Episodic memory as a prerequisite for online updates of model structure

David G. Nagy[1,2], Gergo Orban [1]
{nagy.g.david, orban.gergo}@wigner.mta.hu
[1]Computational Systems Neuroscience Lab, Wigner Research Centre for Physics, Budapest, Hungary
[2]Institute of Physics, Eotvos Lorand University, Budapest, Hungary

Abstract

Human learning in complex environments critically depends on the ability to perform model selection, that is to assess competing hypotheses about the structure of the environment. Importantly, information is accumulated continuously, which necessitates an online process for model selection. While model selection in human learning has been explored extensively, it is unclear how memory systems support learning in an online setting. We formulate a semantic learner and demonstrate that online learning on open model spaces results in a delicate choice between either tracking a possibly infinite number of competing models or retaining experiences in an intact form. Since none of these choices is feasible for a bounded-resource memory system, we propose an episodic learner that retains an optimised subset of experiences in addition to semantic memory. On a simple model system we demonstrate that this normative theory of episodic memory can effectively circumvent the challenge of online model selection.

Keywords: episodic memory; semantic memory; online model selection; Bayesian modeling; bounded-resource-rationality

Introduction

In a complex, structured environment that is capable of providing a practically infinite variety of possible experiences, storing them in all their detail would take a prohibitive amount of memory and would be useless in responding to novel situations. It is more beneficial for an learning agent to extract the structure of the world into a concise model, which enables both compression and generalisation, and store this model instead of the observations. But then what is the benefit of devoting precious mental resources to encoding inconsequential contingencies by storing rich snapshots of actual experience, that is, what use is episodic memory?

We argue that online learning in open-ended hypothesis spaces under realistic resource constraints — similar to what the human brain faces — presents a computational challenge that makes such a memory system necessary. In an online learning scenario, observations arrive sequentially and predictions have to be continuously updated. Iterative updates of a particular model's parameters do not require storing the data, since it is sufficient to retain only the information relevant to the specification of the parameters. However, if the structural form of the model is a priori unknown (Kemp & Tenenbaum, 2008), then only a subset of candidate models can be tracked at any given time, since the memory cost of retaining even such compressed statistics becomes prohibitive for an infinite set of models. The inevitable information loss resulting from this restriction presents the brain with a delicate problem: relevance judgements, that is, decisions about what to forget and what to remember can only be based on the currently tracked models, but the initial guess for which models these should be is likely to be wrong because the initial data will only warrant an overly simple model and because it might be misleading about the correct structure and form. Introducing such a bias in the interpretation of new experiences towards the wrong models means that statistical power required for model updating cannot accumulate, since the evidence for alternative models and the information needed for fitting those models will often be deemed irrelevant and discarded, preventing the discovery of the correct representation.

We propose that an episodic memory can alleviate the fundamental problem of online learning described above, by retaining a selected subset of samples. This mini-batch allows evidence for a novel model to accumulate by retaining the contingent details of observations irrespective of how relevant they appear under the current model. We also argue that to take full advantage of episodic memory, its contents should be chosen selectively, so that the combination of episodic and semantic memories provide an efficient representation of the observations.

We are aware of two prior attempts to provide a normative explanation for an episodic memory based on computational principles. The complementary learning systems account of (McClelland, McNaughton, & O'Reilly, 1995) suggests that a hippocampal learning system is required in order to avoid interference with knowledge stored in a neocortical system where learning occurs via slow changes of synaptic connectivity in a network of neurons. Catastrophic interference can be seen as a special case of the detrimental consequences of an inability to maintain a lossless representation of observations during learning, but in contrast to our treatment, the complementary learning systems approach lacks a normative framework and only concerns parameter estimation within a single model. Lengyel & Dayan (2009) argue that using the data samples directly for control is advantageous at the early stages of learning in a new environment. A different but related question about how the combination of semantic and episodic memories can be used to optimize reconstruction is explored by (Hemmer & Steyvers, 2009).

