Reduction of Bayesian inference to precision-weighted prediction errors

  More on the mathematics of predictive coding.

- Hierarchical Gaussian filtering: theory

  A solution to the learning rate problem in volatile environments.

- Hierarchical Gaussian filtering: experimental studies and results

  A detailed look at some recent studies.

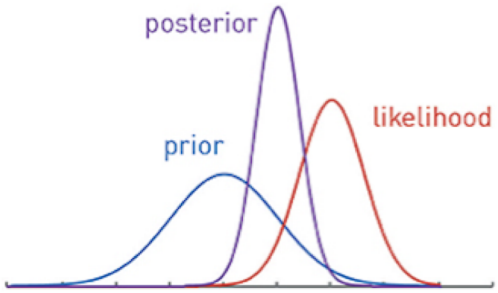
- Active inference

  Instead of updating our beliefs in order to reduce prediction errors, we can use action to make our inputs more like our predictions. Indeed, this is what drives action according to active inference.

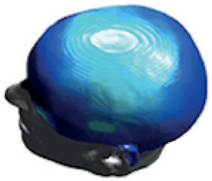
# The predictive (i.e., Helmholtzian) brain

Bayes' Theorem

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




Generative models as computational assays



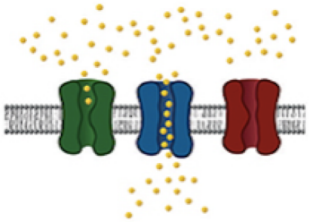
measurement y

forward model

$p(y|x,m)p(x|m)$

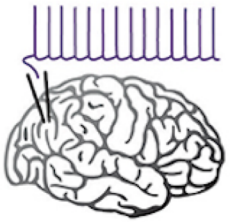
$p(x|y,m)$

inference



hidden system state x

Perception as the inversion of a generative model



neuronal activity y

forward model

$p(y|x,m)p(x|m)$

$p(x|y,m)$

perception

environmental state x



Stephan et al. (2016), Front. Hum. Neurosci., 10:550

# The predictive (i.e., Helmholtzian) brain

‘Objects are always imagined as being present in the field of vision as would have to be there in order to produce the same impression on the nervous mechanism.’

— Helmholtz (1867), Handbuch der physiologischen Optik [Treatise on Physiological Optics], p. 428

In other words: the mind actively infers the objects of its perception in what Helmholtz calls ‘unconscious inference’ (p. 430).

If this is true, it means that **to understand the mind and the brain, we need to understand the mechanics of inference,** i.e. the mechanics of updating predictions.

# Inference: relating uncertain quainties

## For example: a spectacular piece of information

# Inference: relating uncertain quantities

Messerli, F. H. (2012). Chocolate Consumption, Cognitive Function, and Nobel Laureates. *New England Journal of Medicine*, *367*(16), 1562–1564.

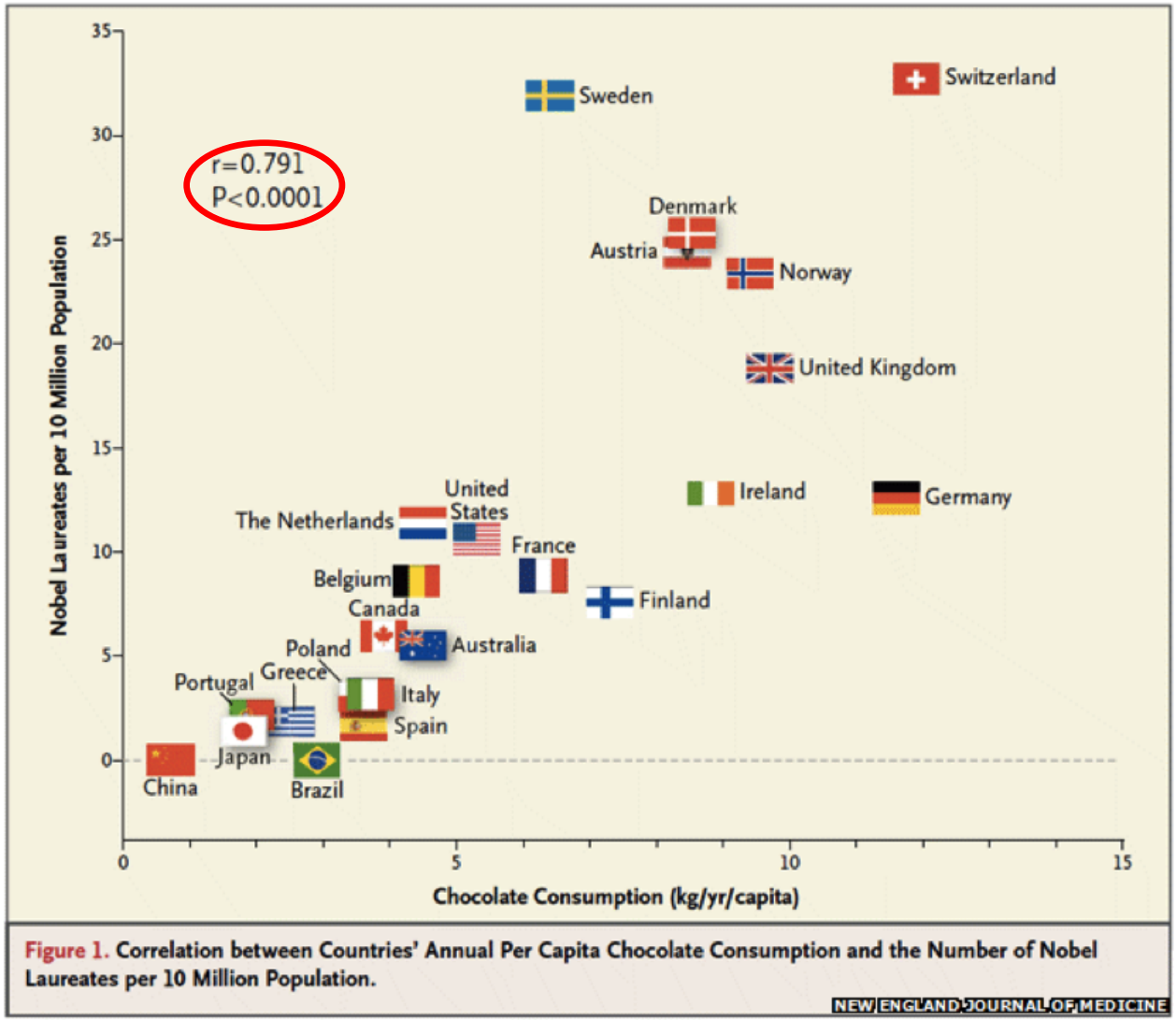



**Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

NEW ENGLAND JOURNAL OF MEDICINE

# So will I win the Nobel prize if I eat lots of chocolate?

This is a question referring to uncertain quantities. Like almost all scientific questions, it cannot be answered by deductive logic. Nonetheless, quantitative answers can be given – but they can only be given in terms of probabilities.

