Sequential Bayesian Learning: A Cognitive Neuroscience Framework

Robert T. Lange *
Einstein Center for Neurosciences Berlin
roberttlange.github.io
robert.lange17@imperial.ac.uk

Miro Grundei
Neurocomputation and Neuroimaging Unit
Free University Berlin
miro.grundei@fu-berlin.de

Sam Gijsen
Neurocomputation and Neuroimaging Unit
Free University Berlin
samgijsen@gmail.com

Contents

*For further details & replication view github.com/RobertTLange/SequentialBayesianLearning.

1 Introduction

The world is inherently sequential. Our very own survival relies on making sense of temporally as well as well as spatially ordered sensory data. Bayesian methods provide a probabilistic framework for integrating such sensory information over time. Within cognitive neuroscience there has been a long-lasting tradition of theoretical frameworks which attempt to formalise this notion: The *Bayesian Brain Hypothesis* (BBH; Knill and Pouget [3]) postulates that the brain updates its posterior belief based on integrating its prior distribution (or previous posterior) with new likelihood evidence from the most recent sensory percept. Furthermore, the *Free Energy Principle* (FEP; Friston [2]) provides an extension to this more general hypothesis. It postulates that every self-organising organism seeks to minimise surprise, i.e. free energy. Based on the log model evidence decomposition into the negative variational free energy and Kullback-Leibler (KL) divergence, this requires a Variational Inference (VI) scheme for approximate Bayesian model inversion. By iteratively minimising the free energy term, one is able to obtain an ever better approximation to the log model evidence, on which we can subsequently perform Bayesian model comparison.

Still cognitive computational neuroscience lacks a general and unifying framework which allows experimentalists to easily generate sequential data, to build expressive regressors, and to analyse the quality of the neural correlates. In this piece of work we try to overcome such limitations. We introduce a general paradigm for generating feature- as well as temporally-dependent sequences of trials based on a simple graphical model. This procedure incorporates the roving paradigm as a special case and allows for flexible design choices such as second-order Markov dependencies. Furthermore, we outline how to model agents that learn the data-generating process in a sequential fashion. Based on the resulting posterior updating rules, we are able to compute different surprise measures which have recently become prominent in the cognitive computational neuroscience literature [1]. As a final contribution we outline a simple log model evidence-based model comparison scheme for the analysis of the resulting regressors.

The following document is structured as follows: First, we introduce the necessary background required in Bayesian modelling in order for us to introduce the special case of *Sequential Bayesian Learning* (SBL). Afterwards, we outline the general data-generating process and introduce agents which learn such sequences. Based on their posterior estimates and realisations one is able to calculate surprise measures. Given neural activity recordings the surprise measures can then be used as regressors in a Bayesian regression model comparison scheme. Finally, we provide an example EEG experiment which utilises the general framework.

2 Background

Probabilistic methods allow one to elegantly formulate belief updating in terms of a simple computational heuristic: Bayes' theorem. Compared to traditional frequentist statistical methods it accounts for the innate uncertainty associated with the statistical relationship between measured observations $y \in \mathbb{R}^d$ and the hidden/unobserved/latent state $s \in \mathbb{R}^d$. At time t, the agent combines their prior over the distribution of the hidden state s_t with the likelihood of the newly observed state y_t:[2]

$$p(s_t|y_t) = \frac{p(y_t, s_t)}{p(y_t)} = \frac{p(y_t|s_t)p(s_t)}{\int_{-\infty}^{\infty} p(y_t|s_t)p(s_t)d\mathcal{S}} \propto p(y_t|s_t)p(s_t)$$

The updated posterior then forms the prior distribution for time $t + 1$:

$$p(s_{t+1}) \coloneqq p(s_t|y_t)$$

The computational procedure can often be formulated as a precision-weighted prediction error correction. The beauty lies within the computational simplicity and its wide applicability.

[2]The theorem follows directly from the product and sum rule of probability.

2.1 Bayesian Operations: Filtering, Smoothing, Decoding and Evaluation

There is a set of classical operations done within the Bayesian paradigm: **Filtering** corresponds to the computation of the posterior distribution over the hidden state s at time t, having observed measurements $y_1, \ldots, y_t$, $p(s_t|y_1, \ldots, y_t)$. **Smoothing**, on the other hand, computes the posterior with evidence accumulated over the entire series of measurements up to a final point T, $p(s_t|y_1, \ldots, y_T)$. **Decoding** corresponds to obtaining the hidden state sequence which maximises the the joint likelihood given the measurements, $y_1, \ldots, y_T$. Hence, $arg\max_{s_1,\ldots,s_T} p(s_1, \ldots, s_T|y_1, \ldots, y_T)$. This operation is not to be confused with the term of decoding in cognitive neuroscience, relating to classifying the identity of a trial based on neural activity representations. Finally, **evaluation** defines the operation of evaluating the joint likelihood of the observations after marginalising out the hidden state sequence. In summary we have:

- Filtering: $p(s_t|y_1, \ldots, y_t)$
- Smoothing: $p(s_t|y_1, \ldots, y_T)$
- Decoding: $arg\max_{s_1,\ldots,s_T} p(s_1, \ldots, s_T|y_1, \ldots, y_T)$
- Evaluation: $p(y_1, \ldots, y_T)$