While this paper is intended primarily as a normative argument for the existence of a cognitive system, the problem explored here is intimately related to the efforts in machine learning to handle the problem of online Bayesian model selection in arbitrarily complex model spaces. There are numerous proposals for methods that deal with online model selection or model selection in infinite model-spaces (Grosse, Salakhutdinov, Freeman, & Tenenbaum, 2012; Hjort, Holmes, Müller, & Walker, 2010) separately.

Recently, there have been attempts to tackle both challenges at once in a similar setting, but these are concerned with a restricted hypothesis space over possible model forms, such as mixture models (Sato, 2001; Fearnhead, 2004; Gomes, Welling, & Perona, 2008). Methods that are specific to a given model form have the potential to be vastly more efficient within their domain, but we are striving to find the principles for a general purpose computational architecture that is flexible enough to accommodate uncertainty in the structural form of the model (Kemp & Tenenbaum, 2008). To the best of our knowledge, such a scenario has not yet been explored.

Learning paradigm

In this paper we aim to study how the computational problem of learning shapes the architecture and dynamics of long-term memory. We assume that the main goal of human learning is the acquisition of a suitable representation of the world and propose that this learning process is characterised by the following fundamental properties: i) it is incremental; ii) it requires an open-ended hypothesis space which incorporates not only an arbitrary amount of complexity but also enables the discovery of the appropriate model form; and iii) it is subject to computational constraints, most notably a limited amount of memory.

Our main argument is agnostic to the choice of learning method, but we are adopting the Bayesian inference framework. This framework provides us with a consistent, general and arguably elegant solution for dealing with uncertainty during learning and is central to many state-of-the-art advances in machine learning (Ghahramani, 2015) while simultaneously being able to capture a large body of knowledge concerning the acquisition of abstract knowledge in humans (Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Orbán, Fiser, Aslin, & Lengyel, 2008). In this framework the problem of learning can be formalised as the continual refinement and updating of a probabilistic generative model, where information about unobservable or currently not observed variables, parameters and candidate world structures can all be expressed as probability distributions over latent variables.

In our treatment the memory constraints are formalised such that after the model has been updated, the observation is discarded and only the sufficient statistics for the best performing model is kept. The two main challenges introduced by these constraints are that the learner needs to both: i) assess the plausibility and ii) approximate the right parameter settings of alternative models based solely on the sufficient statistics of the tracked model, without having access to the data.

We set out with an example learning problem that can demonstrate both the challenges and the power of the proposed approach: a mixture of Gaussians model (MoG) has the benefit of showing non-trivial model-learning dynamics while also providing an opportunity for analytical treatment. Mixture models are also frequently used as cognitive models of human category learning (Sanborn, Griffiths, & Navarro, 2006). We use a version where model selection corresponds to determining the correct number of mixture components based solely on the data; parameter learning consists of finding the means for the components; while mixture weights and variance of mixture components are assumed to be fixed and known. Although a more flexible model would provide richer dynamics, the main challenges stated earlier can be clearly demonstrated on this simplified model.

The rest of the paper is structured as follows: first, we show how incremental Bayesian inference works in a setting without resource constraints; next, we introduce a learning agent that only has access to a semantic memory and demonstrate that it has a propensity to discard the information that would enable model change; finally, we show that the introduction of an episodic memory substantially mitigates this problem.

