Computational Brain & Behavior (2019) 2:95–108
https://doi.org/10.1007/s42113-019-00026-1

Velocity Estimation in Reinforcement Learning

Carlos Velázquez[1] · Manuel Villarreal[1] · Arturo Bouzas[1]

Published online: 6 February 2019

Abstract

The current work aims to study how people make predictions, under a reinforcement learning framework, in an environment that fluctuates from trial to trial and is corrupted with Gaussian noise. We developed a computer-based experiment where subjects were required to predict the future location of a spaceship that orbited around planet Earth. Its position was sampled from a Gaussian distribution with the mean changing at a variable velocity and four different values of variance that defined our signal-to-noise conditions. Three reinforcement learning algorithms using hierarchical Bayesian modeling were proposed as candidates to describe our data. The first and second models are the standard delta-rule and its Bayesian counterpart, the Kalman Filter. The third model is a delta-rule incorporating a velocity component which is updated using prediction errors. The main advantage of the later model over the first two is that it assumes participants estimate the trial-by-trial changes in the mean of the distribution generating the observations. We used leave-one-out cross-validation and the widely applicable information criterion to compare the predictive accuracy of the models. In general, our results provided evidence in favor of the model with the velocity term and showed that the learning rate of velocity and the decision noise change depending on the value of the signal-to-noise ratio. Finally, we modeled these changes using an extension of its hierarchical structure that allows us to make prior predictions for untested signal-to-noise conditions.

Keywords Reinforcement learning · Dynamic environments · Velocity estimation · Bayesian methods · Hierarchical modeling

Introduction

Decisions often take place in environments that change over time. Availability of food in foraging animals may vary gradually as a function of source growth or continuous intake, likewise, the position of objects in space can change at a certain rate. Having accurate estimates of relevant variables under these circumstances allows behavior to be better allocated. For example, by changing to a richer foraging location or predicting the correct position of an object moving towards us. In stable environments, prediction error models accomplish this task by reducing the discrepancy of estimates and outcomes as new observations arrive. A straightforward expression to compute this is the Delta-rule:

$$\hat{V}_{t+1} = \hat{V}_t + \alpha\delta_t \tag{1}$$

where, the estimate at time $t + 1$ ($\hat{V}_{t+1}$), depends on the previous estimate ($\hat{V}_t$) and the prediction error (δ_t), weighted by the learning rate parameter (α). Evidence from Experimental Psychology (Dayan and Nakahara 2018; Miller et al. 1995; Rescorla and Wagner 1972; Bush and Mosteller 1951) and Neuroscience (Daw and Tobler 2014; Niv 2009; Schultz et al. 1997) provides support for this algorithm as a plausible mechanism of learning in mammals, and it has also been implemented as an effective solution in multiple machine learning problems (Sutton 1998). However, one of its limitations is the inability to describe behavior in non-stationary environments, partly, due to the fixed nature of the learning rate parameter (O'Reilly 2013). For example, in change-point problems, having a low α makes predictions during stable periods accurate but causes a slow adaptation after a change. A high

Supplementary material of this article, including code and data, is available as a project page on the Open Science Framework at https://osf.io/d6tjw/. A preliminary version of this work was presented at the 51st Annual Meeting of the Society for Mathematical Psychology in 2018.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s42113-019-00026-1) contains supplementary material, which is available to authorized users.

✉ Carlos Velázquez
carlos.unamlab25@gmail.com

[1] Universidad Nacional Autónoma de México, Avenida Universidad 3004, Coyoacán, Col. Copilco Universidad, 04510 Ciudad de México, México

α has the opposite effect, making inaccurate predictions during stability but having a quick adaptation to changes. Adjusting this parameter after the change-point (Nassar et al. 2010) and using multiple delta-rules with their own learning rates (Wilson et al. 2013) are some of the possible solutions that have been proposed. On the other hand, when the environment changes gradually over trials such as in a random walk process, the learning rate is assumed to vary as function of the relative uncertainty in the estimates and the outcomes as expressed in the Kalman filter equations (Kalman 1960; Navarro et al. 2018; Zajkowski et al. 2017; Speekenbrink and Konstantinidis 2015; Speekenbrink and Shanks 2010; Gershman 2017, 2015; Kakade and Dayan 2002). An important limitation of this approach is that, when trial-to-trial changes are large (i.e., the rate of change is high), the learning rate asymptotes at values close to one (Daw and Tobler 2014), making the model extremely sensitive to outcome noise. This problem is likely to occur because there is not an explicit computation of the rate of change of the environment.

In this work, we show that when the environment is changing at certain rate, a concrete estimation of this variable is necessary to guide decisions. Additionally, we show that the updating process of the rate of change is influenced by the level of noise in the observations as expressed by the signal-to-noise ratio (S/N). Previous research has shown that people are sensitive to higher-order statistics of the environment such as the volatility (O'Reilly 2013; Behrens et al. 2007) or the functions controlling changes (Ricci and Gallistel 2017) and that they are able to adapt their behavior accordingly.

In our experiment, subjects were required to predict the angular location of a spaceship moving around planet Earth. Its position was generated from a Gaussian distribution with the mean changing at a variable velocity (rate of change for position) and four values of variance that defined the S/N conditions. We proposed a reinforcement learning model incorporating a velocity component to describe participants predictions throughout the task. The main assumption of the model is that prediction errors are used to update an estimate of the velocity of change in the outcomes mean, which is later incorporated to the computations of new predictions.

We compared the performance of this model at describing behavior with the standard delta-rule and its Bayesian counterpart, the Kalman Filter. Importantly, all models were built using a hierarchical Bayesian structure where individual parameters were generated from Gaussian distributions defined at the level of conditions. In general, hierarchical modeling allows to specify the generative process of relevant psychological variables rather than assuming they simply exist (Shiffrin et al. 2008; Lee 2018). One of the main advantages of this type of models is their ability to generalize the results to new conditions or participants (Lee 2018). In the current work, we initially assume hierarchies that allow all models to make predictions of new subjects on each condition. After showing that the model with the velocity component outperforms the other two, we extended its structure to allow predictions for untested S/N values. In particular, we assumed that the means of the Gaussian distributions for two of the model parameters (the rate of learning for the velocity component and the decision noise) followed a hyperbolic function of the S/N values.

Our results show that errors between the generative mean of the spaceship and participants' predictions remain close to zero in the four conditions and that accuracy increases with the S/N. The model-based analysis indicates that a prediction error model incorporating a velocity component is better at describing participants' behavior than the standard delta-rule and the Kalman Filter. A formal model comparison, using a recent approach to leave-one-out cross-validation developed by Vehtari et al. (2017) and the Widely Applicable Information Criterion (WAIC), also suggests that this model has the best predictive power. Furthermore, we found that the extended version of the wining model is able to make sensible predictions about a new participant in the four conditions and potentially to new S/N values. We further discuss the implications of our findings for reinforcement learning models and alternative approaches to similar prediction problems.

Learning Models

We evaluated three error-driven algorithms using hierarchical Bayesian modeling (Lee 2018; Shiffrin et al. 2008). Hierarchical models assume that individual parameters are generated from higher-order distributions, e.g., placed at the level of populations or experimental conditions (Lee 2018; Shiffrin et al. 2008). Some of their applications involve modeling individual differences assuming participants are not completely independent (Pratte and Rouder 2011), and predicting behavior of a new subject based on the information of the population (Lee 2018; Shiffrin et al. 2008). As it is frequently done given its simple interpretation for variability (Zajkowski et al. 2017; Matzke et al. 2015; Lee 2018), we assumed that the higher-order distributions were Gaussian. Importantly, these distributions were set at the level of conditions implying that what is learned from one participant in a given condition affects what is learned about the rest in the same condition.