2.2 Measures of Surprise

In the following sections we will introduce a trial sampling procedure based on graphical model formalism. The resulting sequence $o_1, \ldots, o_T$ can be transformed into measurements $y_1, \ldots, y_T$. These may include the actual identity of the trial ($y_t = o_t$), alternations ($y_t = \mathbf{1}_{o_t \neq o_{t-1}}$) and distinct transitions from one stimulus identity to the other ($y^{(i,j)} = \mathbf{1}_{o_t=i, o_{t-1}=j}$). Given such a sequence of measurements $y_1, \ldots, y_T$ and a posterior model $p(s_t|y_t)$ we are interested in the corresponding surprise associated with experiencing the next trial y_{t+1}. In this work we study three different notions of surprise: Predictive, bayesian and confidence-corrected surprise.

2.2.1 Predictive Surprise

The predictive surprise is defined as:

$$PS(y_t) \coloneqq -\ln p(y_t|s_t) = -\ln p(y_t|y_1, \ldots, y_{t-1}).$$

A posterior predictive distribution $p(y_t|s_t)$ that assigns little probability to an event y_t will cause large amounts of unit-less surprise. For a non-zero predictive probability and a bivariate measurement variable $y \in \{0, 1\}$, we can visualise the predictive surprise (see figure 1a).

Intuition: Following Faraji et al. [1] we can group surprise measures into two classes: puzzlement and enlightenment surprise. While puzzlement is related to an initial stage of misalignment it does not capture the idea of belief commitment. Enlightenment surprise, on the other hand, directly relates to the size of the update of the agent's model of the world.

2.2.2 Bayesian Surprise

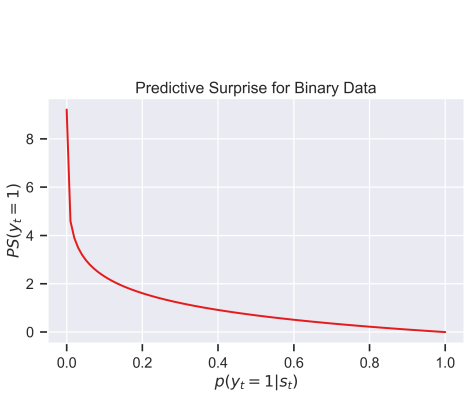
The Bayesian surprise, on the other hand, is defined as the KL divergence between the posterior pre-update and post-update:

$$BS(y_t) \coloneqq KL(p(s_{t-1}|y_{t-1}, \ldots y_1)||p(s_t|y_t, \ldots, y_1))$$

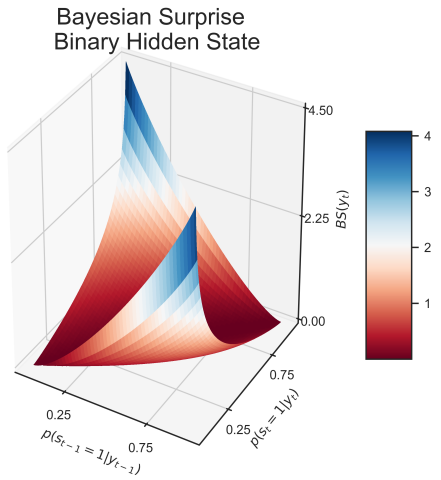
The Kulback-Leibler divergence is an asymmetric measure of how much two probability distributions differ. For a random variable x continuous distributions it is defined as

$$KL(P||Q) \coloneqq \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}$$

It is asymmetric since generally $KL(P||Q) \neq KL(Q||P)$. An example of Bayesian surprise for a binary hidden state and different updates can be seen in figure 1b.



(a) Predictive Surprise



(b) Bayesian Surprise

Figure 1: A Comparison of Three Different Surprise Measures

Intuition: An event that drastically changes the underlying posterior distribution of the agent is more surprising than an event that does not affect the underlying world model.

2.2.3 Confidence Corrected Surprise

$$CS(y_t) \coloneqq KL(p(s_t)||\hat{p}(s_t|y_t))$$

Rob: Talk about intuition and formal background of confidence-corrected surprise

2.2.4 General Remark on Surprise Comparisons

All three reviewed surprise measures are unit-less. Hence, a comparison is only possible across time and after proper standardisation. The explanative power of the individual surprise regressors results from their inter-time variation.

2.3 Conjugacy & Approximate Bayesian Inference

Conjugacy describes a simple mathematical relationship between the prior probability distribution and the likelihood: The updated posterior distribution which combines prior and likelihood information follows the same distribution as the prior distribution.

Popular conjugate pairs include the following:

Conjugate Prior	Likelihood	Posterior
Beta	Bernoulli	Beta
Dirichlet	Categorical	Dirichlet
Univ. Gaussian/Inverse Gamma	Univ. Gaussian	Univ. Gaussian/Inverse Gamma
Multiv. Gaussian/Inverse Wishart	Multiv. Gaussian	Multiv. Gaussian/Inverse Wishart

Please view the supplementary material for a proof of the general conjugate exponential family of conjugacy.