Learning in an unconstrained setting

Bayesian inference provides a consistent framework for learning the form, the structure and the parameters of the model estimating the probability distribution of data. Learning entails the estimation of the posterior probability of parameters (θ) in a given model and/or that of the model (m) itself:

$$\mathrm{P}(\theta \,|\, \mathcal{D}, m) \;\propto\; \mathrm{P}(\mathcal{D} \,|\, \theta, m)\,\mathrm{P}(\theta, m) \tag{1}$$

$$\mathrm{P}(m \,|\, \mathcal{D}) \;\propto\; \mathrm{P}(\mathcal{D} \,|\, m)\,\mathrm{P}(m) \tag{2}$$

Posterior probabilities for alternative model structures, and/or forms need to be assessed individually and the marginal likelihood (mLLH),

$$P(\mathcal{D} \,|\, m) = \int d\theta\, P(\mathcal{D} \,|\, \theta, m)\,\mathrm{P}\,(\theta \,|\, m)\,, \tag{3}$$

plays a critical role in comparing these models: even with a uniform prior probability distribution over alternative models, the mLLH function implements the automatic Occam's razor principle, which ensures that the simplest model that can account for the observed variance in the data has the highest posterior probability. Even when the model prior is flat, the evaluation of mLLH is sufficient to compare the models.

In the analytic treatment of MoG, the posterior over the means μ is a MoG again, in which the number of mixture components grows exponentially with the number of observations T. Whether learning is performed on the whole batch of data at once or is done in an online manner, Bayesian inference yields the posterior distribution of parameters for any particular model structure at any particular time (Fig. 1a-d). This posterior distribution can be used to make predictions on upcoming data and learning helps to disentangle the predictions of different models. While early in the training a complex model that reflects the actual statistics of the data adequately might be discounted because of lack of sufficient evidence, after extended experience the marginal likelihood of the simpler model will be overcome by the model of right complexity (Fig. 1a-d). Switching time in model selection is determined by the actual data samples and is defined by the evolution of the mLLH (Fig. 1e,f). The Automatic Occam's

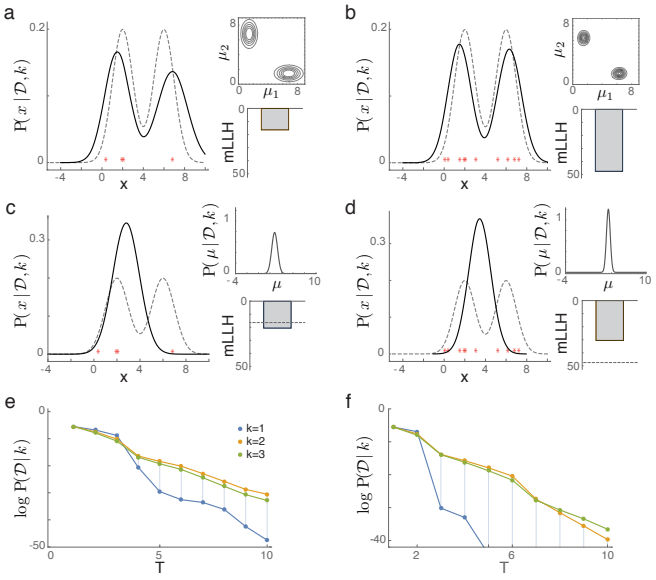

Figure 1: Illustration of model learning on a MoG model. **a**, The goal of learning is the estimation of the probability distribution of the data (*left panel, dashed grey line*) from a limited sample (*asterisks*, $n = 4$). Inference in a given model yields a posterior probability distribution over model parameters (*upper right panel*). The model assumes two mixture components ($k = 2$). Based on the posterior, the predictive posterior distribution (*solid black line*) provides our estimate on how data points are distributed. Marginal likelihood assesses the statistical power of the model (*lower right panel*). **b**, Same as **a** but using a larger data set ($n = 10$). Tighter posterior results in a tighter and more accurate predictive probability distribution and higher average marginal likelihood. **c**, **d**, Same as **a** and **b** but for a $k = 1$ model. **e**, Evolution of mLLH as more data is accumulated from a $k = 2$ model. Colours show models with different number of mixture components. Equality of mLLH at $T = 1$ is a consequence of learning limited to the means. **f**, Same as **e** but for a data set from a $k = 3$ mixture.

razor that is implemented by the mLLH function ensures that no overfitting happens: the learner discovers more complex structures if data statistics justifies such a model but keeps the model as simple as possible.

