Change-Point Detection without Needing to Detect Change-Points?

Chaitanya K Ryali[1,*] and Angela J Yu[2]

[1] Computer Science and Engineering, University of California, San Diego, CA, USA
[2] Cognitive Science, University of California, San Diego, CA, USA

* rckrishna@ucsd.edu

Abstract

Online change-point detection is an important and much-studied problem in both neuroscience and machine learning. While most theoretical analysis has focused on this problem in the context of real-valued data, relatively little attention has been paid to the specific case when the observations are categorical (e.g. binary), even though the latter case is common in both neuroscience experiments and some engineering applications. In this paper, we focus on the latter scenario and demonstrate that, due to the information poverty of categorical data, near-Bayes-optimal data prediction can be achieved using a simple linear-exponential filter for binary data, or, more generally, $m-1$ separate linear-exponential filters for m-nary data. The computations are dramatically simpler than exact Bayesian inference, requiring only $O(m)$ computation per observation instead of $O(e^{km})$, where k depends on representation. We demonstrate how parameters of this approximation depend on the parameters of the generative model, and characterize the parameter regime in which the approximation can be expected to perform well, as well as how its performance degrades away from that regime. Interestingly, our results imply that, under appropriate parameter settings, change-point detection can be done near-optimally without the explicit computation of the probability of a change having taken place. Paradoxically, while detecting a change-point promptly based on sequentially presented categorical data is difficult, making near Bayes-optimal predictions about future data turns out to be quite simple. This work demonstrates that greater attention needs to be paid, in the context of online change-point detection, to a theoretical distinction between the problem of *predicting* future data and that of *deciding* that a change has taken place. With respect to neuroscience, our approximate algorithm is equivalent to the dynamics of an appropriately-tuned leaky integrating neuron with *constant gain*, or a particular variant of the delta learning rule with *fixed learning rate*, with obvious implications for the neuroscientific investigation of human and animal change-point detection.

Introduction

In recent years, there has been much progress in understanding how humans and animals learn about statistical regularities in the environment to make optimal decisions, as well as how they track changes in environmental statistics based on noisy data. Broadly speaking, this problem has been studied using two kinds of data, those that are continuous or ordinal-valued, and those that are binary or categorical. The first kind has often been modeled using variants of the Kalman filter, while the latter kind has been found to be successfully captured by the Dynamic Belief Model (DBM) [1], a

Bayesian hidden Markov model that assumes the observations to be iid distributed from a hidden variable, which itself goes occasional and discrete changes. We previously showed that DBM can explain a variety of behavioral phenomena: sequential adjustment effects in 2AFC discrimination tasks [1], inhibitory control (e.g. stop-signal) tasks [2], [3], and explicit prediction tasks, as well as providing a normative explanation for matching-type behavior in a multi-choice visual search task [4] and exploration stochasticity in multi-arm bandit tasks [5].

The problem of sequential estimation and prediction, while abrupt, un-signaled changes occur in the underlying generative parameters [6] is known as online change-point detection and is an important problem with many applications in both neuroscience and engineering. The DBM has been hypothesized to provide a normative explanation for the brain readily seeking temporal patterns and suggests an inadvertent engaging of online change-point detection mechanisms that would be of behavioural benefit in extracting patterns in a truly volatile world while the cost of such a belief being in error is small. However, the computational and representational complexity of exact inference in DBM and other change-point detection algorithm poses a challenge for explaining how the necessary computations can be implemented by the corresponding neural substrate. For engineering applications, one may also ask how the brain, with limited representational/computational power, manages to solve the online change-point detection problem effectively.

In this work, we identify the categorical nature of the data that DBM deals with as being crucial. Unlike change-point detection tasks in which the observations are real-valued or ordinal, categorical-valued data (e.g. binary) will be shown to be both a blessing and a curse. Based on relatively information-poor categorical data, it is quite difficult to detect a true change in the environment promptly, but it also means that the Bayesian update rule for predicting future outcomes may have a simple update form, such that the "learning rate" based on each new data point is nearly constant, since the posterior probability of a change just having happened is largely driven by the prior and is never very high.

We will show that for the regime of relatively frequent changes in the generative parameters, near-Bayes optimal data prediction can be achieved using a simple, linear-exponential filter for binary data, rigorously justifying a previous specific case [1] and extending it to a more general case using $m-1$ separate linear-exponential filters for m-nary data. The multi-alternative EXP extension demonstrates that near Bayes-optimal statistical inference can be achieved with no recurrent or otherwise complex interactions among the alternatives unlike previous proposals such as leaky competing integrators or MSPRT. The complexity of this approximation scales linearly with the number of categories of observed data, instead of exponential as in the case of the exact algorithm. We will also characterize the parameter regime in which the approximation can be expected to perform well, as well as how its performance degrades away from that regime.