Rob: Provide general proof - See Jordan notes

As soon as we diverge from the simple conjugate-pair case, Bayesian inference faces severe computational challenges. There is no longer a closed-form expression for the posterior distribution, since the integration needed to compute the normalizing constant becomes computationally infeasible.

Instead, one has to revert to an approximate estimation of the posterior. The two most common frameworks for doing so are Markov Chain Monte Carlo (MCMC) and Variational Inference (VI) methods.

3 A General Data Generating Process

Throughout this text we will make use of a simple and general data-generation process. We differentiate between a simple first-order Markov paradigm in which a latent state s_t follows a Markov chain and emits observation o_t which also depend on its precessor o_{t-1}.

3.1 First-Order Markov Dependencies

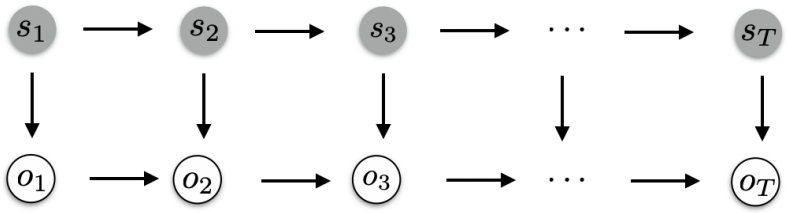


Figure 2: Graphical Model of Data-Generating Process with 1st order Markov Dependency.

$$p(s_{1:T}, o_{1:T}) = p(s_1)p(o_1|s_1)\prod_{t=2}^{T} p(s_t|s_{t-1})p(o_t|o_{t-1}, s_t)$$

- State space: $s \in \mathcal{S} = \{1, \dots, K\}$
- Observation space: $o \in \mathcal{O} = \{1, \dots, M\}$
- Initial state distribution: $p(s_1) = \{\frac{1}{K}, \dots, \frac{1}{K}\} \in [0,1]^K, \sum_{j=1}^{K} p(s_1 = j) = 1$
- Initial obs. distribution: $p(o_1|s_1) = p(o_1) = \{\frac{1}{M}, \dots, \frac{1}{M}\} \in [0,1]^M, \sum_{j=1}^{M} p(o_1 = j) = 1$
- Transitions: $p(s_t|s_{t-1}) = \mathbf{A} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$
  - $\mathbf{A}_{ij} = p(s_t = j|s_{t-1} = i)$; $\sum_{j=1}^{|\mathcal{S}|} \mathbf{A}_{ij} = 1 \; \forall i = 1, \dots, |\mathcal{S}|$ and $\mathbf{A}_{ij} \geq 0 \; \forall i, j = 1, \dots, |\mathcal{S}|$
- Emissions:$p(o_t|o_{t-1}, s_t) = \mathbf{B} \in \mathbb{R}^{|\mathcal{O}|\times|\mathcal{O}|\times|\mathcal{S}|}$
  - $\mathbf{B}_{ijk} = p(o_t = j|o_{t-1} = i, s_t = k) \coloneqq p_{j|i}^{(k)}$
  - $\sum_{j=1}^{|\mathcal{O}|} \mathbf{B}_{ijk} = 1 \; \forall i = 1, \dots, |\mathcal{O}|, k = 1, \dots, |\mathcal{S}|$
  - $\mathbf{B}_{ijk} \geq 0 \; \forall i, j = 1, \dots, |\mathcal{O}|, k = 1, \dots, |\mathcal{S}|$

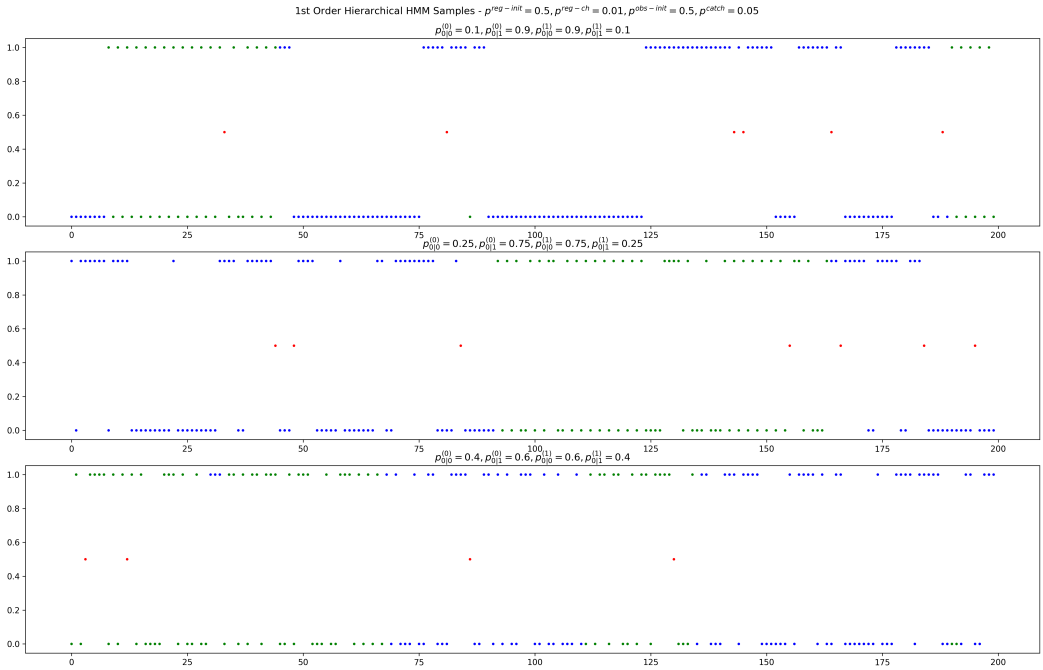


Figure 3: Example Sequences sampled from Graphical Model with first-order Markov dependency structure

