AARHUS UNIVERSITY

# Methods 4 – 7a

**Chris Mathys**


BSc Programme in Cognitive Science

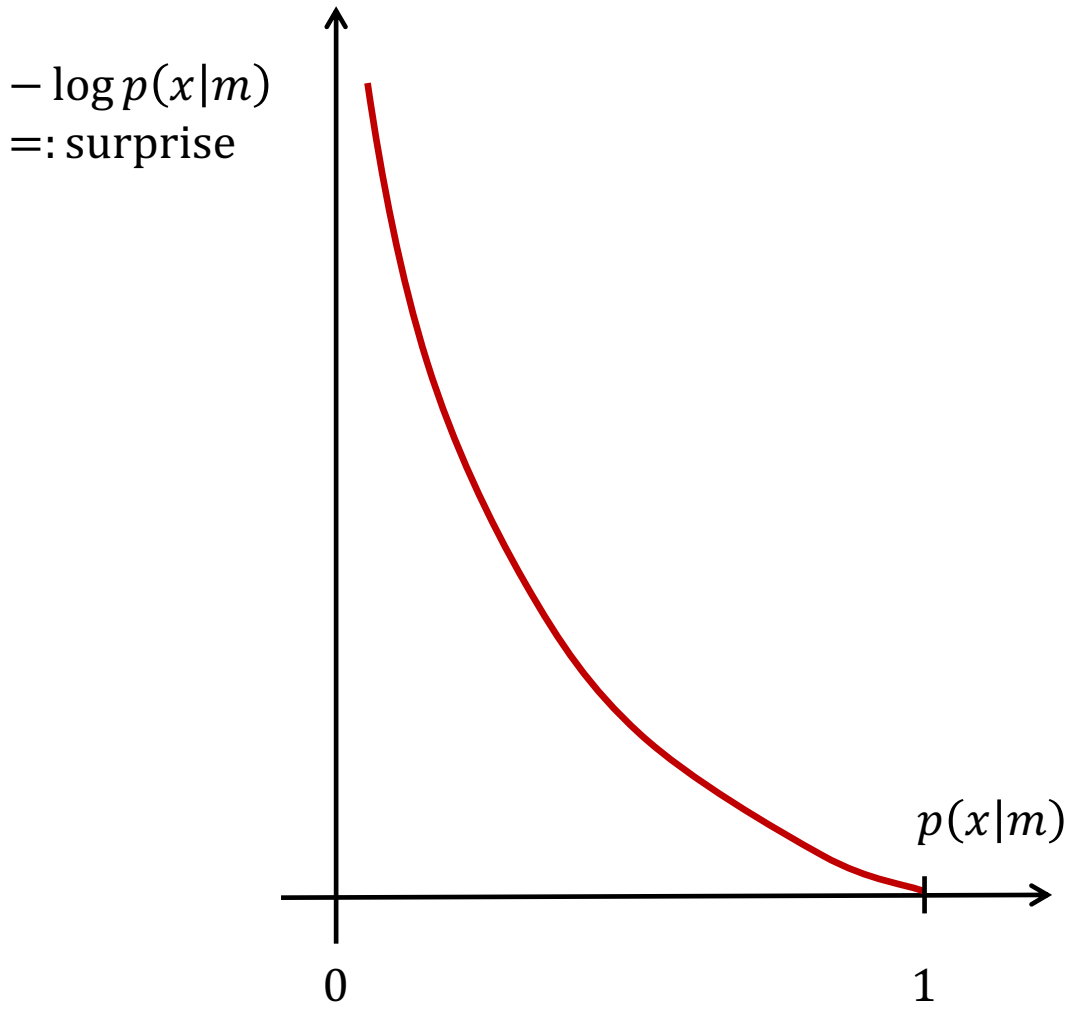
Spring 2024

# General Principles of Model Fitting and Model Comparison

# Optimizing (models, parameters,…) means *minimizing surprise*

- How surprising an event is felt to be depends on its probability.
- It makes intuitive sense to take the negative logarithm of $p(x|m)$ as a measure of surprise.
- If $p(x|m) = 1$, the outcome was certain and there is no surprise at all $(-\log(p(x|m)) = 0)$.
- If $p(x|m) = 0$, the outcome was impossible and surprise is infinite $(-\log(p(x|m)) = \infty)$.
- 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 $p$:

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

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

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

- 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) $x$:

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

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

- The free energy of thermodynamics
- The free energy of statistical physics
- Informational free energy (as above)
- Variational free energy

# Informational free energy

We take the definition of informational free energy and perform a series of algebraic operations on it:

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

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

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

# 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.

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

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

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

# 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 $A_v$.

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

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 $p(\vartheta|x, m)$ by $q(\vartheta)$ because of the theorem above and because $A_v = A$ for $q(\vartheta) = p(\vartheta|x, m)$.

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

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

Jensen's inequality

□

# Three ways to decompose $A_v$

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

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

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

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

# The first decomposition of $A_v$

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

$$= E_v - S_v$$

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

This first decomposition illustrates the mathematical analogy to statistical mechanics.

More importantly, it only contains quantities known to the model-builder: the joint density $p(x, \vartheta|m)$, 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 $A_v$

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

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

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 $p(\vartheta|x,m)$ by $q(\vartheta)$**.