One important theoretical implication of this work is that, as Bayesian prediction can be done nearly perfectly (when changes are relatively frequent) with a fixed learning rate pre-multiplying the prediction error in the update equation, there is no need to modulate the update rule with the estimated probability of a change having taken place, or indeed any need to compute this probability at all. For computational neuroscience, our work shows that Bayesian prediction in changeable contexts can be approximated by appropriately tuned leaky integrator dynamics. Our work also has important implications for experiment design in neuroscience and suggests that it is not ideal to use categorical observations if the goal is to discern behavioral and neural changes (e.g. LC or ACC) in response to detecting a true change in the environment, since the learning rate will not be modulated trial-by-trial based on the observations, as can be

expected for real-valued data. In contrast, exact inference for the Kalman Filter (KF), a popular model in machine learning, is a delta rule. However, a constant (no dependence on data) learning rate delta rule is inadequate as soon as change points are introduced. In [7] a neurally plausible solution for a switching KF that involved detecting changes and modulating the learning rate was proposed, but was for real-valued data and is not applicable to categorical data.

We will also relate the parameters of our approximation to the parameters of the generative model, thus allowing machine learning practitioners to analytically define the appropriate linear exponential filter approximation given the generative parameters; conversely, given human or animal behavioral data, one can fit a linear exponential filter to a subject's behavioral data and then infer the equivalent generative parameters assumed by the subject greatly simplifying data fitting in such tasks. Previously, [1] used both a Bayesian model and a delta-rule to fit the data, but did not analyze the relationship between the two. Finally, and more broadly our work makes an important point for anyone engaged in online data analysis: if the primary interest is in making predictions, then, under certain conditions (delineated here), it's not necessary to track change-points or have a probabilistic representation to make near-Bayes optimal predictions.

The rest of the paper is organized as follows: in section 1, we briefly describe the two learning models, DBM and the linear-exponential filter (EXP), motivate a parametric approximation to exact prediction in the DBM, discuss the implicit calculation of posterior over run length in calculating the predictive probabilities, and discuss why categorical data make both detecting a change hard while predicting the future easy. In section 2, we justify the approximation for the binary case, rigorously extending the specific case suggested in [1], and then extend the approximation to m-nary categorical observations. We also characterize the relationship between DBM and EXP. We conclude (in section 3) with a discussion of broader implications of the work and future directions of research.

1 Learning models for categorical data

1.1 The Dynamic Belief Model (DBM)

1.1.1 Generative Model

The Dynamic Belief Model (DBM) is a Bayesian hidden Markov model that assumes categorical observations to be drawn from parametric distributions whose parameters themselves occasionally change was introduced in [1] and was used to provide a normative explanation for sequential effects in 2AFC discrimination tasks. DBM has found much success in reproducing human data in a variety of behavioral tasks, e.g. inhibitory control (e.g. stop-signal) tasks [2], [3], explicit prediction tasks, multi-choice visual search task [4] and in multi-arm bandit tasks [5]. The framework of switching experts for online learning of non stationary sequences [8] is a closely related model for prediction using expert advice when the the best expert may change at some un-signaled time step, which finds many applications in engineering.

In this section, we assume that observations are binary and the probability of observing unity is equal to the hidden variable γ. Here, the prior belief $p_0(.)$ on γ will be assumed to be a Beta distribution: $p_0(\gamma) = \text{Beta}(\gamma; a, b)$. In later sections, we offer a generalization to categorical data with priors of Dirichlet form. The hidden variable γ_t has a Markovian dependence on γ_{t-1}:

$$p(\gamma_t = \gamma|\gamma_{t-1}) = \alpha\delta(\gamma - \gamma_{t-1}) + (1 - \alpha)p_0(\gamma), \tag{1}$$

i.e., the hidden variable remains the same ($\gamma_t = \gamma_{t-1}$) with a fixed probability α or is redrawn from the prior $p_0(\gamma)$. Binary observations x_t are generated as Bern(γ_t), so that the likelihood is given by $p(x_t|\gamma_t) = \gamma_t^{x_t}(1-\gamma_t)^{\bar{x}_t}$, where $\bar{x}_t = 1 - x_t$.