3.2 Second-Order Markov Dependencies

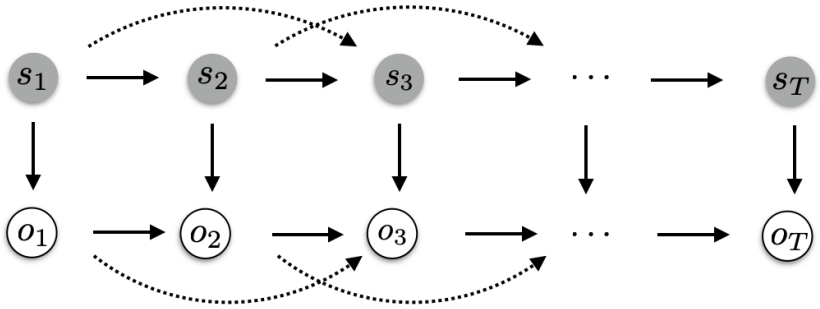


Figure 4: Graphical Model of Data-Generating Process with 2nd order Markov Dependency on the observation and hidden level.

$$p(s_{1:T}, o_{1:T}) = p(s_1)p(o_1|s_1)p(s_2|s_1)p(o_2|s_2, o_1)\prod_{t=3}^{T} p(s_t|s_{t-1}, s_{t-2})p(o_t|o_{t-1}, o_{t-2}, s_t)$$

- State space: $s \in \mathcal{S} = \{1, \ldots, K\}$
- Observation space: $o \in \mathcal{O} = \{1, \ldots, M\}$
- Initial state distribution: $p(s_1) = p(s_2|s_1) = \{\frac{1}{K}, \ldots, \frac{1}{K}\} \in [0,1]^K, \sum_{j=1}^{K} p(s_1 = j) = 1$
- Initial obs. distribution: $p(o_1|s_1) = (o_2|o_1, s_1) = \{\frac{1}{M}, \ldots, \frac{1}{M}\} \in [0,1]^M, \sum_{j=1}^{M} p(o_1 = j) = 1$
- Transitions: $p(s_t|s_{t-1}) = \mathbf{A} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|\times|\mathcal{S}|}$

- $\mathbf{A}_{ijk} = p(s_t = j | s_{t-1} = i, s_{t-2} = k)$
- $\sum_{j=1}^{|\mathcal{S}|} \mathbf{A}_{ijk} = 1 \; \forall i, k = 1, \ldots, |\mathcal{S}|$
- $\mathbf{A}_{ijk} \geq 0 \; \forall i, j, k = 1, \ldots, |\mathcal{S}|$

- Emissions: $p(o_t | o_{t-1}, o_{t-2}, s_t) = \mathbf{B} \in \mathbb{R}^{|\mathcal{O}| \times |\mathcal{O}| \times |\mathcal{O}| \times |\mathcal{S}|}$
  - $\mathbf{B}_{ijkl} = p(o_t = j | o_{t-1} = i, o_{t-2} = k, s_t = l) := p^{(k)}_{j|il}$
  - $\sum_{j=1}^{|\mathcal{O}|} \mathbf{B}_{ijkl} = 1 \; \forall i = 1, \ldots, |\mathcal{O}|, k = 1, \ldots, |\mathcal{S}|$
  - $\mathbf{B}_{ijkl} \geq 0 \; \forall i, j, k = 1, \ldots, |\mathcal{O}|, l = 1, \ldots, |\mathcal{S}|$

