Learning optimal decisions with confidence

Jan Drugowitsch[1,2] and Alexandre Pouget[2,3,4]
[1]Department of Neurobiology, Harvard Medical School, Boston, MA, USA
[2]University of Geneva, Geneva, Switzerland
[3]Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA
[4]Gatsby Computational Neuroscience Unit, University College of London, London, UK

Abstract

Diffusion decision models are immensely successful models for human and animal decisions under uncertainty. The decisions they model require the temporal accumulation of evidence to improve choice accuracy, and thus balance the trade-off between the decisions' speed and their accuracy. Commonly, diffusion models have a one-dimensional abstract input that represents noisy momentary decision-related evidence. However, the nervous system typically uses population codes to represent sensory variables, which implies that the momentary evidence is distributed across multiple inputs. It is currently unknown how decision makers could learn how to combine these multiple inputs to obtain a one-dimensional abstract decision variable. We present here a Bayesian learning rule for learning a near-optimal linear combination of these inputs based on trial-by-trial feedback. The rule is Bayesian in the sense that it learns not only the mean of the weights but also the uncertainty around this mean in the form of a covariance matrix. This yields a rule whose learning rate is strongly modulated by decision confidence, providing a computational role of the latter for every-day decisions. Furthermore, we show that, in volatile environments, the rule predicts a bias towards repeating the same choice after correct decisions, with a bias strength that is modulated by the previous choice' difficulty. Last, we extend our learning rule to cases for which one of the choices is more likely a-priori, which provides new insights into how such biases modulate the mechanisms leading to optimal decisions in diffusion models.

Introduction

Decisions are a ubiquitous component of every-day behavior. For decisions to be efficient, they need to handle the uncertainty arising from the noisy and ambiguous information that the environment provides us (Doya et al. 2006). One reflection of this is the trade-off between the speed and accuracy of decisions. Fast choices rely on little information, and are therefore usually of low accuracy. In contrast, slow choices result in the opportunity to accumulate more evidence and are thus more likely correct, but might come at a higher cost of accumulating this evidence (e.g., cost of attention, loss of time, opportunity cost). Therefore, efficient decisions require not only a mechanism to accumulate evidence, but also one to trigger a choice once enough evidence has been collected. An exceedingly popular model family known as *diffusion models* (or *diffusion decision models*; DDMs) (Ratcliff 1978) provides both mechanisms, and does not only yield surprisingly good fits to human and animal behavior (Bogacz et al. 2006; Ratcliff and McKoon 2008; Ratcliff and Smith 2004), but is also known to implement the Bayes-optimal decision strategy under a wide range of circumstances (Bogacz et al. 2006; Drugowitsch et al. 2012; Drugowitsch, Deangelis, et al. 2014; Frazier and Yu 2008; Gold and Shadlen 2002; Tajima, Drugowitsch, and Pouget 2016).

Diffusion models assume a particle that drifts and diffuses until it reaches one of two boundaries, each triggering a different choice (**Fig. 1a**). Its drift reflects a net evidence towards one of two choices, like the motion direction (drift sign) and coherence (drift magnitude) in a random-dot motion (RDM) task. Its stochastic diffusion reflects the uncertainty in the momentary evidence. The stochasticity of this diffusion does not only cause variability in the time that boundaries are reached and decisions are triggered, but also in which of the two boundaries is reached first, resulting in potential errors. Both of these features are also observed in human in animal decision makers performing tasks that require the trade-off between the decision accuracy and its speed (Ratcliff and McKoon 2008; Ratcliff and Smith 2004). A standard assumption underlying such models is that of a one-dimensional noisy momentary evidence that is accumulated over time to form the particle's location. This momentary evidence is an abstraction of the

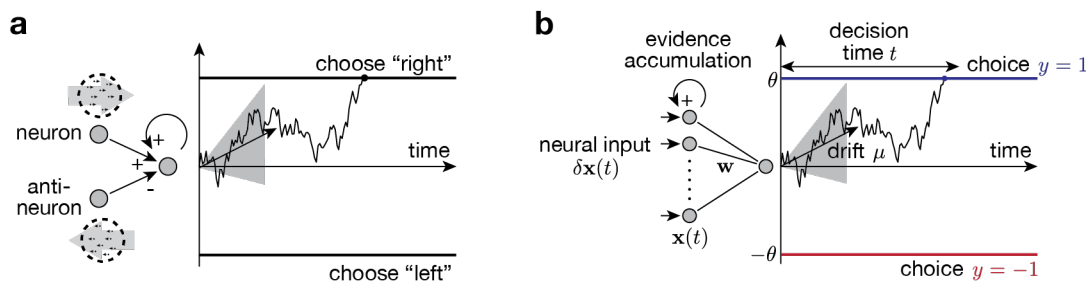

Figure 1. **Diffusion models with single and multiple inputs**. In diffusion models, the input(s) provide at each point in time noisy evidence about the world's true state, here given by the drift μ. The decision maker accumulates this evidence over time (illustrated here by the black traces) to form a belief about μ. As shown in the main text, Bayes-optimal decisions choose according to the sign of the accumulated evidence, justifying the two decision boundaries that trigger opposing choices. (**a**) In standard diffusion models, the momentary evidence either arises directly from noisy samples of μ, or, as illustrated here, from a neuron/anti-neuron pair that codes for opposing directions of evidence. In this example, the task is assumed to be a random dot task, in which the decision maker needs to identify if most of the dots that compose the stimulus are moving either to the left or to the right. The two neurons (or neural pools) are assumed to extract motion energy of this stimulus towards the right (top) and left (bottom), such that their difference forms the momentary evidence towards right-ward motion. A decision is made once the accumulated momentary evidence reaches one of two decision boundaries, triggering opposing choices. (**b**) Our setup differs from that in (**a**) in two ways. First, we assume the input information $\delta\boldsymbol{x}(t)$ to be encoded in a larger neural population whose activity is linearly combined to yield the one-dimensional momentary evidence. Second, we assume that the decision maker is initially unaware of the correct linear combination weights $\boldsymbol{w}$, and needs to learn them from feedback about the correctness of her choices.

decision-related evidence that some stimulus provides at each point in time. In reality, however, this evidence is usually distributed across a larger number of inputs. In the brain, for example, it would be encoded in the activity of large neural populations, rather than in individual neurons (or neuron/anti-neuron pairs; **Fig. 1a**). Furthermore, a-priori, the brain would not know how this distributed encoding provides information about the correctness of either choice. As a consequence, it needs to learn how to interpret neural population activity from feedback provided to previous choices. How such an interpretation can be learned over time is the focus of this work.

There already exist multiple computational models for how humans and animals might learn to improve their decisions (e.g., Courville, Daw, and Touretzky 2006; Dayan, Kakade, and Montague 2000; Dayan and Yu 2003; Körding and Wolpert 2004). They all assume that, for each choice, all evidence is provided at once, without considering the temporal aspect of evidence accumulation. This is akin fixed-duration experiments, in which the evidence accumulation time is determined by the environment rather than the decision maker. We, instead, address a more natural reaction time setting, in which decision times are under the decision maker's control. This makes the interpretation of the accumulated evidence depend on how much time has been spent on accumulating it, which impacts how the decision strategy ought to be updated in the light of feedback. We furthermore extend on previous work by considering a-priori biases towards believing that one option is more correct than the others, and how such biases can be learned over time. This yields a new theoretical understanding of how such biases impact optimal decision-making in diffusion models. Furthermore, it clarifies of how different implementations of this prior bias result in different diffusion model implementations, like the one proposed by Hanks et al. (2011).

We will first review the theory that demonstrates the optimality of diffusion models for evidence accumulation for one-dimensional momentary evidence. We then extend this theory to handle multi-dimensional inputs, whose linear combination results in the same statistics as the previously considered one-dimensional input. Next, we ask how to best learn the linear combination weights from feedback about the correctness of previous choices. Ideally, this learning would correspond to sequentially updating the Bayesian posterior over these weights for every new feedback that is received. However, this turns out to be intractable which is why we instead resort to an approximation scheme, which we show in simulations

to achieve close-to-optimal posterior inference and to significantly outperform simpler heuristics. Last, we show how this learning scheme can be extended to learning a biased prior, and its consequences on the dependency between sequential choices.

Results

Bayes-optimal decision-making with diffusion models

One standard way (Drugowitsch et al. 2012; Drugowitsch, Deangelis, et al. 2014; Moreno-Bote 2010) to interpret diffusion models as mechanistic implementations of Bayes-optimal decision-making is to assume that, in each trial, a latent state μ (or *drift rate* in the terminology of diffusion models) is drawn from a prior distribution, $\mu \sim \mathrm{N}(0, \sigma_\mu^2)$, with zero mean and variance σ_μ^2. The decision maker's aim is to infer whether this latent state is positive or negative (e.g., leftward vs. rightward motion in an RDM task), irrespective of its magnitude (e.g., the RDM task coherence). The latent state itself is unobserved, but the decision maker aims to infer its sign based on a stream of noisy, momentary evidence, $\delta z_1, \delta z_2, \ldots$. In each small time step of size δt, this momentary evidence provides noisy information about μ through $\delta z_i | \mu \sim \mathrm{N}(\mu \delta t, \delta t)$. Here, we have chosen a unit variance, scaled by δt. Any re-scaling of this variance by an additional parameter would result in a global re-scaling of the evidence that can be factored out (Bogacz et al. 2006; Drugowitsch et al. 2012; Palmer, Huk, and Shadlen 2005), thus making such a re-scaling unnecessary.

Having after some time $t \equiv n\delta t$ observed n pieces of such evidence, $\delta z_{1:n}$, the decision maker's best guess about the latent state μ is given by the posterior (see Methods)

$$p(\mu | \delta z_{1:n}) \propto p(\mu) \prod_{i=1}^{n} p(\delta z_i | \mu) \propto \mathrm{N}\left(\mu | \frac{z(t)}{t + \sigma_\mu^{-2}}, \frac{1}{t + \sigma_\mu^{-2}}\right), \tag{1}$$

where all proportionalities are with respect to μ, and where we have defined $z(t) = \sum_{i=1}^{n} \delta z_i$ as the accumulated evidence. From this we find the posterior belief about μ being positive by

$$p(\mu \geq 0 | z(t), t) = \int_0^\infty \mathrm{N}(\mu | \delta z_{1:n}) \mathrm{d}\mu = \Phi\left(\frac{z(t)}{\sqrt{t + \sigma_\mu^{-2}}}\right), \tag{2}$$

where $\Phi(\cdot)$ is the cumulative function of a standard Gaussian. The opposite belief about μ being negative is simply $p(\mu < 0 | z(t), t) = 1 - p(\mu \geq 0 | z(t), t)$.

In the above, the accumulated evidence follows a diffusion process, $z(t) | \mu \sim N(\mu t, t)$, and thus can be interpreted as the location of a drifting and diffusing particle, where latent state μ determines the drift (**Fig. 1a**). The posterior belief about the latent state's sign depends on the sign of the particle's location. If $z(t)$ is positive, we have $p(\mu \geq 0 | z(t), t) > 1/2$, such that the latent state is more likely positive than negative. If $z(t)$ is negative, on the other hand, a negative latent state is more likely. This justifies bounding $z(t)$ from below and above by some (possibly time-varying) boundaries $\pm\theta(t)$ and make one or the other choice depending on which boundary was reached first (**Fig. 1a**). With such a decision mechanism, the posterior belief about having made the correct choice, or decision confidence (Pouget, Drugowitsch, and Kepecs 2016), is then given by Eq. (2) with $z(t)$ replaced by $\theta(t)$. Overall, this demonstrates that diffusion models can be interpreted as mechanisms that, under certain assumptions of prior and evidence, implement the Bayes-optimal decision strategy.

Before proceeding to multi-dimensional evidence, let us justify our choice of the latent state prior, $\mu \sim \mathrm{N}(0, \sigma_\mu^2)$. The amount of information provided about the sign of μ by each piece of momentary evidence depends on the magnitude of μ. If $|\mu|$ is small, then the δz_i's will be close to zero, such that it is

hard to tell if the generating μ was positive or negative. If $|\mu|$ is large, in contrast, the δz_i's will be far away from zero, such that their sign is very informative about the sign of μ. Therefore, the magnitude of μ determines the difficulty of individual trials, and larger $|\mu|$'s imply easier trials. Using a zero-mean Gaussian prior implies a uniform prior over the sign of μ. Furthermore, it implies that small $|\mu|$'s are more likely than large ones, corresponding to harder trials being more likely than easy one – a standard setup in decision-making experiments. Here, σ_μ^2 determines the spread of μ's across trials, and therefore the overall difficulty of the task (larger σ_μ^2 = overall easier task). In general, what is important here is the assumption that the difficulty varies across trials, but not exactly how it varies, as determined by the choice of the prior (Drugowitsch et al. 2012). Qualitatively, our results should be robust to different choices of this prior. Model predictions would change qualitatively if we assume the difficulty to be known a-priori (see Drugowitsch et al. 2012), but we will not consider this case, as it rarely if ever occurs in the real world.

A generative model for high-dimensional diffusion model inputs

Let us extend the above model to the case of a multi-dimensional momentary evidence. That is, let us assume the momentary evidence in each time step is given the a k-dimensional vector $\delta\boldsymbol{x}_i$. This evidence might represent inputs from multiple sensors, or the (abstract) activity of a neuronal population (**Fig. 1b**). We furthermore extend the decision-making mechanism by triggering choices as soon as the linear combination $\boldsymbol{w}^T\boldsymbol{x}(t)$ of the accumulated evidence $\boldsymbol{x}(t) = \sum_{i=1}^{n} \delta\boldsymbol{x}_i$ reaches one of two decision boundaries $\pm\theta(t)$, where $\boldsymbol{w}$ denote the k combination weights (for now assumed known). Thus, the linearly combined accumulated evidence now replaces the diffusion model particle. Nonetheless, to maintain parallels to diffusion models, we assume the momentary evidence to be generated according to

$$\delta\boldsymbol{x}_i|\mu,\boldsymbol{w} \sim \mathrm{N}\left(\frac{\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{w}}\mu\delta t, \frac{\boldsymbol{I}}{\boldsymbol{w}^T\boldsymbol{w}}\delta t\right), \tag{3}$$

such that $\boldsymbol{w}^T\boldsymbol{x}(t)|\mu = z(t)|\mu \sim N(\mu t, t)$, as before. Therefore, the posterior belief about μ is given by

$$p(\mu|\boldsymbol{w},\boldsymbol{x}(t),t) = \mathrm{N}\left(\mu|\frac{\boldsymbol{w}^T\boldsymbol{x}(t)}{t+\sigma_\mu^{-2}}, \frac{1}{t+\sigma_\mu^{-2}}\right), \tag{4}$$

just as in Eq. (1). Furthermore, the posterior belief about its sign is

$$p(\mu \geq 0|\boldsymbol{w},\boldsymbol{x}(t),t) = \Phi(\boldsymbol{w}^T\tilde{\boldsymbol{x}}(t)), \tag{5}$$

where we have defined the time-attenuated accumulated evidence $\tilde{\boldsymbol{x}}(t) = \boldsymbol{x}(t)/\sqrt{t+\sigma_\mu^{-2}}$.

Knowing the expression for the posterior belief makes it easy to derive the decision confidence. We define this confidence as the belief about having made the correct decision. At the upper boundary, at which $\boldsymbol{w}^T\boldsymbol{x}(t) = \theta(t)$, the decision is correct if $\mu \geq 0$. Thus, the decision confidence in this case is the above posterior belief with $\boldsymbol{w}^T\tilde{\boldsymbol{x}}(t)$ replaced by $\theta(t)/\sqrt{t+\sigma_\mu^{-2}}$. At the lower boundary, at which $\boldsymbol{w}^T\boldsymbol{x}(t) = -\theta(t)$, the decision is correct if $\mu < 0$. In this case, the decision confidence is one minus the above posterior belief with $\boldsymbol{w}^T\tilde{\boldsymbol{x}}(t)$ replaced by $-\theta(t)/\sqrt{t+\sigma_\mu^{-2}}$. However, using $1-\Phi(a) = \Phi(-a)$, we return to the same expression as before, such that for either decision, the confidence in choice y is given by

$$p(y|t) = \Phi\left(\frac{\theta(t)}{\sqrt{t + \sigma_\mu^{-2}}}\right). \tag{6}$$

For time-independent decision bounds, $\theta(t) = \theta$, this confidence decreases over time. This decrease reflects the uncertainty about μ, and that late choices are likely due to a low μ, which is associated with a hard trial, and thus low decision confidence.

Using feedback to find the posterior weights

So far we have assumed that the decision maker knows the linear combination weights $\boldsymbol{w}$ to combine the evidence to perform Bayes-optimal decisions. How could she learn these weights if they were unknown? To address this question, consider the following setup. The decision maker has some belief, $p(\boldsymbol{w})$, about these weights, either as a prior or formed through previous experience. She observes new evidence, $\delta\boldsymbol{x}_1, \delta\boldsymbol{x}_2, \ldots$ and uses the mean of the belief over weights, $\langle\boldsymbol{w}\rangle$, as weights to combine this evidence, and to trigger a choice y once the combined evidence reaches one of the decision boundaries. Once decided, she receives feedback y^* about which choice was correct. How could she update her belief $p(\boldsymbol{w})$ in light of this feedback?

Let us be more specific about what we mean with feedback y^* about the correct choice. First, we assume that there is one true set of weight $\boldsymbol{w}^*$ that, together with the current latent state μ, is used generate the momentary evidence according to the above $\delta\boldsymbol{x}_i|\mu, \boldsymbol{w}^*$. Furthermore, we define the correct choice as the choice that correctly identifies the sign of the latent state, $y^* = \mathrm{sgn}(\mu) \in \{-1,1\}$. This is the one that the decision maker is informed about after her own decision. From the decision maker's perspective, who does not know the latent state μ, this correct choice is only stochastically related to the true weights and the accumulated evidence. In fact, the distribution $p(y^*|\boldsymbol{w}^*, \boldsymbol{x}(t), t)$ that describes this stochastic relationship is the weights' likelihood function, as it specifies how likely it is to observe the given feedback y^* for a set of weights. Thus, we now have a prior $p(\boldsymbol{w})$ and a likelihood $p(y^*|\boldsymbol{w}, \boldsymbol{x}(t), t)$, which is all we need to find the posterior weights by Bayes' rule,

$$p(\boldsymbol{w}|y^*, \boldsymbol{x}(t), t) \propto p(y^*|\boldsymbol{w}, \boldsymbol{x}(t), t)p(\boldsymbol{w}), \tag{7}$$

where the proportionality is with respect to the weights.

What remains is to find the form of the likelihood. This likelihood describes how the sign of the latent state μ, which determines the correct choice y^*, relates to a given set of weights $\boldsymbol{w}$ and some accumulated evidence $\boldsymbol{x}(t)$. As it turns out, we already know the form of this likelihood, as it is in fact the same as the posterior belief about the latent state's sign, $p(\mu \geq 0|\boldsymbol{w}, \boldsymbol{x}(t), t)$, which we have computed further above (Eq. (5)). This likelihood is not restricted to performed choices y, but can be equally used to ask how likely the weights $\boldsymbol{w}$ are for hypothetical choices y^* that differ from y if the choice y was wrong. Therefore, we can use it to update our belief about these $\boldsymbol{w}$'s once we have received feedback y^*. Conveniently, it does not matter for this update which combination weights the decision maker used to commit to a choice, as the likelihood is a function over hypothetical rather than actual weights that only depends on feedback and accumulated evidence. Nonetheless, for all of the below simulations we assume the decision maker to use the mean of her current belief over these weights.

Let us now turn to actually computing the posterior over combination weights. To do so, we assume a Gaussian prior, $p(\boldsymbol{w}) = N(\boldsymbol{w}|\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$, which results in the posterior to be given by

$$p(\boldsymbol{w}|y^*, \boldsymbol{x}(t), t) \propto \Phi\big(y^*\boldsymbol{w}^T\tilde{\boldsymbol{x}}(t)\big)\mathrm{N}(\boldsymbol{w}|\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w), \tag{8}$$

where we have used the property that $1 - \Phi(a) = \Phi(-a)$, which allows us to use y^* to modulate the sign of the combined accumulated evidence instead of using different likelihood functions for different values of y^*. As the likelihood parameters are linear within a cumulative Gaussian function, such problems are known as *Probit Regression* and don't have a closed-form expression for the posterior. We could proceed by sampling from the posterior by Markov Chain Monte Carlo methods, but that would not provide much insight into the different factors that modulate learning the posterior weights. Instead, we proceed by deriving a closed-form approximation to this posterior to provide such insight.

Decision confidence with uncertain combination weights

Before finding an expression for the approximate posterior, let us consider the impact of uncertain combination weights on the decision confidence. We have previously derived this confidence under the assumption of knowing $\boldsymbol{w}$. Having uncertainty about these weights introduces uncertainty about if the incoming evidence was accumulated correctly. A consequence is a further drop in the decision confidence. Mathematically, we find this decision confidence by taking the posterior belief for a known $\boldsymbol{w}$ and marginalizing out the distribution over the weights, $p(\boldsymbol{w}) = N(\boldsymbol{w}|\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$, resulting in (see Methods)

$$p(\mu \geq 0|\boldsymbol{x}(t), t) = \int p(\mu \geq 0|\boldsymbol{w}, \boldsymbol{x}(t), t)p(\boldsymbol{w})\mathrm{d}\boldsymbol{w} = \Phi\left(\frac{\boldsymbol{\mu}_w^T\widetilde{\boldsymbol{x}}(t)}{\sqrt{1 + \widetilde{\boldsymbol{x}}^T\boldsymbol{\Sigma}_w\widetilde{\boldsymbol{x}}}}\right). \tag{9}$$

From this marginal posterior belief we can use the same argument as before to find the choice confidence

$$p(y|x(t), t) = \Phi\left(\frac{y\boldsymbol{\mu}_w^T\widetilde{\boldsymbol{x}}(t)}{\sqrt{1 + \widetilde{\boldsymbol{x}}^T\boldsymbol{\Sigma}_w\widetilde{\boldsymbol{x}}}}\right). \tag{10}$$

If we accumulate evidence using the mean $\boldsymbol{\mu}_w$ as an estimate of $\boldsymbol{w}$, then we will have $y\boldsymbol{\mu}_w^T\widetilde{\boldsymbol{x}}(t) = \theta(t)/\sqrt{t + \sigma_\mu^{-2}}$ at the time of the choice (recall that $y = -1$ when deciding at the lower boundary $-\theta(t)$). Even then, and in contrast to Eq. (6), the decision confidence will depend on both time and the accumulated evidence because the influence of the uncertainty about $\boldsymbol{w}$ on the decision confidence depends on the state of the accumulated evidence $\widetilde{\boldsymbol{x}}(t)$, which appears in the denominator.

Decision confidence controls the learning rate

We use *Assumed Density Filtering* (ADF) (Bishop 2006; Murphy 2012) to find a close-form approximation to incrementally updating the moments of the posterior belief. ADF is based on assuming a particular distributional form q to approximate the posterior, and then minimizing the Kullback-Leiber divergence $\mathrm{KL}(p|q)$ between the true posterior p and its approximation q. An incremental update of the posterior moments after each choice requires the approximate posterior to have the same form as the prior, namely to be Gaussian. Denoting this posterior by $q(\boldsymbol{w}|y^*, \boldsymbol{x}(t), t) = \mathrm{N}(\boldsymbol{w}|\boldsymbol{\mu}_w^*, \boldsymbol{\Sigma}_w^*) \approx p(\boldsymbol{w}|y^*, \boldsymbol{x}(t), t)$, the aim becomes to find the posterior moments $\boldsymbol{\mu}_w^*$ and $\boldsymbol{\Sigma}_w^*$ that minimize

$$\mathrm{KL}\left(\frac{1}{Z}\Phi(y^*\boldsymbol{w}^T\widetilde{\boldsymbol{x}}(t))\mathrm{N}(\boldsymbol{w}|\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)\middle|\mathrm{N}(\boldsymbol{w}|\boldsymbol{\mu}_w^*, \boldsymbol{\Sigma}_w^*)\right), \tag{11}$$

where Z is an adequate normalization constant. As the posterior approximation is Gaussian and thus in the exponential family of distributions, this minimization corresponds to choosing the moments of q to match the moments of the true posterior (see Methods), resulting in

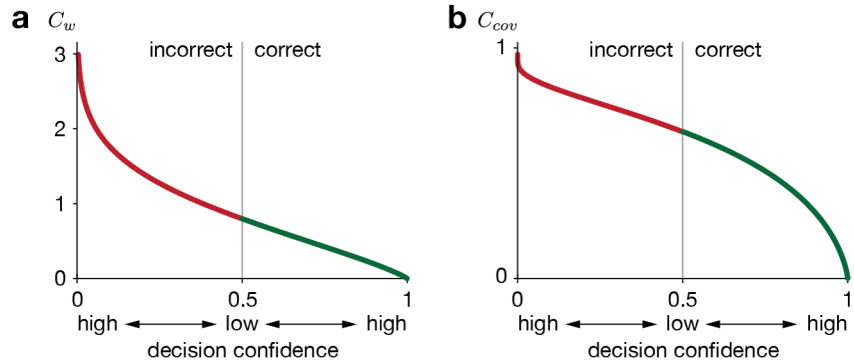

Figure 2. Learning rate modulators for mean and covariance update. (**a**) The learning rate modulator C_w for the combination weight mean as a function of the posterior belief over the feedback. The green part of the curve corresponds to correct choices, whereas the red part corresponds to incorrect choices. For incorrect choices, the belief about the correct choice being correct is below ½. It is close to zero if the choices are made with high confidence. (**b**) The learning rate modulator C_{cov} for the covariance matrix update. The color scheme is the same as in (**a**).

$$\boldsymbol{\mu}_w^* = \boldsymbol{\mu}_w + \frac{C_w}{\sqrt{1 + \tilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}}} y^* \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}, \tag{12}$$

$$\boldsymbol{\Sigma}_w^* = \boldsymbol{\Sigma}_w + C_{cov}((\boldsymbol{\Sigma}_w^{-1} + \tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T)^{-1} - \boldsymbol{\Sigma}_w), \tag{13}$$

with learning rate modulators

$$C_w = \frac{\mathrm{N}\left(\frac{y^* \boldsymbol{\mu}_w^T \tilde{\boldsymbol{x}}}{\sqrt{1 + \tilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}}} \middle| 0,1\right)}{\Phi\left(\frac{y^* \boldsymbol{\mu}_w^T \tilde{\boldsymbol{x}}}{\sqrt{1 + \tilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}}}\right)}, \tag{14}$$

$$C_{cov} = C_w \left(C_w + \frac{y^* \boldsymbol{\mu}_w^T \tilde{\boldsymbol{x}}}{\sqrt{1 + \tilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}}} \right). \tag{15}$$

Note that the factor C_w modulates how strongly the mean $\boldsymbol{\mu}_w$ of the belief over the combination weights is updated. This factor is a deterministic function of the confidence related to the feedback, $p(y^*|\boldsymbol{x}(t), t)$, and a monotonically decreasing function thereof (**Fig. 2a**). For incorrect choices, in which case $y \neq y^*$, this confidence will be less than $p(y^*|\boldsymbol{x}(t), t) = 1 - p(y|\boldsymbol{x}(t), t) < 1/2$. C_w is largest for incorrect choices that were made with high confidence, in which case strong adjustments are necessary. Uncertain choices in general are followed by more moderate updates. For correct choices, in which case feedback and choice confidence are equal, $p(y^*|\boldsymbol{x}(t), t) = p(y|\boldsymbol{x}(t), t) > 1/2$, C_w drops to zero for highly confident choices. For uncertain, correct choices, the adjustments are again moderate, promoting a higher confidence for future choices.

The covariance update is modulated by the learning rate C_{cov}, which is also a function of the decision confidence (**Fig. 2b**). As for C_w, C_{cov} promotes the strongest updates if incorrect high-confidence decisions, only weak updates for correct high-confidence decisions, and moderate updates for decisions of low confidence.

Further modulators of the update strength for both mean and covariance come in through $\tilde{\boldsymbol{x}}$, which is the accumulated evidence $\boldsymbol{x}$ modulated by time. At the decision time, we have $y\boldsymbol{\mu}_w^T \boldsymbol{x} = \theta(t)$, such that the magnitude of $\tilde{\boldsymbol{x}}$ depends on both time and the current size of $\boldsymbol{\mu}_w$. Therefore, the way it modulates the mean and covariance updates is less straight-forward.

Performance comparison to optimal inference and to simpler heuristics

Approximating Bayes-optimal learning by ADF might be associated with a significant performance drop. To quantify this drop, we compared the performance of our ADF rule to that of the Bayes-optimal rule. We found the learning performance of the latter by Gibbs sampling (see Methods), which requires a complete memory of inputs and feedbacks for past decisions, and becomes intractable for longer decision sequences.

Furthermore, we tested if simpler learning rules can match ADF performance. We focused on three rules of increasing complexity. The *delta rule* updates its weight estimate after the nth decision by

$$\boldsymbol{w}^{n+1} = \boldsymbol{w}^n + \alpha(y^{n*}\theta(t) - \boldsymbol{x}^n(t)^T\boldsymbol{w}^n)\boldsymbol{x}^n(t), \tag{16}$$

where $y^{n*} \in \{-1,1\}$ is the feedback about the correct choice provided after this decision. As decisions are triggered at one of the two boundaries, $\boldsymbol{x}^n(t)^T\boldsymbol{w}^n \in \{-\theta(t), \theta(t)\}$, the residual in brackets is zero for correct choices, and $\pm 2\theta(t)$ for incorrect choices. In contrast to ADF, weight adjustments are only performed after incorrect choices, and then with a fixed learning rate α rather than one modulated by confidence. As shown further below, we found that the delta rule did not yield stable weight learning. Therefore, we additionally considered a *normalized delta rule*, that updates the weight estimates as the delta rule, but thereafter normalizes them by $\boldsymbol{w} \leftarrow \boldsymbol{w}/\|\boldsymbol{w}\|$ to ensure a unit weight norm. Lastly, we tested a learning rule that performs stochastic gradient ascent on the feedback log-likelihood,

$$\boldsymbol{w}^{n+1} = \boldsymbol{w}^n + \alpha\nabla_w \log p(y^{n*}|\boldsymbol{w}^n, \boldsymbol{x}^n(t), t) = \boldsymbol{w}^n + \alpha y^{n*}C_w\tilde{\boldsymbol{x}}^n(t). \tag{17}$$

This rule introduces decision confidence weighting through C_w, but differs from ADF in that it does not take the weight uncertainty (in ADF expressed through $\boldsymbol{\Sigma}_w$) into account.

To evaluate the performance of either learning rule, we simulated weight learning across 500 consecutive decisions (called *trials*; see Methods for details), and evaluated learning rule performance by computing at each step the angle between the true weight vector and that estimated by the learning rule. Using this performance measure focuses directly on the weight vector that we aim at learning while disregarding its magnitude, which is important for decision confidence, but irrelevant for making correct choices.

Figure 3 illustrates the average angular error across 10,000 repetitions of this procedure for all learning rules for different numbers of inputs. As can be seen, the performance of ADF is almost indistinguishable from Bayes-optimal weight learning for all tested numbers of inputs. All other learning heuristics perform significantly worse. The delta rule initially improves its weight vector estimate but worsens it again at a later stage, as is particularly noticeable for higher learning rates. The reason for this is that its implicit objective function is to avoid making wrong choices, which is best achieved by accumulating more evidence. This, in turn, can be achieved by down-weighting the momentary evidence by lowering the magnitude of the combination weights (see insets in **Fig. 3**). For a fixed learning rate, however, such lower magnitudes cause relatively stronger learning-induced across-trial changes of the combination weights, with an associated increase in angular errors. The normalized delta rule avoids such weight shrinkage, and thus doesn't feature this angular error inflection. Nonetheless, its angular error is significantly larger than that of the Bayes-optimal learning rule for all learning rates. The only heuristic that comes close to ADF and Bayes-optimal learning is stochastic gradient ascent on the log-likelihood, but with the downside of having to choose a learning rate. Small learning rates lead to slow learning, and an associated slower drop in angular error. Large learning rates initially match close-to-optimal performance, but later over-correct the weight estimates, leading to suboptimal asymptotic performance. Overall, ADF significantly outperformed all tested heuristic learning rules and matched the weight learning performance of the Bayes-optimal estimator.

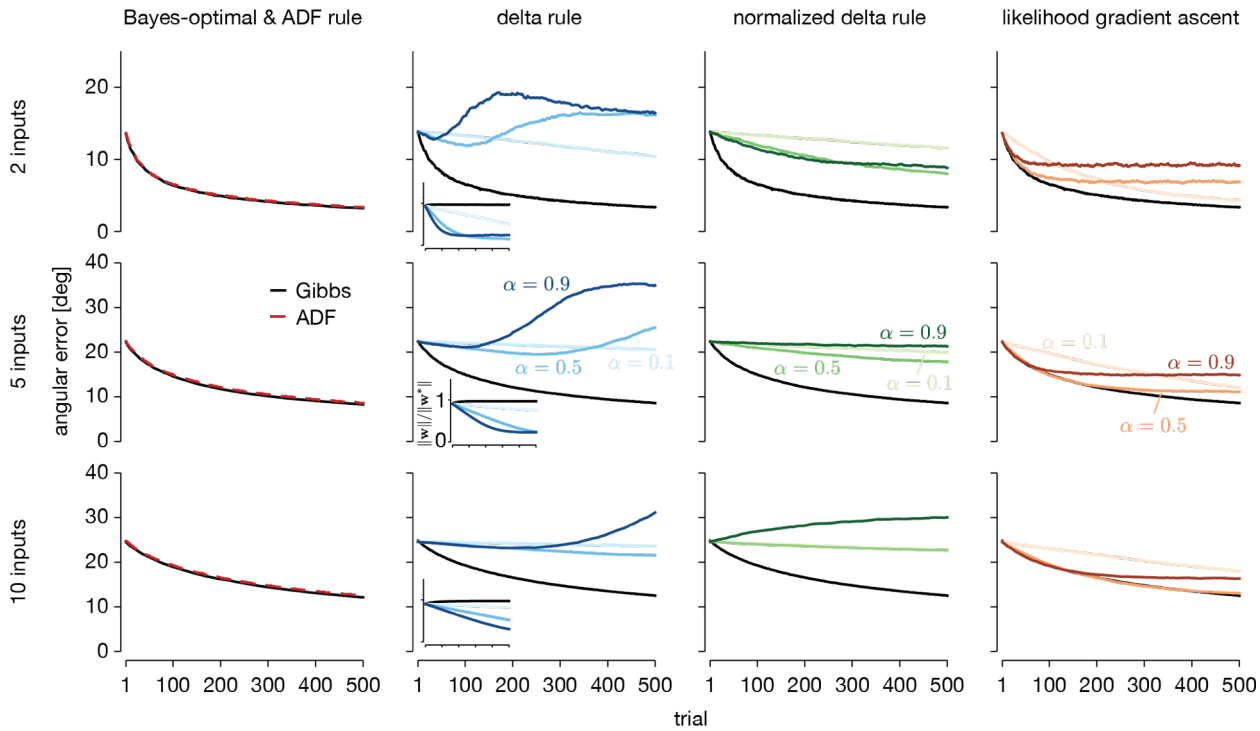

Figure 3. Learning performance of different learning rules. The figure shows the change in angular error (angle between estimated and true weight vector) over trials for different numbers of inputs (top to bottom) and different learning rules (left to right), averaged over 10,000 simulations. The left-most panels illustrate that the automated relevance determination (ADF) rule (red dashed lines) almost perfectly matches the learning performance of the Bayes-optimal rule (solid black line in all panels), here computed by Gibbs sampling. Both methods track a whole weight posterior distribution, and the angular error is computed using the mean of this posterior. The performance of the other learning rules is influenced by their learning rate parameter α, whose value differs across curves of different color. The insets for the delta rule additionally illustrate the size of the weight vector in relation to the true weights (one = same size) over trials, and shows that this size decreases over time. For the normalized delta rule with 10 inputs, the angular error for $\alpha = 0.1$ matches that for $\alpha = 0.5$, such that its associated performance curve is barely visible.

Tracking non-stationary combination weights

So far, we have tested how fast our weight learning rule is able to learn the true, underlying weights from binary feedback about the correctness of the decision maker's choices. For this we assumed that the true weights remained constant across decisions. What would happen if these weights change slowly over time? Such a scenario could occur if, for example, the world around us changes slowly, or if the neural representation of this world changes slowly through neural plasticity or similar. In this case, the true weights would become a moving target that we would never be able to learn perfectly. Instead, we would after some initial transient expect to reach steady-state performance that remains roughly constant across consecutive decisions. We compared this steady-state performance associated with the Bayes-optimal learning to the ADF learning rule and the heuristics introduced in the previous section.

We modeled a change of the true weights by a first-order autoregressive process that is the discrete-time equivalent of an Ornstein-Uhlenbeck process. After each decision, the true weights diffuse according to a multivariate isotropic Gaussian, and subsequently leak towards a pre-set mean (see Methods for details). The combination of diffusion and leak ensures that the weights change slowly across time while remaining bounded. Both the Bayesian model, which in this case is implemented by a particle filter, and ADF are modified to consider the dynamics of the true weights. Specifically, after updating the moments of the weight vector posterior following feedback, its covariance is amplified and its mean leaks in

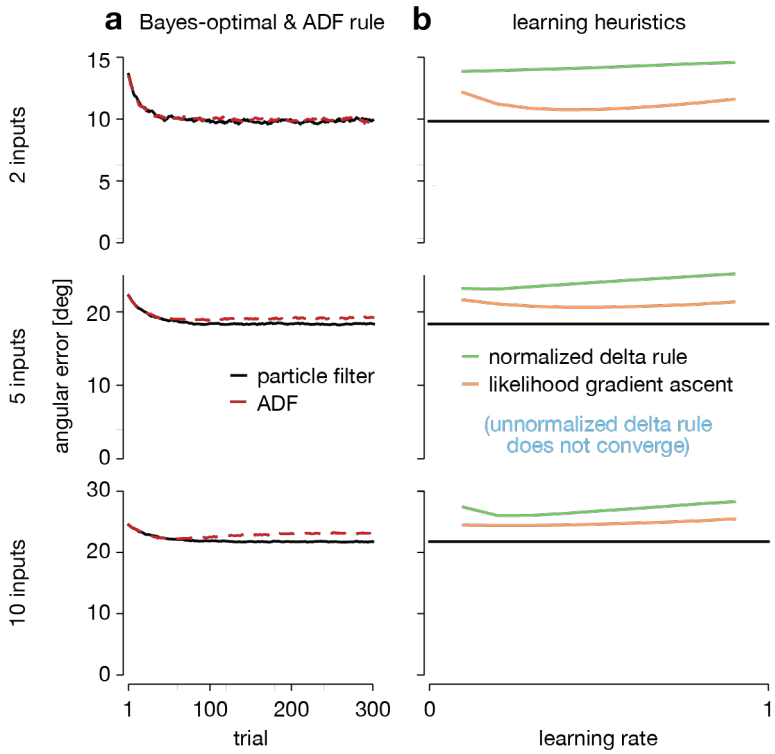

Figure 4. Steady-state learning performance in a volatile world. The figure shows the performance of different learning rules (measured as angle between estimated and true weight vector) in an environment in which the true weight vector continuously changes, for different numbers of inputs (top to bottom), averaged over 10,000 simulations. (**a**) Both Bayes-optimal learning (here, a particle filter) and the ADF rule converge to their respective steady-state performance after approximately 100-200 trials. The performance difference between the ADF rule and Bayes-optimal learning remains small, but increases with the dimensionality of the input. (**b**) The steady-state performance of the heuristic learning rules depends on the chosen learning rate α (horizontal axis). Using a rule that follows the likelihood gradient consistently outperforms the normalized delta rule, and neither rule achieves the same performance as the particle filter or ADF rule, for any choice of the learning rate.

correspondence to diffusion and leak of the true weight vector. The heuristic learning rules remain unmodified, as their use of a constant learning rate already encapsulates the assumption that the true weights change across decisions.

Figure 4 illustrates the performance of the different learning rules, measured by the angle between the estimated and true weight vectors, and averaged over 100,000 simulations. As **Fig. 4a** shows, both the Bayes-optimal rule and ADF achieve steady-state performance after around 100-200 trials. The ADF rule only approximates the Bayes-optimal rule, and the performance drop associated with this approximation slightly grows with an increase in the input (and weight vector) dimensionality. Nonetheless, even for a 10-dimensional input, it remains small. The heuristic learning rules have an additional tunable learning rate parameter, and their steady-state performance is shown in **Fig. 4b** for different choices of this parameter, ranging from 0.1 to 0.9. The normalized delta rule is for all learning rate choices outperformed by the rule that relies performs stochastic gradient ascent on the likelihood gradient. This should come as no surprise, as, due to the inclusion of the likelihood function, the latter relies on significantly more knowledge about the structure of the task. The delta rule never converges, and its performance is therefore not shown. If the only advantage of the ADF rule over the heuristics is a better one-time adjustment of its learning rate, then there should exist a learning rate for the heuristics at which they match the performance of ADF. As **Fig. 4b** shows, this is not the case, illustrating that the difference between ADF and the heuristic learning rules cannot simply be explained by a fine-tuning of the learning rate. Overall, **Fig. 4** shows that, even when

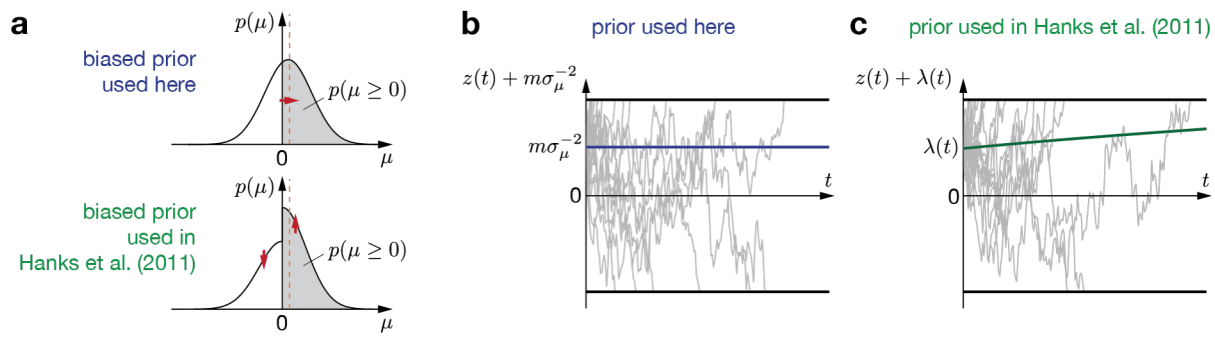

Figure 5. The impact of prior shape on optimal decision-making. (**a**) We assume a biased prior to be implemented by shifting the distribution over μ by m (here, to the right; top panel). Hanks et al. (2011), in contrast, assume the prior to implemented by boosting one half of the distribution, while dampening the other half (bottom panel). Both priors illustrated here implement a prior bias of $p(\mu \geq 0) = 0.6$. (**b**) For a bias resulting from a shift of $p(\mu)$, optimal decision-making is implemented by a shift in the diffusion model starting point by $m\sigma_\mu^{-2}$ (blue line). (**c**) If the prior is instead implemented by a one-sided boost, as used by Hanks et al. (2011), optimal decision-making instead requires an additional signal $\lambda(t)$ that varies over time (green line). All 20 simulated trajectories (grey) in panels (**b**) and (**c**) use a drift rate μ as indicated by the orange dashed line in panel (**a**).

tracking the weight vector in a volatile environment, the ADF learning rule achieves close-to-optimal performance that is unmatched by alternative heuristics.

Learning both weights and a latent state prior bias

In all the above we have assumed that, a priori, both choices are equally likely to be correct. This arises from choosing a Gaussian prior on μ that is centered on zero. We will now relax this assumption by allowing a shift m in this prior that is unknown to the decision maker and thus needs to be learned. With this shift, the prior becomes $\mu \sim \mathrm{N}(m, \sigma_\mu^2)$. Choices $y = 1$ and $y = -1$ remain correct if $\mu \geq 0$ and $\mu < 0$, respectively, such that $y = 1$ is now a-priori correct with probability $p(\mu \geq 0) = \Phi(m/\sigma_\mu)$, which is $\geq 1/2$ if $m \geq 0$ and $< 1/2$ otherwise (**Fig. 5a**). We will first consider the consequence of such a biased prior for single-dimensional inputs, $\delta z_i | \mu \sim \mathrm{N}(\mu\delta t, \delta t)$, and will later generalize to the multi-dimensional case.

Following the same steps that led to Eq. (2), the posterior belief of $\mu \geq 0$ given accumulated evidence $z(t)$ at time t is given by

$$p(\mu \geq 0 | m, z(t), t) = \Phi\left(\frac{z + \sigma_\mu^{-2} m}{\sqrt{t + \sigma_\mu^{-2}}}\right). \tag{18}$$

In contrast to the bias-free posterior, this posterior has an additional term $\sigma_\mu^{-2} m$ whose relative influence wanes over time. This additional term requires us to re-think the association between decision boundaries and choices. Leaving the decision boundaries on $z(t)$ unchanged at $\pm\theta(t)$, would make it sometimes preferable to choose the option opposing this decision boundary. For example, consider a prior mean $m > 0$ that favors choice $y = 1$. If $\sigma_\mu^{-2} m > \theta(t)$, $z(t) + \sigma_\mu^{-2} m \geq 0$ when we reach the lower boundary $z(t) = -\theta(t)$. This, in turn, implies that the posterior $p(\mu \geq 0 | z(t), t, m) > 1/2$, such that choosing $y = 1$ is preferable over $y = -1$, despite having reached the lower boundary. In fact, in this case, $y = -1$ would never be chosen.

We avoid this ambiguity between decision boundaries and choices by shifting the decision boundaries. In particular, we shift both the upper and lower boundaries, $\theta^+(t)$ and $\theta^-(t)$, by $\sigma_\mu^{-2} m$ to

$\theta^+(t) = \theta(t) - \sigma_\mu^{-2}m$ and $\theta^-(t) = -\theta(t) - \sigma_\mu^{-2}m$. This has the following effect on the posterior belief. At the upper decision boundary, at which $z(t) = \theta(t) - \sigma_\mu^{-2}m$, the posterior belief is given by

$$p(\mu \geq 0 | m, z(t) = \theta^+(t), t) = \Phi\left(\frac{\theta(t)}{\sqrt{t + \sigma_\mu^{-2}}}\right). \tag{19}$$

At the lower decision boundary, at which $z(t) = -\theta(t) - \sigma_\mu^{-2}m$, we find the same posterior belief with $\theta(t)$ replaced by $-\theta(t)$. Therefore, making the decision boundaries asymmetric symmetrizes the posterior beliefs. This resolves the ambiguities between boundaries and choices, as choosing $y = 1$ and $y = -1$ is now again optimal at the upper and lower boundary, respectively. The same is achieved by shifting the integration starting point up to $z(0) = m\sigma_\mu^{-2}$ while leaving the boundaries at $\pm\theta(t)$ (**Fig. 5b**). However, to keep the exposition unambiguous, we assume starting point $z(0) = 0$ and prior-shifted decision boundaries.

A consequence of shifting the boundaries is that the decision confidence at both boundaries is the same for some fixed decision time t. This seems at odds with the intuition that a biased prior ought to bias the confidence in favor of the more likely option. However, this mechanism does end up assigning higher average confidence to the more likely option because of reactions times. Indeed, shifting the boundaries causes the boundary associated with the a-priori less likely correct choice to be further away from the accumulation starting point. Thus, it will on average take longer to reach this boundary, which lowers the decision confidence since confidence decreases with elapsed time. Therefore, even though the decision confidence at both boundaries is the same for the same decision time, it will on average take longer to reach the non-preferred decision boundary. As a result, the average decision confidence at the a-priori preferred boundary will be higher than at the non-preferred boundary. We will show this for a simplified model further below.

Armed with these insights, moving to multi-dimensional momentary evidence becomes straightforward. All we need to do is to replace $z(t)$ with $\boldsymbol{w}^T\boldsymbol{x}(t)$, resulting in the posterior belief

$$p(\mu \geq 0 | \boldsymbol{w}, m, \boldsymbol{x}(t), t) = \Phi\left(\frac{\boldsymbol{w}^T\boldsymbol{x}(t) + \sigma_\mu^{-2}m}{\sqrt{t + \sigma_\mu^{-2}}}\right). \tag{20}$$

If we combine combination weights $\boldsymbol{w}$ and prior bias m into a new weight vector, and augment the accumulated evidence $\boldsymbol{x}(t)$ by the constant σ_μ^{-2}, this posterior takes the same form $\Phi(\boldsymbol{w}^T\tilde{\boldsymbol{x}}(t))$ as before. Then, the likelihood of combination weights and prior bias associated with the choice feedback, $p(y^* | \boldsymbol{w}, m, \boldsymbol{x}(t), t)$ can be written as $\Phi(y^*\boldsymbol{w}^T\tilde{\boldsymbol{x}}(t))$, as in Eq. (5). This reveals that learning the prior bias at the same time as the combination weights from choice feedback is achieved by combining them into a single weight vector that is updated with the same learning rules as derived further above.

Difference to Hanks et al. (2011)

In our formulation, the addition of a prior choice bias is handled by a shift in the boundaries, or, equivalently, and change in the accumulation starting point. In Hanks et al. (2011), they instead argue that such a bias needs to be handled by a time-dependent signal that biases the evidence accumulation. How can this difference be explained? It turns out to arise from a different formulation of the prior. In particular, Hanks et al. (2011) do not assume the prior bias to shift the whole distribution over μ. Instead, they assume

this bias to boost the density for $\mu \geq 0$ and suppress it for $\mu < 0$ while leaving the prior shape otherwise unchanged. Formally, we achieve such a boost by assuming the prior over μ to take the form

$$p(\mu) = 2\mathrm{N}\left(\mu \middle| 0, \sigma_\mu^2\right) \begin{cases} P^+, & \mu \geq 0, \\ 1 - P^+, & \mu < 0, \end{cases} \tag{21}$$

where $P^+ \in [0,1]$ parameterizes the bias towards $\mu \geq 0$ by $p(\mu \geq 0) = P^+$ (**Fig. 5a**). With this prior, the posterior over μ becomes

$$p(\mu | z(t), t) \propto \mathrm{N}\left(\mu \middle| \frac{z(t)}{t + \sigma_\mu^{-2}}, \frac{1}{t + \sigma_\mu^{-2}}\right) \begin{cases} P^+, & \mu \geq 0, \\ 1 - P^+, & \mu < 0, \end{cases} \tag{22}$$

adequately normalized. Furthermore, the posterior belief that $\mu \geq 0$ is given by

$$p(\mu \geq 0 | z(t), t) = \frac{P^+ \Phi\left(\frac{z(t)}{\sqrt{t + \sigma_\mu^{-2}}}\right)}{P^+ \Phi\left(\frac{z(t)}{\sqrt{t + \sigma_\mu^{-2}}}\right) + (1 - P^+) \Phi\left(\frac{-z(t)}{\sqrt{t + \sigma_\mu^{-2}}}\right)}. \tag{23}$$

As before, it is optimal to choose $y = 1$ if $p(\mu \geq 0 | z(t), t) \geq 1/2$, and $y = -1$ otherwise. Thus, the decision maker would choose $y = 1$ as long as

$$\log \frac{\Phi\left(\frac{z(t)}{\sqrt{t + \sigma_\mu^{-2}}}\right)}{1 - \Phi\left(\frac{z(t)}{\sqrt{t + \sigma_\mu^{-2}}}\right)} \geq \log \frac{1 - P^+}{P^+}, \tag{24}$$

that is, if the log-odds provided by the evidence exceed that of the prior mass on $\mu < 0$. For the same accumulator value $z(t)$, the evidence log-odds drops over time towards zero. For symmetric bounds $\pm\theta(t)$ on $z(t)$ this causes the same problem as before. Namely, it might be that the optimal choice at one of the boundaries flips at some point in time, at which the prior log-odds exceed that of the evidence.

Hanks et al. (2011) resolve the ambiguity between boundaries and choices by not putting these boundaries on the accumulated evidence $z(t)$ but instead on another decision variable $V(t)$ that includes both prior and evidence. To map between $V(t)$ and probabilities, they assume the same mapping as for $z(t)$ in the absence of any prior biases, Eq. (2). That is, they assume this mapping to follow

$$p(\mu \geq 0 | V(t), t) = \Phi\left(\frac{V(t)}{\sqrt{t + \sigma_\mu^{-2}}}\right). \tag{25}$$

Bounding $V(t)$ instead of $z(t)$ resolves the ambiguity between boundaries and choices, as all positive $V(t)$ correspond to $p(\mu \geq 0 | V(t), t) > 1/2$.

By equating the two expressions for $p(\mu \geq 0 | \ldots, t)$ and expressing $V(t) = z(t) + \lambda(t)$ as a combination of the influence of accumulated evidence $z(t)$ and prior $\lambda(t)$, the prior influence can be expressed as

$$\lambda(t) = \sqrt{t + \sigma_\mu^{-2}} \Phi^{-1} \left(\frac{P^+ \Phi\left(\frac{z(t)}{\sqrt{t + \sigma_\mu^{-2}}} \right)}{P^+ \Phi\left(\frac{z(t)}{\sqrt{t + \sigma_\mu^{-2}}} \right) + (1 - P^+) \Phi\left(\frac{-z(t)}{\sqrt{t + \sigma_\mu^{-2}}} \right)} \right) - z(t), \tag{26}$$

which does not allow for a simple solution. While Hanks et al. (2011) assume a linear log-odds approximation, we will instead use the sigmoidal approximation $\Phi(z) \approx 1/(1 + \exp(-C_\sigma z))$, where $C_\sigma = \pi^2/6$. This yields

$$\lambda(t) \approx \sqrt{t + \sigma_\mu^{-2}} \frac{1}{C_\sigma} \log \frac{P^+}{1 - P^+}, \tag{27}$$

which illustrates that the prior influence to $V(t)$ increases over time. As in Hanks et al. (2011), this influence causes both an initial bias in the starting point, $V(0) = \lambda(0)$, and a time-dependent bias that is added to the accumulated evidence $z(t)$ (**Fig. 5c**).

In summary, the main difference between our use of the prior and that of Hanks et al. (2011) is that we assume the prior bias to shift the whole distribution, whereas Hanks et al. (2011) assume it to boost one half of it, and suppress the other half. For us, this can be implemented by a shift in the accumulation starting point, whereas Hanks et al. (2011) require both such a shift and an additional time-dependent bias signal. In both approaches, however, the ambiguity between decision boundaries and choices is resolved by bounding a combination of prior bias and accumulated evidence rather than accumulated evidence alone. Which of the two approaches is more adequate depends on how well it matches the task. Our approach has the advantage of yielding a simple extension to the previously derived learning rule. How learning prior biases in the framework of Hanks et al. (2011) could be achieved remains unclear.

Confidence calibration with a latent state bias

In both setups, the confidence at the boundaries is the same for some fixed decision time, irrespective of the choice. Here we will revisit how this can be consistent with a choice probability that differs at the different boundaries. To do so, we will consider a simplified setup in which this can be shown analytically. Namely, rather than assuming a prior distribution across multiple decision difficulties $|\mu|$, we assume a single difficulty μ_0 with a prior $p(\mu) = P^+ \delta(\mu - \mu_0)/2 + (1 - P^+)\delta(\mu + \mu_0)/2$, where δ is the Dirac delta function. That is, $\mu = \mu_0$ with probability P^+, and $\mu = -\mu_0$ with probability $1 - P^+$. Assuming again the noisy observations $\delta z_i | \mu \sim \mathrm{N}(\mu \delta t, \delta t)$, it is easy to show that the posterior belief is given by the logistic sigmoid,

$$p(\mu = \mu_0 | z) = \frac{1}{1 + e^{-2\mu_0 \left(z - \frac{1}{2\mu_0} \log \frac{P^+}{1 - P^+} \right)}}. \tag{28}$$

Following the same principle as before, we will fix the decision boundaries at $\theta^+ = \theta - \frac{1}{2\mu_0}\log\frac{P^+}{1-P^+}$ and $\theta^- = -\theta - \frac{1}{2\mu_0}\log\frac{P^+}{1-P^+}$. This results in a constant decision confidence $(1+\exp(-2\mu_0\theta))^{-1}$ at both boundaries.

To show that this decision confidence equals the probability of making the correct choice on average, let us find this probability for each possible latent state value. For either way, we will use the expression for boundary hitting probability for diffusion models with asymmetric boundaries, as given in (Cox and Miller 1977; Palmer et al. 2005). For $\mu = \mu_0$, the upper boundary θ^+ leads to the correct choice. This boundary is reached with probability

$$p(z = \theta^+ | z \in \{\theta^+, \theta^-\}, \mu = \mu_0) = \frac{e^{2\mu_0\theta} - \frac{1-P^+}{P^+}}{e^{2\mu_0\theta} - e^{-2\mu_0\theta}}, \tag{29}$$

yielding the expression for the probability of making correct choices if $\mu = \mu_0$. Note that, unlike the confidence, this probability is modulated by P^+. In particular, it grows with an increase in P^+. In other words, the larger the a-priori probability that the upper boundary leads to the correct choice, the larger the probability that the decision maker chooses correctly in cases in which the upper boundary is indeed the correct choice. For $\mu = -\mu_0$, the lower boundary θ^- leads to correct choices, which happens with probability

$$p(z = \theta^- | z \in \{\theta^+, \theta^-\}, \mu = -\mu_0) = \frac{e^{2\mu_0\theta} - \frac{P^+}{1-P^+}}{e^{2\mu_0\theta} - e^{-2\mu_0\theta}}, \tag{30}$$

where the only difference to the expression for the upper boundary is the impact of the prior. Specifically, this probability shrinks with an increasing P^+. The average probability of choosing correctly is a combination of both bound-hitting probabilities, weighted by the latent state probabilities,

$$\begin{aligned} p(\text{correct}) = {} & p(z = \theta^+ | \in \{\theta^+, \theta^-\}, \mu = \mu_0) p(\mu = \mu_0) \\ & + \; p(z = \theta^- | z \in \{\theta^+, \theta^-\}, \mu = -\mu_0) p(\mu = -\mu_0) = \frac{1}{1 + e^{-2\mu_0\theta}}, \end{aligned} \tag{31}$$

where we have used $p(\mu = \mu_0) = P^+$ and $p(\mu = -\mu_0) = 1 - P^+$, and the above expressions of the bound-hitting probabilities, to find the final result after some algebra. This demonstrates that, even though the decision confidence differs from the probability of making the correct choices for individual choices, it equals the average probability of making correct choices. This unintuitive result follows from conditioning the choice probabilities on the latent state, which is unknown to the decision maker, and thus cannot be reflected in her decision confidence. Once this latent state is marginalized out (by averaging over it in Eq. (31)), consistency with the decision confidence is restored (Drugowitsch, Moreno-Bote, and Pouget 2014). The same principle applies to the more complex priors used further above, but for those, it becomes hard to establish the equivalence between choice probability and decision confidence analytically.

Sequential choice dependencies due to continuous weight tracking

In every-day situations, no two decisions are made under the exact same circumstances. Nonetheless, we need to be able to learn from the outcome of past choices to improve future ones. We can do so if we assume that past choices become increasingly less informative about future choices over time. One way to express this formally is to assume that the world changes slowly over time, and that our aim is to track these changes. In our setup, this corresponds to the assumption of a slow change of the true weight vector across

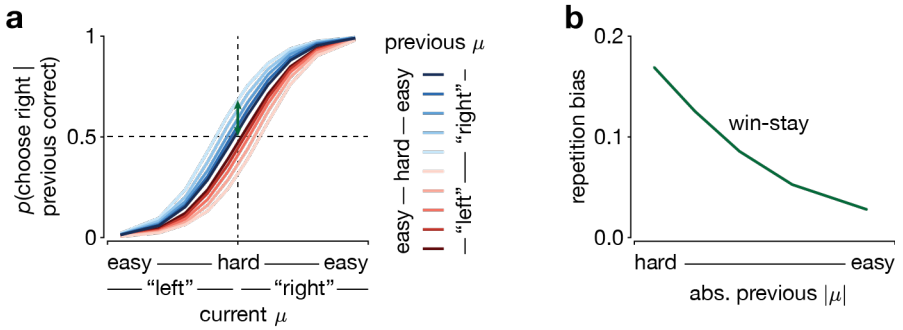

Figure 6. Sequential choice dependencies due to continuous learning. We simulated learning in a stationary environment with an ADF rule that assumed that the both input weights and latent state bias slowly changed across consecutive decisions (here "left" and "right"). (**a**) Conditional on having made correct choices in the previous trial, the psychometric curve was shifted towards repeating the same choice. Specifically, if the previous choice was "left" (or "right"), then the current choice was towards "left" (or "right") as shown by the red (or blue) psychometric curves. The shift was modulated by the choice difficulty of the previous trials, with stronger shifts for harder (bright curves) than easier (dark curves) previous choices. (**b**) If we measure the repetition bias as the probability of repeating the same choice at $\mu = 0$ (green arrow in (**a**) for easy "right" previous choices), we find that this repetition bias is most pronounced if the previous decision was hard, and reduces for easier decisions.

consecutive choices. Here we consider the effect such an assumption has on the dependency between consecutive choices.

Under the assumption of a volatile world, weight learning never ends. Rather, the input combination weights are continuously adjusted to make correct choices more likely in the close future. After correct choices, this means that weights will be adjusted to repeat the same choice upon observing a similar input in the future. After incorrect choices, the aim is to adjust the weights to perform the opposite choice, instead. Furthermore, our model predicts that, after an easy correct choice, in which confidence can be expected to be high, the weight adjustments are lower than after hard correct choices (see **Fig. 1a**, green line). As a consequence, we would expect the model to be more likely to repeat the same choices after correct and hard, than after correct and easy trials.

To test this prediction, we simulated weight and latent state bias learning using the ADF rule under the assumption of both the true weight and a latent state bias that slowly changes across trials. We then measured how likely the model was to repeat the same choice after correct decisions. **Figure 6a** illustrates that this repetition bias manifests itself in a shift of the psychometric curve that makes it more likely to repeat the previous choice. Furthermore, and as predicted, this shift is modulated by the difficulty of the previous choice, and is stronger if the previous choice was easy (i.e., associated with a large $|\mu|$). This can be seen more explicitly by plotting the repetition bias directly against $|\mu|$, as done in **Fig. 6b**. Therefore, if the decision maker expects to operate in a volatile, slowly changing world, our model predicts a repetition bias to repeat the same choices after correct decisions, and that this bias is stronger if the previous choice was easy.

Discussion

Previous work on the normative foundations of diffusion models has mostly focused on evidence accumulation and the optimal speed-accuracy trade-off, but has – to our knowledge – always assumed knowledge of how to interpret the – usually low-dimensional – input in the light of the upcoming decision. This is a strong assumption if this evidence is encoded in larger neural populations or is formed by multiple inputs. In this case, it might be unclear how exactly the multiple inputs encode decision-related information. Here, we asked how such encoding could be learned from feedback about the correctness of decisions alone, if these decisions are generated by diffusion models. This is a fairly natural setup as long as the decision requires some accumulation of evidence over time, which is almost always the case in natural decisions.

Addressed from the normative perspective, the question of how to best linearly combine the information from multiple inputs turns into how to find the posterior over the combination weights given all observed choices and feedback so far. Framing the problem this way, we found the feedback likelihood, which is required to find the posterior, to be closely related to the decision confidence of individual choices. Unfortunately, such a likelihood prevented us from finding an analytical expression for this posterior. However, we could apply an approximation called *assumed density filtering* that, after each choice and consecutive feedback finds the closest approximate, tractable, posterior to its precise, but intractable, counterpart. This yielded analytical and interpretable learning rules for the posterior moments of the combination weights, that – as we showed in simulations – performed close-to-optimal learning.

Interestingly, the resulting learning rule modulated the rate of learning by the decision confidence of the preceding decision. It suppressed learning after high-confidence correct decisions, supported learning for uncertain decisions irrespective of their correctness, and promoted strong change of the combination weights after wrong decisions that were made with high confidence (**Fig. 2**). It should be noted, however, that decision confidence was not the only factor influencing the learning rate. Revisiting the update rules for weight estimate mean and covariance reveals that the strength of learning is additionally influenced by decision time and the magnitude of the accumulated evidence, in ways that potentially counter-act the influence of the decision confidence. Nonetheless, as decision confidence appears as a multiplicative term in these learning rules, it modulates the impact of all additional term, and thus will have a significant impact on over-all learning.

In hindsight, it shouldn't come as a surprise that parameter learning from feedback is modulated by decision confidence. If we more abstractly assume that, in each trial the decision maker observes some stimulus x^n (that might, as in our case, unfold over time), and after a choice receives feedback y^{n*}, then it should update its decision-making parameters θ after each such feedback according to

$$p(\theta|x^{1:n}, y^{1:n*}) \propto p(y^{n*}|x^n, \theta)p(\theta|x^{1:n-1}, y^{1:n-1*}). \tag{32}$$

If we furthermore assume a 2-AFC setup in which the decision maker can for each decision choose amongst two options only, then both choice and feedback are binary variables. This implies that $p(y^{n*}|x^n, \theta)$ is the decision confidence (i.e., the belief that the choice was correct) after correct choices, and one minus the decision confidence for incorrect choices. Thus, in both cases, decision confidence modulates the update of the decision-making parameters (Pouget et al. 2016). The same principle applies to N-AFC setups for N>2 if the feedback is only whether the choice was correct or incorrect, as is commonly the case. We only find deviations from this principle once the decision maker is told after incorrect choices what the correct choice would have been. Therefore, decision confidence is expected to modulate learning of the parameters of the decision-making strategy under a wide range of natural scenarios. The only caveat with this argument is that $p(y^{n*}|x^n, \theta)$ in the above equation is technically a function of θ that returns how confident one would have been in the feedback-promoted choice under different settings of θ. Therefore, it deviates from a notion of "full" decision confidence $p(y^{n*}|x^n)$ that takes parameter uncertainty into account by marginalizing them out.

Moving to higher-dimensional inputs changes the structure of evidence accumulation in diffusion models. Traditionally, stochastic momentary evidence is added to the particle's current location, causing this one-dimensional particle to accumulate all evidence required to make a choice. Once we move to multiple inputs, the accumulation needs to be performed for each input dimension separately (**Fig. 1b**; $\boldsymbol{x}(t)$). The bounded particle location that triggers choices is then a linear projection $\boldsymbol{w}^T\boldsymbol{x}(t)$ of this multi-dimensional accumulator. In fact, we don't even need to keep track of such a one-dimensional particle, but can instead directly bound a linear combination of $\boldsymbol{x}(t)$. The reason why we cannot perform this projection directly on the momentary evidence $\delta\boldsymbol{x}(t)$ to subsequently accumulate evidence in a one-dimensional particle is that of credit assignment. If we don't keep track of how each input dimension contributed to the particle crossing the decision boundary, we are unable to adjust how individual input dimensions are best combined to trigger choices. Mathematically, this is illustrated by $\boldsymbol{x}(t)$ appearing in the update Equations

(12)-(15) for both mean and covariance of the input combination weight estimate. This only occurs because we have uncertainty about the combination weights. Once we know these weights, we can return to performing evidence accumulation in the original one-dimensional diffusion model space. Even if we know these weights, accumulating evidence in all input dimensions separately is natural if we assume the inputs to follow the properties of linear probabilistic population codes (Ma et al. 2006) that encode full distributions across the task-relevant variables. In this case, decisions are again made as in our framework, by bounding the pooled activity of all units that represent the accumulated evidence (Beck et al. 2008).

Throughout the work we have focused on learning the correct weights to combine the accumulated evidence, and have measured performance as deviations from the true underlying weights (through their angular distance). However, is this really the quantity that the decision maker ought to minimize? Wouldn't a more adequate performance measure be the decision maker's choice performance, which we could measure through the amount of accumulated reward? It turns out that, if one were to formulate a learning problem in which the aim is to maximize reward, and in which the input combination weights are unknown, learning these decision weights becomes a sub-problem to maximizing this reward. An additional component of the larger problem of reward maximization involves how to tune the decision boundaries, which we have addressed separately in previous work (Drugowitsch et al. 2012, 2015). However, combining both into one big model might become overly confusing and blur the exposition. Thus, we have instead focused exclusively on the narrower objective of weight learning.

A prediction resulting from this model is that weight learning should introduce sequential choice dependencies. If previous choices were correct, the decision maker should be more likely to repeat these choices. Additionally, this choice bias should be boosted by the difficulty of the previous trial: for easier decision, it should be more pronounced than for hard choices (**Fig. 6**). This prediction arises from the model's property that the learning rate for easier choices should be higher than that for hard choices. Sequential choice dependencies are frequently observed in both humans and animals (e.g., Busse et al. 2011; Yu and Cohen 2009), even if the task promotes independent choices in consecutive trials. Our model provides a potential explanation for when such choice dependencies are, in fact, rational: they could arise from the decision maker's belief that the environment is volatile, promoting a continual adjustment of the decision-making strategy. What remains to be shown is that this adjustment is modulate by choice difficulty, as predicted by our model.

We have furthermore clarified how the a-prior belief that one option is more likely correct than the other ought to impact Bayes-optimal decision-making in diffusion models. Previously, Hanks et al. (2011) have suggested that such a biased prior should impact evidence accumulation through both an initial offset of the accumulator and an additive signal that grows over time. Their argument was based on intuitions about how diffusion models encode posterior beliefs. Here, we confirm their analysis more formally, and reveal that it is specific to the way they introduce biases in the prior distribution. Specifically, in the context of the random-dot motion task that they used, they introduced this bias by boosting the occurrence of one motion direct, while leaving the distribution over coherences unchanged. In our model, this corresponds to boosting $p(\mu)$ for either $\mu \geq 0$ or $\mu < 0$ without otherwise changing the shape of this prior (**Fig. 5a**). If we instead introduce such a bias by shifting the mean of $p(\mu)$ away from zero (**Fig. 5a**), then Bayes-optimal decisions require a simple shift in the diffusion starting point without the additional additive signal. A consequence of this simplification is that learning such a prior bias can be achieved in the same way as learning the input weights. How the prior used in Hanks et al. (2011) could be efficiently learned remains unclear. Overall, our model predicts that a change in how bias is introduced in $p(\mu)$ changes the optimal decision-making strategy. It would be interesting to see if such a change would also be reflected in LIP activity of monkeys performing the random-dot motion task, as recorded in Hanks et al. (2011).

In summary, we have for the first time derived a close-to-optimal learning rule for how to combine multi-dimensional input to decisions that require temporal evidence accumulation before committing to a choice. This learning rule reveals that the rate of learning ought to be modulated by decision confidence, thus highlighting the computational role of such decision confidence in every-day decision-making. Our rule predicts a particular form of dependencies in sequential choices, and impact on prior biases, pending

experimental validation. It would be interesting to see how our theory, which currently does not make any claims of biological realism, could be extended to increase such biological realism. Previous work (Beck et al. 2008) has already suggested how diffusion model-like decision-making mechanisms could be implemented in biological networks, and could act as guidance to our brain could implement mechanisms that learn how to interpret the input to such decision-making mechanisms.

Methods

Bayesian decision-making with diffusion models

When interpreting diffusion models as a mechanistic implementation of Bayesian decision-making, we have assumed the latent state to be drawn from $\mu \sim \mathrm{N}\left(0, \sigma_\mu^2\right)$, and the momentary evidence in each time step δt to provide information about this latent state by $\delta z_i | \mu \sim \mathrm{N}(\mu \delta t, \delta t)$. After having observed this evidence for some time $t \equiv n\delta t$, the posterior μ given all observed evidence $\delta z_{1:n}$ is by Bayes' rule given by

$$\begin{aligned} p(\mu|\delta z_{1:n}) &\propto \mathrm{N}\left(\mu|0, \sigma_\mu^2\right) \prod_{i=1}^{n} \mathrm{N}(\delta z_i|\mu\delta t, \delta t) \\ &\propto \mathrm{e}^{-\left(\frac{1}{\sigma_\mu^2}+t\right)\frac{\mu^2}{2}+z(t)\mu} \\ &\propto \mathrm{N}\left(\mu|\frac{z(t)}{t+\sigma_\mu^{-2}}, \frac{1}{t+\sigma_\mu^{-2}}\right). \end{aligned} \tag{33}$$

In the above, all proportionalities are with respect to μ, and we have defined $z(t) = \sum_{i=1}^{n} \delta z_i$ and have used $t = \sum_{i=1}^{n} \delta t$. How to find the posterior belief about μ's sign is described in Eq. (2) in the main text.

Decision confidence with uncertain combination weights

We find the decision confidence with uncertain combination weights by using the expression of this confidence for known weights, and marginalizing over the weights, $\boldsymbol{w} \sim \mathrm{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$. The confidence itself is involves the standard Gaussian cumulative, and thus can be expressed by the integral

$$p(y|\boldsymbol{w}, \boldsymbol{x}, t) = \Phi(y\boldsymbol{w}^T\widetilde{\boldsymbol{x}}) = \int_{-\infty}^{0} \mathrm{N}(y| - y\boldsymbol{w}^T\widetilde{\boldsymbol{x}}, 1)\mathrm{d}y. \tag{34}$$

We evaluate the marginalization by reversing the order of integration to find the decision confidence with uncertain combination weights, resulting in

$$\begin{aligned} \iint_{-\infty,-\infty}^{0,\infty} \mathrm{N}(y|-y\boldsymbol{w}^T\widetilde{\boldsymbol{x}}, 1)\mathrm{N}(\boldsymbol{w}|\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)\mathrm{d}\boldsymbol{w}\mathrm{d}y &= \int_{-\infty}^{0} \mathrm{N}(y| - y\boldsymbol{\mu}_w^T\widetilde{\boldsymbol{x}}(t), 1 + \widetilde{\boldsymbol{x}}^T\boldsymbol{\Sigma}_w\widetilde{\boldsymbol{x}})dy \\ &= \Phi\left(\frac{y\boldsymbol{\mu}_w^T\widetilde{\boldsymbol{x}}}{\sqrt{1+\widetilde{\boldsymbol{x}}^T\boldsymbol{\Sigma}_w\widetilde{\boldsymbol{x}}}}\right). \end{aligned} \tag{35}$$

Assumed Density Filtering moment updates

As discussed in the main text around Eq. (11), a single step of the Assumed Density Filter (ADF) corresponds to have the posterior approximation $\boldsymbol{w} \sim N(\boldsymbol{\mu}_w^*, \boldsymbol{\Sigma}_w^*)$ match the moments of the true posterior $p(\boldsymbol{w}|y^*, \boldsymbol{x}, t)$ (Bishop 2006; Murphy 2012), given by

$$p(\boldsymbol{w}|y^*, \boldsymbol{x}, t) = \frac{1}{Z}\, \Phi(y^*\boldsymbol{w}^T\widetilde{\boldsymbol{x}}(t))\mathrm{N}(\boldsymbol{w}|\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w), \tag{36}$$

where Z is the normalization constant. This normalization turns out to be given by the decision confidence with uncertain combination weights, as given by Eq. (25). It allows us to express the posterior mean and covariance by

$$\langle \boldsymbol{w} | y^*, \boldsymbol{x}, t \rangle = \boldsymbol{\mu}_w + \boldsymbol{\Sigma}_w \nabla_{\mu_w} \log Z, \tag{37}$$

$$\mathrm{cov}(\boldsymbol{w} | y^*, \boldsymbol{x}, t) = \boldsymbol{\Sigma}_w - \boldsymbol{\Sigma}_w \left(\nabla_{\mu_w} \log Z \left(\nabla_{\mu_w} \log Z \right)^T - 2 \nabla_{\Sigma_w} \log Z \right) \boldsymbol{\Sigma}_w. \tag{38}$$

Using $\nabla_u \log \Phi(f(\boldsymbol{u})) = \Phi\big(f(\boldsymbol{u})\big)^{-1} \mathrm{N}(f(\boldsymbol{u}) | 0,1) \nabla_u f(\boldsymbol{u})$, the required derivatives of the log-normalizer are given by

$$\nabla_{\mu_w} \log Z = C_w \frac{y^* \widetilde{\boldsymbol{x}}}{\sqrt{1 + \widetilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \widetilde{\boldsymbol{x}}}}, \tag{39}$$

$$\nabla_{\Sigma_w} \log Z = -C_w \frac{y^* \boldsymbol{\mu}_w^T \widetilde{\boldsymbol{x}}}{2\sqrt{1 + \widetilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \widetilde{\boldsymbol{x}}}} \widetilde{\boldsymbol{x}} \widetilde{\boldsymbol{x}}^T, \tag{40}$$

with pre-factor C_w defined by Eq. (14) in the main text. Substituting these gradients into the above expressions for the mean results in Eq. (12) in the main text. The covariance results in

$$\mathrm{cov}(\boldsymbol{w} | y^*, \boldsymbol{x}, t) = \boldsymbol{\Sigma}_w - \left(C_w^2 + C_w \frac{y^* \boldsymbol{\mu}_w^T \widetilde{\boldsymbol{x}}}{\sqrt{1 + \widetilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \tilde{x}}} \right) \frac{\boldsymbol{\Sigma}_w \widetilde{\boldsymbol{x}} \widetilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w}{1 + \widetilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \widetilde{\boldsymbol{x}}}. \tag{41}$$

Applying the matrix inversion lemma to $(\boldsymbol{\Sigma}_w^{-1} + \widetilde{\boldsymbol{x}} \widetilde{\boldsymbol{x}}^T)^{-1}$ allows us to re-express the last fraction in terms of this inverse. As a last step, we define C_{cov} to be given by the above term in brackets (that is, Eq. (15)), which then results in the covariance Eq. (13) provided in the main text.

Sampling from the weight posterior by Gibbs sampling

After n decisions, we have observed n times some accumulated evidence $\boldsymbol{x}^1(t^1), \dots, \boldsymbol{x}^n(t^n)$ and receive n feedbacks, $y^{1*}, \dots, y^{n*}$. At this point, the posterior weight is given by

$$p(\boldsymbol{w} | \boldsymbol{x}^{1:n}, y^{1:n*}, t^{1:n}) \propto \mathrm{N}(\boldsymbol{w} | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \prod_{i=1}^{n} \Phi\big(y^{n*} \boldsymbol{w}^T \widetilde{\boldsymbol{x}}^n(t^n)\big), \tag{42}$$

where we have assumed a Gaussian prior with mean $\boldsymbol{\mu}_0$ and covariance $\boldsymbol{\Sigma}_0$. We computed the sampling covariance $\boldsymbol{\Sigma}_w$ by

$$\boldsymbol{\Sigma}_w = \left(\boldsymbol{\Sigma}_0^{-1} + \sum_{i=1}^{n} \widetilde{\boldsymbol{x}}^n \widetilde{\boldsymbol{x}}^{nT} \right)^{-1}, \tag{43}$$

which can be incrementally updated by the Sherman-Morrison formula. Then, we alternated between sampling the auxiliary variables $a^1, \dots, a^n$ from a truncated Gaussian

$$a_n \sim \mathrm{N}(y^{n*} \boldsymbol{w}^T \widetilde{\boldsymbol{x}}^n) I(a_n \geq 0), \tag{44}$$

where $I(\cdot)$ is the indicator function, and the posterior weight vector $\boldsymbol{w}$,

$$\boldsymbol{w} \sim N\left(\boldsymbol{\Sigma}_w \left(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \sum_{i=1}^{n} y^{n*} \widetilde{\boldsymbol{x}}^n a^n \right), \boldsymbol{\Sigma}_w \right). \tag{45}$$

For all simulations, we used 10 burn-in samples that we discarded, and an additional 50 samples to estimate the posterior mean.

Learning rule performance comparison

In order to compare the weight learning performance of ADF to the Bayes-optimal weight update and to that of alternative heuristics, we performed 10,000 learning simulations of 500 trials each, for different dimensionalities of $\boldsymbol{w}$. In each of these simulations we initially drew the true weights from $\boldsymbol{w}^* \sim \mathrm{N}(2\boldsymbol{e}, \boldsymbol{I})$, where $\boldsymbol{e}$ is a vector of ones. These true weights remained constant across all trials. Within each trial, we drew the drift rate according to $\mu^n \sim \mathrm{N}(0,4)$, which determines the correct choice by $y^{n*} = 2I(\mu^n \geq 0) - 1$. Here, $I(a)$ is the identifier function that returns $I(a) = 0$ if a is true, and $I(a) = 0$ otherwise. Based on this drift rate μ^n, we generated the momentary evidence by

$$\delta \boldsymbol{x}_i^n | \mu^n, \boldsymbol{w}^* \sim \mathrm{N}\left(\frac{\boldsymbol{w}^*}{\boldsymbol{w}^{*T}\boldsymbol{w}^*}\mu^n \delta t, \frac{\boldsymbol{I}}{\boldsymbol{w}^{*T}\boldsymbol{w}^*}\delta t\right), \tag{46}$$

which was accumulated over time to yield $\boldsymbol{x}^n(t)$. A decision was triggered as soon as $|\boldsymbol{w}^T\boldsymbol{x}^n(t)| \geq \theta$, where $\boldsymbol{w}$ is the learning rule's current weight estimate (or the posterior weight mean for ADF / Gibbs sampling / the particle filter), and where we chose $\theta = 0.5$. The decision maker's simulated choice was determined by $y^n = 2I(\boldsymbol{w}^T\boldsymbol{x}^n(t^n) \geq 0) - 1$. After each update we computed the angular error between the weight estimate and its true value by $\cos^{-1}\left(\frac{\boldsymbol{w}^{*T}\boldsymbol{w}}{\|\boldsymbol{w}^*\|\|\boldsymbol{w}\|}\right)$, as reported in **Fig. 3**. For this we again used the posterior mean if both mean and covariance were available.

We sped up the simulations by simulating the diffusion directly in the one-dimensional $\boldsymbol{w}^T\boldsymbol{x}^n(t)$ space. This resulted in a one-dimensional diffusion model whose first-passage time distribution is known and can be efficiently drawn from (Drugowitsch 2016). The final $\boldsymbol{x}^n(t^n)$ was recovered by drawing it from $\boldsymbol{x}^n(t^n) \sim \mathrm{N}\left(\frac{\mu^n t^n}{\|\boldsymbol{w}^*\|^2}\boldsymbol{w}^*, \frac{t^n}{\|\boldsymbol{w}^*\|^2}\boldsymbol{I}\right)$, subject to the constraint $\boldsymbol{w}^T\boldsymbol{x}^n(t^n) = y^n\theta$.

Performance comparison when tracking non-stationary weights

To evaluate the performance of tracking non-stationary weights, we assumed the true weights to evolve across consecutive decisions according to

$$\boldsymbol{w}^{n+1*} | \boldsymbol{w}^{n*} \sim \mathrm{N}\left(\boldsymbol{w}^{n*} + \psi_{dw}(\boldsymbol{\mu}_w - \boldsymbol{w}^{n*}), \sigma_{dw}^2\boldsymbol{I}\right), \tag{47}$$

which is a first-order autoregressive process with leak coefficient ψ_{dw}, diffusion variance σ_{dw}^2, and mean $\boldsymbol{\mu}_w$. For all simulation, these were set to $\psi_{dw} = 0.01$, $\sigma_{dw}^2 = 1 - (1 - \psi_{dw})^2$, and $\boldsymbol{\mu}_w = 2\boldsymbol{e}$. The diffusion variance was chosen to achieve an identity steady-state covariance of the process.

The Bayes-optimal model was augmented to take into account the dynamics of the weight vector. This resulted in a state space model with probit likelihoods. We used a particle filter with 1,000 particles to approximate Bayes-optimal inference in this model. The ADF rule was similarly modified to include the latent weight vector dynamics. Specifically, after each update of the estimate mean and covariance, we added $\sigma_{dw}^2\boldsymbol{I}$ to the covariance matrix, and performed a leak of the estimate mean according to the above dynamics. All other components of the simulations were performed as in the previous section.

Sequential dependencies in non-stationary environments

Before each trial sequence of 1,000 decisions each, we drew the true weight vector from $\boldsymbol{w}^{1*} \sim \mathrm{N}(2\boldsymbol{e}, \boldsymbol{I})$ and the latent state bias from $m^{1*} \sim \mathrm{N}(0,1)$. Across consecutive decisions, the weight vector evolved according to the same dynamics as described in the previous section. The latent state bias followed the first-order autoregressive process

$$m^{n+1*}|m^{n*} \sim N\left(m^{n*} + \psi_{dm}(\mu_m - m^{n*}), \sigma_{dm}^2\right), \tag{48}$$

with leak coefficient $\psi_{dm} = 0.01$, diffusion variance $\sigma_{dm}^2 = 1 - (1 - \psi_{dm})^2$, and mean $\mu_m = 0$. To track both the weight vector and the latent state bias, we augmented the weight vector estimate by an estimate of the latent state bias, as described in Section "Learning both weights and a latent state prior bias" in the main text. We then used the ADF learning rule with this augmented weight vector to simulate learning and decision-making. This procedure was repeated 10,000, and the results reported in **Fig. 6** were computed on all but the first 200 trials for each trial sequence, to avoid initial transients to impact the analysis.

References

Beck, Jeffrey M. et al. 2008. "Probabilistic Population Codes for Bayesian Decision Making." *Neuron* 60(6):1142–52. Retrieved (http://dx.doi.org/10.1016/j.neuron.2008.09.021).

Bishop, Christopher. 2006. *Pattern Recognition and Machine Learning*. Springer.

Bogacz, Rafal, Eric Brown, Jeff Moehlis, Philip Holmes, and Jonathan D. Cohen. 2006. "The Physics of Optimal Decision Making: A Formal Analysis of Models of Performance in Two-Alternative Forced-Choice Tasks." *Psychological Review* 113(4):700–765. Retrieved (http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.113.4.700).

Busse, L. et al. 2011. "The Detection of Visual Contrast in the Behaving Mouse." *Journal of Neuroscience* 31(31):11351–61. Retrieved (http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.6689-10.2011).

Courville, Aaron C., Nathaniel D. Daw, and David S. Touretzky. 2006. "Bayesian Theories of Conditioning in a Changing World." *Trends in Cognitive Sciences* 10(7):294–300.

Cox, David Roxbee and Hilton David Miller. 1977. *The Theory of Stochastic Processes*. CRC Press.

Dayan, P., S. Kakade, and P. R. Montague. 2000. "Learning and Selective Attention." *Nature Neuroscience* 3 Suppl(november):1218–23.

Dayan, Peter and A. J. Yu. 2003. "Uncertainty and Learning." *IETE Journal of Research* 49(2–3):171–81. Retrieved (http://discovery.ucl.ac.uk/185387/).

Doya, Kenji, Shin Ishii, Alexandre Pouget, and Rajesh P. N. Rao. 2006. *Bayesian Brain: Probabilistic Approaches to Neural Coding*. MIT Press.

Drugowitsch, J., R. Moreno-Bote, A. K. Churchland, M. N. Shadlen, and A. Pouget. 2012. "The Cost of Accumulating Evidence in Perceptual Decision Making." *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience* 32(11):3612–28. Retrieved (http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&DbFrom=pubmed&Cmd=Link&LinkName=pubmed_pubmed&LinkReadableName=Related Articles&IdsFromResult=22423085&ordinalpos=3&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_RVDocSum).

Drugowitsch, Jan. 2016. "Fast and Accurate Monte Carlo Sampling of First-Passage Times from Wiener Diffusion Models." *Scientific Reports* 6(August 2015):20490. Retrieved (http://www.nature.com/articles/srep20490).

Drugowitsch, Jan, Gregory C. Deangelis, Dora E. Angelaki, and Alexandre Pouget. 2015. "Tuning the Speed-Accuracy Trade-off to Maximize Reward Rate in Multisensory Decision-Making." *eLife* 4(JUNE2015):1–11.

Drugowitsch, Jan, Gregory C. Deangelis, Eliana M. Klier, Dora E. Angelaki, and Alexandre Pouget. 2014. "Optimal Multisensory Decision-Making in a Reaction-Time Task." *eLife* 2014(3):1–19.

Drugowitsch, Jan, Rubén Moreno-Bote, and Alexandre Pouget. 2014. "Relation between Belief and Performance in Perceptual Decision Making." *PLoS ONE* 9(5).

Frazier, Peter I. and Angela J. Yu. 2008. "Sequential Hypothesis Testing under Stochastic Deadlines." *Advances in Neural Information Processing Systems* 1–8. Retrieved (http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2007_706.pdf%5Cnhttps://papers.nips.cc/paper/3355-sequential-hypothesis-testing-under-stochastic-deadlines.pdf%5Cnhttp://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2007_706.pdf).

Gold, Joshua I. and Michael N. Shadlen. 2002. "Banburismus and the Brain: Decoding the Relationship between Sensory Stimuli, Decisions, and Reward." *Neuron* 36(2):299–308.

Hanks, Timothy D., Mark E. Mazurek, Roozbeh Kiani, Elisabeth Hopp, and Michael N. Shadlen. 2011. "Elapsed Decision Time Affects the Weighting of Prior Probability in a Perceptual Decision Task." *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience* 31(17):6339–52.

Körding, Konrad P. and Daniel M. Wolpert. 2004. "Bayesian Integration in Sensorimotor Learning." *Nature* 427(6971):244–47. Retrieved (http://www.ncbi.nlm.nih.gov/pubmed/14724638).

Ma, Wei Ji, Jeffrey M. Beck, Peter E. Latham, and Alexandre Pouget. 2006. "Bayesian Inference with

Probabilistic Population Codes." *Nature Neuroscience* 9(11):1432–38. Retrieved (http://www.ncbi.nlm.nih.gov/pubmed/17057707).

Moreno-Bote, Rubén. 2010. "Decision Confidence and Uncertainty in Diffusion Models with Partially Correlated Neuronal Integrators." *Neural Computation* 22(7):1786–1811. Retrieved (http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=20141474&retmode=ref&cmd=prlinks%5Cnpapers2://publication/doi/10.1162/neco.2010.12-08-930).

Murphy, Kevin Patrick. 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press.

Palmer, John, Alexander C. Huk, and Michael N. Shadlen. 2005. "The Effect of Stimulus Strength on the Speed and Accuracy of a Perceptual Decision." *Journal of Vision* 5(5):376–404. Retrieved (http://www.journalofvision.org/content/5/5/1.long).

Pouget, Alexandre, Jan Drugowitsch, and Adam Kepecs. 2016. "Confidence and Certainty: Distinct Probabilistic Quantities for Different Goals." *Nature Neuroscience* 19(3):366–74. Retrieved (http://www.nature.com/doifinder/10.1038/nn.4240%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26906503).

Ratcliff, Roger. 1978. "A Theory of Memory Retrieval." *Psychological Review* 85(2):59–108.

Ratcliff, Roger and Gail McKoon. 2008. "The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks." *Neural Computation* 20(4):873–922.

Ratcliff, Roger and Philip L. Smith. 2004. "A Comparison of Sequential Sampling Models for Two-Choice Reaction Time." *Psychological Review* 111(2):333–67. Retrieved (http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.111.2.333).

Tajima, Satohiro, Jan Drugowitsch, and Alexandre Pouget. 2016. "Optimal Policy for Value-Based Decision-Making." *Nature Communications* 7:12400. Retrieved (http://www.nature.com/ncomms/2016/160818/ncomms12400/full/ncomms12400.html).

Yu, Angela J. and Jonathan D. Cohen. 2009. "Sequential Effects: Superstition or Rational Behavior?" *Advances in Neural Information Processing Systems* 21:1873–80. Retrieved (http://www.pni.princeton.edu/ncc/publications/2009/YuNIPS09.pdf).