1.1.2 Recognition

Given the observations, the prior probability for trial t and the posterior upon observing x_t may be recursively computed as

$$p(\gamma_t = \gamma|x_{1:t-1}) = \alpha p(\gamma_{t-1} = \gamma|x_{1:t-1}) + (1-\alpha)p_0(\gamma_t = \gamma), \tag{2}$$

$$p(\gamma_t|x_{1:t}) \propto p(x_t|\gamma_t)p(\gamma_t|x_{1:t-1}). \tag{3}$$

1.1.3 Exact prediction

The predictive probability for trial $t+1$, given the past observations $x_{1:t}$ is computed as

$$P_{t+1} \triangleq P(x_{t+1} = 1|x_{1:t}) = \int \gamma p(\gamma_{t+1} = \gamma|x_{1:t})d\gamma = \langle\gamma\rangle_{p(\gamma_{t+1}|x_{1:t})}. \tag{4}$$

As previously discussed, computing the exact predictive probabilities typically involves calculating the posteriors on change points either directly (as in [6]) or indirectly, followed by marginalization to make predictions. This will be evident in calculations that follow, which will allow us to recursively calculate exact predictive probabilities.

Direct Calculation The calculation is similar to one in [6] and maybe consulted for more detail. The run length at time t is denoted by r_t and is defined as the time since last change in the hidden bias and can range from 0 to $t-1$. A run length zero, i.e $r_t = 0$, means that a change occurred at time t and x_t is a sample from a new bias γ_t redrawn from $p_0(\gamma)$. We decompose the prior distribution $p(\gamma_{t+1}|x_{1:t})$ as follows:

$$p(\gamma_{t+1}|x_{1:t}) = \sum_{r_{t+1}=0}^{t} p(\gamma_{t+1}|r_{t+1}, x_{1:t})P(r_{t+1}|x_{1:t}) \tag{5}$$

$$= (1-\alpha)p_0(\gamma_{t+1}) + \sum_{r_{t+1}=1}^{t} p(\gamma_{t+1}|r_{t+1}, x_{t-r_t:t}) \sum_{r_t=0}^{t-1} P(r_{t+1}|r_t)P(r_t|x_{1:t}) \tag{6}$$

$$= (1-\alpha)p_0(\gamma_{t+1}) + \alpha \sum_{r_t=0}^{t-1} p(\gamma_{t+1}|r_{t+1}, x_{t-r_t:t})P(r_t|x_{1:t}). \tag{7}$$

where $p(\gamma_{t+1}|r_{t+1}, x_{t-r_t:t}) = \text{Beta}(\gamma_{t+1}; a + \sum_{i=t-r_t}^{t} x_i, b + \sum_{i=t-r_t}^{t} \bar{x}_i)$. The prior probability distribution is a mixture of beta distributions with the posterior probability on run length, $P(r_t|x_{1:t})$, determining the mixture weights, which makes intuitive sense. An unchanging hidden bias corresponds to $\alpha = 1$ and it's easily seen that the prior probability distribution is $\text{Beta}(\gamma_{t+1}; a + \sum_{i=1}^{t} x_i, b + \sum_{i=1}^{t} \bar{x}_i)$ and the predictive probability $P_{t+1} = \frac{\sum_{\tau=1}^{t} x_\tau + a}{t+a+b}$. On the other extreme, $\alpha = 0$ corresponds to change on every trial and the prediction can be no better than the forecast by the prior $p_0(\gamma)$ giving $P_{t+1} = \frac{a}{a+b}$ on every trial.

The posterior on run length can be calculated as $p(r_t|x_{1:t}) = \frac{p(r_t, x_{1:t})}{\sum_{r_t} p(r_t, x_{1:t})}$ and the joint recursively as:

$$P(r_t, x_{1:t}) = \sum_{r_{t-1}} P(r_{t-1}, x_{1:t-1}) P(r_t|r_{t-1}) P(x_t|r_t, r_{t-1}, x_{1:t-1}).$$

Simplifying, the recursion maybe be written as

$$P(r_t, x_{1:t}) = \begin{cases} \alpha P(r_{t-1} = r_t - 1, x_{1:t-1}) \langle\gamma\rangle_{q(\gamma_t)}^{x_t} \langle\bar{\gamma}\rangle_{q(\gamma_t)}^{\bar{x}_t}, & r_t \neq 0 \\ (1-\alpha) \sum_{r_{t-1}} P_0(x_t) P(r_{t-1}, x_{1:t-1}), & r_t = 0 \end{cases} \tag{8}$$

where $q(\gamma_t) = \text{Beta}(\gamma; a + \sum_{\tau=t-r_t}^{t} x_\tau, b + \sum_{\tau=t-r_t}^{t} \bar{x}_\tau)$ and $P_0(x_t) = P_0^{x_t}(\bar{P}_0)^{\bar{x}_t}$.