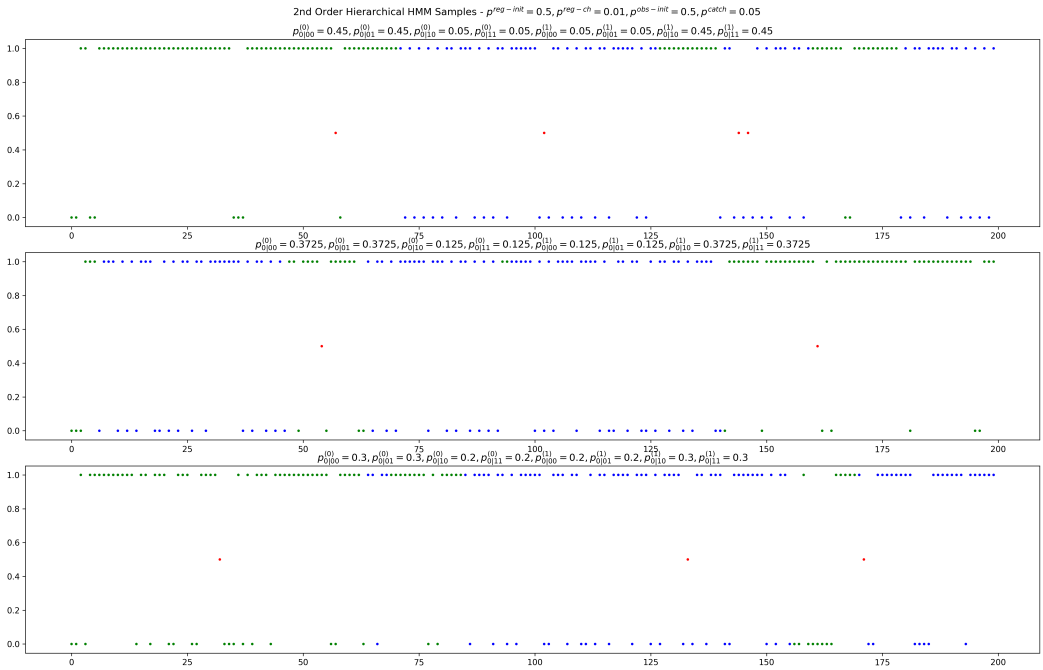


Figure 5: Example Sequences sampled from Graphical Model with second-order Markov dependency structure

4 Sequential Bayesian Agents for Categorical Data

In the following section we will derive updating equations as well as analytically tractable expressions for the surprise measures of interest. More specifically, we introduce two different classes of agents: A conjugate Categorical-Dirichlet (CD) agent as well as a Hidden Markov Model (HMM) agent. They differ in their degree of complexity and allow for different types of hidden state tracking:

- **CD agent**: The hidden state $\mathbf{s}_t$ is assumed to be shared across time and hence static:

$$p(\mathbf{s}_t | \mathbf{s}_{t-1}) = \delta_{\mathbf{s}_{t-1}}(\mathbf{s}_t) \Leftrightarrow \mathbf{s}_t = \mathbf{s}_{t-1} = \mathbf{s} \; \forall t = 2, \ldots, T.$$

- **HMM agent:** The hidden state s_t is a discrete variable and assumed to follow a first-order Markov chain. For a set of discrete hidden states K ($s_t \in 1, \ldots, K$), the transition dynamics are given by the row-stochastic matrix $\mathbf{A} \in \mathbb{R}^{K \times K}$ with $a_{ij} \geq 0$ and $\sum_{j=1}^{K} a_{ij} = 1$:

$$p(s_t | s_{t-1}) = \mathbf{A} \Leftrightarrow p(s^j_t | s^j_{t-1}) = a_{ij} \; \forall t = 1, \ldots, T.$$

We differentiate between three model settings:

1. **Stimulus Probability (SP) Inference**: The model of the agent does not capture any Markov dependency. The current observation o_t only depends on the hidden state s.

2. **Alternation Probability (AP) Inference**: The model captures a limited form of first-order Markov dependency, where the probability the event of altering observations $d_t = \mathbf{1}_{o_t \neq o_{t-1}}$ is estimated given the hidden state s and o_{t-1}.
3. **Transition Probability (TP) Inference**: The model accounts for full first-order Markov dependency and estimates separate alternation probabilities depending on the previous state o_{t-1} and s_t, i.e. $p(o_t|o_{t-1}, s_t)$.

Hence, the degrees of freedom and level of abstractions with which the agent is able to model the received sequence differs between the agents as well as the tracked statistic of interest.

4.1 Categorical-Dirichlet Agent

The Categorical-Dirichlet agent is part of the Bayesian conjugate pairs. It models the likelihood of the observations with the help of the Categorical distribution with $\{1, \dots, M\}$ different possible realizations per sample y_t. Given the probability vector $\mathbf{s} = \{s_1, \dots, s_M\}$ defined on the $M-1$ dimensional simplex $\mathcal{S}_{M-1}$ with $s_i > 0$ and $\sum_{j=1}^{M} s_j = 1$, the probability density function is given by

$$p(y_t = j | s_1, \dots, s_M) = s_j$$

Furthermore, the prior distribution over the hidden state $\mathbf{s}$ is given by the Dirichlet distribution which is parametrized by the probability vector $\alpha = \{\alpha_1, \dots, \alpha_M\}$:

$$p(s_1, \dots, s_M | \alpha_1, \dots \alpha_M) = \frac{\Gamma(\sum_{j=1}^{M} \alpha_j)}{\prod_{j=1}^{M} \Gamma(\alpha_j)} \prod_{j=1}^{M} s_j^{\alpha_j - 1}.$$

Hence, we have

$$\begin{aligned} &\textbf{Prior: } s_1, \dots, s_M \sim Dir(\alpha_1, \dots, \alpha_M) \\ &\textbf{Likelihood: } y \sim Cat(s_1, \dots, s_M) \end{aligned}$$