Semantic-only learner under constraints

While Eq. 1 provides a general recipe for adjusting the model parameters to data, learning can be formulated in two markedly different ways. i), In order to obtain a posterior at a particular time T, the whole data set $\mathcal{D}^T$ is evaluated according to Eq. 1. ii), Online learning relies on a parameter posterior obtained at an earlier time point $T-1$ to provide a prior for the evaluation of novel data:

$$\mathrm{P}\left(\theta \,|\, \mathcal{D}^T, m\right) \propto \mathrm{P}\left(x^T \,|\, \theta, m\right) \mathrm{P}\left(\theta \,|\, \mathcal{D}^{T-1}, m\right) \qquad (4)$$

While online learning has the same power as batch learning, it has the benefit that it is explicitly formulated such that the effect of the earlier data points is summarized in the posterior calculated for $\mathcal{D}^{T-1}$. As a consequence, online learning liberates us from the need to retain the whole data set: once the posterior has been updated the data can be discarded. As long as both parameters and models are updated, this procedure provides a consistent method to update and compare alternative hypotheses on how the model was generated without needing to keep a growing data set in memory. In contrast, if we track only a limited number of models (one model being an extreme but valid approach), discarding data prevents the consistent assessment of alternative models.

The unavailability of the original data leads to an uncertainty as to the possible past data sets that could lead to the same available statistics. An ideal learner represents this uncertainty by means of a probability distribution over possible past data sets. The learner needs a method for constructing such a distribution based solely on the posterior of the current model, since this contains all the information that it has retained. Given such a distribution, a method is required to compare alternative models (i.e. estimate the mLLHs, Eq. 3) and to assess what the parameters of the alternative models would have been had those been tracked from the beginning (i.e. estimate parameter posteriors of novel models Eq. 1). We propose that a natural approximation of the current model's estimate of the distribution of possible past data sets can be obtained by the assessment of the posterior predictive distribution, $\mathrm{P}(x \,|\, \mathcal{D}, m) = \int d\theta \, \mathrm{P}(x \,|\, \theta, m) \, \mathrm{P}(\theta \,|\, \mathcal{D}, m)$, of the tracked model. This choice is conceptually related to using "pseudopatterns" to transfer knowledge between different models (French, 1999). It has the benefit that while the parameter posteriors of different models in general span very different spaces and are thus not comparable, all models give predictions over the same data space (Fig. 1a-d). Another benefit is that the predictive distribution is presumably available for the learner in any case, since it is a fundamental component of numerous other cognitive computations as well.

Inferring the posterior of a novel model

In a given model, the posterior distribution of parameters summarises the model's knowledge about the statistics of the data. Since the predictive distribution of the tracked model carries information about the uncertainty of the parameters this can be used to approximate the posterior of the parameters in a novel model by minimising the dissimilarity of the predictive posterior distributions. Minimising the KL divergence solves exactly this problem:

$$\mathrm{P}\left(\theta \,|\, \mathcal{D}, m'\right) \approx \underset{\mathrm{P}(\theta|\mathcal{D},m')}{\operatorname{argmin}} \; \mathrm{KL}\left[\mathrm{P}(x \,|\, \mathcal{D}, m) \,||\, \mathrm{P}\left(x \,|\, \mathcal{D}, m'\right)\right]. \qquad (5)$$

Calculating the KL divergence analytically is in most cases unfeasible, therefore two approximations have been made. First, inspired by Snelson and Ghahramani (2005) we were looking for a compact representation of the predictive posterior, but instead of achieving this by simply taking a likely set of parameter settings, we've assumed that the posterior comes

