Learning optimal decisions with confidence

Jan Drugowitsch[a,1], André G. Mendonça[b], Zachary F. Mainen[b], and Alexandre Pouget[c]

[a]Department of Neurobiology, Harvard Medical School, Boston, MA 02115; [b]Champalimaud Research, Champalimaud Centre for the Unknown, 1400-038 Lisbon, Portugal; and [c]Department of Basic Neuroscience, University of Geneva, CH-1211 Geneva, Switzerland

Edited by Paul W. Glimcher, New York University, New York, NY, and accepted by Editorial Board Member Thomas D. Albright October 18, 2019 (received for review April 19, 2019)

Diffusion decision models (DDMs) are immensely successful models for decision making under uncertainty and time pressure. In the context of perceptual decision making, these models typically start with two input units, organized in a neuron–antineuron pair. In contrast, in the brain, sensory inputs are encoded through the activity of large neuronal populations. Moreover, while DDMs are wired by hand, the nervous system must learn the weights of the network through trial and error. There is currently no normative theory of learning in DDMs and therefore no theory of how decision makers could learn to make optimal decisions in this context. Here, we derive such a rule for learning a near-optimal linear combination of DDM 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. In this rule, the rate of learning is proportional (respectively, inversely proportional) to confidence for incorrect (respectively, correct) decisions. Furthermore, we show that, in volatile environments, the rule predicts a bias toward repeating the same choice after correct decisions, with a bias strength that is modulated by the previous choice's difficulty. Finally, we extend our learning rule to cases for which one of the choices is more likely a priori, which provides insights into how such biases modulate the mechanisms leading to optimal decisions in diffusion models.

decision making | diffusion models | optimality | confidence

Decisions are a ubiquitous component of everyday behavior. To be efficient, they require handling the uncertainty arising from the noisy and ambiguous information that the environment provides (1). This is reflected in the trade-off between speed and accuracy of decisions. Fast choices rely on little information and may therefore sacrifice accuracy. In contrast, slow choices provide more opportunity to accumulate evidence and thus may be more likely to be correct, but are more costly in terms of attention or effort and lost time and opportunity. Therefore, efficient decisions require not only a mechanism to accumulate evidence but also one to trigger a choice once enough evidence has been collected. Drift-diffusion models (or diffusion decision models) (DDMs) are a widely used model family (2) that provides both mechanisms. Not only do DDMs yield surprisingly good fits to human and animal behavior (3–5), but they are also known to achieve a Bayes-optimal decision strategy under a wide range of circumstances (4, 6–10).

DDMs assume a particle that drifts and diffuses until it reaches one of two boundaries, each triggering a different choice (Fig. 1*A*). The particle's drift reflects the net surplus of evidence toward one of two choices. This is exemplified by the random-dot motion task, in which the motion direction and coherence set the drift sign and magnitude. The particle's stochastic diffusion reflects the uncertainty in the momentary evidence and is responsible for the variability in decision times and choices widely observed in human and animal decisions (3, 5). A standard assumption underlying DDMs is that the noisy momentary evidence that is accumulated over time is one-dimensional—an abstraction of the momentary decision-related evidence of some stimulus. In reality, however, evidence would usually be distributed across a larger number of inputs, such as a neural population in the brain, rather than individual neurons (or neuron/antineuron pairs; Fig. 1*A*). Furthermore, the brain would not know a priori 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 the success and failure of previous choices. How such an interpretation can be efficiently learned over time, both normatively and mechanistically, is the focus of this work.

The multiple existing computational models for how humans and animals might learn to improve their decisions from feedback (e.g., refs. 11–14) do not address the question we are asking, as they all assume that all evidence for each choice is provided at once, without considering the temporal aspect of evidence accumulation. This is akin to fixed-duration experiments, in which the evidence accumulation time is determined by the environment rather than the decision maker. We, instead, address a more general and natural case in which decision times are under the decision maker's control. In this setting, commonly studied using "reaction time" paradigms, the temporal accumulation of evidence needs to be treated explicitly, and—as we will show—the time it took to accumulate this evidence impacts how the decision strategy is updated after feedback. Some models for both choice and reaction times have addressed the presence of high-dimensional inputs (e.g., refs. 15–17). However, they usually assumed as many choices as inputs, were mechanistic rather than normative, and did not consider how interpreting the input could be learned. We furthermore extend on previous work by considering the effect of a priori biases toward believing that one option is more correct than the other, and how such biases can be learned. This yields a theoretical understanding of how choice biases impact optimal decision making in diffusion models. Furthermore, it clarifies of how different implementations of this bias result in different diffusion model implementations, like the one proposed by Hanks et al. (18).

Significance

Popular models for the trade-off between speed and accuracy of everyday decisions usually assume fixed, low-dimensional sensory inputs. In contrast, in the brain, these inputs are distributed across larger populations of neurons, and their interpretation needs to be learned from feedback. We ask how such learning could occur and demonstrate that efficient learning is significantly modulated by decision confidence. This modulation predicts a particular dependency pattern between consecutive choices and provides insight into how a priori biases for particular choices modulate the mechanisms leading to efficient decisions in these models.

Author contributions: J.D., A.G.M., Z.F.M., and A.P. designed research; J.D. and A.P. performed research; J.D. analyzed data; and J.D., A.G.M., Z.F.M., and A.P. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. P.W.G. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: jan_drugowitsch@hms.harvard.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1906787116/-/DCSupplemental.

First published November 15, 2019.

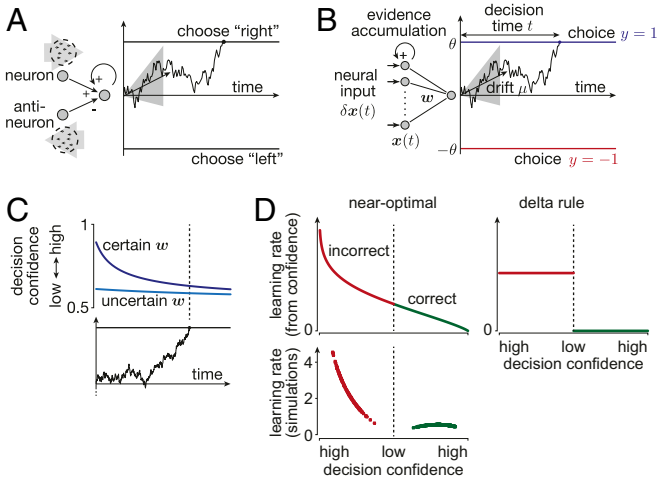

Fig. 1. Learning the input weights from feedback in diffusion models. In diffusion models, the input(s) provide at each point in time noisy evidence about the world's true state, here given by the drift μ. The decision maker accumulates this evidence over time (e.g., black example traces) to form a belief about μ. 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/antineuron pair that codes for opposing directions of evidence. The illustrated example assumes a random-dot task, in which the decision maker needs to identify whether 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 toward the right (*Top*) and left (*Bottom*), such that their difference forms the momentary evidence toward rightward 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 that we assume the input information $\delta \boldsymbol{x}(t)$ to be encoded in a larger neural population whose activity is linearly combined with weights $\boldsymbol{w}$ to yield the one-dimensional momentary evidence, and that the decision maker aims to learn these weights from feedback about the correctness of her choices. (*C*) Decision confidence (i.e., the belief that the made choice was correct) in this kind of diffusion model drops as a function of time (horizontal axis) and with increased uncertainty about the input weights (different shades of blue). (*D*) For near-optimal learning, the learning rate (the term ξ_w in Eq. 6) is modulated by decision confidence (*Top Left*). High-confidence decisions lead to little learning if correct (green, *Right*), and strong learning if incorrect (red, *Left*). Low-confidence decisions result in a moderate confidence-related learning rate term (*Top* and *Center*). The learning rate in 1,000 simulated trials (*Bottom*) shows that the overall learning rate preserves this trend, with an additional suppression of learning for low-confidence decisions. Other learning heuristics (e.g., the delta rule, *Right*) do not modulate their learning by confidence.

Results

Bayes-Optimal Decision Making with Diffusion Models. A standard way (8, 10, 19) to interpret diffusion models as mechanistic implementations of Bayes-optimal decision making is to assume that, in each trial, an unobservable latent state μ (called "drift rate" in 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., rightward vs. leftward motion in the random-dot motion task), irrespective of its magnitude (e.g., the dot coherence level). The latent state itself is not directly observed but is indirectly conveyed via a stream of noisy, momentary evidence values $\delta z_1, \delta z_2, \ldots$, that, in each small time step of size δt, provide independent and identically distributed 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 rescaling of this variance by an additional parameter would result in a global rescaling of the evidence that can be factored out (4, 8, 20), thus making such a rescaling unnecessary.

Having after some time $t \equiv n\delta t$ observed n pieces of such evidence, $\delta z_{1:n}$, the decision maker's posterior belief about μ, $p(\mu | \delta z_{1:n})$, turns out to be fully determined by the accumulated evidence, $z(t) = \sum_{i=1}^{n} \delta z_i$, and time t (*Materials and Methods*). Then, the posterior belief about μ being positive (e.g., leftward motion) results in the following (8):

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

where $\Phi(\cdot)$ is the cumulative function of a standard Gaussian. The opposite belief about μ being negative is simply $p(\mu < 0 | z(t), t) = 1 - p(\mu \geq 0 | z(t), t)$ (see Fig. 3*A*). 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 with drift μ and unit diffusion variance (Fig. 1*A*). By Eq. 1, the posterior belief about $\mu \geq 0$ is $>1/2$ for positive $z(t)$, and $<1/2$ for negative $z(t)$. To make Bayes-optimal decisions, Bayesian decision theory (21) requires that these decisions are chosen to maximize the expected associated reward (or, more formally, to minimize the expected loss). Assuming equally rewarding correct choices, this implies choosing the option that is considered more likely correct. Given the above posterior belief, this makes $y = \mathrm{sign}(z(t)) \in \{-1, 1\}$ the Bayes-optimal choice, which can be implemented mechanistically by (possibly time-varying) boundaries $\pm\theta(t)$ on $z(t)$, associated with the two choices. At these boundaries, the posterior belief about having made the correct choice, or decision confidence (22), is then given by Eq. 1 with $z(t)$ replaced by $\theta(t)$. The sufficient statistics, $z(t)$ and t, of this posterior remain unchanged by the introduction of such decision boundaries, such that Eq. 1 remains valid even in the presence of these boundaries (8). Thus, under the above assumptions of prior and evidence, diffusion models implement the Bayes-optimal decision strategy (see Fig. 3*B*).