Given a sequence of observations $y_1, \dots, y_t$ the agent then combines the likelihood evidence with their prior beliefs in order refine their posterior estimates over the latent variable space:

$$\begin{aligned} p(s_1, \dots, s_M | y_1, \dots, y_t) &\propto p(s_1, \dots, s_M |\ \alpha_1, \dots, \alpha_M) \prod_{i=1}^{t} p(y_i | s_1, \dots, s_M) \\ &= \prod_{j=1}^{M} s_j^{\alpha_j - 1} \prod_{i=1}^{t} \prod_{j=1}^{M} s_j^{\mathbf{1}\{y_i = j\}} \\ &= \prod_{j=1}^{M} s_j^{\alpha_j - 1 + \sum_{i=1}^{t} \mathbf{1}\{y_i = j\}} \end{aligned}$$

Hence, we see that the Dirchlet prior and Categorical likelihood pair follow the concept of conjugacy. More specifically, given an initial $\alpha^0 = \{\alpha_1^0, \dots, \alpha_M^0\}$ (set as a hyperparameter or optimized in an Empirical Bayes or Bayesian non-parametrics setting) we can compute the **filtering** step:

$$p(\mathbf{s}_t | y_1, \dots, y_t) = p(s_1, \dots, s_M | y_1, \dots, y_t) = Dir(\alpha^t) \text{ with } \alpha_j^t = \alpha_j^0 + \sum_{i=1}^{t} \mathbf{1}\{y_i = j\}.$$

Furthermore, we can easily obtain the **posterior predictive** distribution needed to compute the predictive surprise measure by integrating over the space of latent states:

$$
\begin{aligned}
p(y_t = x|y_1, \ldots, y_{t-1}) &= \int p(y_t = x|s_1, \ldots, s_M) p(s_1, \ldots, s_M|y_1, \ldots, y_{t-1}) d\mathcal{S}_M \\
&= \int s_x \frac{\Gamma(\sum_{j=1}^M \alpha_j^t)}{\prod_{j=1}^M \Gamma(\alpha_j^t)} \prod_{j=1}^M s_j^{\alpha_j^t - 1} d\mathcal{S}_M \\
&= \frac{\Gamma(\sum_{j=1}^M \alpha_j^t)}{\prod_{j=1}^M \Gamma(\alpha_j^t)} \int \prod_{j=1}^M s_j^{\mathbf{1}\{x=j\} + \alpha_j^t - 1} d\mathcal{S}_M \\
&= \frac{\Gamma(\sum_{j=1}^M \alpha_j^t)}{\prod_{j=1}^M \Gamma(\alpha_j^t)} \frac{\prod_{j=1}^M \Gamma(\mathbf{1}\{x=j\} + \alpha_j^t)}{\Gamma(1 + \sum_{j=1}^M \alpha_j^t)} \\
&= \frac{\alpha_x^t}{\sum_{j=1}^M \alpha_j^t}
\end{aligned}
$$

Finally, we can **evaluate** the likelihood of a specific sequence of events:

$$
p(y_1, \ldots, y_t) = p(y_1) \prod_{i=2}^t p(y_i|y_{1:i}) = \frac{1}{M} \prod_{i=2}^t \prod_{j=1}^M \frac{\alpha_j^i}{\sum_{k=1}^M \alpha_k^i}
$$

Rob: Not 100 percent certain about marginal distribution derivation

We are now able to model different agents which track different statistics of the incoming sequence. We differ between three cases:

1. The stimulus probability model: $y_t = o_t \; \forall t = 1, \ldots, T$.
2. The alternation probability model: $y_t = d_t \; \forall t = 2, \ldots, T$
3. The transition probability model: $y_t = o_t \; \forall t = 1, \ldots, T$ with a set of hidden parameters $\mathbf{s}^{(i)}$ for each transition from $o_{t-1} = i$.

The most simple agent models the identity of the stimulus, i.e. $o_1, \ldots, o_t$.

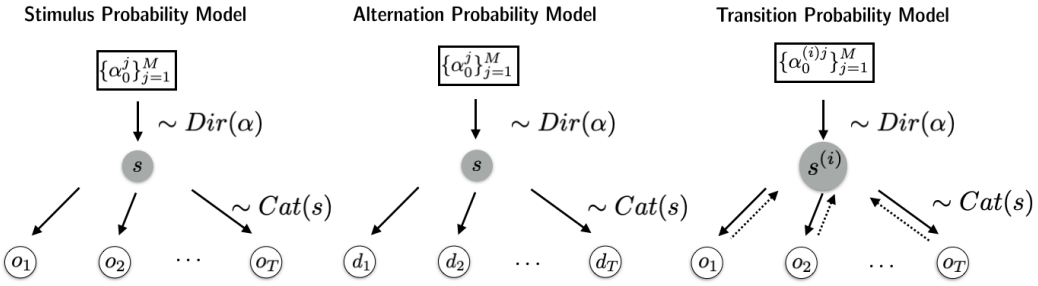


Figure 6: Categorical-Dirichlet Agent as a graphical model. **Left.** The stimulus probability model in which the agent tracks the hidden state vector which determines the sampling process of the raw observations. **Middle.** The transition probability model in which the agent infers the hidden state distribution based on alternations of the observations. **Right.** The transition probability in which the agent assumes a different data-generating process based on the previous observations. Hence, he infers M sets of probability vectors α^i.