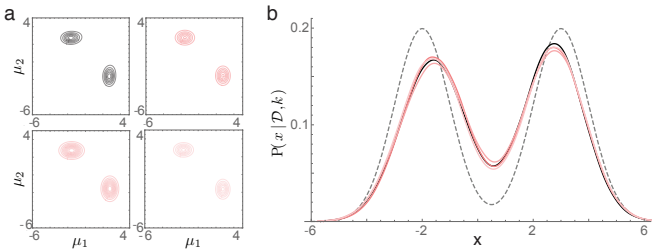

Figure 2: Reconstruction of the parameter posterior from the predictive posterior distribution. **a**, Posterior distribution of the component means in a $k = 2$ model after observing $n = 10$ data points. *Black contour plot*: posterior obtained by analytical calculation; *coloured contour plots*: posterior reconstructions. **b**, Comparison of the true predictive posterior distribution (*black line*) and its approximations. Colours are matched across panels, *dashed line*: data distribution.

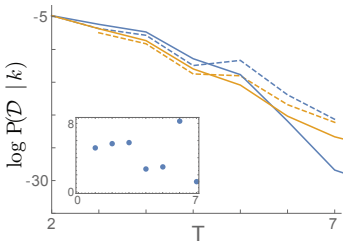

Figure 3: Inability of the memory-constrained learner to increase model complexity. Evolution of the true mLLH (analytic batch learner) of different models (*continuous lines*) and the mLLH of a constrained semantic learner (*dashed lines*). *Inset*:the data set used.

from a simple parametric distribution family:

$$\mathrm{P}\big(\theta \,|\, \mathcal{D}, m'\big) \to \mathrm{P}\big(\theta|\eta, m'\big)\,, \tag{6}$$

where η provides a parametrisation of the approximate posterior. As a result, the former functional optimization problem in (Eq. 5) reduces to

$$\hat{\eta} = \underset{\eta}{\operatorname{argmin}}\ \mathrm{KL}\left[\mathrm{P}(x|\mathcal{D}, m)\ ||\,\mathrm{P}\big(x|\eta, m'\big)\right], \tag{7}$$

where $\mathrm{P}(x|\eta, m') = \int d\theta\, \mathrm{P}(x|\theta, m')\,\mathrm{P}(\theta|\eta, m')$ is the approximate predictive posterior distribution. Eq. 7 is equivalent to minimising the cross entropy, which can be approximated using a Monte Carlo integral. After sampling $\hat{x}_i \sim \mathrm{P}(x \,|\, \mathcal{D}, m)$ we have to choose the η for which the expected value of $\log\left(\mathrm{P}(x|\eta, m')\right)$ is maximal, concluding to a maximum likelihood estimation over the generated 'fake data'

$$\hat{\eta} = \underset{\eta}{\operatorname{argmax}} \sum_i \log\left(\mathrm{P}\big(\hat{x}_i|\eta, m'\big)\right) \tag{8}$$

The resulting $\mathrm{P}(\theta|\hat{\eta}, m')$ is our estimate of the parameter posterior on the original data. The parameter posterior in our implementation of MoG is a MoG again, hence a convenient and effective parametrisation of the posterior uses a single mixture component. This approximate posterior effectively reproduces both the true posterior and the true predictive posterior distribution of the model (Fig. 2).

Model comparison in constrained learners

Model comparison requires the assessment of the mLLH function for alternative models (Eq. 3). However, even if we have access to the marginal likelihood of the tracked model, discarding the original data points renders the construction of the mLLH for the novel model impossible. Again, the posterior of the tracked model summarises our knowledge of the data and and therefore we rely on the predictions that can be drawn from the model posterior in order to assess the possible data sets. This can be achieved by calculating the expected value of the marginal likelihood over the predictive posterior:

$$\left\langle \mathrm{P}\big(\mathcal{D}^* \,|\, m'\big) \right\rangle_{\mathcal{D}^* \sim \mathrm{P}(\mathcal{D}^* \,|\, \mathcal{D}, m)}, \tag{9}$$