Our question here can be rephrased in terms of a conditional probability:

$$p(Nobel \mid lots\ of\ chocolate) = ?$$

To answer it, we have to learn to calculate such quantities. The tool for this is Bayesian inference.

# «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 $\boldsymbol{a}$ **given** $\boldsymbol{b}$ 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:

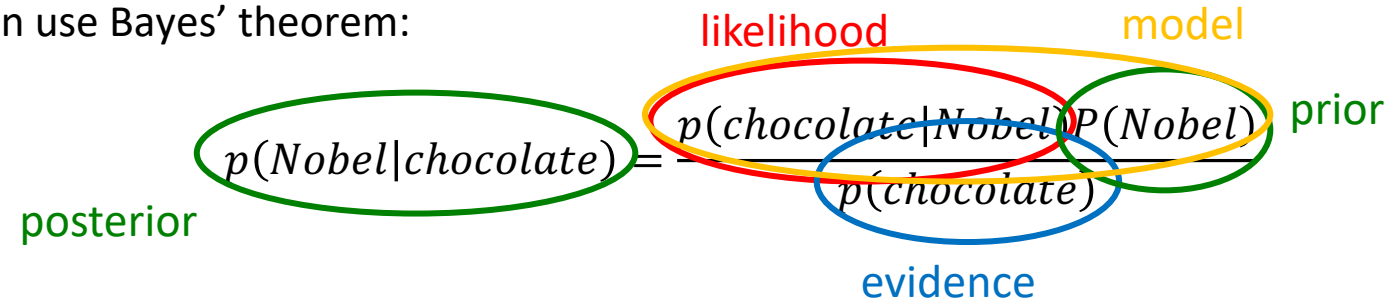
$$p(a) = \sum_b p(a, b) = \sum_b p(a|b)p(b)$$

Because of the product rule, we also have the following rule (**Bayes' theorem**) for going from $p(a|b)$ to $p(b|a)$:

$$p(b|a) = \frac{p(a|b)p(b)}{p(a)} = \frac{p(a|b)p(b)}{\sum_{b'} p(a|b')p(b')}$$

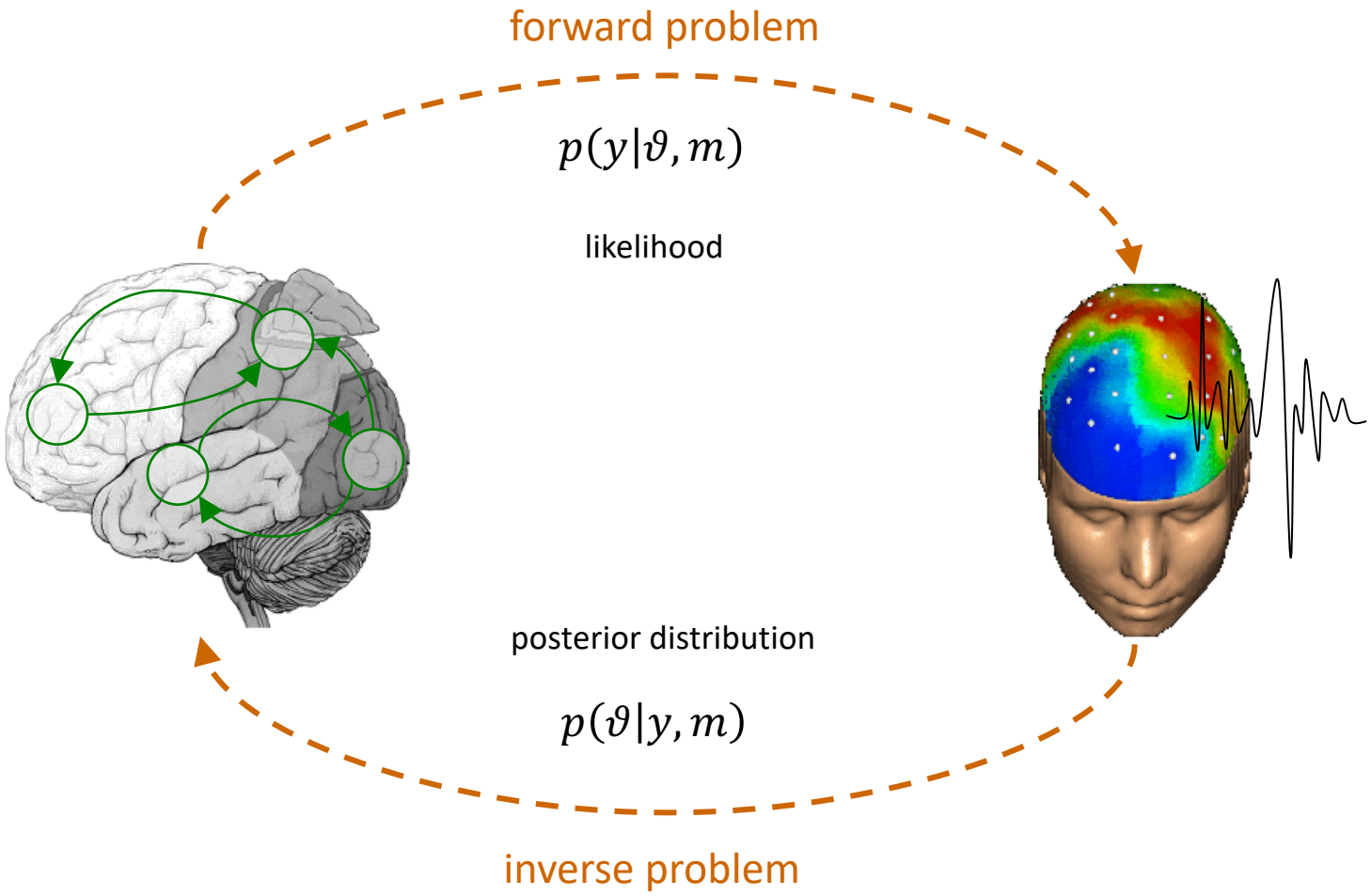
# The chocolate example

In our example, it is immediately clear that $P(Nobel|chocolate)$ is very different from $P(chocolate|Nobel)$. While the first is hopeless to determine directly, the second is much easier to find out: ask Nobel laureates how much chocolate they eat. Once we know that, we can use Bayes' theorem:

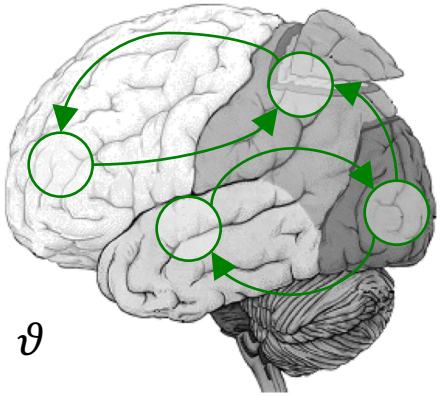



Inference on the quantities of interest in neuroimaging studies has exactly the same general structure.