Standard Delta-Rule (SD) As specified in Eq. 1, this model updates a variable $\hat{V}$ using prediction errors weighted by a learning rate. We now assumed a decision rule in which the behavior for trial $t+1$, B_{t+1}, is generated from a Gaussian

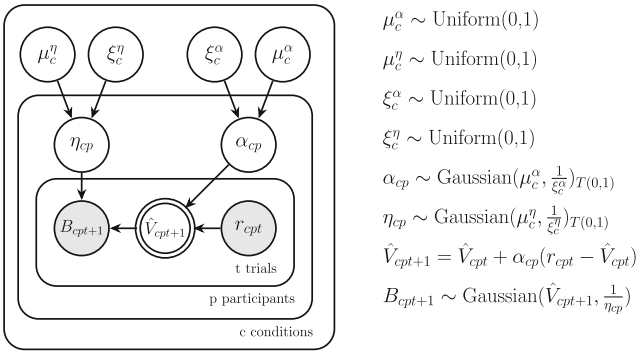


Fig. 1 Graphical representation of a hierarchical delta-rule

distribution with mean $\hat{V}_{t+1}$ and precision $\frac{1}{\eta}$ (where η is the variance of the distribution and represents decision noise).[1] Formally:

$$\begin{aligned} \hat{V}_{t+1} &= \hat{V}_t + \alpha\left(r_t - \hat{V}_t\right) \\ B_{t+1} &\sim \mathit{Gaussian}\left(\hat{V}_{t+1}, \frac{1}{\eta}\right) \end{aligned} \tag{2}$$

where r_t is the observed outcome in trial t. The learning rate α and the decision noise η are generated from Gaussian distributions with hyperparameters $\left(\mu_c^\alpha, \frac{1}{\xi_c^\alpha}\right)$ and $\left(\mu_c^\eta, \frac{1}{\xi_c^\eta}\right)$, respectively, for each experimental condition c. Figure 1 is the graphical representation of this model. In this notation, nodes correspond to variables and arrows connecting them refer to dependencies. Shaded nodes are observed variables, whereas unshaded nodes are latent variables. Stochastic and deterministic variables are represented using single- and double-boarded nodes, respectively, and continuous variables are represented using circular nodes. Plates refer to replications of the process inside them. On the right-hand side of the graphic, we show the detailed relations among variables and the prior distributions of the hyperparameters.

Although a useful model of animal and machine learning, the core structure of Eq. 2 has difficulties performing under changing conditions (Ritz et al. 2018; Ricci and Gallistel 2017; Gallistel et al. 2014; Wilson et al. 2013; Nassar et al. 2010). In particular, for the purpose of this work, we will emphasize that it is unable to track potential trends (e.g., a velocity) underlying the data.

Delta-Rule with Velocity Term (VD) This model incorporates to SD an estimate of the trial-by-trial change (labeled as velocity) in the generative process. VD model consists of an update equation and a prediction equation. The update equation is expressed as:

$$\underbrace{\mathbf{v}_t}_{\text{Updated vector}} = \overbrace{\hat{\mathbf{v}}_t}^{\text{Prediction vector}} + \underbrace{\mathbf{a}}_{\text{Learning rate vector}} \overbrace{(r_t - \mathbf{H}\hat{\mathbf{v}}_t)}^{\text{Prediction error}} \tag{3}$$

where $\mathbf{H} = [1\ 0]$, $\mathbf{v}_t = \begin{bmatrix} V_t \\ V_t' \end{bmatrix}$, $\hat{\mathbf{v}}_t = \begin{bmatrix} \hat{V}_t \\ \hat{V}_t' \end{bmatrix}$ and $\mathbf{a} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$. V_t and V_t' are the updated values for the position and velocity, respectively, after the outcome r_t is observed. $\hat{V}_t$ and $\hat{V}_t'$ are *predicted* values for the position and velocity before outcome r_t is observed. α and β correspond to the learning rates for position and velocity, respectively. The prediction equation is computed following:

$$\hat{\mathbf{v}}_{t+1} = \mathbf{F}\mathbf{v}_t \tag{4}$$

where $\mathbf{F} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, which gives:

$$\begin{bmatrix} \hat{V}_{t+1} \\ \hat{V}'_{t+1} \end{bmatrix} = \begin{bmatrix} V_t + V_t' \\ V_t' \end{bmatrix} \tag{5}$$

By expressing (5) in terms of Eq. 3, we have:

$$\begin{bmatrix} \hat{V}_{t+1} \\ \hat{V}'_{t+1} \end{bmatrix} = \begin{bmatrix} \hat{V}_t + \hat{V}_t' + (\alpha+\beta)(r_t - \hat{V}_t) \\ \hat{V}_t' + \beta(r_t - \hat{V}_t) \end{bmatrix} \tag{6}$$

Which can be rearranged as:

$$\begin{bmatrix} \hat{V}_{t+1} \\ \hat{V}'_{t+1} \end{bmatrix} = \begin{bmatrix} \hat{V}_t + \hat{V}'_{t+1} + \alpha(r_t - \hat{V}_t) \\ \hat{V}_t' + \beta(r_t - \hat{V}_t) \end{bmatrix} \tag{7}$$

Finally, we assume the same response rule as in SD model, where the behavior for trial $t+1$, B_{t+1}, is generated from a Gaussian distribution with mean $\hat{V}_{t+1}$ and precision $\frac{1}{\eta}$:

$$B_{t+1} \sim \mathit{Gaussian}\left(\hat{V}_{t+1}, \frac{1}{\eta}\right) \tag{8}$$

The key difference between VD over SD model, is the incorporation of the velocity component $\hat{V}'$ along with its update equation. If observations are generated from a Gaussian distribution (as we do in this work), this term tracks the trial-by-trial changes of the mean. In the Method section, we will detail that such variations are controlled by the v term of Eq. 13. In the absence of an estimate for this variable, SD model can only adapt to changes using the learning rate, where faster adaptation occurs when this parameter approaches one. However, in this case, the model predictions would resemble the just-observed outcome and not the generative mean. Top panels of Fig. 2 show simulations (gray lines) of the SD and VD

[1] Throughout the text, we use the parametrization of the Gaussian distribution in terms of a mean and a precision, where the precision is the reciprocal of the variance. This is largely because the software used for our model-based analysis (JAGS) adopts this convention.

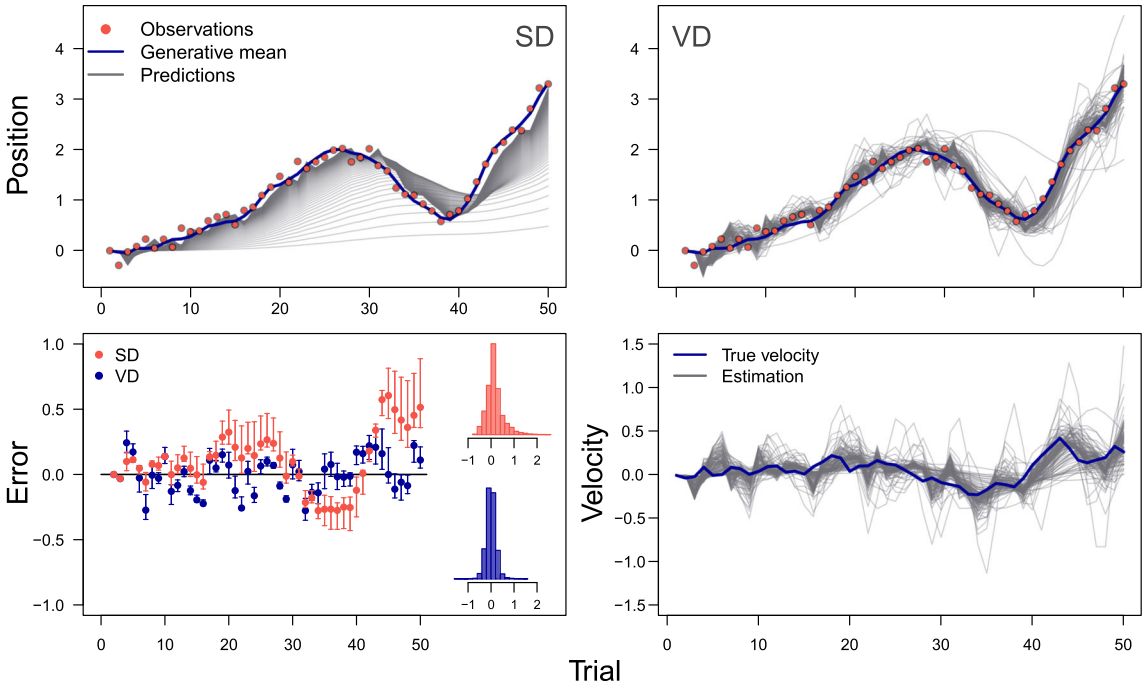


Fig. 2 Simulations of the SD (left) and VD (right) models tracking the mean of a Gaussian distribution changing at a variable velocity. In the top panels red points represent the observations in the environment (r in Eq. 13), the blue line is the true mean of the generating process (x) and the gray lines represent the mean predicted by each model given the observations with different parameter values. The bottom panels show the interquartile range and the median error of the simulations (left) and the true velocity v (right) along with the estimation of the VD model ($\hat{V}'$)

models tracking the moving mean (blue line) of a Gaussian distribution based on samples from it (red dots). Changes of the mean occur at a variable velocity represented by the blue line of the bottom right panel. Each gray line in the top plots corresponds to a simulation using a different value of α for SD, and of α and β for the VD model. It can be observed that SD makes poor predictions for many values of α. In particular, the lower the learning rate the worse the predictions of SD. On the other hand, by incorporating an estimate of changes in the mean, VD model makes better predictions with different values of α and β. The bottom left panel shows the errors between the generative mean and the simulations on every trial. It is important to note that, as the mean begins to increase (around trial 15 and 40) or decrease (around trial 30) errors for SD are considerably greater compared to the ones of VD. The bottom right panel shows the estimate of the changes in the mean by the velocity component of VD compared to the actual velocity of the mean.