where $\mathcal{D}^*$ denotes fake data sets obtained from the predictive posterior distribution. This expected value can be evaluated by Monte Carlo sampling. Upon the arrival of a novel data point x^T, fake data sets are sampled from the predictive distribution. The novel data point is then appended to the fake dataset and the marginal likelihoods are calculated and averaged. In general, a single experience does not constitute adequate evidence for switching to an alternative model, since it lacks sufficient statistical power (Fig. 3). Note, that this claim is not true in extreme cases: there always exist outliers such that the marginal likelihood's automatic Occam's Razor effect will be overpowered by the unlikeliness of the new data (data not shown). If the present model estimate is correct, and the observed data corroborates this model then it can be integrated without information loss. We argue however, that models of differing form and complexity have different kinds of regularities that they can capture, and it is exactly the recurring appearance of features of the data that the current model is unable to represent that necessitates model change. Consequently, when a novel data point arrives which pushes the learner toward a change of model form but is insufficient in itself to force a switch, then the information loss prevents any subsequent model change (Fig. 3). This results in an inability to switch models for the memory-constrained learner even after observing arbitrary amount of evidence that supports a different one.

Episodic learner

The episodic learner differs from the semantic learner only in an additional limited capacity storage for observations. Since the semantic learner's inability to change models is a result of loss of information about past data, it is reasonable to expect that providing a buffer for data points is bound to help. However, we also require that the capacity of episodic memory necessary to enable model change should be small relative to the memory demands of a batch learner. Simply using this

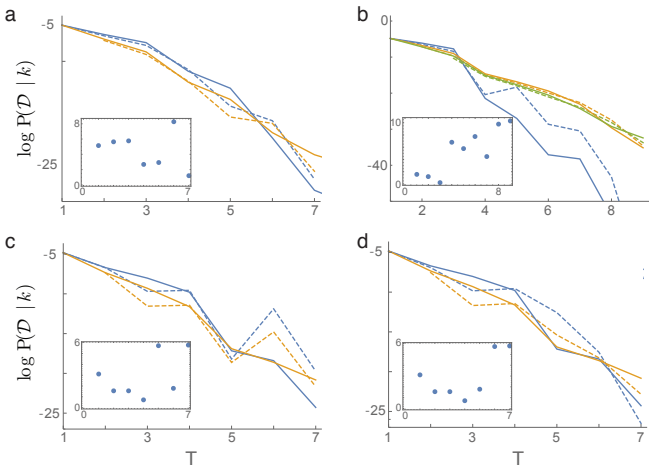

Figure 4: Effect of introducing episodic memory. mLLHs of the analytic batch learners (*solid lines)* approximate learners (*dashed lines*). **a, b**, Using ordered data and retaining the last two data points, the more complex model can obtain sufficient statistical power to overcome the Occam's razor effect at transitions $k = 1 \rightarrow 2$ and $k = 2 \rightarrow 3$, respectively. **c**, For sampled data (unordered) a sliding window for two data points is insufficient to induce model switch. **d**, Episodic memory effectively rearranges data points (compare with panel **c**) such that the arrival of a subsequent data point(s) incompatible with the simple model induces model switch.

storage indiscriminately as a sliding window is inefficient for enabling model change (Fig. 4) since the experiences that taken together would provide the necessary statistical power for model change might not arrive consecutively. Taking full advantage of episodic storage requires the learner to optimise its contents and use it selectively. Thus, given a bounded capacity, the selection criterion for determining which data points to store in episodic and which in semantic memory is expected to be optimised to support the learner in dealing with online model selection.

In order to retain statistics necessary for model transitions, an episodic learner needs to identify points that have a large information content with respect to fitting the models. The Shannon definition of surprise $-log(P(\mathcal{D}|m))$ has been criticised as being unfit for this purpose because low predictive probably does not guarantee that the observation is informative with respect to the appropriateness of the model. Therefore we adopt the Bayesian definition of surprise (Itti & Baldi, 2005), which characterises the extent to which the posterior is different from the prior expectations

$$S(\mathcal{D},m) = KL(P(m|\mathcal{D})||P(m)). \tag{10}$$

Ideally, episodes that are maximally informative regarding the model form would be sought but that would require evaluating the model posterior, $P(m|\mathcal{D}_{T-1})$, which is not accessible, since the learner doesn't necessarily evaluate the same set of models at different steps. Instead, we use the surprise in the model parameters as a proxy: this selects observations that change the learner's beliefs about the parameters the most. A large change in the parameter posterior signifies a difficulty in explaining the new observations and previously seen data under the current model which suggests that a change of models might be appropriate. Another insight can shed further light on the motivation behind our choice of selection criterion. Adopting the perspective that the memory trace is a lossily compressed form of the data, it should be optimised so that the distribution over past data – used in approximating the mLLH and alternative posteriors – is going to be as accurate as possible. We can view the combination of episodic and semantic memories as jointly providing a representation of the agent's past experiences $P_{SM}(x|\eta) + \sum_{x_m \in EM} \delta(x - x_m)$. In order to achieve the best compression the learner needs to use each kind of memory system to store the information it is most suited to reconstruct. Performing such an optimisation would be relatively straightforward by comparing the combined representation with the data, but the data was previously discarded. The learner can, however, select the data points that would change the reconstruction to a large extent, by seeing how much the posterior would change if the given experience was stored in semantic memory. Taken together, we formulated the criterion for selecting a data point for storing in episodic memory by assessing whether the dissimilarity of posteriors with the novel data exceeds a fixed threshold:

$$KL(P(\mu|x_T,\eta,k)||P(\mu|\eta,k)) > \tau. \tag{11}$$

Threshold τ is measured in units of surprise and its value was determined empirically, but performance is relatively robust to its choice. At low threshold values the learner becomes non-selective, which results in accumulating sequential mini-batches. On the other hand, at high threshold levels the learner will be reluctant to store anything in episodic memory and is thus asymptotically equivalent to the semantic learner. When episodic memory is saturated the learner "consolidates" the episodes by performing batch learning on its content. Upon triggering a model change the episodes also serve to find the parameter posterior of the novel model. For demonstration we have set the maximal size of episodic memory to one and used it to show that the problems of a constrained semantic learner can be effectively alleviated (Fig.4).

We have directly contrasted the performance of learning models in the model selection task on random data sets of length $T = 12$ where the generating distribution had $k = 2$ or $k = 3$ components (Fig. 5). Besides the unconstrained learner and the semantic learner, we set up models for an episodic learner with a memory capacity of one and two items, and also a pseudo-episodic learner that does not perform optimisation on the items to be stored in episodic memory. The episodic learner can demonstrate a remarkable increase in performance even with an extremely limited capacity. In order to make a fair comparison between $k = 1 \rightarrow 2$ and $k = 2 \rightarrow 3$ switches we balanced the difficulty of model switch. Our analysis on $k = 2 \rightarrow 3$ switch revealed an even more pronounced advantage of the episodic learner over the semantic learner, doubling the probability of a correct switch.

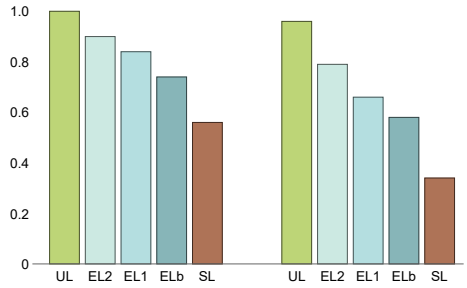

Figure 5: Comparison of model learning in different learners. UL:unconstrained; EL2: episodic with capacity 2; EL1: episodic with capacity 1; ELb: pseudo episodic with no selectivity; SL: semantic. Probability of $k = 1 \rightarrow 2$ and $k = 2 \rightarrow 3$ model switch when data comes from a MoG with $k = 2$ and $k = 3$ (*left* and *right panels*, respectively) estimated from a thousand model runs each.

Discussion

We have offered a normative argument for the existence of episodic memory by analysing a computational problem that the brain has to solve, namely online model selection in an open-ended model space. We used a simple minimal model to demonstrate that the introduction of memory constraints has dire consequences for a semantic-only learner and showed that these problems are substantially mitigated by an episodic memory, the contents of which are selected based on the Bayesian formalisation of surprise.

Our choice of model was motivated by analytical tractability which helped us to set a benchmark to model learning. While this choice constrained the form of the model and the size of the data set, the demonstrated problem is fundamental. These restrictions can be lifted by allowing the iterative posterior updates to be approximate, for example by using particle filters. Importantly, we strove to only use principles and approximations that are agnostic to the model class, so that the episodic learner can straightforwardly be extended to richer hypothesis spaces.

The overall goal of our normative account is to shed light on the dynamics underlying the organisation of long-term memory: from a continuous stream of experience, how does the human brain determine what parts to remember and what to forget? It is extensively documented that humans are prone to systematic biases in these decisions. We share the widely-held belief that these systematic memory errors reflect rational adaptations to computational resource constraints. In our assessment, a comprehensive explanation and detailed predictions on how these processes work requires an understanding of both the computational function and the constraints that shape the dynamics of long-term memory. In this paper, we aimed to provide the computational backbone for such a normative understanding: although some aspects of the current treatment are reminiscent of the characteristics of the dynamics of human memory (e.g. storing detailed representations of surprising events), a more direct comparison between model predictions and human performance will require the analysis of model classes that can be related to available human data.

Acknowledgements.

We thank Zoubin Ghahramani and Szabolcs Káli for discussions and Máté Lengyel and Balázs Ujfalussy for comments on earlier versions of the manuscript. We thank the anonymous reviewers for useful suggestions. This work was supported by an MTA Lendület Fellowship.

References

Fearnhead, P. (2004). Particle filters for mixture models with an unknown number of components. *Stat Comput*.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, *3*(4), 128–135.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*, 452–459.

Gomes, R., Welling, M., & Perona, P. (2008). Incremental learning of nonparametric Bayesian mixture models. *2008 IEEE Conf on Comput Vis and Pattern Recogn*.

Grosse, R. B., Salakhutdinov, R., Freeman, W. T., & Tenenbaum, J. B. (2012). Exploiting compositionality to explore a large space of model structures. *Conference on Uncertainty in Artificial Intelligence*.

Hemmer, P., & Steyvers, M. (2009). Integrating episodic and semantic information in memory for natural scenes. *Proc 31st Ann Conf of the Cogn Sci Soc*, 1557–1562.

Hjort, N. L., Holmes, C., Müller, P., & Walker, S. G. (2010). *Bayesian nonparametrics* (Vol. 28). Cambridge Univ. Press.

Itti, L., & Baldi, P. F. (2005). Bayesian surprise attracts human attention. In *Adv neur inf proc sys* (pp. 547–554).

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proc Natl Acad Sci*, *105*, 10687–92.

Lengyel, M., & Dayan, P. (2009). Hippocampal contributions to control: The third way. *Adv Neur Inf Proc Sys*.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psych Rev*, *102*(3), 419–57.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proc Natl Acad Sci*, *105*, 2745–50.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A More Rational Model of Categorization. *Proc 28th Ann Conf of the Cogn Sci Soc*, 1–6.

Sato, M.-a. (2001). Online Model Selection Based on the Variational Bayes. *Neural Comput*, *13*, 1649–81.

Snelson, E., & Ghahramani, Z. (2005). Compact approximations to Bayesian predictive distributions. *Proc 22nd Int Conf on Machine Learning*.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science*, *331*, 1279–85.