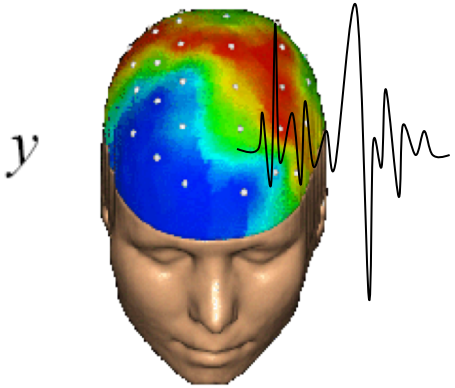
# Inference in neuroimaging



# Inference in neuroimaging



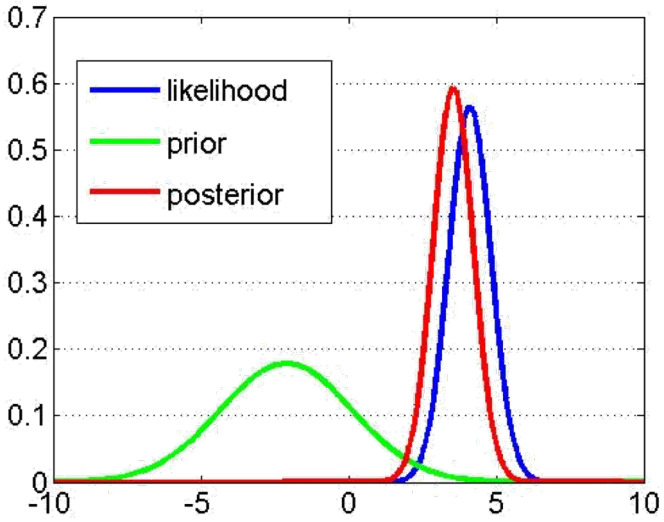

generative model $m$



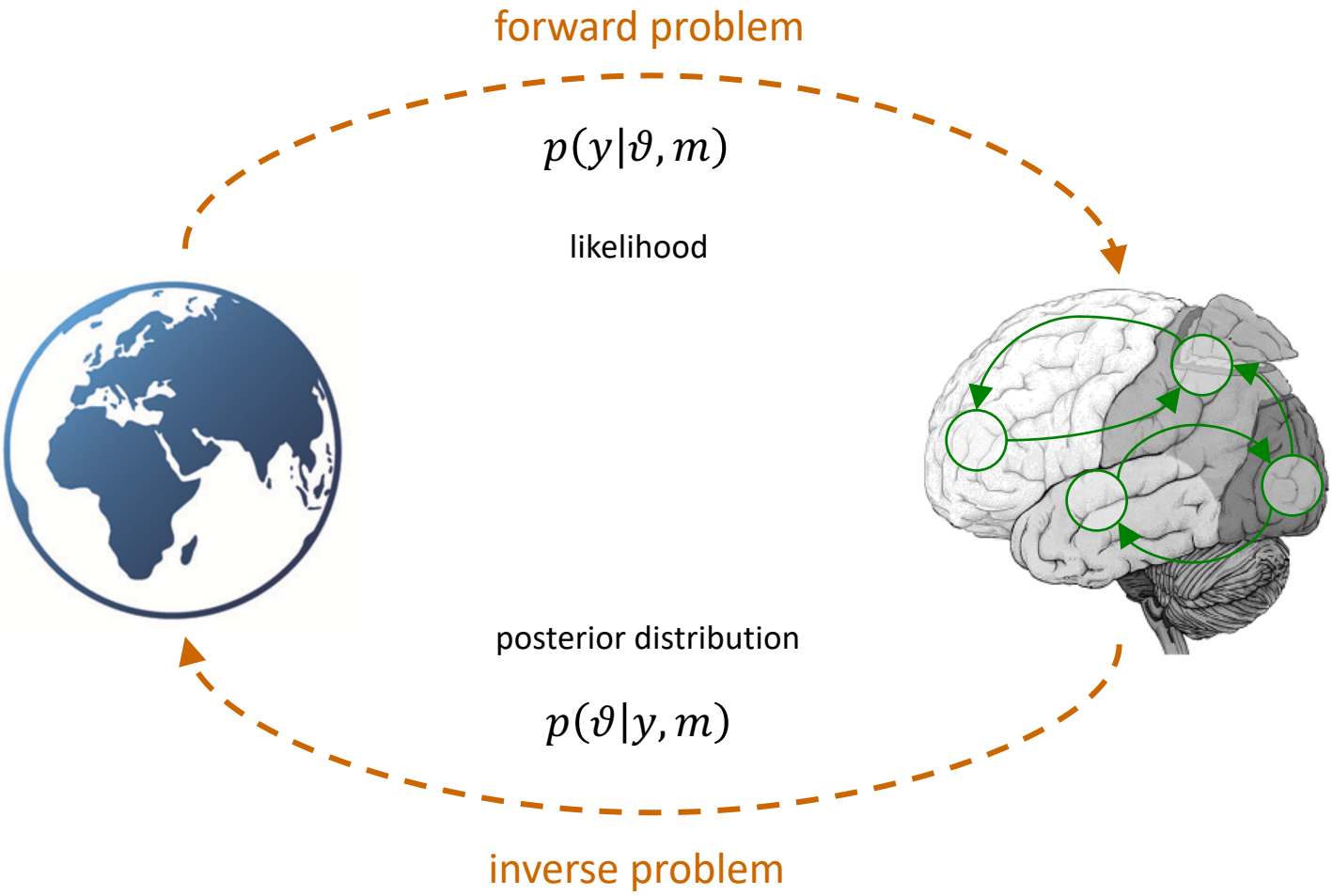

Likelihood: $p(y|\vartheta, m)$

Prior: $p(\vartheta|m)$

Bayes' theorem: $p(\vartheta|y, m) = \frac{p(y|\vartheta, m)p(\vartheta|m)}{p(y|m)}$



# Inference by the brain



# “Every good regulator of a system must be a model of that system” (Conant & Ashby, 1970)

Abstract:

«*The design of a complex regulator often includes the making of a model of the system to be regulated. The making of such a model has hitherto been regarded as optional, as merely one of many possible ways.*

*In this paper a theorem is presented which shows, under very broad conditions, that any regulator that is maximally both successful and simple* must *be isomorphic with the system being regulated. (The exact assumptions are given.) Making a model is thus necessary.*

*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.***»

# What the inferential brain does: updating beliefs

- «Bayesian inference» simply means inference on uncertain quantities according to the rules of probability theory (i.e., according to logic).
- Agents who use Bayesian inference will make better predictions (provided they have a good model of their environment), which will give them an evolutionary advantage.
- We may therefore assume that evolved biological agents use Bayesian inference, or a close approximation to it.
- Predictions are predictive probability distributions. We refer to these distributions as ‘beliefs’. Bayesian inference then amounts to updating of beliefs.
- But how can we reduce Bayesian inference to a simple algorithm that can be implemented by neurons?

# Before we return to Bayes: a very simple example of updating in response to new information

Imagine the following situation:

You're on a boat, you're lost in a storm and trying to get back to shore. A lighthouse has just appeared on the horizon, but you can only see it when you're at the peak of a wave. Your GPS etc., has all been washed overboard, but what you can still do to get an idea of your position is to measure the angle between north and the lighthouse. These are your measurements (in degrees):

76, 73, 75, 72, 77

What number are you going to base your calculation on?

Right. The mean: 74.6. How do you calculate that?

# Updating the mean of a series of observations

The usual way to calculate the mean $\bar{x}$ of $x_1, x_2, \ldots, x_n$ is to take

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

This requires you to remember all $x_i$, which can become inefficient. Since the measurements arrive sequentially, we would like to update $\bar{x}$ sequentially as the $x_i$ come in – without having to remember them.

It turns out that this is possible. After some algebra (see next slide), we get

$$\bar{x}_{n+1} = \bar{x}_n + \frac{1}{n+1}(x_{n+1} - \bar{x}_n)$$

## Updating the mean of a series of observations

Proof of sequential update formula:

$$\bar{x}_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} x_i = \frac{x_{n+1}}{n+1} + \frac{1}{n+1}\sum_{i=1}^{n} x_i = \frac{x_{n+1}}{n+1} + \frac{n}{n+1}\underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{=\bar{x}_n} =$$

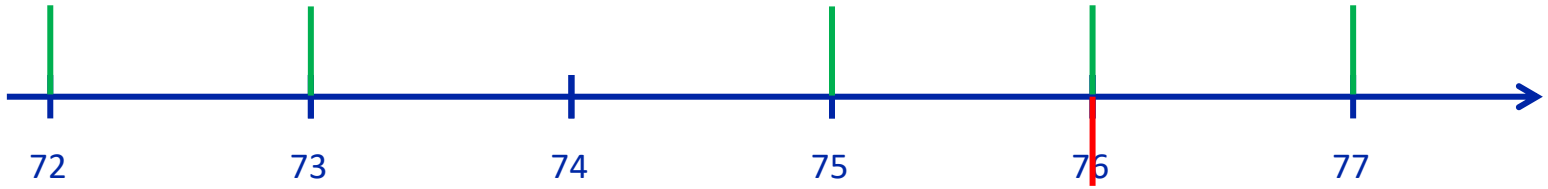
$$= \frac{x_{n+1}}{n+1} + \frac{n}{n+1}\bar{x}_n = \bar{x}_n + \frac{x_{n+1}}{n+1} + \frac{n}{n+1}\bar{x}_n - \frac{n+1}{n+1}\bar{x}_n =$$

$$= \bar{x}_n + \frac{1}{n+1}(x_{n+1} + (n - n - 1)\bar{x}_n) = \bar{x}_n + \frac{1}{n+1}(x_{n+1} - \bar{x}_n)$$

q.e.d.

# Updating the mean of a series of observations

The seqential updates in our example now look like this:




$$\bar{x}_1 = 76$$

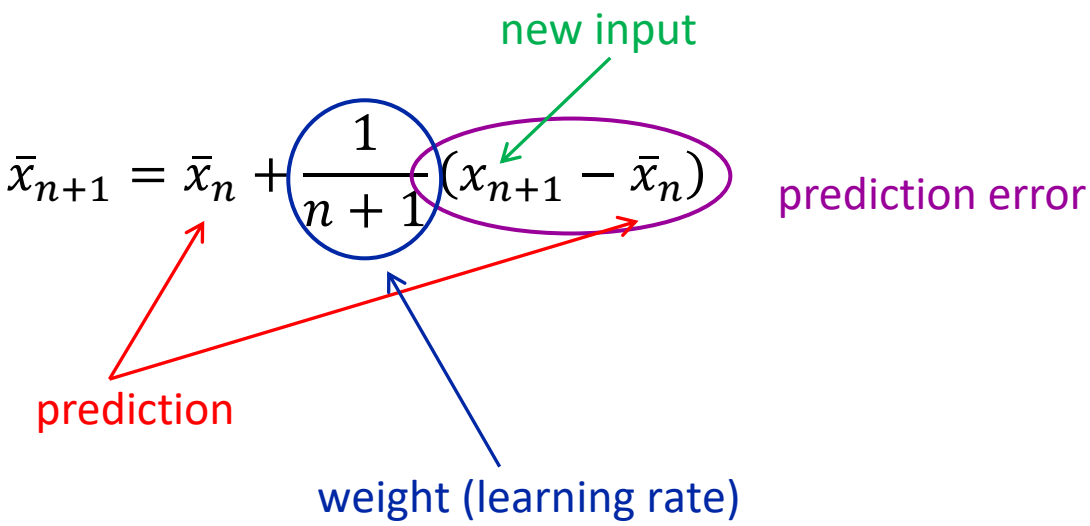
$$\bar{x}_2 = 76 + \frac{1}{2}(73 - 76) = 74.5$$

$$\bar{x}_3 = 74.5 + \frac{1}{3}(75 - 74.5) = 74.\overline{6}$$

$$\bar{x}_4 = 74.\overline{6} + \frac{1}{4}(72 - 74.\overline{6}) = 74$$

$$\bar{x}_5 = 74 + \frac{1}{5}(77 - 74) = 74.6$$

# What are the building blocks of the updates we've just seen?




Is this a general pattern?

More specifically, does it generalize to Bayesian inference?

Indeed, it turns out that in many cases, Bayesian inference can be based on parameters that are updated using **precision-weighted prediction errors**.

# Updates in a simple Gaussian model

Think boat, lighthouse, etc., again, but now we're doing Bayesian inference.

Before we make the next observation, our belief about the true value of the parameter $\vartheta$ can be described by a Gaussian prior:

$$p(\vartheta) \sim \mathcal{N}(\mu_\vartheta, \pi_\vartheta^{-1})$$

The likelihood of an observation $y$ is also Gaussian, with precision $\pi_\varepsilon$ :

$$p(y|\vartheta) \sim \mathcal{N}(\vartheta, \pi_\varepsilon^{-1})$$

Bayes' rule now tells us that the posterior is Gaussian again:

$$p(\vartheta|x) = \frac{p(y|\vartheta)p(\vartheta)}{\int p(y|\vartheta')p(\vartheta')\mathrm{d}\vartheta'} \sim \mathcal{N}\left(\mu_{\vartheta|y}, \pi_{\vartheta|y}^{-1}\right)$$

# Updates in a simple Gaussian model

Here's how the updates to the sufficent statistics $\mu$ and $\pi$ describing our belief look like:

$$\pi_{\vartheta|y} = \pi_{\vartheta} + \pi_{\varepsilon}$$

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

prediction error

prediction

weight (learning rate)=$\frac{\text{how much we're learning here}}{\text{how much we already know}}$

The mean is updated by an uncertainty-weighted (more specifically: precision-weighted) prediction error.

The size of the update is proportional to the likelihood precision and inversely proportional to the posterior precision.

This pattern is not specific to the univariate Gaussian case, but generalizes to Bayesian updates for all exponential families of likelihood distributions with conjugate priors (i.e., to all formal descriptions of inference you are ever likely to need).

# Generalization to all exponential families of distributions

Many of the most widely used probability distributions are families of exponential distributions.

For example, the Gaussian distribution is an exponential family of distributions (and so are the beta, gamma, binomial, Bernoulli, multinomial, categorical, Dirichlet, Wishart, Gaussian-gamma, log-Gaussian, multivariate Gaussian, Poisson, and exponential distributions, among others). This means it can be written the following way:

$$p(\boldsymbol{x}|\boldsymbol{\vartheta}) = h(\boldsymbol{x}) \exp(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\vartheta})) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma}\right)$$

with

$$\boldsymbol{x} = x, \qquad \boldsymbol{\vartheta} = (\mu, \sigma)^{\mathrm{T}}, \qquad h(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}}, \qquad \boldsymbol{\eta}(\boldsymbol{\vartheta}) = \left(\frac{\mu}{\sigma}, -\frac{1}{2\sigma}\right)^{\mathrm{T}}, \qquad \boldsymbol{T}(\boldsymbol{x}) = (x, x^2)^{\mathrm{T}}, \qquad A(\boldsymbol{\vartheta}) = \frac{\mu^2}{\sigma} + \frac{\ln \sigma}{2}$$

This allows us to look at Bayesian belief updating in a very general way for all exponential families of distributions.

# Generalization to all exponential families of distributions (Mathys, 2016)

Our likelihood is an exponential family in its general form:

$$p(\boldsymbol{x}|\boldsymbol{\vartheta}) = h(\boldsymbol{x}) \exp(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\vartheta}))$$

The vector $\boldsymbol{T}(\boldsymbol{x})$ (a function of the observation $\boldsymbol{x}$) is called the sufficient statistic.

For the prior, we may assume that we have made $\nu$ observations with sufficient statistic $\boldsymbol{\xi}$:

$$p(\boldsymbol{\vartheta}|\boldsymbol{\xi}, \nu) = z(\boldsymbol{\xi}, \nu) \exp\big(\nu(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{\xi} - A(\boldsymbol{\vartheta}))\big)$$ (where $z(\xi, \nu)$ is a normlization constant)

It then turns out that the posterior has the same form, but with an updated $\boldsymbol{\xi}$ and $\nu$ replaced with $\nu + 1$:

$$p(\boldsymbol{\vartheta}|\boldsymbol{x}, \boldsymbol{\xi}, \nu) = z(\boldsymbol{\xi}', \nu + 1) \exp\big((\nu + 1)(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{\xi}' - A(\boldsymbol{\vartheta}))\big)$$

$$\boldsymbol{\xi}' = \boldsymbol{\xi} + \frac{1}{\nu + 1}(\boldsymbol{T}(\boldsymbol{x}) - \boldsymbol{\xi})$$

# Proof of the update equation

$$\overbrace{p(\boldsymbol{\vartheta}|\boldsymbol{x},\boldsymbol{\xi},\nu)}^{\text{posterior}} \propto \overbrace{p(\boldsymbol{x}|\boldsymbol{\vartheta})}^{\text{likelihood}}\ \overbrace{p(\boldsymbol{\vartheta}|\boldsymbol{\xi},\nu)}^{\text{prior}}$$

$$= h(\boldsymbol{x}) \exp(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\vartheta}))z(\boldsymbol{\xi},\nu) \exp\big(\nu(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{\xi} - A(\boldsymbol{\vartheta}))\big)$$

$$\propto \exp\big(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot (\boldsymbol{T}(\boldsymbol{x}) + \nu\boldsymbol{\xi}) - (\nu+1)A(\boldsymbol{\vartheta})\big)$$

$$= \exp\left((\nu+1)\left(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \frac{1}{\nu+1}(\boldsymbol{T}(\boldsymbol{x}) + \nu\boldsymbol{\xi}) - A(\boldsymbol{\vartheta})\right)\right)$$

$$= \exp\left((\nu+1)\left(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \left(\boldsymbol{\xi} + \frac{1}{\nu+1}(\boldsymbol{T}(\boldsymbol{x}) + \nu\boldsymbol{\xi} - (\nu+1)\boldsymbol{\xi})\right) - A(\boldsymbol{\vartheta})\right)\right)$$

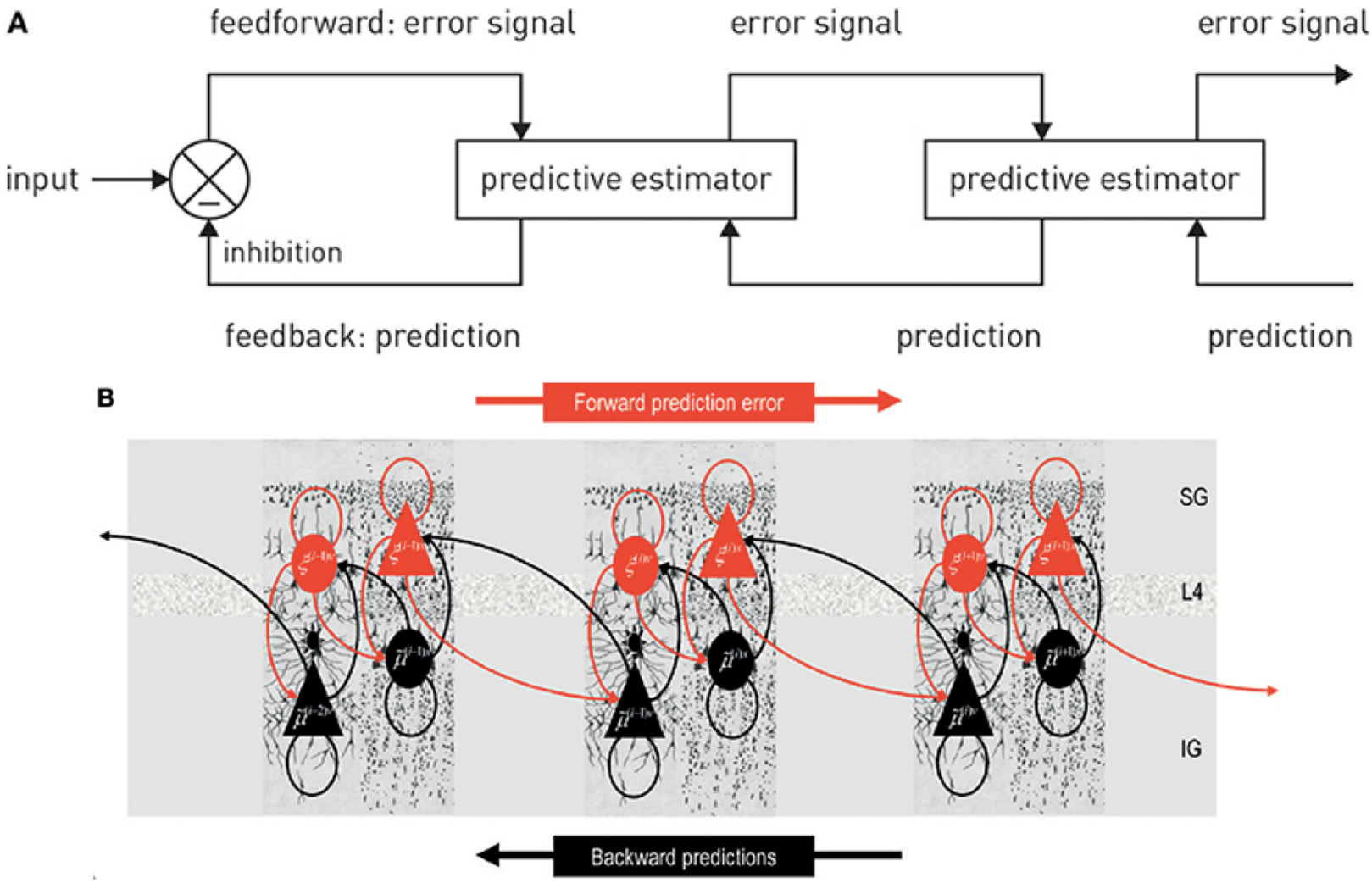
$$= \exp\left((\nu+1)\left(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \left(\underbrace{\boldsymbol{\xi} + \frac{1}{\nu+1}(\boldsymbol{T}(\boldsymbol{x}) - \boldsymbol{\xi})}_{=:\boldsymbol{\xi}'}\right) - A(\boldsymbol{\vartheta})\right)\right)$$

$$\implies \quad p(\boldsymbol{\vartheta}|\boldsymbol{x},\boldsymbol{\xi},\nu) = z(\boldsymbol{\xi}',\nu') \exp\Big(\nu'\big(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{\xi}' - A(\boldsymbol{\vartheta})\big)\Big)$$

$$\text{with } \nu' := \nu + 1, \qquad \boldsymbol{\xi}' := \boldsymbol{\xi} + \frac{1}{\nu+1}(\boldsymbol{T}(\boldsymbol{x}) - \boldsymbol{\xi})$$

q.e.d.

# Neurobiology: predictive coding (e.g., Friston, 2005)



# A different take on Bayesian inference: minimizing surprise/free energy

- Bayesian inference makes optimal use of available information
- It therefore makes sense to assume that Bayesian inference leads to a minimization of surprise at new input
- In what follows, we shall see that minimizing surprise is equivalent to doing Bayesian inference
- Often it is simpler to do inference implicitly by minimizing surprise

# Surprise

- How surprising an event is felt to be depends on its probability.

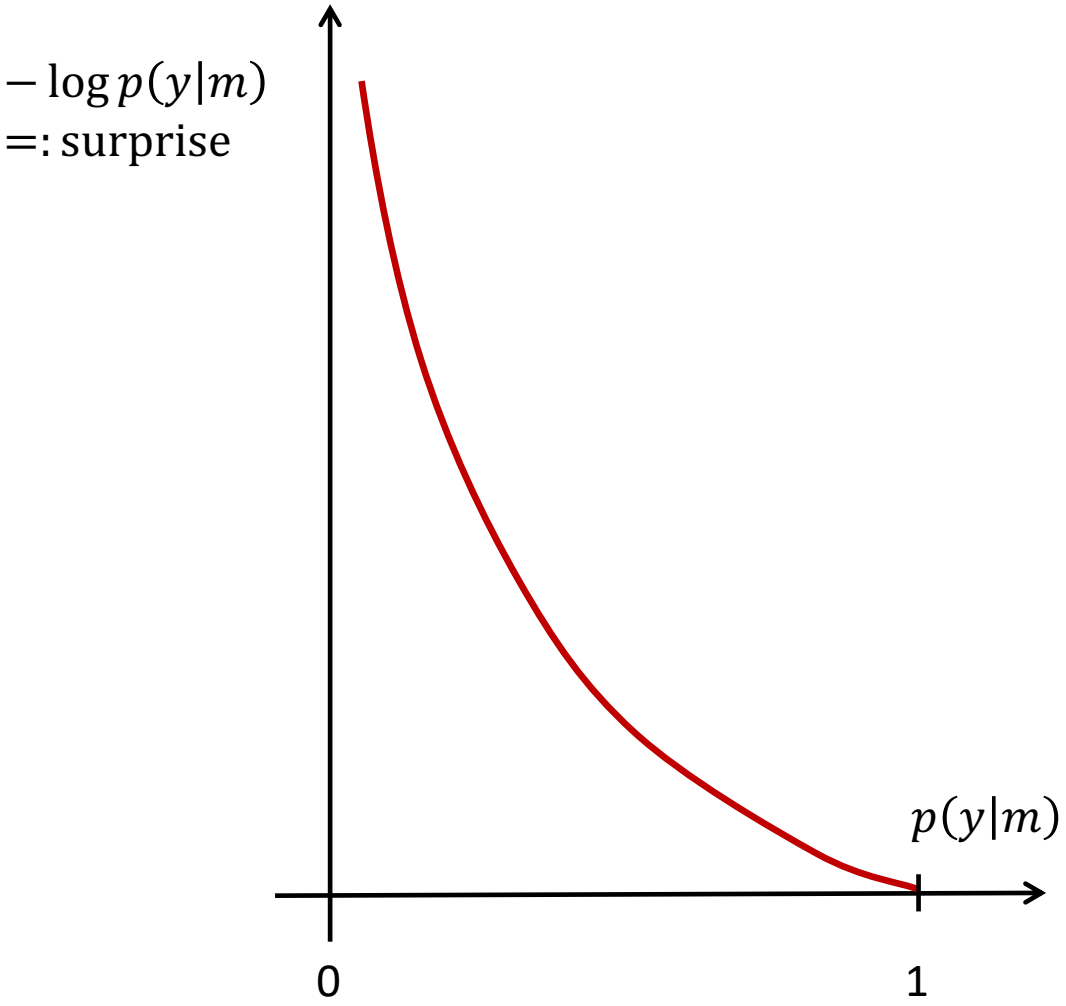
- It makes intuitive sense to take the negative logarithm of $p(y|m)$ as a measure of surprise.

- If $p(y|m) = 1$, the outcome was certain and there is no surprise at all ($-\log\big(p(y|m)\big) = 0$).

- If $p(y|m) = 0$, 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.

# Surprise in a graph



# Entropy

- A concept closely related to surprise is 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 $p$:

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

- 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: $y \in \{\text{heads}, \text{tails}\}$
- Since outcomes are discrete and binary, we use a sum instead of an integral and the binary logarithm to define the entropy:

$$S[p] \coloneqq -\sum_{y} p(y) \log_2 p(y)$$

- 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.

# Varieties of free energy

In information theory, free energy $A$ is the surprise of given model $m$ at a particular (set of) observation(s) $y$:

$$A \coloneqq -\log p(y|m)$$

However, at least three kinds of free energy have to be kept apart:

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

# Free energy in thermodynamics and statistical mechanics

Two kinds: Gibbs and Helmholtz (him again!)

Helmholtz free energy:

$$A \coloneqq U - TS$$

$U$: internal energy; $T$: temperature; $S$: entropy

[Gibbs free energy: $G \coloneqq U + pV - TS$]

In **statistical mechanics**, the Boltzmann distribution describes the **relation between energy and probability**. If particles in a system can be in states $s_1, s_2, s_3, \ldots$ corresponding to energy levels $E_1, E_2, E_3, \ldots$, then the probability $p_i$ of finding a particle in state $s_i$ is

$$p_i = \frac{\exp(-E_i/kT)}{\sum_j \exp\left(-E_j/kT\right)} = \frac{\exp(-E_i/kT)}{Z},$$

where $T$ is *temperature* and $k$ is *Boltzmann's constant*, and $Z \coloneqq \sum_j \exp\left(-E_j/kT\right)$ is the *partition function*.

# Free energy in thermodynamics and statistical mechanics

Taking the logarithm on both sides and rearranging gives us

$$-kT \log Z = E_i + kT \log p_i$$

Taking the expectation value on both sides, we get

$$\sum_i p_i \, (-kT \log Z) = \sum_i p_i \, E_i + \sum_i p_i \, kT \log p_i$$

$$-kT \log Z = \langle E \rangle - T \left( -k \sum_i p_i \log p_i \right) = \langle E \rangle - TS$$

with entropy $\mathrm{S} \coloneqq -k \sum_i p_i \log p_i$.

In analogy to the definition $A \coloneqq U - TS$ of Helmholtz free energy from thermodynamics, this motivates the definition

$$A \coloneqq -kT \log Z$$

of free energy in statistical mechanics.

# Informational free energy

Returning to information theory, we take the definition of informational free energy and perform a series of algebraic operations on it:

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

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

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

This gives us an **information theoretic analogon** to the definition of Helmholtz free energy in statistical mechanics. Compare:

$$A \coloneqq -kT\log Z = -kT\log\left(\sum_j \exp\left(-E_j/kT\right)\right)$$

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

This is the same if we set $kT = 1$ and **interpret the negative logarithm of the joint probability distribution as an energy**:

$$E_\vartheta \equiv -\log p(y,\vartheta|m)$$

# 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.

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

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

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

# 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 $A_v$

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

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

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

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

# The first decomposition of $A_v$

$$
\begin{aligned}
A_v &= -\int q(\vartheta) \log p(y, \vartheta|m) \, \mathrm{d}\vartheta - \left( -\int q(\vartheta) \log q(\vartheta) \, \mathrm{d}\vartheta \right) \\
&= E_v - S_v \\
&= \text{Expected energy} - \text{Entropy}
\end{aligned}
$$

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 $A_v$

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

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

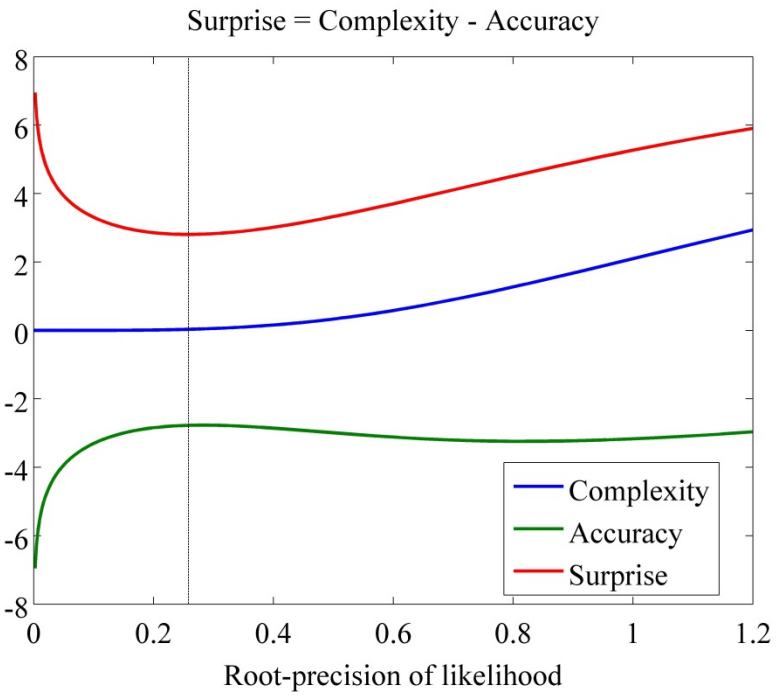
The principled reason why $A_v$ is a good measure of model quality is that the difference in $A_v$ is an approximation to the log-Bayes factor.

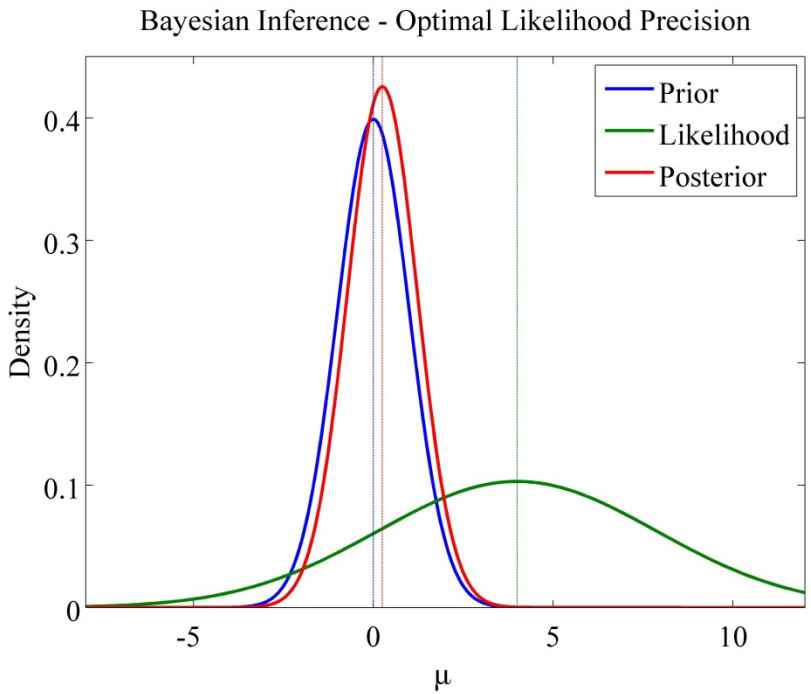
AIC (the Akaike Information Criterion) and BIC (the Bayesian Information Criterion) are approximations to $A_v$ where the complexity term is replaced by a function of the number of parameters.

# The third decomposition of $A_v$

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

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



# 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 $\vartheta$:

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

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

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

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

# 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.