Note that $|\mu|$ (i.e., the momentary evidence's signal-to-noise ratio) controls the amount of information provided about the sign of μ, and thus the difficulty of individual decisions. Thus, the used prior $\mu \sim \mathrm{N}(0, \sigma_\mu^2)$, which has more mass on small $|\mu|$, reflects that the difficulty of decisions varies across trials and that harder decisions are more frequent than easier ones. The prior width, σ_μ^2 determines the spread of μ values across trials, and therefore the overall difficulty of the task (larger σ_μ^2 = overall easier task). We chose a Gaussian prior for mathematical convenience, and also because hard trials are more frequent than easy ones in many experiments (e.g., ref. 20), even though they do not commonly use Gaussian priors. In general, the important assumption is that the difficulty varies across trials, but not exactly how it does so, which is to say that the shape of the prior distribution is not critical (8). Different prior choice will not qualitatively change our results but would make it hard or impossible to derive interpretable closed-form expressions. Model predictions would change qualitatively if we assume the difficulty to be fixed, or known a priori (8), but we will not consider this case, as it rarely if ever occurs in the real world.

Using High-Dimensional Diffusion Model Inputs. To extend diffusion models to multidimensional momentary evidence, we assume it to be given the k-dimensional vector $\delta \boldsymbol{x}_i$. This evidence might represent inputs from multiple sensors, or the (abstract) activity of a neuronal population (Fig. 1*B*). As the activity of neurons in a population that encodes limited information about the latent state μ is likely correlated across neurons (23, 24), we chose the

momentary evidence statistics to also feature such correlations (*Materials and Methods*). In general, we choose these statistics such that $\boldsymbol{w}^T\delta\boldsymbol{x}_i = \delta z_i$, where the vector $\boldsymbol{w}$ denotes the k input weights (for now assumed known). Defining the high-dimensional accumulated evidence by $\boldsymbol{x}(t) = \sum_{i=1}^{n} \delta\boldsymbol{x}_i$, this implies $z(t) = \boldsymbol{w}^T\boldsymbol{x}(t)$, such that it is again Bayes-optimal to trigger decisions as soon as $\boldsymbol{w}^T\boldsymbol{x}(t)$ equals one of two decision boundaries $\pm\theta(t)$. Furthermore, the posterior belief about $\mu \geq 0$ is, similar to Eq. 1, given by the following:

$$p(\mu \geq 0 \,|\, \boldsymbol{w}, \boldsymbol{x}(t), t) = \Phi\left(\boldsymbol{w}^T\tilde{\boldsymbol{x}}(t)\right), \qquad [2]$$

where we have defined the time-attenuated accumulated evidence $\tilde{\boldsymbol{x}}(t) = \boldsymbol{x}(t)/\sqrt{t+\sigma_\mu^{-2}}$. As a consequence, the decision-confidence for either choice, is, as before, given by Eq. 2, with $\boldsymbol{w}^T\tilde{\boldsymbol{x}}(t)$ replaced by $\theta(t)/\sqrt{t+\sigma_\mu^{-2}}$. For time-independent decision bounds, $\theta(t) = \theta$, this confidence decreases over time (Fig. 1*C*), reflecting 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. This counterintuitive drop in confidence with time has been previously described for diffusion models with one-dimensional inputs (8, 25) and is a consequence of a trial difficulty that varies across trials. Specifically, it arises from a mixture of easy trials associated with large $|\mu|$ that lead to rapid, high-confidence choices, and hard trials associated with small $|\mu|$ that lead to slow, low-confidence choices. Therefore, it does not depend on our choice of Gaussian prior but is present for any choice of symmetric prior over μ (*SI Appendix*). The confidence remains constant over time only when the difficulty is fixed across trials (i.e., $\mu \in \{-\mu_0, \mu_0\}$ for some fixed μ_0).

Using Feedback to Find the Posterior Weights. So far, we have assumed the decision maker knows the linear input weights $\boldsymbol{w}$ to make Bayes-optimal choices. If they were not known, how could they be learned? Traditionally, learning has been considered an optimization problem, in which the decision maker tunes some decision-making parameters (here, the input weights $\boldsymbol{w}$) to maximize their performance. Here, we will instead consider it as an inference problem in which the decision maker aims to identify the decision-making parameters that are most compatible with the provided observations. These two views are not necessarily incompatible. For example, minimizing the mean squared error of a linear model (an optimization problem) yields the same solution as sequential Bayesian linear regression (an inference problem) (26). In fact, as we show in *SI Appendix*, our learning problem can also be formulated as an optimization problem. Nonetheless, we here take the learning-by-inference route, as it provides a statistical interpretation of the involved quantities, which provides additional insights. Specifically, we focus on learning the weights while keeping the diffusion model boundaries fixed. The decision maker's reward rate (i.e., average number of correct choices per unit time), which we use as our performance measure, depends on both weights and the chosen decision boundaries. However, to isolate the problem of weight learning, we fix the boundaries such that a particular set of optimal weights $\boldsymbol{w}^*$ maximize this reward rate. The aim of weight learning is to find these weights. Weight learning is a problem that needs to be solved even if the decision boundaries are optimized at the same time. We have addressed how to best tune these boundaries elsewhere (8, 27).

To see how learning can be treated as inference, consider the following scenario. Before having observed any evidence, the decision maker has some belief, $p(\boldsymbol{w})$, about the input weights, either as a prior or formed through previous experience. They now observe new evidence, $\delta\boldsymbol{x}_1, \delta\boldsymbol{x}_2, \ldots$ and use the mean of the belief over weights, $\langle\boldsymbol{w}\rangle$ (or any other statistics), to combine this evidence and to trigger a choice y once the combined evidence reaches one of the decision boundaries. Upon this choice, they receive feedback y^* about which choice was the correct one. Then, the best way to update the belief about $\boldsymbol{w}$ in light of this feedback is by Bayes' rule:

$$p(\boldsymbol{w}|\boldsymbol{x}(t), t, y^*) \propto p(y^*|\boldsymbol{w}, \boldsymbol{x}(t), t)p(\boldsymbol{w}), \qquad [3]$$

where we have replaced the stream of evidence $\delta\boldsymbol{x}_1, \delta\boldsymbol{x}_2, \ldots$ by the previously established sufficient statistics $\boldsymbol{x}(t)$ and t.

The likelihood $p(y^*|\boldsymbol{w}, \boldsymbol{x}(t), t)$ expresses for any hypothetical weight vector $\boldsymbol{w}$ the probability that the observed evidence makes y^* the correct choice. To find its functional form, consider that, for a known weight vector, we have shown that $p(\mu \geq 0|\boldsymbol{w}, \boldsymbol{x}(t), t)$, given by Eq. 2, expresses the probability that $y = 1$ (associated with $\mu \geq 0$) is the correct choice. Therefore, $1 - p(\mu \geq 0|\boldsymbol{w}, \boldsymbol{x}(t), t)$ corresponds to the probability that $y = -1$ (associated with $\mu < 0$) is the correct choice. Therefore, it can act as the above likelihood function, which, by Eq. 2, is given by $p(y^*|\boldsymbol{w}, x(t), t) = \Phi(y^*\boldsymbol{w}^T\tilde{\boldsymbol{x}}(t))$, where we have used $1 - \Phi(a) = \Phi(-a)$. In summary, the decision maker's belief is optimally updated after each choice by the following:

$$p(\boldsymbol{w}|\boldsymbol{x}(t), t, y^*) \propto \Phi\left(y^*\boldsymbol{w}^T\tilde{\boldsymbol{x}}(t)\right)p(\boldsymbol{w}). \qquad [4]$$

This update equation only requires knowing the accumulated evidence $\boldsymbol{x}(t)$, decision time t, and feedback y^*, but is independent of the chosen option y, and how the decision maker came to this choice. As a matter of fact, the decision maker could make random choices, irrespective of the accumulated evidence, and still learn $\boldsymbol{w}$ according to the above update equation, as long as they keep track of $\boldsymbol{x}(t)$ and t, and acknowledge the feedback y^*. Therefore, learning and decision making are not necessarily coupled. Nonetheless, we assume for all simulations that decision makers perform decisions by using the mean estimate $\langle\boldsymbol{w}\rangle$, which is an intuitively sensible choice if the decision maker's aim is to maximize their reward rate (*SI Appendix*).

As in Eq. 4, the likelihood parameters, $\boldsymbol{w}$, are linear within a cumulative Gaussian function, such problems are known as "probit regression" and do not 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, as well as a potential mechanistic implementation.

Confidence Controls the Learning Rate. To find an approximation to the posterior in Eq. 4, let us assume the prior to be given by the Gaussian distribution, $p(\boldsymbol{w}) = N(\boldsymbol{w}\,|\,\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$, with mean $\boldsymbol{\mu}_w$ and covariance $\boldsymbol{\Sigma}_w$, which is the maximum entropy distribution that specifies the mean and covariance (28). First, we investigated how knowing $\boldsymbol{w}$ with limited certainty, as specified by $\boldsymbol{\Sigma}_w$, impacts the decision confidence. Marginalizing over all possible $\boldsymbol{w}$ values (*Materials and Methods*) resulted in the choice confidence to be given by the following:

$$p(y\,|\,\boldsymbol{x}(t), t) = \Phi\left(\frac{y\boldsymbol{\mu}_w^T\tilde{\boldsymbol{x}}}{\sqrt{1+\tilde{\boldsymbol{x}}^T\boldsymbol{\Sigma}_w\tilde{\boldsymbol{x}}}}\right). \qquad [5]$$

Compared to Eq. 2, the choice confidence is additionally attenuated by $\boldsymbol{\Sigma}_w$. Specifically, higher weight uncertainty (i.e., an overall larger covariance $\boldsymbol{\Sigma}_w$) results in a lower decision confidence, as one would intuitively expect (Fig. 1*C*).

Next, we found a closed-form approximation to the posterior (Eq. 4). For repeated learning across consecutive decisions, the posterior over the weights after the previous decision becomes the prior for the new decision. Unfortunately, a direct application of this principle would lead to a posterior that changes its functional form after each update, making it intractable. We instead used assumed density filtering (ADF) (26, 29) that posits a fixed functional form $q(\boldsymbol{w}|y^*,\boldsymbol{x}(t),t)=\mathrm{N}(\boldsymbol{w}|\boldsymbol{\mu}_w^*,\boldsymbol{\Sigma}_w^*)$ of the posterior density—in our case, Gaussian for consistency with the prior—and then finds the posterior parameters $\boldsymbol{\mu}_w^*$ and $\boldsymbol{\Sigma}_w^*$ that make this approximate posterior best match the "true" posterior $p(\boldsymbol{w}\,|\,y^*,\boldsymbol{x}(t),t)$ (Eq. 4). Performing this match by minimizing the Kullback–Leiber divergence $\mathrm{KL}(p\|q)$ results in the posterior mean (30, 31):

$$\boldsymbol{\mu}_w^*=\boldsymbol{\mu}_w+\frac{\xi_w}{\sqrt{1+\tilde{\boldsymbol{x}}^T\boldsymbol{\Sigma}_w\tilde{\boldsymbol{x}}}}y^*\boldsymbol{\Sigma}_w\tilde{\boldsymbol{x}}, \qquad [6]$$

and a similar expression for the posterior covariance (*Materials and Methods*). Choosing $\mathrm{KL}(p\|q)$ to measure the distance between p and q is to some degree arbitrary, but has beneficial properties, such as that it causes the first two moments of q to match those of p (*SI Appendix*). In Eq. 6, the factor ξ_w modulates how strongly this mean is updated toward $y^*\boldsymbol{\Sigma}_w\tilde{\boldsymbol{x}}$, and turns out to be a monotonically decreasing function of decision confidence (Fig. 1*D*, *Top*; see *Materials and Methods* for mathematical expression). For incorrect choices, for which the decision confidence is $p(y^*|\boldsymbol{x}(t),t)<1/2$, ξ_w is largest for choices made with high confidence, promoting significant weight adjustments. For low-confidence choices, it only promotes moderate adjustments, notably irrespective of whether the choice was correct or incorrect. High-confidence, correct choices yield a low ξ_w, and thus an intuitively minor strategy update. The update of the posterior covariance follows a similar confidence-weighted learning rate modulation (*SI Appendix*, Fig. S1, and *Materials and Methods*).

Decision confidence is not the only factor that impacts the learning rate in Eq. 6. For instance, $\tilde{\boldsymbol{x}}$ shrinks for longer, less confidence choices (because it is inversely proportional to time) and results in overall less learning. Less certain weights, associated with larger magnitudes of $\boldsymbol{\Sigma}_w$, have a similar effect. To investigate the overall impact of all of these factors combined on the learning rate, we simulated a long sequence of consecutive choices and plotted the learning rate for a random subset of these trials against the decision confidence (Fig. 1 *D*, *Bottom*). This plot revealed a slight down-weighting of the learning rate for low-confidence choices when compared to ξ_w, but left the overall dependency on ξ_w otherwise unchanged.

Performance Comparison to Optimal Inference and to Simpler Heuristics. The intuitions provided by near-optimal ADF learning are only informative if its approximations do not cause a significant performance drop. We quantified this drop by comparing ADF performance to that of the Bayes-optimal rule, as found by Gibbs sampling (*Materials and Methods*). Gibbs sampling is biologically implausible as it requires a complete memory of inputs and feedbacks for past decisions and is intractable for longer decision sequences, but nonetheless provides an optimal baseline to compare against. We furthermore tested the performance of two additional approximations. One was an ADF variant that assumes a diagonal covariance matrix $\boldsymbol{\Sigma}_w$, yielding a local learning rule that could be implemented by the nervous system. This variant furthermore reduced the number of parameters from quadratic to linear in the size of $\boldsymbol{w}$. The second was a second-order Taylor expansion of the log-posterior, resulting in a learning rule similar to ADF, but with a lower impact of weight uncertainty on the learning rate (*Materials and Methods*).

Furthermore, we tested whether simpler learning heuristics can match ADF performance. We focused on three rules of increasing complexity. The delta rule, which can be considered a variant of temporal-difference learning, or reinforcement learning (32), updates its weight estimate after the nth decision by the following:

$$\boldsymbol{w}_{n+1}=\boldsymbol{w}_n+\frac{\alpha}{2\theta(0)}\left(y_n^*\theta(t)-\boldsymbol{x}_n(t)^T\boldsymbol{w}_n\right)\boldsymbol{x}_n(t), \qquad [7]$$

where $y_n^*\in\{-1,1\}$ is the feedback about the correct choice provided after this decision, and we have chosen to normalize the learning rate α by the initial bound height $\theta(0)$ to make it less sensitive to this chosen height. 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 $\pm2\theta(t)$ for incorrect choices. As a result, and in contrast to ADF, weight adjustments are only performed after incorrect choices, and with a fixed learning rate α rather than one modulated by confidence (Fig. 1 *D*, *Right*). Our simulations revealed that the delta rule excessively and suboptimally decrease in the weight size $\|\boldsymbol{w}\|$ over time, leading to unrealistically long reaction times and equally unrealistic near-zero weights. To counteract this problem, we designed a normalized delta rule, that updates the weight estimates as the delta rule, but thereafter normalizes them by $\boldsymbol{w}\leftarrow\boldsymbol{w}\|\boldsymbol{w}^*\|/\|\boldsymbol{w}\|$ to ensure that its size matches that of the true weights $\boldsymbol{w}^*$. Access to these true weights, $\boldsymbol{w}^*$, makes it an omniscient learning rule that cannot be implemented by a decision maker in practice. Last, we tested a learning rule that performs stochastic gradient ascent on the feedback log-likelihood:

$$\boldsymbol{w}_{n+1}=\boldsymbol{w}_n+\alpha\nabla_w\log p\left(y_n^*\,|\,\boldsymbol{w}_n,\boldsymbol{x}_n(t),t\right)=\boldsymbol{w}_n+\alpha y_n^*\xi_w\tilde{\boldsymbol{x}}_n(t). \qquad [8]$$

This rule introduces decision confidence weighting through ξ_w, but differs from ADF in that it does not take the weight uncertainty ($\boldsymbol{\Sigma}_w$ in ADF) into account, and requires tuning of the learning rate parameter α.

We evaluated the performance of these learning rules by simulating weight learning across 1,000 consecutive decisions (called "trials"; see *Materials and Methods* for details) in a task in which use of the optimal weight vector maximizes the reward rate. This reward rate was the average reward for correct choices minus some small cost for accumulating evidence over the average time across consecutive trials and is a measure we would expect rational decision makers to optimize. For each learning rule, we found its reward rate relative to random behavior and optimal choices.

Fig. 2*A* shows this relative reward rate for all learning rules and different numbers of inputs. As can be seen, the performance of ADF and the other probabilistic learning rules is indistinguishable from Bayes-optimal weight learning for all tested numbers of inputs. Surprisingly, the ADF variant that ignores off-diagonal covariance entries even outperformed Bayes-optimal learning for a large number of inputs (Fig. 2*A*, yellow line for 50 inputs). The reason that a simpler learning rule could outperform the rule deemed optimal by Bayesian decision theory is that this simpler rule has less parameters and a simpler underlying model that was nonetheless good enough to learn the required weights. Learning fewer parameters with the same data resulted in initially better parameter estimates, and better associated performance. Conceptually, this is similar to a linear model outperforming a quadratic model when fitting a quadratic function if little data are available, and if the function is sufficiently close to linear (as illustrated in *SI Appendix*, Fig. S2). Once more data are available, the quadratic model will outperform the linear one. Similarly, the Bayes-optimal learning rule will outperform the simpler one once more feedback has been observed. In our simulation, however, this does not occur within the 1,000 simulated trials.

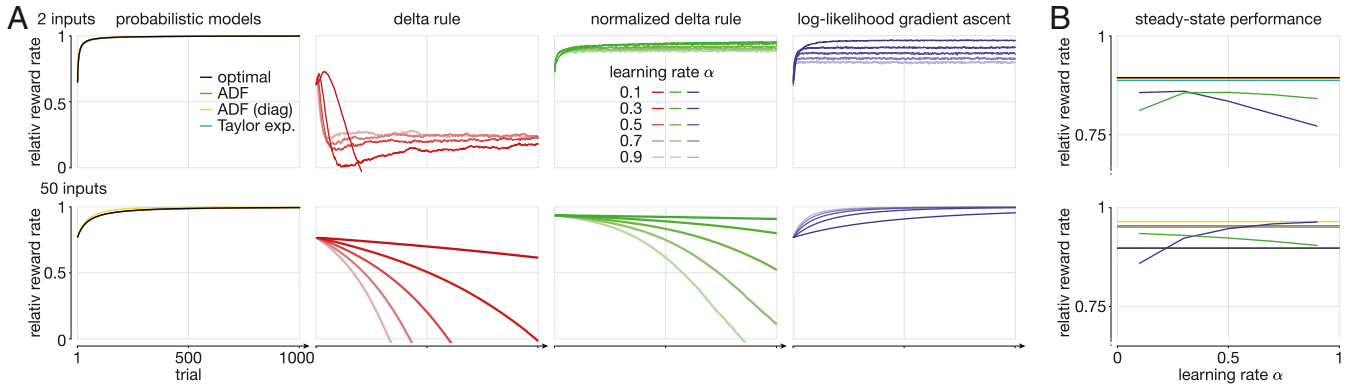

Fig. 2. Input weight learning and tracking performance of different learning rules. All plots show the relative reward rate (0 = immediate, random choices; 1 = optimal) averaged over 5,000 simulations with different true, underlying weights, and for 2 (*Top*) and 50 (*Bottom*) inputs. (*A*) The relative reward rate for probabilistic and heuristic learning rules. The probabilistic learning rules include the optimal rule (Gibbs sampling), assumed density filtering (ADF), ADF with a diagonal covariance matrix (ADF [diag]), and a learning rule based on a second-order Taylor expansion of the log-posterior (Taylor exp.). For both 2 and 50 inputs, all rules perform roughly equally. For the heuristic rules, different color shadings indicate different learning rates. The initial performance shown is that after the first application of the learning rule, such that initial performances can differ across learning rules. (*B*) The steady-state performance across different heuristic rule learning rates. Steady-state performance was measured as an average across 5,000 simulations, averaging over the last 100 of 1,000 simulated trials in which the true weights slowly change across consecutive trials. An optimal relative reward rate of 1 corresponds to knowing the true weight in each trial, which, due to the changing weight, is not achievable in this setup. The color scheme is the same as in *A*, but the vertical axis has a different scale. The delta rule did not converge and was not included in *B*.

All other learning heuristics performed significantly worse. For low-dimensional input, the delta rule initially improved its reward rate but worsens it again at a later stage across all learning rates. The normalized delta rule avoided such performance drops for low-dimensional input, but both delta rule variants were unable to cope with high-dimensional inputs. Only stochastic gradient ascent on the log-likelihood provided a stable learning heuristic for high-dimensional inputs, 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. Overall, the probabilistic learning rules significantly outperformed all tested heuristic learning rules and matched (and in one case even exceeded) the weight learning performance of the Bayes-optimal estimator.

Tracking Nonstationary Input Weights. So far, we have tested how well 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 of Bayes-optimal learning (now implemented by a particle filter) to that of the probabilistic and heuristic learning rules introduced in the previous section. The probabilistic rules were updated to take into account such a trial-by-trial weight change, as modeled by a first-order autoregressive process (*Materials and Methods*). The heuristic rules remained unmodified, as their use of a constant learning rate already encapsulates the assumption that the true weights change across decisions.

Fig. 2*B* illustrates the performance of the different learning rules. First, it shows that, for low-dimensional inputs the various probabilistic models yield comparable performances, but for high-dimensional inputs the approximate probabilistic learning rules outperform Bayes-optimal learning. In case of the latter, these approximations were not actually harmful, but instead beneficial, for the same reason discussed further above. In particular, the more neurally realistic ADF variant that only tracked the diagonal of the covariance matrix again outperformed all other probabilistic models. Second, only the heuristic learning rule that performed gradient ascent on the log-likelihood achieved steady-state performance comparable to the approximate probabilistic rules, and then only for high input dimensionality and a specific choice of learning rate. This should come as no surprise, as its use of the likelihood function introduces more task structure information than the other heuristics use. The delta rule did not converge and therefore never achieved steady-state performance. Overall, the ADF variant that focused only on the diagonal covariance matrix achieved the best overall performance.

Learning Both Weights and a Latent State Prior Bias. Our learning rule can be generalized to learn prior biases in addition to the input weights. The prior we have used so far for the latent variable, $\mu \sim \mathrm{N}(0, \sigma_\mu^2)$, is unbiased, as both $\mu \geq 0$ and $\mu < 0$ are equally likely. To introduce a prior bias, we instead used $\mu \sim \mathrm{N}(m, \sigma_\mu^2)$, where m controls the bias through $P^+ \equiv p(\mu \geq 0) = \Phi(m/\sigma_\mu)$. A positive (or negative) m causes $P^+ > 1/2$ (or $<1/2$), thus making $y = 1$ (or $y = -1$) the more likely correct choice even before evidence is accumulated. After evidence accumulation, such a prior results in the posterior:

$$p(\mu \geq 0 | \boldsymbol{w}, \boldsymbol{x}(t), t) = \Phi\left(\frac{\boldsymbol{w}^T \boldsymbol{x}(t) + \sigma_\mu^{-2} m}{\sqrt{t + \sigma_\mu^{-2}}}\right). \quad [9]$$

Comparing this to the unbiased posterior, Eq. 2, reveals the additional term $\sigma_\mu^{-2} m$ whose relative influence wanes over time.

This additional term has two consequences. First, appending the elements m and σ_μ^{-2} to the vectors $\boldsymbol{w}$ and $\boldsymbol{x}(t)$, respectively, shows that $\boldsymbol{w}$ and m can be learned jointly by the same learning rule we have derived before (*Materials and Methods*). Second, the term requires us to rethink the association between decision boundaries and choices. As Fig. 3*C* illustrates, such a prior causes a time-invariant shift in the association between the

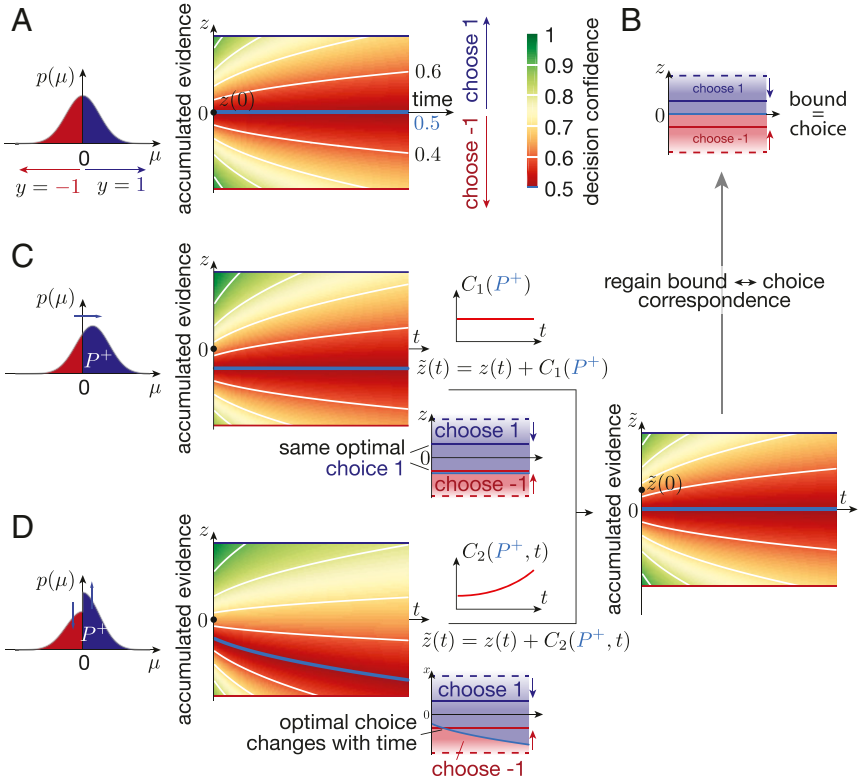

Fig. 3. Decision confidence, prior biases, and the relation between decision boundary and choice. (*A*) For an unbiased prior [i.e., $P^+ \equiv p(\mu \geq 0) = 1/2$], the decision confidence (color gradient) is symmetric around $z = 0$ for each fixed time t. The associated posterior belief $p(\mu \geq 0 | z(t), t)$ (numbers above/below "time" axis label; constant along white lines; ½ along light blue line) promote choosing $y = 1$ and $y = -1$ above (blue area in *B*) and below (red area in *B*) $z = 0$. (*B*) As a result, different choices are Bayes-optimal at the blue/red decision boundaries, as long as they are separated by $z = 0$, irrespective of the boundary separation (solid vs. dashed blue red lines). (*C*) If the prior is biased by an overall shift, the decision confidence is countershifted by the same constant across all t. In this case, both decision boundaries might promote the same choice, which can be counteracted by a time-invariant shift of z by $C_1(P^+)$. (*D*) If the prior is biased by boosting one side while suppressing the other, the decision confidence shift becomes time dependent, such that the optimal choice at a time-invariant boundary might change over time. Counteracting this effect requires a time-dependent shift of z by $C_2(P^+, t)$. In both *C* and *B*, we have chosen $P^+ = 0.6$, for illustration.

accumulated evidence, $z(t) = \boldsymbol{w}^T \boldsymbol{x}(t)$, and the posterior belief of $\mu \geq 0$ and corresponding decision confidence. This shift makes it possible to have the same Bayes-optimal choice at both decision boundaries (Fig. 3*C*, blue/red decision areas). Hence, we have lost the mechanistically convenient unique association between decision boundaries and choices. We recover this association by a boundary counter shift, such that these boundaries come to lie at the same decision confidence levels for opposite choices, making them asymmetric around $z = 0$. Mathematically, this is equivalent to shifting the evidence accumulation starting point, $\tilde{z}(0)$ away from zero in the opposite direction [Fig. 3*C*, shift by $C_1(P^+) = \sigma_\mu^{-2} m$; *SI Appendix*]. Therefore, a prior bias is implemented by a bias-dependent simple shift of the accumulation starting point, leading to a mechanistically straightforward implementation of Bayes-optimal decision making with biased priors.

A consequence of the shifted accumulation starting point is that, for some fixed decision time t, the decision confidence at both boundaries is the same (Fig. 3 *C*, *Right*). This seems at odds with the intuition that a biased prior ought to bias the decision 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. As the starting point is now further away from the less likely correct boundary, 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 given decision time, it will on average across decision times be lower for the a priori non-preferred boundary, faithfully implementing this prior (see *SI Appendix* for a mathematical demonstration).

Our finding that a simple shift in the accumulation starting point is the Bayes-optimal strategy appears at odds with previous work that suggested that the optimal shift of the accumulator variable $z(t)$ varies with time (18). This difference stems from a different implementation of the bias. While we have chosen an overall shift in the prior by its mean (Fig. 3*C*), an alternative implementation is to multiply $p(\mu \geq 0)$ by P^+, and $p(\mu < 0)$ by $1 - P^+$ (Fig. 3*D*), again resulting in $P^+ = p(\mu \geq 0)$. A consequence of this difference is that the associated shift of the posterior belief of $\mu \geq 0$ in the evidence accumulation space becomes time dependent. Then, the optimal choice at a time-invariant boundary in that space might change over time (Fig. 3*D*). Furthermore, undoing this shift to regain a unique association between boundaries and choices not only requires a shifted accumulation starting point, but additionally a time-dependent additive signal [$C_2(P^+, t)$ in Fig. 3*D*; *SI Appendix*], as was proposed in ref. 18. Which of the two approaches is more adequate depends on how well it matches the prior implicit in the task design. Our approach has the advantage of a simpler mechanistic implementation, as well as yielding a simple extension to the previously derived learning rule. How learning prior biases in the framework of ref. 18 could be achieved remains unclear (but see ref. 33).

Sequential Choice Dependencies due to Continuous Weight Tracking. In everyday 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. A common assumption is 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. By "slow," we mean that we can consider it constant over a single trial but that it is unstable over the course of an hour-long session. We implemented this tracking of the moving world, as in Fig. 2*B*, by slowly allowing the weights mapping evidence to decisions to change. With such continuously changing weights, weight learning never ends. Rather, the input 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. 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 (Fig. 1 *D*, *Top*, 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 relied on the same simulation to generate Fig. 2*B* to measure how likely the model repeated the same choice after correct decisions. Fig. 4*A* 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 $|\mu|$; Fig. 4*B*). 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.

Unreliable Feedback Reduces Learning. What would occur if choice feedback is less-than-perfectly reliable? For example, the feedback itself might not be completely trustworthy, or hard to interpret. We simulated this situation by assuming that the feedback is inverted with probability β. Here, $\beta = 0$ implies the so far assumed perfectly reliable feedback, and $\beta = 1/2$ makes the feedback completely uninformative. This change impacts how decision confidence modulates the learning rate (Fig. 4*C*) as follows. First, it reduces the overall magnitude of the correction, with weaker learning for higher feedback noise. Second, it results in no learning for highly confident choices that we are told are incorrect. In this case, one's decision confidence overrules the unreliable feedback. This stands in stark contrast to the optimal learning rule for perfectly reliable feedback, in which case the strongest change to the current strategy ought to occur.

Discussion

Diffusion models are applicable to model decisions that require some accumulation of evidence over time, which is almost always the case in natural decisions. We extended previous work on the normative foundations of these models to more realistic situations in which the sensory evidence is encoded by a population of neurons, as opposed to just two neurons, as has been typically assumed in previous studies. We have focused on normative and mechanistic models for learning the weights from the sensory neurons to the decision integrator without additionally adjusting the decision boundaries, as weight learning is a problem that needs to be solved even if the decision boundaries are optimized at the same time.

From the Bayesian perspective, weight learning corresponds to finding the weight posterior given the provided feedback, and resulted in an approximate learning rule whose learning rate was strongly modulated by decision confidence. 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. 1*D*). Evidence for such confidence-based learning has already been identified in human experiments (34), but not in a task that required the temporal accumulation of evidence in individual trials. Indeed, as we have previously suggested (22), such a modulation by decision confidence should arise in all scenarios of Bayesian learning in N-AFC tasks in which the decision maker only receives feedback about the correctness of their choices, rather than being told which choice would have been correct. In the 2-AFC task we have considered, being told that one's choice was incorrect automatically reveals that the other choice was correct, making the two cases coincide. Moving from one-dimensional to higher-dimensional inputs requires performing the accumulation of evidence for each input dimension separately [Fig. 1*B*; Eqs. 6 and 12 require $\boldsymbol{x}(t)$ rather than only $\boldsymbol{w}^T\boldsymbol{x}(t)$], even if triggering choices only requires a linear combination of $\boldsymbol{x}(t)$. This is because uncertain input weights require keeping track of how each input dimension contributed to the particle crossing the decision boundary in order to correctly improve these weights upon feedback (i.e., proper credit assignment). The multidimensional evidence accumulation

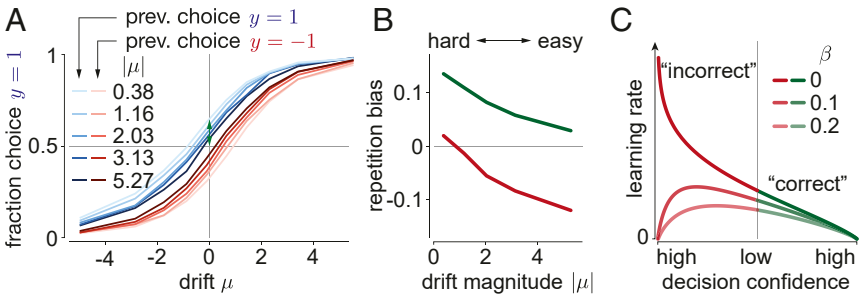

Fig. 4. Sequential choice dependencies due to continuous learning, and effects of noisy feedback. Bayes-optimal learning in a slowly changing environment predicts sequential choice dependencies with the following pattern. (*A*) After hard, correct choices (low prev. $|\mu|$; light colors), the psychometric curve is shifted toward repeating the same choice (blue/red = choice $y = 1/-1$). This shift decreases after easier, correct choices (high prev. $|\mu|$; dark colors). (*B*) We summarize these tuning curve shifts in the repetition bias, which is the probability of repeating the same choice to a $\mu = 0$ stimulus (example green arrow for $\mu = -0.38$ in *A*). After correct/incorrect choices (green/red curve), this leads to a win–stay/lose–switch strategy. Only the win–stay strategy is shown in *A*. (*C*) If choice feedback is noisy (inverted with probability β), the learning rate becomes overall lower. In particular for high-confidence choices with "incorrect" feedback, the learning rate becomes zero, as the learners trust their choice more than the feedback.

predicted by our work arises naturally if inputs encode full distributions across the task-relevant variables, such as in linear probabilistic population codes (35) that trigger decisions by bounding the pooled activity of all units that represent the accumulated evidence (36).

Continual weight learning predicts sequential choice dependencies that make the repetition of a previous, correct choice more likely, in particular if this choice was difficult (Fig. 4). Thus, based on assuming a volatile environment that promotes a continual adjustment of the decision-making strategy, we provide a rational explanation for sequential choice dependencies that are frequently observed in both humans and animals (e.g., refs. 37 and 38). In rodents making decisions in response to olfactory cues, we have furthermore confirmed that these sequential dependencies are modulated by choice difficulty, and that the exact pattern of this modulation depends on the stimulus statistics, as predicted by our theory (39) (but consistency with ref. 40 is unclear).

Last, we have clarified how prior biases ought to impact Bayes-optimal decision making in diffusion models. Extending the work of Hanks et al. (18), we have demonstrated that the exact mechanisms to handle these biases depend on the specifics of how these biases are introduced through the task design. Specifically, we have suggested a variant that simplifies these mechanisms and the learning of this bias. This variant predicts that the evidence accumulation offset, that has previously been suggested to be time-dependent, to become independent of time, and it would be interesting to see whether the lateral intraparietal cortex activity of monkeys performing the random-dot motion task, as recorded by Hanks et al. (but see ref. 41), would change accordingly.

Materials and Methods

We here provide an outline of the framework and its results. Detailed derivations are provided in *SI Appendix*.

Bayesian Decision Making with One and Multidimensional Diffusion Models. We assume the latent state to be drawn from $\mu \sim \mathrm{N}(m, \sigma_\mu^2)$, and the momentary evidence in each time step δt to provide information about this latent state by $\delta z_i | \mu \sim \mathrm{N}(\mu \delta t, \delta t)$. The aim is to infer the sign of μ, and choose $y = 1$ if $\mu \geq 0$, and $y = -1$ otherwise. 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 the following:

$$p(\mu \,|\, \delta z_{1:n}) \propto \mathrm{N}\left(\mu \,|\, m, \sigma_\mu^2\right) \prod_{i=1}^{n} \mathrm{N}(\delta z_i \,|\, \mu \delta t, \delta t) \propto \mathrm{N}\left(\mu \,|\, \frac{\sigma_\mu^{-2} m + z(t)}{\sigma_\mu^{-2} + t}, \frac{1}{\sigma_\mu^{-2} + t}\right). \quad [11]$$

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 for μ's sign with $m = 0$ is described around Eq. 1.

We extend diffusion models to multidimensional inputs with momentary evidence $\delta \boldsymbol{x}_i \,|\, \mu, \boldsymbol{w} \sim N((\boldsymbol{a}\mu + \boldsymbol{b})\delta t, \boldsymbol{\Sigma}\delta t)$, with $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{\Sigma}$ chosen such that $\boldsymbol{w}^T \boldsymbol{x}(t) | \mu = z(t) | \mu \sim N(\mu t, t)$, as before. The posterior over μ and $\mu \geq 0$ is the same as for the one-dimensional case, with $z(t)$ replaced by $\boldsymbol{w}^T \boldsymbol{x}(t)$. Defining $\tilde{\boldsymbol{x}}(t) = \boldsymbol{x}(t) / \sqrt{\sigma_\mu^{-2} + t}$, we find $p(\mu \geq 0 | \boldsymbol{w}, \boldsymbol{x}(t), t) = \Phi(\boldsymbol{w}^T \tilde{\boldsymbol{x}}(t))$. As $y = 1$ and $y = -1$ correspond to $\mu \geq 0$ and $\mu < 0$, and $y = 1$ is only chosen if $p(\mu \geq 0 | \boldsymbol{w}, \boldsymbol{x}(t), t) \geq 1/2$, the decision confidence for $m = 0$ at some boundary $\boldsymbol{w}^T \boldsymbol{x}(t) = \pm\theta(t)$ is given by $\Phi(\theta(t) / \sqrt{\sigma_\mu^{-2} + t})$. If input weights are unknown, and the decision maker holds belief $\boldsymbol{w} \sim \mathrm{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$ about these weights, the decision confidence needs to additionally account for weight uncertainty by marginalizing over $\boldsymbol{w}$, resulting in Eq. 5.

Probabilistic and Heuristic Learning Rules. We find the approximate posterior $q(\boldsymbol{w}) = N(\boldsymbol{w} | \boldsymbol{\mu}_w^*, \boldsymbol{\Sigma}_w^*)$ that approximates the target posterior p Eq. 4 by ADF. This requires minimizing the Kullback–Leiber divergence $KL(p|q)$ (26, 29), resulting in Eq. 6 for the posterior mean, and the following:

$$\boldsymbol{\Sigma}_w^* = \boldsymbol{\Sigma}_w + \xi_{cov}(\gamma) \left(\left(\boldsymbol{\Sigma}_w^{-1} + \tilde{\boldsymbol{x}} \tilde{\boldsymbol{x}}^T \right)^{-1} - \boldsymbol{\Sigma}_w \right), \quad [12]$$

with learning rate modulators $\xi_w(\gamma) = \mathrm{N}(\gamma | 0, 1) / \Phi(\gamma)$ and $\xi_{cov}(\gamma) = \xi_w(\gamma)^2 + \xi_w(\gamma)\gamma$, and where we have defined $\gamma \equiv y^* \boldsymbol{\mu}_w^T \tilde{\boldsymbol{x}} / \sqrt{1 + \tilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}}$, which is monotonic in the decision confidence (Eq. 5). Noisy choice feedback (Fig. 4*C*) changes the likelihood to assume reversed feedback with probability β, and follow the same procedure as above to derive the posterior moments (*SI Appendix*). The ADF variant that only tracks the diagonal covariance elements assumes $\boldsymbol{\Sigma}_w$ to be diagonal, and only computes the diagonal elements of $\boldsymbol{\Sigma}_w^*$. A second-order Taylor expansion of the log of Eq. 4 leads to update equations similar to Eqs. 6 and 12, but without the normalization by weight uncertainty (see *SI Appendix* for details). All heuristic learning rules are described in the main text.

We modeled nonstationary input weights by $\boldsymbol{w}_n | \boldsymbol{w}_{n-1} \sim \mathrm{N}(\boldsymbol{A}\boldsymbol{w}_{n-1} + \boldsymbol{b}, \boldsymbol{\Sigma}_d)$ after a decision in trial $n - 1$. This weight transition is taken into account by the probabilistic learning rules by setting the parameter priors to $\boldsymbol{\mu}_{w,n} = \boldsymbol{A}\boldsymbol{\mu}_{w,n-1}^* + \boldsymbol{b}$ and $\boldsymbol{\Sigma}_{w,n} = \boldsymbol{A}\boldsymbol{\Sigma}_{w,n-1}^* \boldsymbol{A}^T + \boldsymbol{\Sigma}_d$. For stationary weights, we have $\boldsymbol{A} = \boldsymbol{I}$, $\boldsymbol{b} = 0$, and $\boldsymbol{\Sigma}_d = 0$.

Bayes-optimal weight inference was for stationary weights performed by Gibbs sampling for probit models, and for nonstationary weights by particle filtering (*SI Appendix*).

Simulation Details. We used parameters $\boldsymbol{a} = \boldsymbol{w} / \|\boldsymbol{w}\|^2$ and $\boldsymbol{b} = 0$ for the momentary evidence $\delta \boldsymbol{x}$. Its covariance $\boldsymbol{\Sigma}$ was generated to feature eigenvalues that drop exponentially from $\sigma_x^2 = 2 / \|\boldsymbol{w}\|^2$ to zero until it reaches a constant $\sigma_0^2 = 0.001 / \|\boldsymbol{w}\|^2$ noise baseline, as qualitatively observed in neural populations. It additionally contains an eigenvector $\boldsymbol{w}$ with eigenvalue set to guarantee $\boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w} = 1$, limiting the information that $\delta \boldsymbol{x}$ provides about μ. For nonstationary weights, all momentary evidence parameters are adjusted after each weight change (*SI Appendix*). The diffusion model bounds $\pm\theta$ were time-invariant and tuned to maximize the reward rate when using the correct weights. The reward rate is given by $(p(correct) - c_{accum} t) / (t_{iti} + t)$, where averages were across trials, and we used evidence accumulation cost $c_{accum} = 0.01$ and intertrial interval $t_{iti} = 2s$. We used $\sigma_\mu^2 = 3^2$ to draw μ in each trial, and drew $\boldsymbol{w}$ from $\boldsymbol{w} \sim \mathrm{N}(1, \boldsymbol{I})$ before each trial sequence. For nonstationary weights, we resampled weights after each trial according to $\boldsymbol{w}_n | \boldsymbol{w}_{n-1} \sim \mathrm{N}(\lambda \boldsymbol{w}_{n-1} + (1 - \lambda), \sigma_d^2 \boldsymbol{I})$, with decay factor $\lambda = 1 - 0.01$ and $\sigma_d^2 = 1 - \lambda^2$ to achieve steady-state mean $\boldsymbol{1}$ and identity covariance.

To compare the weight learning performance of ADF to alternative models (Fig. 2*A*), we simulated 1,000 learning trials 5,000 times, and reported the reward rate per trial averaged across these 5,000 repetitions. To assess steady-state performance (Fig. 2*B*), we performed the same procedure with nonstationary weights and reported reward rate averaged over the last 100 trials, and over 5,000 repetitions. The same 100 trials were used to compute the sequential choice dependencies in Fig. 4 *A* and *B*. To simulate decision making with diffusion models and uncertain weights, we used the current mean estimate $\langle \boldsymbol{w} \rangle$ of the input weights to linearly combine the momentary evidence. The probabilistic learning rules were all independent of the specific choice of this estimate. The learning rate in Fig. 1*D* shows the prefactor to $y^* \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}$ in Eq. 6 over decision confidence for a subsample of the last 10,000 trials of a single 15,000 trial simulation with nonstationary weights. For the Gibbs sampler, we drew 10 burn-in samples, followed by 200 samples in each trial. For the particle filter, we simulated 1,000 particles.

ACKNOWLEDGMENTS. This work was supported by a James S. McDonnell Foundation Scholar Award (220020462) (J.D.) and grants from the National Institute of Mental Health (R01MH115554) (J.D.), the Swiss National Science Foundation (www.snf.ch) (31003A_143707 and 31003A_165831) (A.P.), the Champalimaud Foundation (Z.F.M.), the European Research Council (Advanced Investigator Grants 250334 and 671251) (Z.F.M.), the Human Frontier Science Program (Grant RGP0027/2010) (Z.F.M. and A.P.), the Simons Foundation (Grant 325057) (Z.F.M. and A.P.), and Fundação para a Ciência e a Tecnologia (A.G.M.).

1. K. Doya, S. Ishii, A. Pouget, R. P. N. Rao, *Bayesian Brain: Probabilistic Approaches to Neural Coding* (MIT Press, 2006).
2. R. Ratcliff, A theory of memory retrieval. *Psychol. Rev.* **85**, 59–108 (1978).
3. R. Ratcliff, G. McKoon, The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Comput.* **20**, 873–922 (2008).
4. R. Bogacz, E. Brown, J. Moehlis, P. Holmes, J. D. Cohen, The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113**, 700–765 (2006).
5. R. Ratcliff, P. L. Smith, A comparison of sequential sampling models for two-choice reaction time. *Psychol. Rev.* **111**, 333–367 (2004).

6. P. I. Frazier, A. J. Yu, Sequential hypothesis testing under stochastic deadlines. *Adv. Neural Inf. Process. Syst.* **2008**, 1–8 (2008).
7. S. Tajima, J. Drugowitsch, A. Pouget, Optimal policy for value-based decision-making. *Nat. Commun.* **7**, 12400 (2016).
8. J. Drugowitsch, R. Moreno-Bote, A. K. Churchland, M. N. Shadlen, A. Pouget, The cost of accumulating evidence in perceptual decision making. *J. Neurosci.* **32**, 3612–3628 (2012).
9. J. I. Gold, M. N. Shadlen, Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron* **36**, 299–308 (2002).
10. J. Drugowitsch, G. C. DeAngelis, E. M. Klier, D. E. Angelaki, A. Pouget, Optimal multisensory decision-making in a reaction-time task. e*Life* **3**, 1–19 (2014).
11. P. Dayan, S. Kakade, P. R. Montague, Learning and selective attention. *Nat. Neurosci.* **3** (suppl), 1218–1223 (2000).
12. P. Dayan, A. J. Yu, Uncertainty and learning. *IETE J. Res.* **49**, 171–181 (2003).
13. K. P. Körding, D. M. Wolpert, Bayesian integration in sensorimotor learning. *Nature* **427**, 244–247 (2004).
14. A. C. Courville, N. D. Daw, D. S. Touretzky, Bayesian theories of conditioning in a changing world. *Trends Cogn. Sci.* **10**, 294–300 (2006).
15. R. Ratcliff, A theory of order relations in perceptual matching. *Psychol. Rev.* **88**, 552–572 (1981).
16. P. Gomez, R. Ratcliff, M. Perea, The overlap model: A model of letter position coding. *Psychol. Rev.* **115**, 577–600 (2008).
17. R. Ratcliff, J. J. Starns, Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychol. Rev.* **120**, 697–719 (2013).
18. T. D. Hanks, M. E. Mazurek, R. Kiani, E. Hopp, M. N. Shadlen, Elapsed decision time affects the weighting of prior probability in a perceptual decision task. *J. Neurosci.* **31**, 6339–6352 (2011).
19. R. Moreno-Bote, Decision confidence and uncertainty in diffusion models with partially correlated neuronal integrators. *Neural Comput.* **22**, 1786–1811 (2010).
20. J. Palmer, A. C. Huk, M. N. Shadlen, The effect of stimulus strength on the speed and accuracy of a perceptual decision. *J. Vis.* **5**, 376–404 (2005).
21. J. O. Berger, *Statistical Decision Theory and Bayesian Analysis* (Springer, ed. 2, 1993).
22. A. Pouget, J. Drugowitsch, A. Kepecs, Confidence and certainty: Distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19**, 366–374 (2016).
23. B. B. Averbeck, P. E. Latham, A. Pouget, Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* **7**, 358–366 (2006).
24. R. Moreno-Bote *et al*., *Information-Limiting Correlations* (Nature Publishing Group, 2014).
25. R. Kiani, M. N. Shadlen, Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* **324**, 759–764 (2009).
26. C. Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006).
27. J. Drugowitsch, G. C. Deangelis, D. E. Angelaki, A. Pouget, Tuning the speed-accuracy trade-off to maximize reward rate in multisensory decision-making. *Elife* **4**, 1–11 (2015).
28. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (Wiley, ed. 2, 2006).
29. K. P. Murphy, *Machine Learning: A Probabilistic Perspective* (MIT Press, 2012).
30. T. Graepel, J. Quiñonero-Candela, T. Borchert, R. Herbrich, "Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine" in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, J. Fürnkranz, T. Joachims, Eds. (ACM Press, New York, 2010), pp. 13–20.
31. W. Chu, M. Zinkevich, L. Li, A. Thomas, B. Tseng, "Unbiased online active learning in data streams" in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '11* (ACM Press, New York, 2011), pp. 195–203.
32. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, ed. 2, 2018).
33. A. Zylberberg, D. M. Wolpert, M. N. Shadlen, Counterfactual reasoning underlies the learning of priors in decision making. *Neuron* **99**, 1083–1097.e6 (2018).
34. F. Meyniel, S. Dehaene, Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E3859–E3868 (2017).
35. W. J. Ma, J. M. Beck, P. E. Latham, A. Pouget, Bayesian inference with probabilistic population codes. *Nat. Neurosci.* **9**, 1432–1438 (2006).
36. J. M. Beck *et al*., Probabilistic population codes for Bayesian decision making. *Neuron* **60**, 1142–1152 (2008).
37. L. Busse *et al*., The detection of visual contrast in the behaving mouse. *J. Neurosci.* **31**, 11351–11361 (2011).
38. A. J. Yu, J. D. Cohen, Sequential effects: Superstition or rational behavior? *Adv. Neural Inf. Process. Syst.* **21**, 1873–1880 (2008).
39. A. G. Mendonça *et al*., The impact of learning on perceptual decisions and its implication for speed-accuracy tradeoffs. bioRxiv:10.1101/501858 (19 December 2018).
40. A. E. Urai, A. Braun, T. H. Donner, Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nat. Commun.* **8**, 14637 (2017).
41. V. Rao, G. C. DeAngelis, L. H. Snyder, Neural correlates of prior expectations of motion in the lateral intraparietal and middle temporal areas. *J. Neurosci.* **32**, 10063–10074 (2012).

PNAS
www.pnas.org

Supplementary Information for

Learning Optimal Decisions with Confidence

Jan Drugowitsch, André G. Mendonça, Zachary F. Mainen and Alexandre Pouget

Corresponding Author Name.
E-mail: jan_drugowitsch@hms.harvard.edu

This PDF file includes:

Supporting Information Text

We here provide a self-consistent and extended derivation of all the results provided in the main text. We will use standard font for scalars, lower-case bold symbols for vectors, and upper-case bold symbols for matrices. All vectors are column vectors, unless transposed. $x \sim \mathcal{N}\left(\mu, \sigma^2\right)$ denotes that x is a normal random variable with mean μ and variance σ^2.

Optimal decision making with high-dimensional momentary evidence

One-dimensional momentary evidence. Within each individual trial, we assume the latent state μ to be drawn from $\mu \sim \mathcal{N}\left(m, \sigma_\mu^2\right)$. $m = 0$ corresponds to the case of an unbiased prior for which $p(\mu \geq 0) = p(\mu < 0)$. In each small time step n of size δt from trial-onset at $t = 0$ (i.e., $n = 1$), the decision maker observes the momentary evidence $\delta z_n|\mu \sim \mathcal{N}(\mu\delta t, \delta t)$ that provides noisy information about the value of μ. The decision-maker's aim is to infer the sign of μ from the sequence $\delta z_1, \delta z_2, \dots$ of momentary evidence, to make choice $y = 1$ (for $\mu \geq 0$) or $y = -1$ (for $\mu < 0$).

For Bayes-optimal choices, we find the posterior $p(\mu \geq 0|\delta z_1, \delta z_2, \dots)$ in two steps. First, for N pieces (i.e., $t = N\delta t$ seconds) of accumulated evidence, the posterior μ is give by Bayes' rule,

$$p(\mu|\delta z_{1:N}) \propto p(\mu) \prod_{n=1}^{N} p(\delta z_n|\mu) \propto e^{-\frac{\mu^2}{2}\left(\frac{1}{\sigma_\mu^2}\right)+\mu\left(\frac{m}{\sigma_\mu^2}+z\right)} \propto \mathcal{N}\left(\mu|\frac{\sigma_\mu^{-2}m + z}{\sigma_\mu^{-2} + t}, \frac{1}{\sigma_\mu^{-2} + t}\right), \quad [1]$$

where all proportionalities are with respect to μ, and the second proportionality results from substituting the respective normal distributions, and defining $t = \sum_n \delta t$ and $z(t) = \sum_n \delta z_n$. This shows that the sufficient statistics of the posterior are $z(t)$ and t. For the second step we integrate this posterior over the non-negative half-line to find

$$p(\mu \geq 0|z, t) = \int_0^\infty p(\mu|z, t)\mathrm{d}\mu = \Phi\left(\frac{\sigma_\mu^{-2}m + z}{\sqrt{\sigma_\mu^{-2} + t}}\right), \quad [2]$$

where $\Phi(\cdot)$ is the normal cumulative function. This posterior is more certain (i.e., closer to zero or one) for larger $|\sigma_\mu^{-2}m + z|$ and smaller times t.

Using the correspondence between $\mu \geq 0$ (and $\mu < 0$) and $y = 1$ (and $y = -1$), the fact that $1 - \Phi(a) = \Phi(-a)$, and $p(\mu < 0|z, t) = 1 - p(\mu \geq 0|z, t)$, the more generic posterior over y is given by

$$p(y|z, t) = \Phi\left(y\frac{\sigma_\mu^{-2}m + z}{\sqrt{\sigma_\mu^{-2} + t}}\right). \quad [3]$$

This posterior captures both $y = 1$ and $y = -1$. If y is the made decision, then the expression is the belief that this decision was correct, and hence the *decision confidence* (1).

So far we have assumed a prior over μ with arbitrary mean m. With this prior, the a-priori belief that $y = 1$ is correct is given by $P^+ \equiv p(\mu \geq 0) = \Phi(m/\sigma_\mu)$. The prior is thus unbiased for $m = 0$, in which case $P^+ = 1/2$. In this case, the posterior Eq. [3] prefers $y = 1$ for all $z > 0$ and $y = -1$ for all $z < 0$. Therefore, we can bound evidence accumulation from above and below by the (potentially time-dependent) $\pm\theta(t)$ to make Bayes-optimal choices. In particular, once z reaches $\theta(t)$ (or $-\theta(t)$), it would trigger choice $y = 1$ (or $y = -1$). Observing that the unbounded accumulated evidence follows a Wiener process with drift μ, that is, $z(t)|\mu \sim \mathcal{N}(\mu t, t)$, supports the use of drift-diffusion models for Bayes-optimal decision making. Biased priors, which we discuss in a later section, require additional attention to achieve Bayes-optimal choices.

High-dimensional momentary evidence. To move to J-dimensional momentary evidence $\delta\boldsymbol{x}$ while preserving parallels to the one-dimensional case, we assume that there exist some (for now, known) combination weights $\boldsymbol{w}$ such that $\delta z_n = \boldsymbol{w}^T\delta\boldsymbol{x}_n$. We achieve this by the generative model,

$$\delta\boldsymbol{x}|\mu \sim \mathcal{N}\left((\boldsymbol{a}\mu + \boldsymbol{b})\,\delta t, \boldsymbol{\Sigma}\delta t\right), \quad [4]$$

for vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ that satisfy $\boldsymbol{a}^T\boldsymbol{w} = 1$ and $\boldsymbol{b}^T\boldsymbol{w} = 0$, and a covariance matrix $\boldsymbol{\Sigma}$ for which $\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} = 1$. With these properties it becomes easy to show that $\boldsymbol{w}^T\delta\boldsymbol{x}|\mu \sim \mathcal{N}(\mu\delta t, \delta t)$, as required. We will discuss our specific choices for $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{\Sigma}$ for the simulations shown in the main text further below.

Using the same steps as before, we find the posterior μ given N steps (i.e., $t = N\delta t$ seconds) of momentary evidence to be given by

$$p(\mu|\delta\boldsymbol{x}_{1:N}, \boldsymbol{w}) = \mathcal{N}\left(\frac{\sigma_\mu^{-2}m + \boldsymbol{w}^T\boldsymbol{x}}{\sigma_\mu^{-2} + t}, \frac{1}{\sigma_\mu^{-2} + t}\right), \quad [5]$$

where we have defined the accumulated evidence $\boldsymbol{x}(t) = \sum_n \delta\boldsymbol{x}_n$. The posterior over $\mu \geq 0$ is correspondingly given by

$$p(\mu \geq 0|\boldsymbol{x}, t, \boldsymbol{w}) = \Phi\left(\frac{\sigma_\mu^{-2}m + \boldsymbol{w}^T\boldsymbol{x}}{\sqrt{\sigma_\mu^{-2} + t}}\right). \quad [6]$$

Both expressions differ from the one-dimensional case by replacing z by $\boldsymbol{w}^T\boldsymbol{x}$. Expressed as a posterior over y, the above turns into

$$p\left(y|\boldsymbol{x},t,\boldsymbol{w}\right)=\Phi\left(y\frac{\sigma_\mu^{-2}m+\boldsymbol{w}^T\boldsymbol{x}}{\sqrt{\sigma_\mu^{-2}+t}}\right). \quad [7]$$

If y is the made decision, the above is again the decision confidence. Note that the unbounded accumulated evidence follows the multi-dimensional drifting Wiener process, $\boldsymbol{x}(t)|\mu\sim\mathcal{N}\left(\left(\boldsymbol{a}\mu+\boldsymbol{b}\right)t,\boldsymbol{\Sigma}t\right)$, whose $\boldsymbol{w}$-weighted linear combination reduces to the same one-dimensional process $\boldsymbol{z}(t)=\boldsymbol{w}^T\boldsymbol{x}(t)\sim\mathcal{N}\left(\mu t,t\right)$ as before.

Assuming again an unbiased prior, $m=0$, Bayes-optimal decisions are by the same logic as for the one-dimensional case cast by the boundaries $\pm\theta(t)$ on $\boldsymbol{w}^T\boldsymbol{x}$. Here, the positive (negative) boundary correspond to choice $y=1$ ($y=-1$). We will discuss Bayes-optimal choices for biased priors in a later section.

If difficulty $|\mu|$ varies across trials, the decision confidence at a constant decision boundary drops over time. As the previous sections have shown, the decision confidence is the same for one- and high-dimensional momentary evidence as long as the decision boundary is on $z(t)$ and $\boldsymbol{w}^T\boldsymbol{x}(t)$, respectively. Furthermore, for time-invariant decision boundaries, $\pm\theta(t)=\pm\theta$, this decision confidence drops as a function of time. Here we show that this drop is a general property of symmetric priors over μ for which the difficulty $|\mu|$ can vary across trials, that extends beyond the Gaussian $p\left(\mu\right)$ we assume in other parts of this supplement. To show this, let us redefine $p\left(\mu\right)$ — in this section only — to be (as in (2)) given by

$$p(\mu)=\sum_{i=1}^{L}\frac{p_i}{2}\left(\delta(\mu-\mu_i)+\delta(\mu+\mu_i)\right), \quad [8]$$

which features L point masses at $\pm\mu_1,\pm\mu_2,\ldots,\pm\mu_L$, each weighted by $p_i/2$, and where we have assumed positive p_i that satisfy $\sum_i p_i=1$. Furthermore, without loss of generality, we assume the μ_i's to be positive, ordered, and unique, that is $0<\mu_1<\mu_2<\cdots<\mu_L$. Here, we disallow $\mu_1=0$ for notational convenience, but our argument can be easily extended to include this possibility. Assuming the same one-dimensional momentary evidence as before, $\delta z_n|\mu\sim\mathcal{N}\left(\mu\delta t,\delta t\right)$, it follows from Bayes' rule that

$$p\left(\mu=\mu_i|z,t\right)=\frac{p_i e^{-\frac{t}{2}\mu_i^2+z\mu_i}}{\sum_j p_j e^{-\frac{t}{2}\mu_j^2}\left(e^{z\mu_j}+e^{-z\mu_j}\right)}. \quad [9]$$

Therefore, the belief that $\mu\geq 0$ ($y=1$) at the upper boundary $z=\theta$ is given by

$$p\left(y=1|z=\theta,t\right)=\sum_i p\left(\mu=\mu_i|z=\theta,t\right)=\frac{\sum_i p_i e^{-\frac{t}{2}\mu_i^2+\theta\mu_i}}{\sum_j p_j e^{-\frac{t}{2}\mu_j^2}\left(e^{\theta\mu_j}+e^{-\theta\mu_j}\right)}. \quad [10]$$

For our symmetric prior, this belief at the upper boundary equals the decision confidence at both boundaries. Therefore, we will use it as a proxy for decision confidence.

In what follows, we will show that this belief is a mixture of two components. The first is the belief that $\mu=\mu_i$ given some fixed difficulty, $\mu\in\{-\mu_i,\mu_i\}$, and the second is the probability that this is indeed the current difficulty. The first part turns out to be independent of time, whereas the second changes. In particular, as we will show, the probability that the difficulty is high (i.e., that $|\mu|$ is small) increases over time, resulting in a re-weighting of the per-difficulty beliefs. This re-weighting causes the overall belief to drop, as we argue in the main text.

Mathematically, this mixture can be written as

$$p\left(y=1|z=\theta,t\right)=\sum_i p\left(\mu=\mu_i|z=\theta,t\right)=\sum_i p\left(\mu=\mu_i|z=\theta,t,\mu=\pm\mu_i\right)p\left(\mu=\pm\mu_i|z=\theta,t\right). \quad [11]$$

In the right-most sum, the first probability is the per-difficulty belief g_i for assumed difficulty μ_i, and the second is the probability that μ_i is indeed the current difficulty. Both follow from [9], and are given by

$$g_i\equiv p\left(\mu=\mu_i|z=\theta,t,\mu=\pm\mu_i\right)=\frac{e^{\theta\mu_i}}{e^{\theta\mu_i}+e^{-\theta\mu_i}}=\frac{1}{1+e^{-2\theta\mu_i}}, \quad [12]$$

$$p\left(\mu=\pm\mu_i|z=\theta,t\right)=\frac{p_i e^{-\frac{t}{2}\mu_i^2}\left(e^{\theta\mu_i}+e^{-\theta\mu_i}\right)}{\sum_j p_j e^{-\frac{t}{2}\mu_j^2}\left(e^{\theta\mu_j}+e^{-\theta\mu_j}\right)}=\frac{w_i(t)}{\sum_j w_j(t)}, \quad [13]$$

where we have defined $w_i(t)=p_i e^{-\frac{t}{2}\mu_i^2}\left(e^{\theta\mu_i}+e^{-\theta\mu_i}\right)$ as the unnormalized, time-dependent, per-difficulty weights. This allows us to write the overall belief as the weighted mixture

$$p\left(y=1|z=\theta,t\right)=\sum_i\frac{w_i(t)}{\sum_j w_j(t)}g_i, \quad [14]$$

which is a weighted mixture of per-difficulty beliefs, g_i, in which only the mixture weights are time-dependent. Note that the per-difficulty beliefs are strictly increasing in μ_i, such that are also ordered, that is, $g_1<g_2<\cdots<g_L$. Furthermore,

increasing time by $\delta t > 0$ results in a drop in unnormalized weights, $w_i(t+\delta t) = a_i(\delta t) w_t(t)$ with $a_i(\delta t) = e^{-\frac{\delta t}{2}\mu_i^2} \in [0,1]$. This drop is larger for larger μ_i, that is $a_1(\delta t) > a_2(\delta t) > \cdots > a_L(\delta t)$. Therefore, increasing time results in putting proportionally more weight on per-difficulty beliefs associated with lower μ_i's with lower associated g_i, such that the overall belief, and equally the decision confidence, drops.

Once we stop varying the difficulty (i.e., $L = 1$), the belief reduces to the per-difficulty belief g_1, which does not drop over time. Therefore, the only condition required for the decision confidence to drop at a time-invariant boundary is for the difficulty $|\mu|$ to vary across trials.