# The third decomposition of $A_v$

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

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

The expected log-likelihood $\log p(x|\vartheta, m)$ 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 $x$ 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 $A_v$

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

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

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 $A_v$

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

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

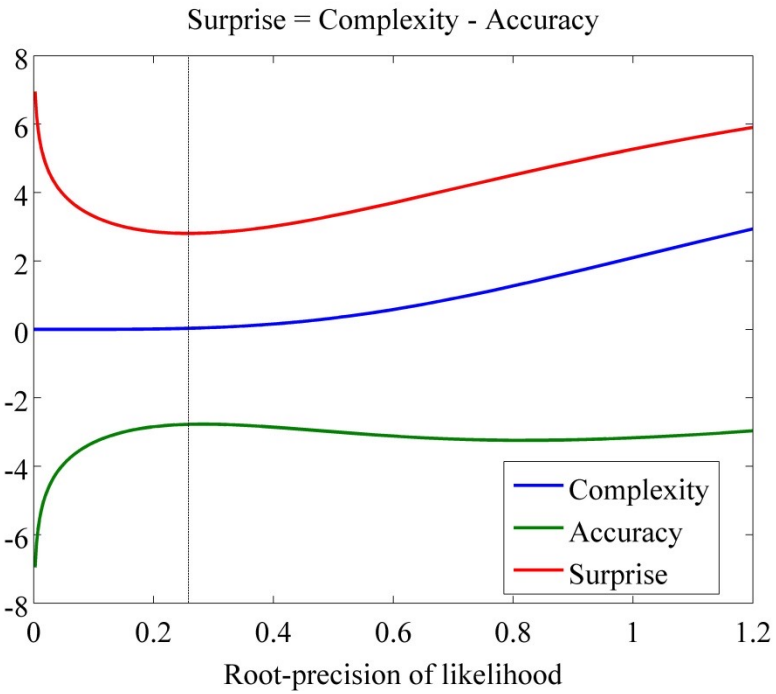
The principled reason why $A_v$ is a good measure of model quality is that the difference in $A_v$ is an approximation to the log-Bayes factor.

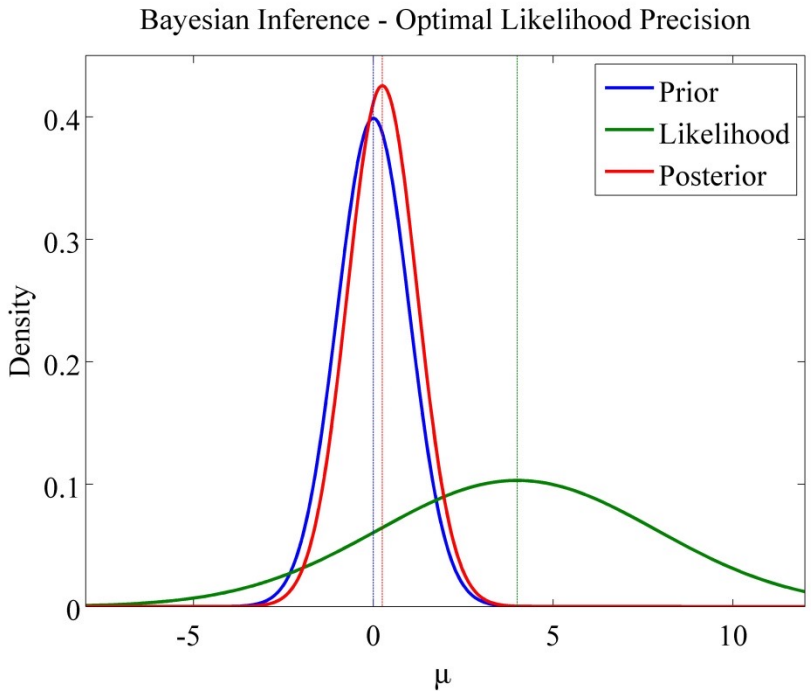
AIC (the Akaike Information Criterion) and BIC (the Bayesian Information Criterion) are approximations to $A_v$ where the complexity term is replaced by a function of the number of parameters.

# The third decomposition of $A_v$

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

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



# How do you compare models?

Related questions:

- How do you quantify the goodness of a model?
- What are the trade-offs involved in improving model fit?
- Do models have an inherent degree of complexity?
- Is complexity good or bad?

# Model comparison: $A_v$ in relation to Bayes factors, AIC, BIC

$$\textbf{Bayes factor} := \frac{p(x|m_1)}{p(x|m_0)} = \exp\left(\log\frac{p(x|m_1)}{p(x|m_0)}\right) = \exp(\log p(x|m_1) - \log p(x|m_0))$$

$$\approx \exp\left(A_{v_0} - A_{v_1}\right)$$

Bayes factor

Posterior odds

Prior odds

[Meaning of the Bayes factor: $\frac{p(m_1|x)}{p(m_0|x)} = \frac{p(x|m_1)}{p(x|m_0)}\frac{p(m_1)}{p(m_0)}$]

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

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

Number of parameters

$$\textbf{AIC} := 2(\mathrm{p} - \text{Accuracy})$$

Number of data points

$$\textbf{BIC} := 2\left(\frac{p}{2}\log N - \text{Accuracy}\right)$$