Exponentially weighted forgetting

$$
p(\mathbf{s}_t|y_1, \ldots, y_t) = p(s_1, \ldots, s_M|y_1, \ldots, y_t) = Dir(\alpha^t) \;\; \text{with} \;\; \alpha_j^t = \alpha_j^0 + \sum_{i=1}^t e^{-\tau(t-i)} \mathbf{1}\{y_i = j\}.
$$

- SP: $p(o_t|s_t)$, AP: $p(d_t|s_t)$, TP: $p(o_t|o_{t-1}, s_t)$
- Limited memory via exponential decay in parameter updates
- Closed-form posterior/suprise via Beta-Bernoulli conjugacy

In summary, the CD provides a simple conjugate model which assumes that the latent dynamics are static. By incorporating the exponential memory-decay parameter $\tau \in [0, 1]$ we are able to model the subject-specific cognitive capabilities. Let us now turn to the computation of surprise regressors for the Categorical-Dirichlet model.

Predictive Surprise

$$\begin{aligned} PS(y_t) &= -\ln p(y_t|y_1, \ldots, y_{t-1}) \\ &= -\ln \left(\prod_{j=1}^{M} \left(\frac{\alpha_x^t}{\sum_{j=1}^{M} \alpha_j^t} \right)^{\mathbf{1}\{y_t=j\}} \right) \end{aligned}$$

Bayesian Surprise

$$BS(o_t) \coloneqq KL(p(s_{t-1}|y_1, \ldots y_{t-1})||p(s_t|y_1, \ldots, y_t))$$

The general KL divergence for two Dirichlet distributions P and Q parametrized by $\{\alpha_m\}_{m=1}^M$ and $\{\alpha'_m\}_{m=1}^M$ is given by

$$\begin{aligned} KL(P||Q) =& \mathbb{E}_{p(x)}[\log P(x) - \log Q(x)] \\ =& \mathbb{E}_{p(x)}[\log \Gamma(\sum_m \alpha_m) - \sum_m \log \Gamma(\alpha_m) + \sum_m (\alpha_m - 1) \log x_m \\ & - \log \Gamma(\sum_m \alpha'_m) + \sum_m \log \Gamma(\alpha'_m) - \sum_m (\alpha'_m - 1) \log x_m] \\ =& \log \Gamma(\sum_m \alpha_m) - \sum_m \log \Gamma(\alpha_m) - \log \Gamma(\sum_m \alpha'_m) + \sum_m \log \Gamma(\alpha'_m) \\ & - \sum_m (\alpha_m - \alpha'_m) \left(\psi(\alpha_m) - \psi(\sum_m \alpha_m) \right) \end{aligned}$$

where $\psi(.)$ denotes the digamma function.

Confidence-Corrected Surprise

$$CS(o_t) \coloneqq KL(p(s_{t-1}|y_1, \ldots, y_{t-1})||\hat{p}(s_t|y_t))$$

The flat prior is can be written as $Dir(\alpha_1, \ldots, \alpha_m)$ where $\alpha_m = 1$ for all $m = 1, \ldots, M$. The naive observer posterior simply updates the flat prior based on only the most recent observation y_t. Hence, we have that $\hat{p}(s_t|y_t) = Dir(\alpha'_1, \ldots, \alpha'_m)$ with $\hat{\alpha}_m = 1 + \mathbf{1}_{y_t=m}$. Hence at a given point in time t, we have:

$$\begin{aligned} KL(p(s_{t-1}|y_1, \ldots, y_{t-1})||\hat{p}(s_t|y_t)) =& \log \Gamma(\sum_m \alpha_m^{t-1}) - \sum_m \log \Gamma(\alpha_m^{t-1}) - \log \Gamma(\sum_m \alpha_m^t) \\ & + \sum_m \log \Gamma(\alpha_m^t) - \sum_m (\alpha_m^{t-1} - \alpha_m^t) \left(\psi(\alpha_m^{t-1}) - \psi(\sum_m \alpha_m^{t-1}) \right) \end{aligned}$$

4.2 Hidden Markov Model Agent

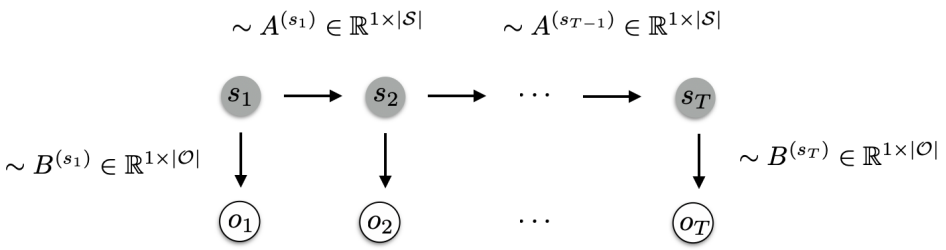


Figure 7: Hidden Markov Model Agent as Graphical Model.