Indirect Calculation The exact, non-linear Bayesian update rule for the predictive probability P_{t+1} may also be written as:

$$P_{t+1} = (1-\alpha)\langle\gamma\rangle_{p_0(\gamma)} + \alpha x_t \frac{Q_t - P_t^2}{P_t(1-P_t)} + \alpha P_t \frac{P_t - Q_t}{P_t(1-P_t)} \tag{9}$$

$$= (1-\alpha)P_0 + \alpha x_t G_t + \alpha P_t(1-G_t), \tag{10}$$

where $Q_t = \langle\gamma^2\rangle_{p(\gamma_t|x_{1:t-1})}$ and $G_t = \frac{\text{var}(\gamma_t|x_{1:t-1})}{\text{var}(\text{Bern}(P_t))} = \frac{Q_t - P_t^2}{P_t(1-P_t)}$. The marginalization over change points, which was explicit in the previous calculation, occurs implicitly in this calculation.

We pay particular attention to the term G_t which governs the trade-off between new data x_t and (a statistic of) past data, P_t. Intuitively, the trade-off G_t, is modulated by how "surprising" recent observations are found to be. Noting that $0 \leq G_t \leq 1$ for any $x_{1:t-1}$ (for all t) and rewriting the update rule as

$$P_{t+1} = (1-\alpha)P_0 + \alpha(P_t + G_t(x_t - P_t)), \tag{11}$$

we may interpret the update rule as a delta rule with learning rate G_t. We expect to see an increase in the learning rate G_t if recent observations are found to be "surprising", i.e have high prediction error, which could signal a switch in environment statistics, promting an increase in the learning rate. Since categorical data are information-poor, it is quite difficult to detect a true change in the environment promptly, but it also means that the Bayesian update rule for predicting future outcomes may have a simple update form, since the posterior probability of a change which drives the learning rate is largely driven by the prior and is never very high. The key idea is that even though the distribution $p(\gamma_{t+1}|x_{1:t})$ is a messy, mixture distribution with the mixture weights $P(r_t|x_{1:t})$ being difficult to compute, G_t is well behaved and can be approximated by a constant in the regime of frequent changes. In Figure 2C, we see an increase in G_t following true and "perceived" (spikes in the posterior probability of change at certain time) changes in underlying generative parameters in a sample run. The inset shows an increase in average G_t following a true change.

We will now briefly motivate an approximation for small α that will be rigorously justified in a later section. As discussed, the exact DBM model employs an adaptive learning rate G_t that is a function of the model parameters α, a, b and modulated by the data $x_{1:t}$. If we wished to approximate the adaptive learning rate by a fixed constant for the regime of frequent changes that is dependent on the model parameters alone, one natural path is to approximate G_t by the constant term in its series expansion around $\alpha = 0$. Setting $\alpha = 0$ and inserting moments of the prior into the expression for G_t, we see that $G_t = \frac{1}{a+b+1} + O(\alpha)$.

1.2 Linear Exponential Filtering (EXP)

EXP is a simple, non-Bayesian algorithm that linearly sums past observations, while exponentially discounting into the past, to predict the probability of encountering different trial types on the next trial. This model was introduced in relation with the DBM in [1], inspired by related work showing that monkeys choices, when tracking reward biases that change at un-signaled times, depend linearly on previous observations and are discounted in an approximately exponential fashion.

This model may be written as

$$P_{t+1} = P(x_{t+1} = 1|x_{1:t}) = C + \eta \sum_{\tau=0}^{t-1} \beta^\tau x_{t-\tau} = C(1-\beta) + \eta\beta x_t + \beta P_t,$$

where the three parameters of the model (C, η, β) are constrained as $0 \leq C, \eta \leq 1$, $0 \leq \beta < 1$, $C + \frac{\eta\beta}{1-\beta} < 1$. In the next section, we will relate the three parameters of the DBM (α, a, b) to those of EXP.

2 Approximation and Implications

2.1 Relationship of EXP to DBM

We will show that the approximate update rule is in the form of the linear exponential filter, which can be thought of as approximately implementing DBM with a constant learning rate. We will show that the equivalent parameters are $\beta = \alpha\frac{a+b}{a+b+1}$, $\eta = \frac{1}{a+b}$ and $C = \frac{(1-\alpha)P_0}{(1-\beta)}$. In particular, for $a = 1, b = 1$ which is the the uniform prior, we have $\beta \approx \frac{2}{3}\alpha$, which matches the previous known empirical result in [1].

The parameter C is the lower bound on P_t determined by the α and the prior $p_0(\gamma)$, which would be attained asymptotically in the limit of observing infinite 0's. Similarly, the upper bound in the limit of observing infinite 1's is $C + \frac{\eta\beta}{1-\beta}$. This sort of bounded behavior is not specific to the linear exponential filter and and is carried over from the DBM.

2.2 2-alternative approximation

We rewrite the exact, non linear, Bayesian update rule for the predictive probability P_{t+1} as:

$$P_{t+1} = (1-\alpha)P_0 + \alpha x_t G_t + \alpha P_t(1-G_t) = (1-\alpha)P_0 + \alpha L_t.$$

Similar to the derivation of the exact update rule for P_{t+1}, we obtain

$$Q_{t+1} = (1-\alpha)Q_0 + \alpha x_t \frac{R_t - Q_t P_t}{P_t(1-P_t)} + \alpha Q_t \frac{Q_t - R_t}{Q_t(1-P_t)},$$

where $R_t = \langle\gamma^3\rangle_{p(\gamma_t|x_{1:t-1})}$.

We focus on the term $1 - G_t$ and expand it around $\alpha = 0$. First, note that

$$P_t(1-P_t) = P_0\bar{P}_0 + \alpha[\bar{P}_0 L_{t-1} + P_0\bar{L}_{t-1} - 2P_0\bar{P}_0] + \alpha^2[L_{t-1}\bar{L}_{t-1} - L_{t-1}\bar{P}_0 - \bar{P}_0 L_{t-1}],$$

where $\bar{P}_0 = 1 - P_0$ and $\bar{L}_{t-1} = 1 - L_{t-1}$. We set
$P_{t-1} = P_0 + O(\alpha), Q_{t-1} = Q_0 + O(\alpha), R_{t-1} = R_0 + O(\alpha)$ and obtain

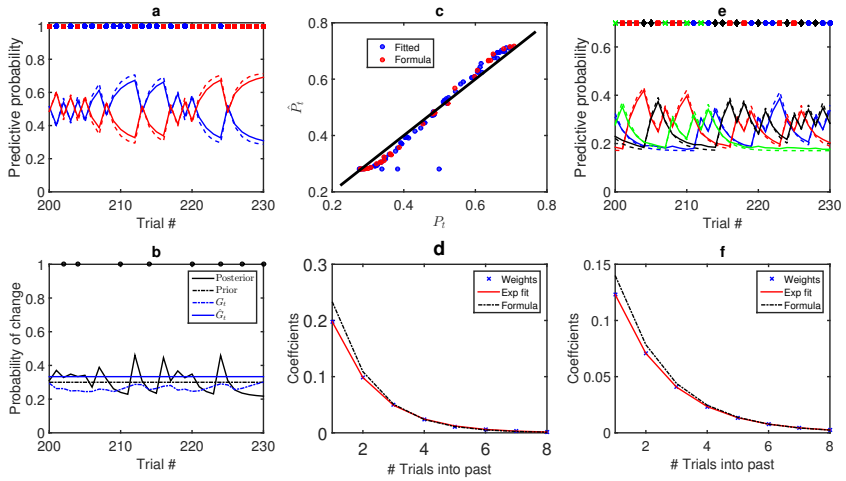

Fig 1. a. Visualization of the approximation and exact predictive probability for $M = 2, \alpha = 0.7$ and $p_0(\gamma) = \text{Beta}(\gamma; 1, 1)$, data was generated by DBM with the same parameters b. Visualization of the prior and posterior probability of change at each time. (Black circles mark actual changes). We also plot exact G_t and $\hat{G}_t$ the approximation by $\frac{1}{a+b+1}$. c. Scatter plot of the approximate predictive probability $\hat{P}_t$ vs exact P_t. We note that the quality of the approximation by the best fitted exponential and the linear exponential filter determined by the relationships we derived are similar. d. The weights were determined by regression and an exponential was fitted to the regression weights. For small α, the best fitted exponential filter is close to the one determined by the derived relationship between DBM and EXP. e. We demonstrate the validity of the generalization of our approximation by visualizing the exact and approximate predictive probabilities for $M = 4, \alpha = 0.75$ and $p_0(\bar{\gamma}) = \text{Dir}(\gamma; 1, 1, 1, 1)$ where data was generated from DBM using the same parameters. f. We note once again that the best fitted exponential closely matches the parameters determined by our approximation. We wish to point out that although the data for these examples was drawn from DBM with the same parameters being used to approximate the exact calculations of predictive probability, our approximation remains valid for all observation sequences.

$$P_t(1-P_t) = P_0\bar{P}_0 + \alpha[(\frac{x_{t-1}+a}{a+b+1})\frac{b}{a+b} + (\frac{\bar{x}_{t-1}+b}{a+b+1})\frac{a}{a+b}] + O(\alpha^2).$$

Once again, setting $P_{t-1} = P_0 + O(\alpha), Q_{t-1} = Q_0 + O(\alpha), R_{t-1} = R_0 + O(\alpha)$, we obtain

$$P_t - Q_t = (P_0 - Q_0) + \alpha\frac{(a-b)}{(a+b)(a+b+1)(a+b+2)}(a\bar{x}_{t-1} - bx_{t-1}) + O(\alpha^2).$$

We conclude that

$$1-G_t = \frac{P_t - Q_t}{P_t(1-P_t)} = \frac{a+b}{a+b+1} + \alpha\frac{a^2-b^2}{ab(a+b+1)^2(a+b+2)}(-a\bar{x}_{t-1} + bx_{t-1}) + O(\alpha^2). \quad (12)$$

Linear approximation Observe that absolute value of the coefficient of α in G_t may be upper bounded by a small value of 0.062 allowing us to approximate G_t by a constant for small α. We also note that the $O(\alpha)$ coefficient in G_t only depends on the

sample x_{t-1} which makes intuitive sense. Approximating G_t by the constant $\frac{1}{a+b+1}$ gives us the approximate, *linear* update rule

$$P_{t+1} = (1-\alpha)P_0 + \frac{\alpha x_t}{a+b+1} + \alpha P_t \frac{a+b}{a+b+1}, \tag{13}$$

which is exact to $O(\alpha^2)$. We observe that for $a = b$, in particular, for the uniform prior $a = b = 1$, the order α term in $1 - G_t$ is exactly zero, so that the approximate linear update rule in the predictive probability is exact to $O(\alpha^3)$. Obviously, this small α approximation gets progressively worse as α approaches 1. In fact, at $\alpha = 1$ observe that $G_t = \frac{1}{a+b+t}$.

Normalization We can come up with a similar, linear, approximate update rule for the other categorical variable. We shall use a superscript $i \in \{0, 1\}$ to denote the update rules for the respective categories. At $t = 1$, $P_1^{(1)} = P_0$ and $P_1^{(0)} = \bar{P}_0$, so $P_1^{(0)} + P_1^{(1)} = 1$. Suppose that at some time step $t = \tau$ that $P_\tau^{(0)} + P_\tau^{(1)} = 1$ then,

$$P_{\tau+1}^{(0)} + P_{\tau+1}^{(1)} = (1-\alpha)(P_0 + \bar{P}_0) + \alpha\frac{x_{\tau+1}^{(0)} + x_{\tau+1}^{(1)}}{a+b+1} + \alpha\frac{a+b}{a+b+1}(P_\tau^{(0)} + P_\tau^{(1)}) = 1.$$

Interestingly, even though we though we maintain estimates for each category separately, normalization is preserved.

"Frequent" Changes: Non-triviality A few comments are in order on the non-triviality of the regime of frequent changes. "Frequent" change regime does not mean draws from the prior on (nearly) every iteration. As shown in Figure 2c;d, while the EXP is exact for alpha=0, it works well for alpha as large as 0.8-0.95 (Fig. 2d), meaning on average a change every 5-20 trials. Secondly, the approx. model (as well as DBM) can be seen to depend systematically on recent history (Figure 1a;e: alpha=0.7; Figure 2a: alpha=0.9), and is very different from solely relying on the prior. This is a key result of the paper: not needing (or being able) to detect change-points promptly does not mean ignoring recent data when making future predictions; it means the way that data is used should (or can) be kept constant over time.

Equivalence to Leaky Integration The EXP model is exactly equivalent to a correctly tuned leaky integrating neuron, with $\eta\beta x_t$ as the input, βP_t as the recurrent term and $C(1-\beta)$ as the bias term. The computational and representational complexity of the DBM posed a challenge in terms of neural implementability which is solved by the approximation. The EXP approximation to the DBM shows that a leaky integrating neuron can implement near Bayes optimal inference cheaply and effectively in a regime of frequent changes.

Humans have been shown to internalize the volatility of the environment and modulate their learning rate accordingly [9] which prompts the question of how subjects learn the volatility of the environment and adopt an appropriate learning rate. The exact Bayes optimal computation to update α

$$p(\alpha, \gamma_t | x_{1:t-1}) \propto p(\alpha | x_{1:t-1}) p(x_t | \gamma_t) p(\gamma_t | \alpha, x_{1:t-1}), \tag{14}$$

is plagued by the same problems as the DBM in terms of neural implementation. However, the EXP approximation to the DBM permits an approximate update rule for α via stochastic gradient descent:

$$\hat{\alpha} \leftarrow \hat{\alpha} + \epsilon(x_t - \hat{P}_t)\frac{d\hat{P}_t}{d\hat{\alpha}} \tag{15}$$

$$\frac{d\hat{P}_t}{d\hat{\alpha}} = \hat{P}_{t-1} + \hat{G}_{t-1}(x_t - \hat{P}_{t-1}) - P_0. \tag{16}$$

Implications to Experiment Design and Data Fitting As near optimal Bayesian prediction in changeable contexts can be approximated by appropriately tuned leaky integrator dynamics with constant gain, our work has important implications for experiment design in neuroscience and suggests that it is not ideal to use categorical observations if the goal is to discern behavioral and neural changes (e.g. LC or ACC) in response to detecting a true change in the environment, since the learning rate will not be modulated trial-by-trial based on the observations, as can be expected for real-valued data.

Our work also simplifies model fitting in behavioural data: it's not necessary to fit the complicated DBM and the easy-to-fit EXP model can be fitted instead. The parametric relationship between the DBM and EXP approximation allows us to infer the parameters of the appropriate DBM model.

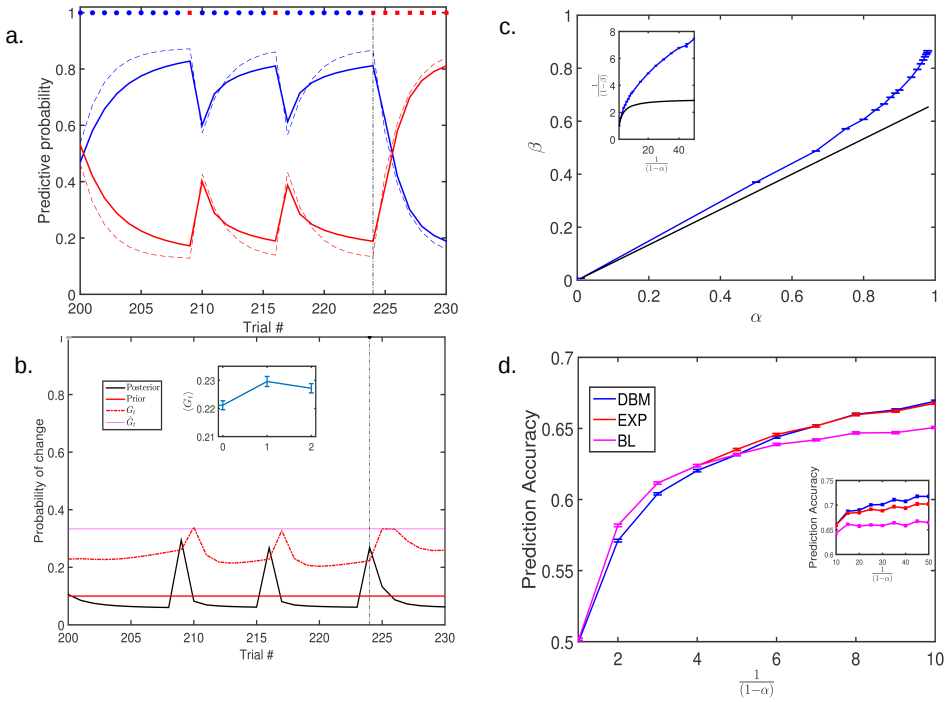

Fig 2. a. We visualize the slight degradation of performance of the approximation for larger α, here $\alpha = 0.9$, true change point at $t = 224$ is marked by the vertical dashed line. b. We visualize the prior and posterior probabilities of change at each time step and see that the posterior probability of change frequently spikes, even when there is no true change. The learning rate G_t rises in response to these true or perceived changes. Inset shows an increase in the average value of G_t following a true change. c. We charecterize how the best exponential fit deviates from our approximation for larger α. Inset show the same plot in the scale of $\frac{1}{(1-\alpha)}$ which corresponds to the average run length before a change occurs. d. We characterize the degradation of our approximation in terms of prediction accuracy and see that even for larger values of α, the prediction accuracy of the approximation remains competitive with DBM, the exact generative model. BL is a baseline comparison and just predicts the previous observation.

2.3 M-Alternative approximation

First, we note that the DBM is easily and naturally extended to m-categorical case as follows: $p_0(\gamma) = \mathrm{Dir}(\mathbf{a})$, where $\mathbf{a} = (a_0, \ldots, a_{m-1})$. The likelihood function is the categorical distribution, $p(x_t|\gamma_t) = \prod_{i=0}^{m} \gamma_{i,t}^{[x_t=i]}$.

For each category i, the variable of interest is $x_t^{(i)}$, which is the binary random variable such that $x_t^{(i)} = 1$ only if $x_t = i$ and zero otherwise. We see that

$$1-G_t^{(i)} = \frac{\sum a_k}{\sum a_k + 1} + \alpha \frac{a_i^2 - (\sum_{k\neq i} a_k)^2}{a_i(\sum_{k\neq i} a_k)(\sum a_k + 1)^2(\sum a_k + 2)}(-a_i\bar{x}_{t-1} + (\sum_{k\neq i} a_k)x_{t-1}) + O(\alpha^2) \tag{17}$$

and the predictive probability update rule is given by

$$P_{t+1}^{(i)} = (1-\alpha)P_0^{(i)} + \frac{\alpha x_t^{(i)}}{\sum a_k + 1} + \alpha P_t^{(i)} \frac{\sum a_k}{\sum a_k + 1}, \tag{18}$$

is exact to $O(\alpha^2)$ and $P_0^{(i)} = a_i / \sum a_k$.

Novelty of M-alternative EXP We note once again that even though we maintain estimates for each $P_t^{(i)}$ separately, normalization is preserved, $\sum_i P_t^{(i)} = 1$. Since nearly identical bounds on coefficient of $O(\alpha)$ in G_t hold for $m > 2$, the performance of the approximation with be similar to $M = 2$. In the regime of relatively frequent changes, near-Bayes optimal data prediction can be achieved simply using $m - 1$ separate linear-exponential filters with no recurrent or otherwise complex interactions among the alternatives unlike previous proposals such as leaky competing integrators or MSPRT [cite]. The complexity of this approximation scales linearly with the number of categories of observed data, instead of exponential as in the case of the exact algorithm.

3 Discussion

The change-point detection problem has been extensively studied in many neuroscience and machine learning contexts. However, not much attention has been paid to the distinction between categorical (e.g. binary) observations and real-valued data, or other kinds of data (e.g. ordinal) that reveal with a single data point *how surprising* an observation is given the prior expectations. With the case of categorical data, the only information when an unusual observation occurs is that it differs from the most probable outcome and there is no differentiation in the *degree* of unusualness. In this paper, we focus on the categorical data and present a very simple linear exponential filter that requires only $O(m)$ computations for sequential predictions, instead of $O(e^{km})$ in the exact model. The approximation is inspired by the insight that since the online estimate of probability of change can't fluctuate very much due to the lack of information in the data, then the update rule might as well be a simple linear form that uses a constant learning rate to weigh the prediction error. We demonstrated the relationship between parameters of the approximation to the parameters of the generative model, and characterized the parameter regime in which the approximation can be expected to perform well, as well as how its performance degrades away from that regime.

Interestingly, the approximation is exactly equivalent to the dynamics of an appropriately tuned leaky integration model. Since Bayesian learning is achieved to a very good approximation, or equivalently a fixed gain in the leaky integrating equation, with no need to explicitly compute the probability of a change having taken place, this implies that not much can be gained by tracking the posterior probability of change, nor

is it ideal to use categorical data in a change-point detection task if the goal is to discern behavioral and neural changes in response to detecting a true change in the environment since the learning rate will not be modulated trial-by-trial based on the observations, like we would expect when the data is real-valued (e.g. [10]).

This work has obvious implications for both neuroscience and machine learning applications.

References

1. Yu AJ, Cohen JD. Sequential effects: Superstition or rational behavior? Advances in Neural Information Processing Systems. 2009;21:1873–80.

2. Shenoy P, Yu AJ. Rational decision-making in inhibitory control. Frontiers in Human Neuroscience. 2011;.

3. Ide JS, Shenoy P, Yu* AJ, Li* CSR. Bayesian prediction and evaluation in the anterior cingulate cortex. Journal of Neuroscience. 2013;33:2039–2047.

4. Yu AJ, Huang H. Maximizing masquerading as matching: Statistical learning and decision-making in choice behavior. Decision. 2014;1(4):275–287.

5. Zhang S, Yu AJ. Forgetful Bayes and myopic planning: Human learning and decision-making in a bandit setting. Advances in Neural Information Processing Systems. 2013;26.

6. Adams RP, MacKay AJ. Bayseian Online changepoint detection. Tech Report. 2007;.

7. Yu AJ, Dayan P. Expected and Unexpected Uncertainty: ACh and NE in the Neocortex. In: S Becker ST, Obermayer K, editors. Advances in Neural Information Processing Systems 15. Cambridge, MA: MIT Press; 2003. p. 157–164.

8. Jaakkola T, Monteleoni C. Online Learning of Non-stationary sequences. Advances in Neural Information Processing Systems. 2004;.

9. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS. Learning the value of information in an uncertain world. Nature Neurosci. 2007;10(9):1214–21.

10. Nassar MR, Rumsey KM, Wilson RC, Parikh K, Heasly B, Gold JI. Rational regulation of learning dynamics by pupil-linked arousal systems. Nature Neuroscience. 2012;15(7):1040–1046.