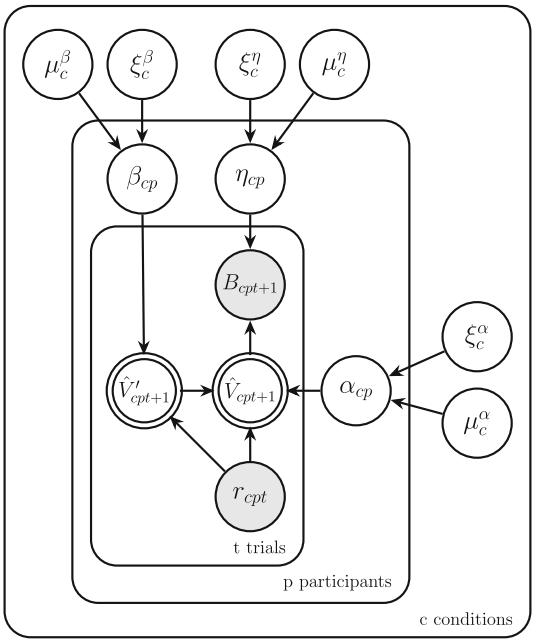
In contrast to dynamic models that propose that the learning rate changes over trials (Nassar et al. 2010), we assumed that α and β of VD are free parameters for each condition. Our analysis was based on this assumption given that in non-stationary environments like ours (see Method section), learning rates stabilize in values that asymptotically correspond to the free parameters (Daw and Tobler 2014). Additionally, the performance of subjects remained stable within conditions indicating that they weighted prediction errors similarly over trials (see Online Resource 1). Figure 3 shows the graphical representation of VD (based on Eqs. 7 and 8) using hierarchical modeling. In the same way as α and η in SD model, β is generated from a Gaussian distribution with hyperparameters $(\mu_c^{\beta}, \frac{1}{\xi_c^{\beta}})$ for each experimental condition c. Apart from that, and the update equation for the velocity component, specifications of the graphical model in Fig. 3 are the same as in Fig. 1.

Kalman Filter This model is a Bayesian form of SD which updates the learning rate on a trial-by-trial basis depending on the current level of uncertainty (Kalman 1960; Navarro et al. 2018; Zajkowski et al. 2017; Speekenbrink and Konstantinidis 2015; Speekenbrink and Shanks 2010; Gershman 2017, 2015; Kakade and Dayan 2002). The Kalman Filter estimates the mean of the Gaussian distribution generating the observations following:

$$\hat{V}_{t+1} = \hat{V}_t + \alpha_t(r_t - \hat{V}_t) \tag{9}$$

where α_t is known as the Kalman gain and is computed as:

$$\alpha_t = \frac{\eta_t + \zeta}{\eta_t + \zeta + \omega} \tag{10}$$



$$\mu_c^\alpha \sim \text{Uniform}(0,1)$$
$$\mu_c^\beta \sim \text{Uniform}(0,1)$$
$$\mu_c^\eta \sim \text{Uniform}(0,1)$$
$$\xi_c^\alpha \sim \text{Uniform}(0,1)$$
$$\xi_c^\beta \sim \text{Uniform}(0,1)$$
$$\xi_c^\eta \sim \text{Uniform}(0,1)$$
$$\alpha_{cp} \sim \text{Gaussian}(\mu_c^\alpha, \tfrac{1}{\xi_c^\alpha})_{T(0,1)}$$
$$\beta_{cp} \sim \text{Gaussian}(\mu_c^\beta, \tfrac{1}{\xi_c^\beta})_{T(0,1)}$$
$$\eta_{cp} \sim \text{Gaussian}(\mu_c^\eta, \tfrac{1}{\xi_c^\eta})_{T(0,1)}$$
$$\hat{V}_{cpt+1} = \hat{V}_{cpt} + \hat{V}'_{cpt+1} + \alpha_{cp}(r_{cpt} - \hat{V}_{cpt})$$
$$\hat{V}'_{cpt+1} = \hat{V}'_{cpt} + \beta_{cp}(r_{cpt} - \hat{V}_{cpt})$$
$$B_{cpt+1} \sim \text{Gaussian}(\hat{V}_{cpt+1}, \tfrac{1}{\eta_{cp}})$$

Fig. 3 Graphical representation of a hierarchical delta-rule with a velocity term

η is the variance of the Gaussian distribution and is updated following:

$$\eta_{t+1} = (1 - \alpha_t)(\eta_t + \zeta) \tag{11}$$

where ζ corresponds to the innovation variance and refers to non-directional changes assumed by the model from trial to trial. ω is the error variance, and corresponds to the estimated noise in the observations. Both the innovation and the error variance are free parameters of this model. Finally, we assumed a response rule where the behavior for trial $t+1$, B_{t+1}, is generated from a Gaussian distribution with mean $\hat{V}_{t+1}$ and precision $\frac{1}{\eta_{t+1}+\zeta}$:

$$B_{t+1} \sim Gaussian\left(\hat{V}_{t+1}, \frac{1}{\eta_{t+1} + \zeta}\right) \tag{12}$$

Note that the precision includes the term ζ (therefore increasing the variance of the prediction), as the model considers that for the next trial the position can change. In contrast to VD model, this estimate of change has no direction, and only modulates the influence of new outcomes in the updating process. Figure 4 is the graphical representation of the Kalman Filter using hierarchical modeling. Parameters ζ and ω are generated from Gaussian distributions with hyperparameters $\left(\mu_c^\zeta, \frac{1}{\xi_c^\zeta}\right)$ and $\left(\mu_c^\omega, \frac{1}{\xi_c^\omega}\right)$, respectively, for each condition c.

Method

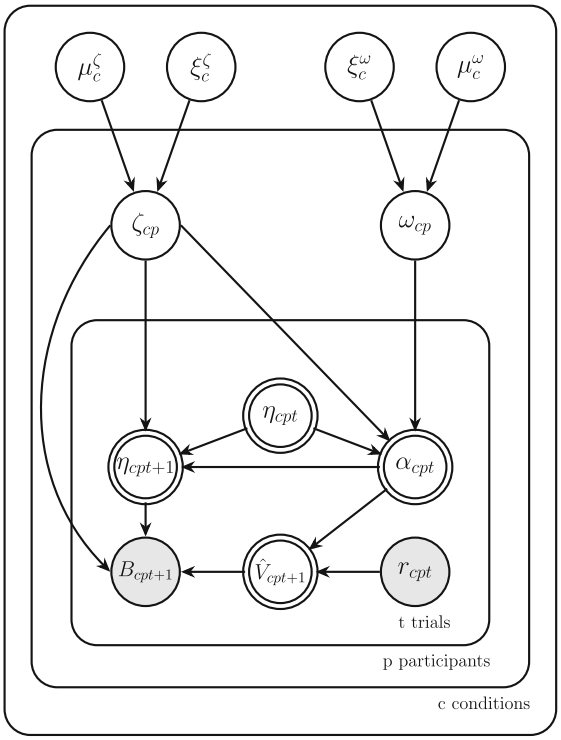
Participants

Seventy two undergraduate students (55 female, mean (SD) age = 19.8 (2.03)) from the School of Psychology at the National Autonomous University of Mexico participated in the study after providing informed consent.

Behavioral Task

The experiment was programmed using the commercial software Matlab and the extension Psychophysics Toolbox (Kleiner et al. 2007; Pelli 1997; Brainard 1997) to create visual stimuli. A standard mouse and keyboard, and a screen of 1920 × 1080 pixels were used.

Figure 5 displays a representation of the task. In this scenario, participants were required to predict the future position of a spaceship moving around planet Earth along a specified orbit (blue dotted circle). Its location was given in radians (rad) using the center of the orbit as reference. On every trial, the spaceship appeared for half a second on a given point of its orbit, after which it disappeared. Participants had to click on the position where they thought it would reappear in the next trial. After their choice, a red dot indicated the location they selected and a red circle a fixed margin of error. At the same time, the spaceship showed up in its new position. If it fell inside the red circle, it turned red indicating the new location was accurately predicted or remained the same color otherwise. There



$$\mu_c^\zeta \sim \text{Gaussian}(0,0.001)_{T(0,\infty)}$$
$$\mu_c^\omega \sim \text{Gaussian}(0,0.001)_{T(0,\infty)}$$
$$\xi_c^\zeta \sim \text{Gaussian}(0,0.001)_{T(0,\infty)}$$
$$\xi_c^\omega \sim \text{Gaussian}(0,0.001)_{T(0,\infty)}$$
$$\zeta_{cp} \sim \text{Gaussian}(\mu_c^\zeta, \tfrac{1}{\xi_c^\zeta})_{T(0,\infty)}$$
$$\omega_{cp} \sim \text{Gaussian}(\mu_c^\omega, \tfrac{1}{\xi_c^\omega})_{T(0,\infty)}$$
$$\alpha_{cpt} = \tfrac{\eta_{cpt}+\zeta_{cp}}{\eta_{cpt}+\zeta_{cp}+\omega_{cp}}$$
$$\eta_{cpt+1} = (1-\alpha_{cpt})(\eta_{cpt}+\zeta_{cp})$$
$$\hat{V}_{cpt+1} = \hat{V}_{cpt} + \alpha_{cpt}(r_{cpt} - \hat{V}_{cpt})$$
$$B_{cpt+1} \sim \text{Gaussian}(\hat{V}_{cpt+1}, \tfrac{1}{\eta_{cpt+1}+\zeta_{cp}})$$

Fig. 4 Graphical representation of a hierarchical Kalman filter

was no time limit for participants to emit their responses. This sequence was repeated throughout the experiment. Moreover, if the spaceship gave a full lap in a counterclockwise direction, i.e., moving 2π rad, the second lap would continue from 2π rad to 4π rad, and so on. Similarly, if the spaceship completed a full lap in a clockwise direction, its next position was given according to values of the previous lap. We followed the same logic to register participant's responses. This transformation allowed the range of possible values of observations and responses to span from

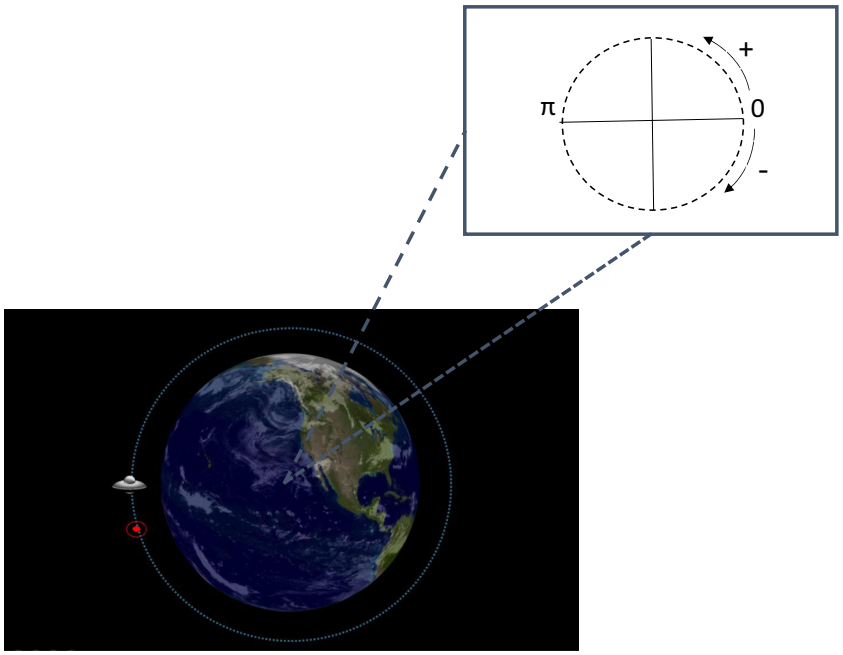


Fig. 5 Representation of the behavioral task. Participants predicted the position of a spaceship that moved around planet Earth along the blue dotted line. Selected positions were indicated with a red point surrounded by a red circle representing a margin of error. Graphics of the task were developed using the on-line site http://planetmaker.wthr.us/. See the main text for a detailed explanation of the task

Table 1 Experimental conditions. Units for S and N are given in rad^2

Condition	$\frac{S}{N}$	S	N	Trials
1	0.05	0.0049	0.098	300
2	0.5	0.0049	0.0098	300
3	1	0.0049	0.0049	300
4	2	0.0049	0.00245	300

$-\infty$ to ∞, and was particularly useful to avoid sudden changes of position from 2π rad to 0 every time a full lap was completed.

Experimental Design

On every trial t the position of the spaceship r was generated from a moving Gaussian distribution following:

$$\begin{aligned} v_{t+1} &\sim Gaussian\left(v_t, \sigma_v^2\right) \\ x_{t+1} &= x_t + v_t \\ r_{t+1} &\sim Gaussian\left(x_{t+1}, \sigma_r^2\right) \end{aligned} \tag{13}$$

where $\times$ is the mean of the distribution, v is a velocity term following a Gaussian random walk with variance σ_v^2, and σ_r^2 is the variance in the actual observations. Our experiment consisted of four conditions that varied the S/N values represented by $\frac{\sigma_v^2}{\sigma_r^2}$. Intuitively, this quantity indicates how easy it is to discriminate changes due to velocity, relative to changes due to random noise; smaller ratios indicate noisier observations. We fixed the numerator of the ratio σ_v^2 so participants faced the same generative process for the velocity component, but varied the denominator σ_r^2 to change the noise in their observations. Table 1 shows the values of S/N used in the experiment. An experimental session consisted of four conditions with 300 trials each and order of presentation was randomized for all participants. Before the experimental task started, a practice phase was completed that consisted of at least 30 trials, after which participants decided whether to continue acquiring more practice or begin the experiment. After completing each condition, there was a time break and participants decided when they were ready to start the next round of trials.

Results

As a measurement of performance, we report the error between the generative mean of the spaceship and participants' predictions. These values are shown in the bottom panels of Fig. 6 for all subjects on the four S/N conditions. It is clear that, in all conditions, most errors remain close to zero, however, accuracy increases with the S/N. This is evident when observing the proportion of values around zero for each condition and the corresponding standard deviation. In other words, participants had greater errors for noisier observations. Top panels are graphical representations of bottom plots where dark and light blue represent $\pm\sigma$ and $\pm 2\sigma$, respectively, and the figure of the spaceship represents the generative mean.

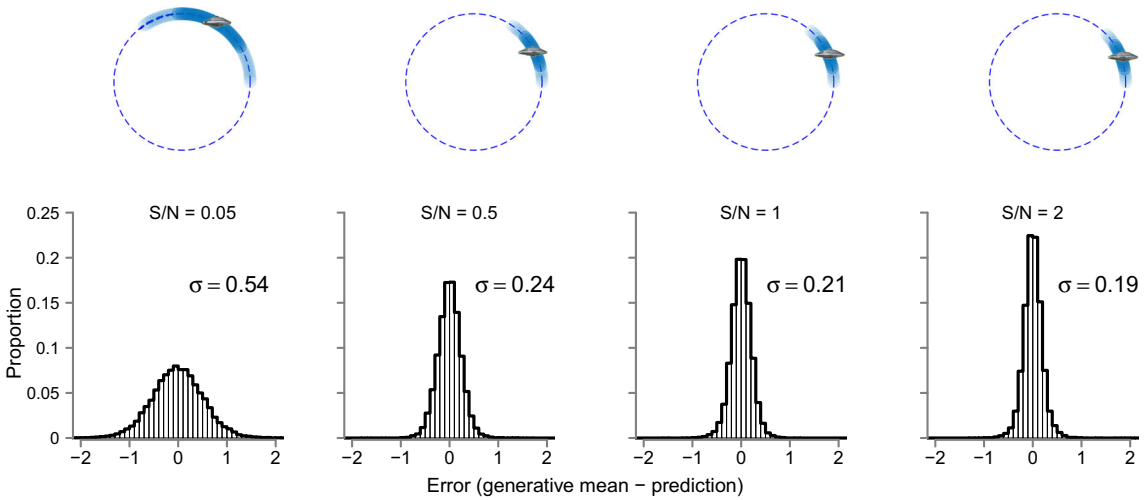


Fig. 6 Overall performance of subjects in the experiment. Bottom panels show the error between the generative mean of the spaceship and the predictions of participants for all trials on each of the conditions indicated with their corresponding S/N values. A graphical representation of the errors is shown in the top panels where dark and light blue indicate $\pm\sigma$ and $\pm 2\sigma$, respectively, and the spaceship represents the generative mean

Bayesian Inference

Posterior distributions of parameters and hyperparameters were approximated using the software JAGS (Just Another Gibbs Sampler; Plummer 2003) implemented in R code. This procedure uses a sampling method known as Markov chain Monte Carlo (MCMC) to estimate parameters in a model. For our three graphical models we used three independent chains with 10^5 samples each and a burn-in period (samples that were discarded in order for the algorithm to adapt) of 8×10^4. A thinning of 10 was used (i.e., values were taken every 10 samples of the chain) to reduce autocorrelation within chains. Convergence was verified by computing the $\hat{R}$ statistic (Gelman and Rubin 1992) which is a measure of between-chain to within-chain variance where values close to 1 indicate convergence (Lee and Wagenmakers 2013). In general, values between 1 and 1.05 are considered as reliable evidence for convergence. All of the nodes for our three models had $\hat{R}$ values within this interval.

Figure 7 shows the results of Bayesian inference about parameters for the SD, VD, and Kalman Filter models. All panels, except for the one at the bottom right, display maximum posterior values (modes) of parameters for each participant, ordered by experimental condition (S/N values). Error bars correspond to the interquartile range and dots represent the medians. It can be observed that for SD and VD models, the learning rate α has values close to one in all conditions (for SD model however values are not visually different from one and do not show any variability between participants). According to both models, this means that the just-observed outcome highly influenced participants' predictions. For SD, the above implies that a new prediction for participants equalled the just-observed outcome (as the learning rate is not visually different from one), and for VD, that the new prediction equalled a value close to the just-observed outcome (as the learning rate is high but visually different from one) *plus* the estimation of velocity (see Eqs. 5 and 7). Importantly, values for η for SD were higher in all conditions compared to the ones of the VD, indicating

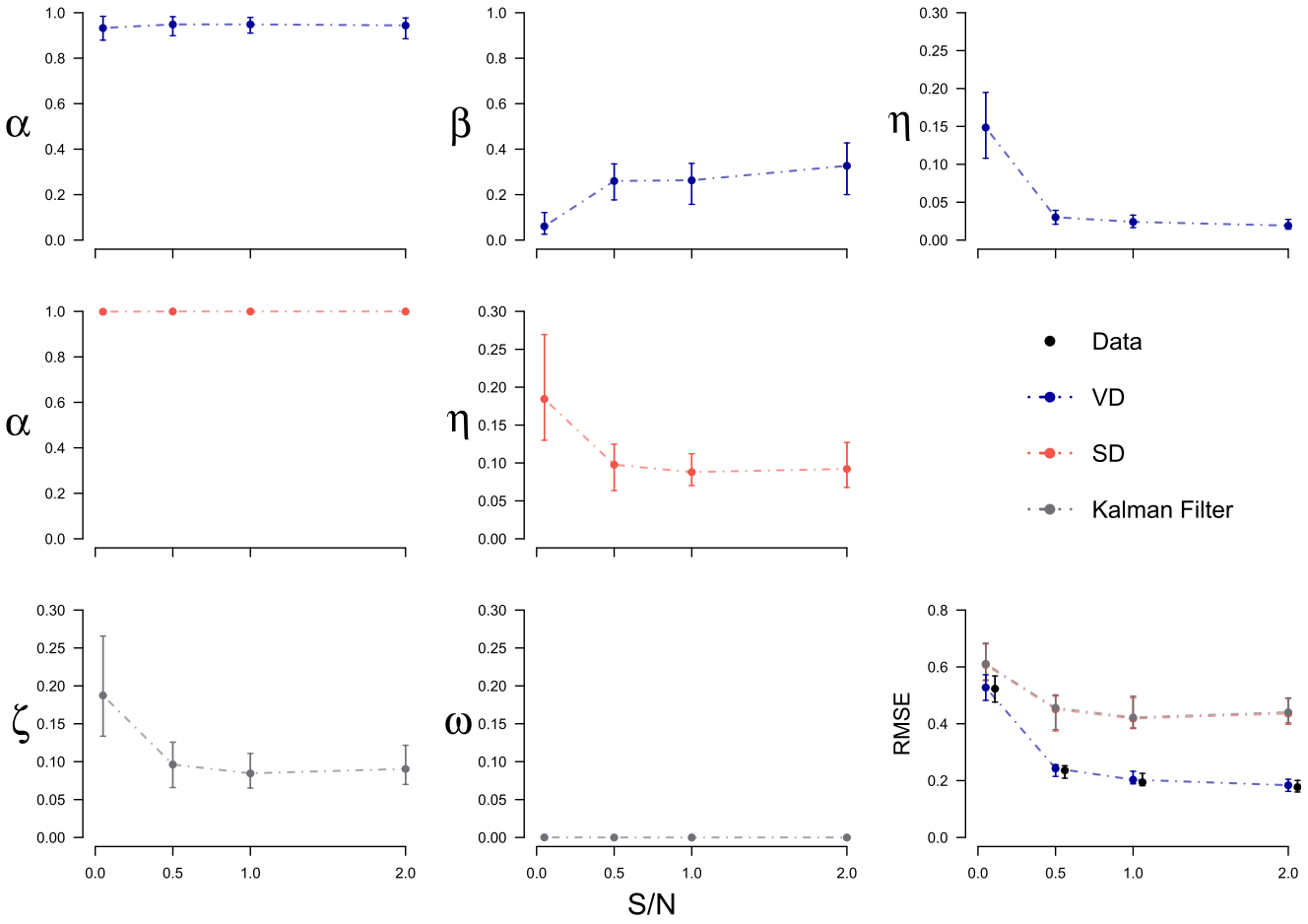


Fig. 7 Bayesian inference results for SD, VD, and Kalman Filter models. Error bars represent the interquartile range and dots the medians of the Maximum posterior (MAP) for each subject. The bottom right panel displays the RMSE generated with the posterior predictions of each model compared to the actual RMSE of participants

that, according to SD, a higher degree of noise influenced participants' decisions. However, η values for both SD and VD models decrease for higher S/N. This relation indicates that when observations were less noisy so were the predictions of participants. In the case of the learning rates for velocity β, we observe a gradual increment of values as the S/N increases, which suggests that the velocity term was updated faster for less noisy observations. Interestingly, the error variance ω of the Kalman Filter is not different from zero in any condition. If we look at Eq. 10, this means that the Kalman gain approximates one which makes the model closely similar to SD. Additionally, values for the innovation variance ζ tightly resemble the behavior of η in SD, which probably arises as the variance η_t of the Kalman Filter approaches zero over trials and the precision in Eq. 12, approximates $\frac{1}{\zeta}$. The bottom right panel of Fig. 7 shows the root mean squared error (RMSE) generated from the posterior predictions of each model compared to the actual RMSE of participants. Posterior predictions were obtained by simulating data with 300 samples with replacement from the joint posterior distribution of parameters and the actual observations of participants. The similarity between the RMSE of the simulated data and the actual RMSE is an indicator of the descriptive adequacy of the models. Note that in all conditions the model incorporating the velocity component recovers the actual RMSE better than the SD and the Kalman Filter. As expected from the parameter values of SD and the Kalman Filter, both models have almost identical RMSE (overlapping gray and red dots and error bars).

Model Comparison

Although RMSE provides a useful measurement of how accurately models can recover actual data, it cannot be used as a metric of model comparison as it ignores complexity, and generally more complex models will capture data better. In order to overcome this limitation, we implemented two standard methods in Bayesian modeling that incorporate model complexity into their computation. The first, is leave-one-out cross-validation (LOO), and the second, the Widely Applicable Information Criterion (WAIC). These techniques compute a pointwise estimator of the predictive accuracy of models for all data points taking one at a time. On the one hand, LOO is a type of cross-validation where the training dataset (the one used to tune the model) consist of all observations but one, which forms the validation dataset (the one to be predicted). In particular, we used a new approach to LOO developed by Vehtari et al. (2017) where a Pareto distribution is implemented to smooth weights in an importance sampling procedure, and that the authors termed PSIS-LOO (for Pareto smoothed importance sampling). On the other hand, WAIC works as an estimate of out-of-sample deviance which overcomes previous limitations of the deviance information criterion (DIC). Unlike DIC, WAIC is based on the entire posterior distribution and is valid under non-Gaussian assumptions. In practice, PSIS-LOO and WAIC can be easily computed using the R package loo (Vehtari et al. 2017) and the log-likelihood evaluated at the posterior simulations of the parameter values (for more details on the computation of PSIS-LOO and WAIC, see equations 3, 10, and 11 of Vehtari et al. (2017)). These two methods return a measurement of deviance at predicting a new dataset and penalize model complexity. Differences of PSIS-LOO and WAIC between each model and the model with the lowest values for each metric are reported in Table 2. More positive values indicate worse predictive performance. It is clear that VD outperforms the other models according to both metrics. As expected from parameter estimation and the descriptive adequacy of models, SD and the Kalman Filter have a very close predictive performance. It is important to note that model complexity in PSIS-LOO and WAIC is not defined in terms of parameter counts like in AIC or BIC. Instead, these metrics consider the variability of posterior predictions. When models make a wide range of predictions they are automatically penalized for the poor ones as in Bayes Factors. This is a relevant feature as in hierarchical models usually increasing the number of parameters reduces the variability of the predictions and, therefore, the complexity of the model (Lee and Vanpaemel 2018).

Table 2 Differences of PSIS-LOO ($\Delta_{PSIS-LOO}$) and WAIC (Δ_{WAIC}) between each model and the model with the lowest value for each metric. Higher values indicate worse predictive performance

Model	$\Delta_{PSIS-LOO}$	Δ_{WAIC}
SD	85730	85534
Kalman	85715	85517
VD	0	0

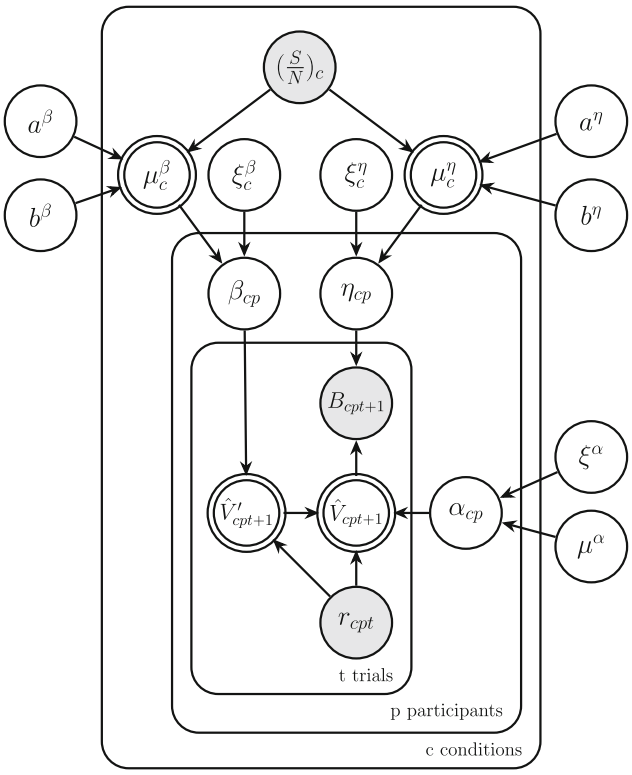
Predictions for New S/N Values

The three models evaluated in this work are able to make predictions for new subjects for each experimental condition given their hierarchical structure. However, neither of them inform us of what would be expected for different values of S/N. This limitation is overcome by extending their hierarchical structure formalizing the relation of parameters between conditions. Given that VD showed the best descriptive adequacy and predictive performance,

we will use this model for the hierarchical extension. By visual inspection of Fig. 7, we can tell that values of α are invariant for the evaluated conditions. Thus, we can simplify the model by assuming they are generated by a Gaussian distributions for the whole experiment. However, this is not the case for β and η. These parameters appear to gradually increase and decrease, respectively, as the S/N increases. To formalize this pattern, we assumed that β and η are generated from Gaussian distributions with means following a hyperbolic function of the S/N values. Each hyperbola takes the S/N values as argument and two parameters with positive value control the shape of the function (a^β and b^β for the hyperbola of μ_c^β and a^η and b^η for the hyperbola μ_c^η). The graphical model of Fig. 8 (labeled as HVD) specifies each hyperbola.

Figure 9 shows the results of Bayesian inference for the HVD model. Top panels corresponds to hyperbolas for the hierarchical means of the learning rates for velocity μ^β (left) and decision noise μ^η (right). These functions were generated within the interval (0,2) using 300 samples with replacement from the joint posterior distribution of the parameters that constitute each hyperbola. In the bottom left panel we show the posterior samples of the mean of learning rates for position μ^α. As there is a single distributions for the whole experiment, values of the S/N were omitted. In the bottom right panel, we show the descriptive adequacy of the HVD model using the same sampling procedure as in the previous models. It can be observed that model HVD is able to recover the actual RMSE of subjects just as accurately as VD in Fig. 7.

Importantly, given that the functions are continuous, we are now able to make predictions about the average behavior of parameters for untested S/N values. In Fig. 10, we show simulations of VD and HVD for a new participant in our experimental conditions, and for two new S/N values (0.35 and 1.5). Predictions for the new conditions were generated taking 100 samples with replacement from the joint posterior distribution of the parameters of the hyperbolas and from the single hierarchical mean of α. On the other hand, predictions for the known conditions (S/N values of 0.05, 0.5, 1, and 2) were generated using the same number of samples from the joint posterior distribution of the model parameters (α, β, and η). Top panels show the RMSE of the simulations on each condition. It can be noted that both models have similar RMSE for the known conditions but VD generates large RMSE values compared to HVD for the untested ones. Bottom panels show the trial-by-trial predictions of the models for each of the simulations. It is evident that the variance of the predictions of VD for the new conditions is considerably



$$\mu^\alpha \sim \text{Uniform}(0,1)$$
$$a^\beta \sim \text{Uniform}(0,1)$$
$$b^\beta \sim \text{Gaussian}(0,0.001)_{T(0,\infty)}$$
$$a^\eta \sim \text{Gaussian}(0,0.001)_{T(0,\infty)}$$
$$b^\eta \sim \text{Gaussian}(0,0.001)_{T(0,\infty)}$$
$$\mu_c^\beta = \frac{a^\beta (S/N)_c}{(S/N)_c + b^\beta}$$
$$\mu_c^\eta = \frac{1}{a^\eta (S/N)_c + b^\eta}$$
$$\xi^\alpha \sim \text{Uniform}(0,1)$$
$$\xi_c^\beta \sim \text{Uniform}(0,1)$$
$$\xi_c^\eta \sim \text{Uniform}(0,1)$$
$$\alpha_{cp} \sim \text{Gaussian}(\mu^\alpha, \tfrac{1}{\xi^\alpha})_{T(0,1)}$$
$$\beta_{cp} \sim \text{Gaussian}(\mu_c^\beta, \tfrac{1}{\xi_c^\beta})_{T(0,1)}$$
$$\eta_{cp} \sim \text{Gaussian}(\mu_c^\eta, \tfrac{1}{\xi_c^\eta})_{T(0,1)}$$
$$\hat{V}_{cpt+1} = \hat{V}_{cpt} + \hat{V}'_{cpt+1} + \alpha_{cp}(r_{cpt} - \hat{V}_{cpt})$$
$$\hat{V}'_{cpt+1} = \hat{V}'_{cpt} + \beta_{cp}(r_{cpt} - \hat{V}_{cpt})$$
$$B_{cpt+1} \sim \text{Gaussian}(\hat{V}_{cpt+1}, \tfrac{1}{\eta_{cp}})$$

Fig. 8 Graphical representation of the extended hierarchy of VD model using hyperbolic functions

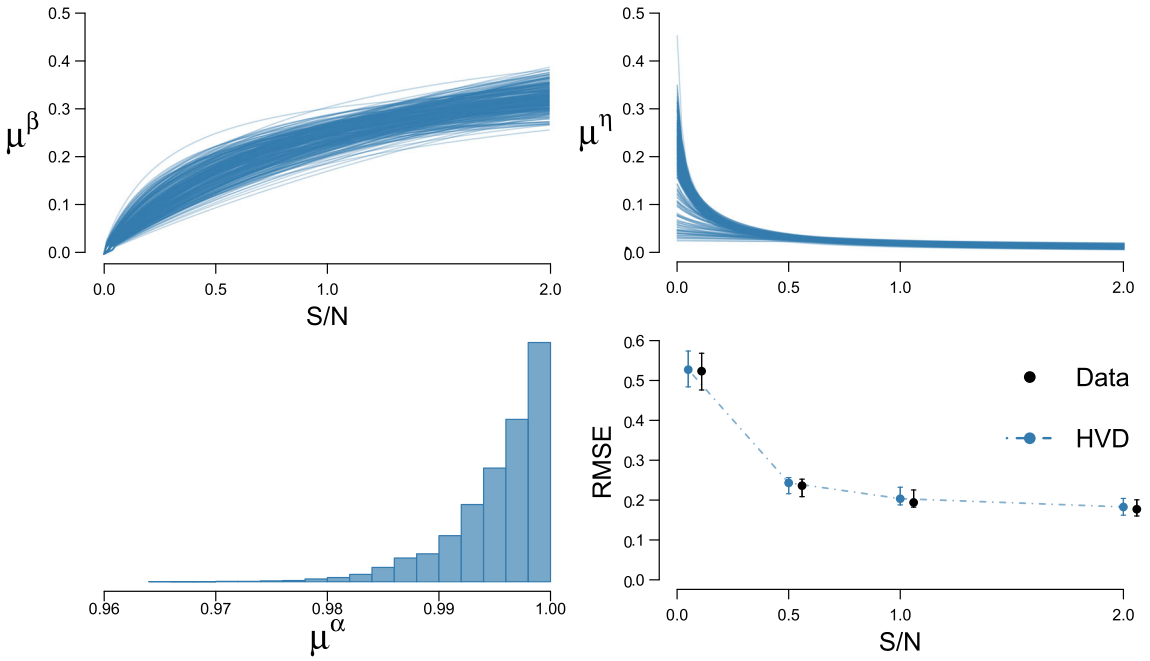


Fig. 9 Results of Bayesian inference for the HVD model. Top panels correspond to the hyperbolas of the hierarchical means of β (left) and η as a function of the S/N. Each hyperbola was generated by sampling with replacement from the joint posterior distribution of a^{β} and b^{β}, for learning rates of velocity, and a^{η} and b^{η}, for decision noise. The bottom left panel shows posterior samples of the hierarchical mean for α. The bottom right panel shows the RMSE generated with the posterior predictions of the model for all subjects compared to the actual RMSE of participants. Error bars represent the interquartile range and points the median

high compared to the one of HVD, which results from a lack of information of parameter values that a participant would use for those S/N values. Importantly, HVD is able to make predictions about new conditions without loosing the descriptive adequacy of VD in the known conditions as shown in Fig. 9. Additionally, it has similar values of PSIS-LOO and WAIC ($\Delta_{PSIS-LOO} = 27$ and $\Delta_{WAIC} = -20$, where values of VD are subtracted from the ones of HVD for both metrics. In other words, for PSIS-LOO VD is a better model, but WAIC favors HVD).

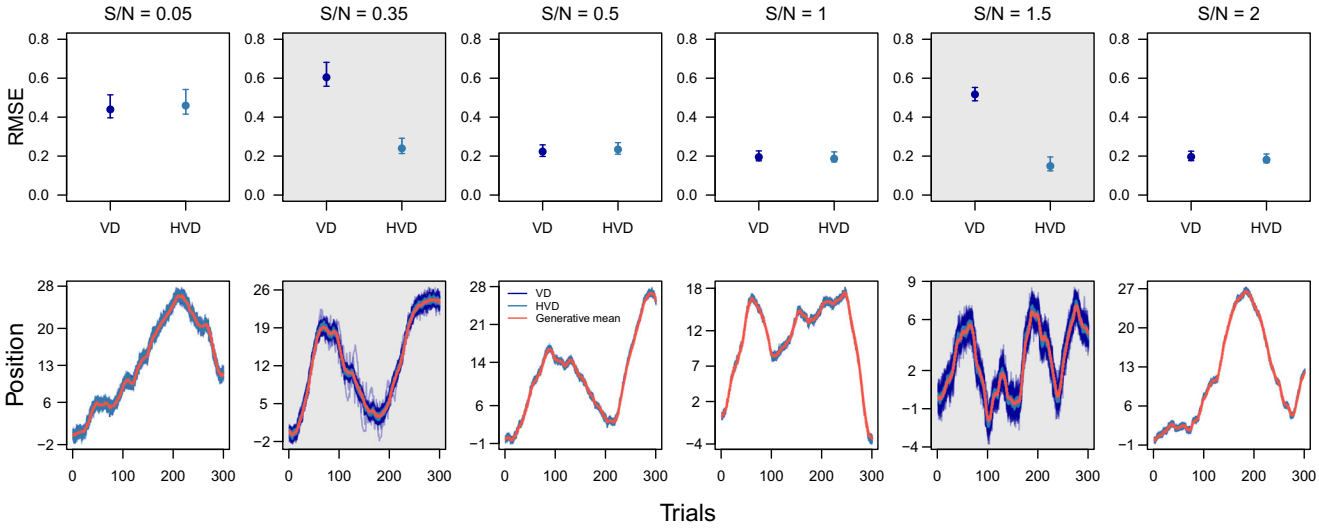


Fig. 10 Predictions for a new subject on each experimental condition. Top panels represent the predicted RMSE for VD and HVD models. Error bars correspond to the interquartile range and dots to the medians from the 100 simulations. Bottom panels correspond to trial-by-trial predictions of VD and HVD models for a new sequence of observations. The red line represents the mean of the generative process. The gray panels represent simulations for untested values of S/N ratio

Discussion

Humans and other animals often face environments that change over time. In some situations, these changes may occur gradually following a rate, and, in order to make accurate predictions, individuals should have a good estimate of this variable. However, the rate of change may not be readily inferred when peoples' observations are corrupted with random fluctuations. In this paper, we tested people's predictions in an environment with these characteristics by using a perceptual-decision making task. In our experiment, subjects predicted the future location of a spaceship that moved at a variable velocity and was corrupted with different levels of Gaussian noise. Our results show participants were able to predict the most likely future location of the spaceship with accuracy increasing for less noisy conditions. A standard reinforcement learning model (SD) was unable to qualitatively describe these results, and Bayesian inference showed learning rates for this model are not visually different from one in all conditions. This strategy is optimal only for a deterministic task and useful after the environment suffers abrupt and unpredictable changes, but inaccurate in a probabilistic setting that is changing gradually over trials. In an attempt to capture deviations from participants' predictions, SD model assumes decision noise is high. This is likely to happen when a model is ignoring a crucial signal in data (namely the velocity component) and construing it as random variations. By incorporating a velocity term to the standard delta-rule (VD model), we were able to describe data in a more reasonable fashion. Furthermore, Bayesian inference showed that learning rates for velocity and decision noise increase and decrease, respectively, with the S/N. The above suggests that, in general, subjects updated their estimate of velocity faster and had less noisy predictions when observations were less corrupted by noise.

Furthermore, the modeling results showed that the Kalman filter, a Bayesian alternative to the delta-rule (Speekenbrink and Konstantinidis 2015; Gershman 2015), was unable to capture participants' behavior. In this case, the posterior distributions from the innovation variance (ζ) showed a behavior similar to the decision noise η in the SD, while the value of the error variance ω was indistinguishable from zero. These results imply that the Kalman gain would approximate one for almost all of the trials of the conditions, which would explain why the RMSE between the SD and the Kalman Filter are indistinguishable from one another. Previously, it has been noted that the SD model can be interpreted as the Kalman Filter with a fixed learning rate (Speekenbrink and Konstantinidis 2015). A formal model comparison of these three modes using PSIS-LOO and WAIC shows that VD model has the best predictive performance overall.

An extension of VD model suggests that the overall behavior of learning rates for velocity and decision noise can be modeled using a hyperbolic function. This model was able to capture participants' errors as accurately as its not-constrained counterpart and to make reasonable predictions about the expected behavior of a new participant under the same experimental conditions. The hyperbolic function inferred for the learning rates of velocity and decision noise can take practically any positive value of S/N as input and provide an overall prediction of parameter values. The results show that this extension can be used to make predictions about the expected behavior of participants under untested conditions of S/N. Further work could show whether this predictions can account for the behavior of participants in a similar task. A formal model comparison between the VD model and the hyperbolic extension using PSIS-LOO and WAIC shows that both models have a similar predictive performance.

This work accords with other studies that propose humans are sensitive to higher-order variables that control the dynamics of the environment (Meder et al. 2017; Behrens et al. 2007; Ricci and Gallistel 2017; McGuire et al. 2014; Yu and Dayan 2005; Courville et al. 2006; Wittmann et al. 2016). In particular, our model suggests that when the environment is changing smoothly at a variable velocity, subjects have an estimate of this quantity and use it to make predictions as suggested in Fig. 2. Furthermore, we showed that this process is influenced by the level of noise in the observations, which enables faster learning for higher S/N values.

Although reinforcement learning models are common in tasks of belief updating in changing environments (Wilson et al. 2013; Nassar et al. 2010; Behrens et al. 2007; Speekenbrink and Shanks 2010), some studies suggest that this process may not take place on a trial-by-trial basis as suggested by delta-rule models either when the environment suffers abrupt (Gallistel et al. 2014; Robinson 1964) or gradual changes (Ricci and Gallistel 2017). Instead, these works suggest that people follow a step-like pattern, some times updating their estimates after hundreds of trials have passed. This is true when people infer the parameter of a Bernoulli distribution (Gallistel et al. 2014; Ricci and Gallistel 2017), however, the conditions under which people follow this pattern or a trial-by-trial update are not clear yet. Of particular interest to our paper is a recent error-driven approach to adaptive behavior using a control theoretic model known as PID (proportional-integral-derivative controller, Ritz et al. 2018). This model incorporates to the standard delta-rule (proportional part), a weighted sum of the history of errors (integral part) and the difference between the current and previous error (derivative part). It is worth noting that the PI part of the model is algebraically equivalent to our VD model (without

hierarchical modeling) when there is perfect integration. A simple rearrangement of the integral part as a Markov process provides the update equation of the velocity term in the VD model (when the memory persistence parameter of PI equals one). We believe our approach is computationally less expensive as it does not require estimating the full history of errors and their corresponding weights on every trial, but only the previous estimate of change and the current error. However, it is important to note that the model proposed here would perform poorly in task with abrupt changes of position or velocity as it suffers from the same pitfalls of models with fixed learning rates. PID model ameliorates this concern by incorporating the derivative part, which allows for sudden corrections when the model estimates depart from the generative process.

In summary, in this work, we have provided evidence that people can use prediction errors to update an estimate of the rate of change (velocity) when the environment is varying gradually over trials, and to update this quantity faster when observations are more reliable. Additionally, we have shown that a hierarchical Bayesian approach provides benefits in terms of predictive power and generalization. Finally, our results are in line with evidence that people and other animals can learn about higher-order statistics of their environment and use that information to guide predictions.

Funding Information This research was supported by the project PAPIIT IG120818.

References

Behrens, T.E., Woolrich, M.W., Walton, M.E., Rushworth, M. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*, 1214–1221.

Brainard, D.H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Bush, R.R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, *58*, 313–323.

Courville, A.C., Daw, N.D., Touretzky, D.S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*, 294–300.

Daw, N.D., & Tobler, P.N. (2014). Value learning through reinforcement: the basics of dopamine and reinforcement learning. In *Neuroeconomics: decision making and the brain*. 2nd edn. (pp. 283–298): Elsevier.

Dayan, P., & Nakahara, H. (2018). Models and methods for reinforcement learning. In *Stevens' handbook of experimental psychology and cognitive neuroscience*. 4th edn. New York: Wiley.

Gallistel, C.R., Krishan, M., Liu, Y., Miller, R., Latham, P.E. (2014). The perception of probability. *Psychological Review*, *121*, 96–123.

Gelman, A., & Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472.

Gershman, S.J. (2015). A unifying probabilistic view of associative learning. *PLOS Computational Biology*, *11*, e1004567.

Gershman, S.J. (2017). Dopamine, inference, and uncertainty. *Neural Computation*, *29*, 3311–3326.

Kakade, S., & Dayan, P. (2002). Acquisition and extinction in autoshaping. *Psychological Review*, *109*, 533–544.

Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*, 35–45.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, *36*, 1–16.

Lee, M.D. (2018). Bayesian methods in cognitive modeling. In *The Stevens' handbook of experimental psychology and cognitive neuroscience*. 4th edn. New York: Wiley.

Lee, M.D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, *25*, 114–127.

Lee, M.D., & Wagenmakers, E.J. (2013). *Bayesian cognitive modeling: a practical course*. Cambridge: Cambridge University Press.

Matzke, D., Dolan, C.V., Batchelder, W.H., Wagenmakers, E.J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, *80*, 205–235.

McGuire, J.T., Nassar, M.R., Gold, J.I., Kable, J.W. (2014). Functionally dissociable influences on learning rate in a dynamic environment. *Neuron*, *84*, 870–881.

Meder, D., Kolling, N., Verhagen, L., Wittmann, M.K., Scholl, J., Madsen, K.H., Hulme, O.J., Behrens, T.E.J., Rushworth, M. (2017). Simultaneous representation of a spectrum of dynamically changing value estimates during decision making. *Nature Communications*, *8*, 1942.

Miller, R.R., Barnet, R.C., Grahame, N.J. (1995). Assessment of the rescorla-wagner model. *Psychological Bulletin*, *117*, 363–386.

Nassar, M.R., Wilson, R.C., Heasly, B., Gold, J.I. (2010). An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, *30*, 12366–12378.

Navarro, D.J., Tran, P., Baz, N. (2018). Aversion to option loss in a restless bandit task. *Computational Brain & Behavior*.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*, 139–154.

O'Reilly, J.X. (2013). Making predictions in a changing world—inference, uncertainty, and learning. *Frontiers in Neuroscience*, *7*, 105.

Pelli, D.G. (1997). The videotoolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Plummer, M. (2003). Jags: a program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing, Vienna, Austria, vol 124*.

Pratte, M.S., & Rouder, J.N. (2011). Hierarchical single-and dual-process models of recognition memory. *Journal of Mathematical Psychology*, *55*, 36–46.

Rescorla, R.A., & Wagner, A.R. (1972). A theory of pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current Research and Theory*, *2*, 64–99.

Ricci, M., & Gallistel, R. (2017). Accurate step-hold tracking of smoothly varying periodic and aperiodic probability. *Attention, Perception, & Psychophysics*, *79*, 1480–1494.

Ritz, H., Nassar, M.R., Frank, M.J., Shenhav, A. (2018). A control theoretic model of adaptive learning in dynamic environments. *Journal of Cognitive Neuroscience*, *30*, 1405–1421.

Robinson, G.H. (1964). Continuous estimation of a time-varying probability. *Ergonomics*, *7*, 7–21.

Schultz, W., Dayan, P., Montague, P.R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Shiffrin, R.M., Lee, M.D., Kim, W.J., Wagenmakers, E.J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, *7*, 351–367.

Speekenbrink, M., & Shanks, D.R. (2010). Learning in a changing environment. *Journal of Experimental Psychology: General*, *139*, 266–298.

Sutton, R.S. (1998). *Reinforcement learning: an introduction*. Cambridge: MIT Press.

Vehtari, A., Gelman, A., Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*, 1413–1432.

Wilson, R.C., Nassar, M.R., Gold, J.I. (2013). A mixture of delta-rules approximation to Bayesian inference in change-point problems. *PLOS Computational Biology*, *9*, e1003150.

Wittmann, M.K., Kolling, N., Akaishi, R., Chau, B., Brown, J.W., Nelissen, N., Rushworth, M.F. (2016). Predictive decision making driven by multiple time-linked reward representations in the anterior cingulate cortex. *Nature Communications*, *7*, 12327.

Yu, A.J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*, 681–692.

Zajkowski, W.K., Kossut, M., Wilson, R.C. (2017). A causal role for right frontopolar cortex in directed, but not random, exploration. *eLife*, *6*, e27430.