Predictive Surprise

$$\begin{aligned} PS(y_t) &= -\ln p(y_t|y_1,\dots,y_{t-1}) \\ &= -\ln\left(\prod_{j=1}^{M}\left(\frac{\alpha_x^t}{\sum_{j=1}^{M}\alpha_j^t}\right)^{\mathbf{1}\{y_t=j\}}\right) \end{aligned}$$

Bayesian Surprise

$$BS(o_t) \coloneqq KL(p(s_{t-1}|o_{t-1},\dots o_1)||p(s_t|o_t,\dots,o_1))$$

Confidence-Corrected Surprise

$$CS(o_t) \coloneqq KL(p(s_t)||\hat{p}(s_t|o_t))$$

5 Bayesian Model Comparison

VI[3]

5.1 Variational Inference for Conjugate Exponential Family Models

5.2 Auto-Differentiation Variational Inference

Now that we have established the theoretical framework, we are putting the above frameworks to the test.

- Catch trial (intermediate) if $o_t = 2$ - Probability: p^{catch}
- Regime change if $o_t = 3$ - Probability: $p^{reg-change}$
- Let the 2-nd order dependencies be denoted by $p^0_{01} = p(o_t = 0|o_{t-1} = 0, o_{t-2} = 1)$.
- Matrix is fully specified by the four probabilities p^0_{00}, p
- E.g. given p^0_{00}, we have $p^1_{00} = 1 - p^0_{00} - p^{catch}/4 - p^{reg-change}/4$.

[3]Original work on VI originates from classical spin glass models in statistical physics. In the machine learning (ML) community the negative free energy term has been better known as the evidence lower bound (ELBO). Hence, the FEP community refers to the minimization of the free energy, while the ML community, on the other hand, speaks about the maximization of the ELBO. In order to avoid confusion: The two are operationally equivalent.

$$\mathbb{P}_{s_t} = p(o_t|o_{t-1}, o_{t-2}, s_t) = \begin{array}{c} \\ o_{t-1}{=}0,\, o_{t-2}{=}0 \\ o_{t-1}{=}0,\, o_{t-2}{=}1 \\ o_{t-1}{=}1,\, o_{t-2}{=}0 \\ o_{t-1}{=}1,\, o_{t-2}{=}1 \\ o_{t-1}{=}2,\, o_{t-2}{=}0 \\ o_{t-1}{=}2,\, o_{t-2}{=}1 \\ o_{t-1}{=}3,\, o_{t-2}{=}0 \\ o_{t-1}{=}3,\, o_{t-2}{=}1 \\ o_{t-1}{=}0,\, o_{t-2}{=}2 \\ o_{t-1}{=}1,\, o_{t-2}{=}2 \\ o_{t-1}{=}2,\, o_{t-2}{=}2 \\ o_{t-1}{=}3,\, o_{t-2}{=}2 \\ o_{t-1}{=}2,\, o_{t-2}{=}3 \\ o_{t-1}{=}0,\, o_{t-2}{=}3 \\ o_{t-1}{=}1,\, o_{t-2}{=}3 \\ o_{t-1}{=}3,\, o_{t-2}{=}3 \end{array} \begin{pmatrix} o_t{=}0 & o_t{=}1 & o_t{=}2 & o_t{=}3 \\ p^0_{00} & p^1_{00} & p^{catch}/4 & p^{reg-change}/4 \\ p^0_{01} & p^1_{01} & p^{catch}/4 & p^{reg-change}/4 \\ p^0_{10} & p^1_{01} & p^{catch}/4 & p^{reg-change}/4 \\ p^0_{11} & p^1_{11} & p^{catch}/4 & p^{reg-change}/4 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

- A^{s_t} is row-stochastic.
- Marginalizing yields the following relationships 0-th order dependency:

$$p(o_t = i) = \sum_{j,k} p(o_t = i|o_{t-1} = j, o_{t-2} = k) = \sum_{j,k} p^i_{jk}$$

- Furthermore, 1-st order dependencies are the following:

$$p(o_t = i|o_{t-1}) = \begin{cases} p^i_{10} + p^i_{11}, & \text{for } o_{t-1} = 1 \\ p^i_{00} + p^i_{01}, & \text{for } o_{t-1} = 0 \end{cases}$$

- What does slow or fast changing mean in this framework (1-st or 2-nd order)?
- How to go about catch trial? Go to third order? $p(o_t|o_{t-1}, o_{t-2} = 2) = p(o_t|o_{t-1}, o_{t-3})$

6 Conclusions

Add a table which compares the different model specifications!

References

[1] FARAJI, M., K. PREUSCHOFF, AND W. GERSTNER (2018): "Balancing new against old information: the role of puzzlement surprise in learning," *Neural computation*, 30, 34–83.

[2] FRISTON, K. (2010): "The free-energy principle: a unified brain theory?" *Nature reviews neuroscience*, 11, 127.

[3] KNILL, D. C. AND A. POUGET (2004): "The Bayesian brain: the role of uncertainty in neural coding and computation," *TRENDS in Neurosciences*, 27, 712–719.

Todo list

Supplementary Material

Mathematical Derivation

Rob: Add log model evidence decomposition

Notes on Reproduction

Please clone the repository `https://github.com/RobertTLange/SequentialBayesianLearning` and follow the instructions outlined below: