Sequential Bayesian Learning:
A Cognitive Neuroscience Framework

Robert T. Lange [*]
Einstein Center for Neurosciences Berlin
www.rob-lange.com
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

Abstract

This report summarizes efforts to make Sequential Bayesian Learning (SBL) accessible to the general neuroscience community. We provide a unifying framework for modeling the surprise of a Bayesian agent who observes a discrete-time and discrete-event sequence of tri- als. More specifically, we derive expressions for update, filtering as well as defined surprise measures.

1 Introduction

The world is inherently sequential due to its temporal nature. Hence, survival relies on making sense of ordered sensory data.

1.1 Bayesian Reasoning

Probabilistic methods allow one to elegantly formulate this form of 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 $o \in \mathbf{R}^d$ and the hidden/unobserved/latent state $s \in \mathbf{R}^d$.

At time t, the agent combines his prior over the distribution of the hidden state s_t with the likelihood of the observed state o_t:

$$p(s_t|o_t) = \frac{p(o_t, s_t)}{p(o_t)} = \frac{p(o_t|s_t)p(s_t)}{\int_{-\infty}^{\infty} p(o_t|s_t)p(s_t)ds_t} \propto p(o_t|s_t)p(s_t)$$

The computational procedure can often be formulated as a precision-weighted prediction error correction. The updated posterior then forms the prior distribution for time $t+1$:

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

The beauty lies within the computational simplicity and its wide applicability.[2]

[*] This report got drafted during a lab rotation during the Winter of 2018/2019 and connects with an EEG study modeling mismatch-negativity in the somatosensory cortex.

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

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

1.2 Filtering and Smoothing

1.3 Surprise

$$PS(o_t) \coloneqq -\ln p(o_t|s_t)$$

$$BS(o_t) \coloneqq KL(p(s_t)||p(s_t|o_t))$$

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

1.4 Conjugacy

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 (differently parametrized) distribution as the prior distribution.

Popular conjugate pairs include the following:

In the following we will quickly exercise a derivation for the Beta-Bernoulli (Binomial) case:

1.5 Approximate Bayesian Inference

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.

2 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}.

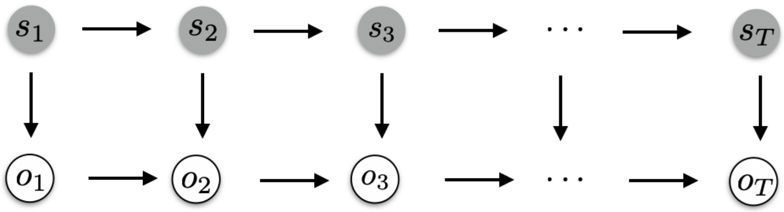

Figure 1: Graphical Model of Data-Generating Process with 1st order Markov Dependency.

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

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

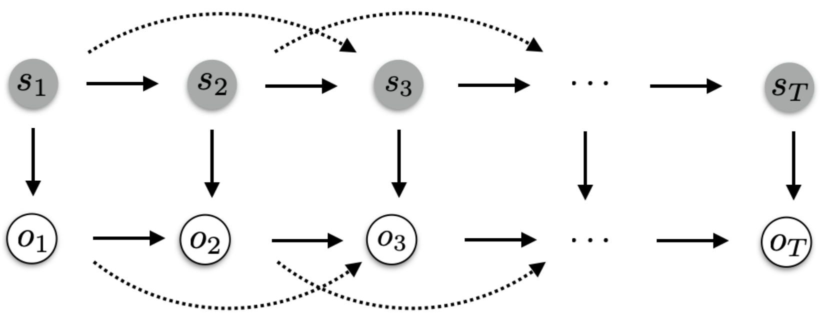

Figure 2: Graphical Model of Data-Generating Process with 2nd order Markov Dependency on the observation and hidden level.

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

- 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: $p(o_1|s_1) = (o_2|o_1, s_1) = \{\frac{1}{M}, \ldots, \frac{1}{M}\} \in [0, 1]^M, \sum\limits_{j=1}^{M} p(o_1 = j) = 1$
- 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: $p(o_t|o_{t-1}, o_{t-2}, s_t) = \mathbf{B} \in \mathbb{R}^{|\mathcal{O}|\times|\mathcal{O}|\times|\mathcal{O}|\times|\mathcal{S}|}$
  - $\mathbf{B}_{ijkl} = p(o_t = j|o_{t-1} = i, o_{t-2} = k, s_t = l) := p^{(k)}_{j|il}$
  - $\sum_{j=1}^{|\mathcal{O}|} \mathbf{B}_{ijkl} = 1 \; \forall i = 1, \ldots, |\mathcal{O}|, k = 1, \ldots, |\mathcal{S}|$
  - $\mathbf{B}_{ijkl} \geq 0 \; \forall i, j, k = 1, \ldots, |\mathcal{O}|, l = 1, \ldots, |\mathcal{S}|$

3 Sequential Bayesian Agents

In the following we will derive updating equations as well as analytically tractable expressions for the three surprise measures introduced above for the simple Beta-Bernoulli agent. A major assumption that we make about the agent is that his hidden state estimate s_t is static, hence

$$p(s_t|s_{t-1}) = \delta_{s_{t-1}}(s_t) \Leftrightarrow s_t = s_{t-1} = s \; \forall t = 1, \dots, T.$$

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 of altering observations $o_t \neq o_{t-1}$ 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.

Hence, the complexity of the modeled learning procedure increases.

3.1 Categorical-Dirichlet Agent

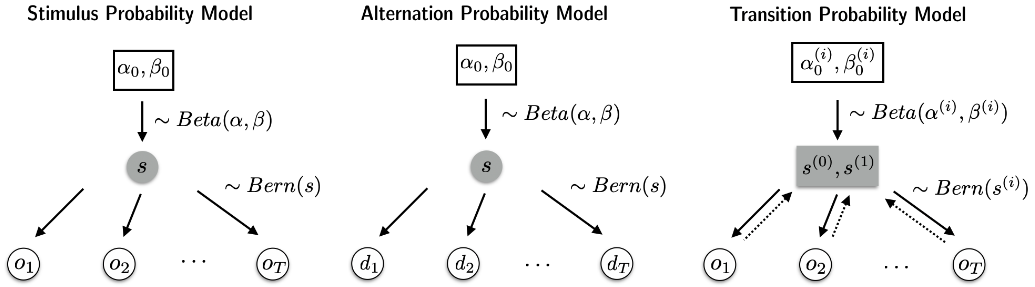
Stimulus Probability Model
α_0, β_0
$\sim Beta(\alpha, \beta)$
s
$\sim Bern(s)$
o_1
o_2
$\dots$
o_T
Alternation Probability Model
α_0, β_0
$\sim Beta(\alpha, \beta)$
s
$\sim Bern(s)$
d_1
d_2
$\dots$
d_T
Transition Probability Model
$\alpha_0^{(i)}, \beta_0^{(i)}$
$\sim Beta(\alpha^{(i)}, \beta^{(i)})$
$s^{(0)}, s^{(1)}$
$\sim Bern(s^{(i)})$
o_1
o_2
$\dots$
o_T

Figure 3: Beta-Bernoulli Agent as Graphical Model. **Left.Middle.Right.**

Exponentially weighted forgetting

- $\mathcal{O} = \{0, 1\}, \mathcal{S} = [0, 1]$
- $p(s_t|s_{t-1}) = \delta_{s_{t-1}}(s_t)$
- $d_t = \mathbf{1}_{o_t \neq o_{t-1}}$
- 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

Stimulus Probability Model

Alternation Probability Model

Transition Probability Model

3.2 Hidden Markov Model Agent

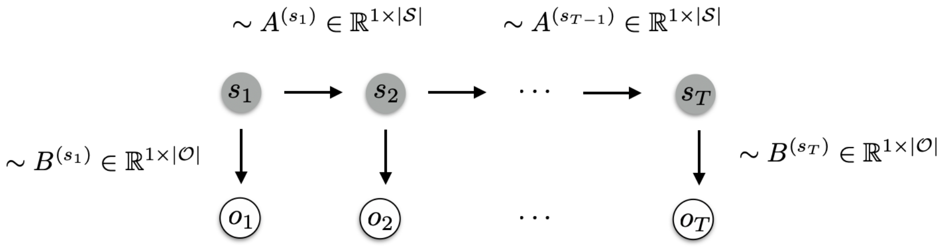
$\sim A^{(s_1)} \in \mathbb{R}^{1\times|\mathcal{S}|}$
$\sim A^{(s_{T-1})} \in \mathbb{R}^{1\times|\mathcal{S}|}$
s_1
s_2
$\cdots$
s_T
$\sim B^{(s_1)} \in \mathbb{R}^{1\times|\mathcal{O}|}$
$\sim B^{(s_T)} \in \mathbb{R}^{1\times|\mathcal{O}|}$
o_1
o_2
$\cdots$
o_T

Figure 4: Hidden Markov Model Agent as Graphical Model.

4 Simulations

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

4.1 The Data-Generating Process

4.1.1 2nd Order Markov Dependencies

- 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 $p^1_{00} = 1 - p^0_{00} - p^{catch}/4 - p^{reg-change}/4$.

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

- 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? $p(o_t|o_{t-1}, o_{t-2} = 2) = p(o_t|o_{t-1}, o_{t-3})$

5 Conclusions

Add a table which compares the different model specifications!

References

Supplementary Material

Notes on Reproduction

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