Learning the input combination weight w from choice feedback

So far we have assumed $\boldsymbol{w}$ to be known. Here we derive learning rules for $\boldsymbol{w}$ based on feedback on the correctness of a choice. Specifically, we assume that the decision maker accumulated evidence $\boldsymbol{x}$ for some time t and (potentially, but not necessarily) made decision y, after which feedback about the correct choice y^* is provided. Before evidence accumulation we assume the decision maker to hold belief $p(\boldsymbol{w})$ about the input combination weights $\boldsymbol{w}$. Our aim is to find the posterior $p(\boldsymbol{w}|\boldsymbol{x}, t, y^*)$ given all the available evidence. We focus here on a feedback after a single choice. The same principles apply to choice sequences, by turning the posterior after a choice into the prior for the subsequent choice.

The desired posterior can be found by Bayes' rule

$$p\left(\boldsymbol{w}|\boldsymbol{x}, t, y^*\right) \propto p\left(y^*|\boldsymbol{x}, t, \boldsymbol{w}\right) p(\boldsymbol{w}), \tag{15}$$

where the likelihood $p\left(y^*|\boldsymbol{x}, t, \boldsymbol{w}\right)$ is conditional on all observed quantities, $\boldsymbol{x}$ and t, and some hypothetical weights $\boldsymbol{w}$, and specifies the probability that y^* is the correct choice given these weights. This likelihood turns out to correspond to the previously derived decision-making posterior, Eq. [7], which is a normal cumulative function with argument linear in $\boldsymbol{w}$. In general, problems with such a likelihood function structure are known as *Probit regression*. Such problems don't yield solutions for which the posterior has the same functional form as the prior — which is a desirable property to support efficient Bayesian input weight learning across longer sequences of choices, and to gain insight into the learning rule. Therefore, we derive below different approximations to such Bayes-optimal learning.

All of the below assumes an unbiased prior over μ by fixing m to $m = 0$. We can extend the below rules to also learn the prior bias m by extending the accumulated evidence vector $\boldsymbol{x}$ by one element fixed to σ_μ^{-2}, and the weight vector $\boldsymbol{w}$ by one element containing m. Learning this extended weight vector then correspond to simultaneously learning the input weight and the prior bias.

The marginal decision confidence. Before deriving approximate weight learning rules, let us consider the consequences of uncertain weights on the decision confidence $p(y|\boldsymbol{x}, t)$ with these weights marginalized out. To do so, we assume our prior weight belief to be normal, $\boldsymbol{w} \sim \mathcal{N}\left(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w\right)$ with mean $\boldsymbol{\mu}_w$ and covariance $\boldsymbol{\Sigma}_w$. Then, we find this marginal decision confidence by first finding the marginal posterior over μ, which is given by

$$p(\mu|\boldsymbol{x}, t) = \int p(\mu|\boldsymbol{x}, t, \boldsymbol{w}) p(\boldsymbol{w}) \mathrm{d}\boldsymbol{w} = \mathcal{N}\left(\frac{\boldsymbol{\mu}_w^T \boldsymbol{x}}{\sigma_\mu^{-2} + t}, \frac{1}{\sigma_\mu^{-2} + t} + \frac{\boldsymbol{x}^T \boldsymbol{\Sigma}_w \boldsymbol{x}}{\left(\sigma_\mu^{-2} + t\right)^2}\right), \tag{16}$$

were we have used Eq. [5] with $m = 0$. We find the marginal decision confidence $p(y|\boldsymbol{x}, t)$ by integrating the above over the non-negative halfline, which results after some simplification in

$$p(y|\boldsymbol{x}, t) = \Phi\left(y \frac{\boldsymbol{\mu}_w^T \boldsymbol{x}}{\sqrt{\sigma_\mu^{-2} + t + \boldsymbol{x}^T \boldsymbol{\Sigma}_w \boldsymbol{x}}}\right) = \Phi\left(y \frac{\boldsymbol{\mu}_w^T \tilde{\boldsymbol{x}}}{\sqrt{1 + \tilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}}}\right), \tag{17}$$

where we have defined $\tilde{\boldsymbol{x}} \equiv \boldsymbol{x} / \sqrt{\sigma_\mu^{-2} + t}$ for the second equality. Comparing this expression to Eq. [7] reveals the additional term $\boldsymbol{x}^T \boldsymbol{\Sigma}_w \boldsymbol{x}$ that lowers the overall posterior confidence (i.e., moving it towards 1/2) due to uncertainty in $\boldsymbol{w}$. If y is the made choice, the above is the decision confidence that takes into account weight uncertainty.

Weight learning by Assumed Density Filtering. Assumed Density Filtering (ADF; (3–6)) approximates the posterior by assuming a particular functional form of the approximate posterior $q(\boldsymbol{w}|\boldsymbol{x}, t, y^*)$ and finding the parameters of this functional form by minimizing the Kullback-Leiber divergence $\mathrm{KL}\left(p\left(\boldsymbol{w}|\boldsymbol{x}, t, y^*\right) \| q\left(\boldsymbol{w}|\boldsymbol{x}, t, y^*\right)\right)$ between the true posterior and its approximation. To minimize this divergence we again assume a normally distributed prior $\boldsymbol{w} \sim \mathcal{N}\left(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w\right)$ with mean $\boldsymbol{\mu}_w$ and covariance $\boldsymbol{\Sigma}_w$. To support sequential choices, we assume the approximate posterior to also be normal, $q\left(\boldsymbol{w}|\boldsymbol{x}, t, y^*\right) = \mathcal{N}\left(\boldsymbol{w}|\boldsymbol{\mu}_w^* \boldsymbol{\Sigma}_w^*\right)$, with updated moments $\boldsymbol{\mu}_w^*$ and $\boldsymbol{\Sigma}_w^*$.

To find these updated moments, we use the fact that the KL-divergence is in our case minimized by matching the moments of the Gaussian sufficient statistics $\boldsymbol{w}$ and $\boldsymbol{w}\boldsymbol{w}^T$ (7). For the source distribution, $p\left(\boldsymbol{w}|\boldsymbol{x}, t, y^*\right)$, these moments can be found by the gradients of the log-normalizing constant of this source distribution, $\nabla \log p(y^*|x, t)$ (7, 8), where we will use the already derived marginal likelihood $p(y|\boldsymbol{x}, t)$ in Eq. [17]. Using these principles, the updated moments of the approximate posterior can be found by

$$\boldsymbol{\mu}_w^* = \boldsymbol{\mu}_w + \boldsymbol{\Sigma}_w \nabla_{\mu_w} \log p(y^*|\boldsymbol{x}, t), \tag{18}$$

$$\boldsymbol{\Sigma}_w^* = \boldsymbol{\Sigma}_w - \boldsymbol{\Sigma}_w \left(\nabla_{\mu_w} \log p(y^*|\boldsymbol{x}, t) \left(\nabla_{\mu_w} \log p(y^*|\boldsymbol{x}, t)\right)^T - 2 \nabla_{\Sigma_w} \log p(y^*|\boldsymbol{x}, t)\right) \boldsymbol{\Sigma}_w. \tag{19}$$

The required gradients are given by

$$\nabla_{\mu_w} \log p(y^*|\boldsymbol{x}, t) = \xi_w(\gamma) y^* \frac{\tilde{\boldsymbol{x}}}{\sqrt{1 + \tilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}}}, \tag{20}$$

$$\nabla_{\Sigma_w} \log p(y^*|\boldsymbol{x}, t) = -\xi_w(\gamma) y^* \frac{\boldsymbol{\mu}_w^T \tilde{\boldsymbol{x}}}{2\left(1 + \tilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}\right)^{3/2}} \tilde{\boldsymbol{x}} \tilde{\boldsymbol{x}}^T, \tag{21}$$

$$\tag{22}$$

where $\xi_w(\gamma)$ is given by

$$\xi_w(\gamma) = \frac{\partial}{\partial \gamma} \log \Phi(\gamma) = \frac{\mathcal{N}(\gamma|0,1)}{\Phi(\gamma)}, \tag{23}$$

and we have defined γ as

$$\gamma \equiv y^* \frac{\boldsymbol{\mu}_w^T \tilde{\boldsymbol{x}}}{\sqrt{1 + \tilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}}}. \tag{24}$$

Overall, this leads to the moments update equations,

$$\boldsymbol{\mu}_w^* = \boldsymbol{\mu}_w + y^* \frac{\xi_w(\gamma)}{\sqrt{1 + \tilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}}} \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}, \tag{25}$$

$$\boldsymbol{\Sigma}_w^* = \boldsymbol{\Sigma}_w + \xi_{cov}(\gamma) \left(\left(\boldsymbol{\Sigma}_w^{-1} + \tilde{\boldsymbol{x}} \tilde{\boldsymbol{x}}^T \right)^{-1} - \boldsymbol{\Sigma}_w \right), \tag{26}$$

where the covariance learning rate is given by

$$\xi_{cov}(\gamma) = \xi_w(\gamma)^2 + \xi_w(\gamma)\gamma. \tag{27}$$

As illustrated in Fig. S1, both mean and covariance updates are modulated by the marginal decision confidence in the feedback, y^*, given by Eq. [17]. To see how ξ_w and ξ_{cov} are a function of the marginal decision confidence about the actual choice y (rather than the feedback y^*) let us first focus on correct choices. For correct choices, $y = y^*$, such that the marginal decision confidence about y equals that of y^*, that is, $p(y|\boldsymbol{x}, t) = p(y^*|\boldsymbol{x}, t)$. Furthermore, by the definition of $p(y^*|\boldsymbol{x}, t)$, it can be written as $p(y^*|\boldsymbol{x}, t) = \Phi(\gamma)$ (where γ is defined in Eq. [24]). This function is strictly increasing in γ, such that small/large γ's corresponds to low/high confidence. Therefore, as $\xi_w(\gamma)$ and $\xi_{cov}(\gamma)$ are functions of only γ, they are in turn functions of the marginal decision confidence $p(y|\boldsymbol{x}, t)$.

For incorrect choices we have $y \neq y^*$, such that $p(y|\boldsymbol{x}, t) = 1 - p(y^*|\boldsymbol{x}, t) = 1 - \Phi(\gamma)$, which is strictly decreasing in γ. Therefore, we can again assign a unique decision confidence $p(y|\boldsymbol{x}, t)$ to each γ, such that $\xi_w(\gamma)$ and $\xi_{cov}(\gamma)$ are again functions of the decision confidence about the made decision y.

Assumed Density Filtering with a diagonal covariance matrix. The above update equations require tracking of the full covariance matrix, making these updates scale badly with the size of the input space, J, and require non-local interactions. To find alternative, local update equations, we here assume that both the prior covariance, as well as the approximate posterior covariance are given by diagonal matrices, given by $\boldsymbol{\Sigma}_w = \text{diag}\left(\sigma_{w,1}^2, \ldots, \sigma_{w,k}^2\right)$ and $\boldsymbol{\Sigma}_w^* = \text{diag}\left(\sigma_{w,1}^{2*}, \ldots, \sigma_{w,k}^{2*}\right)$. Following the same derivation as before, this leads to the update equations

$$\mu_{w,i}^* = \mu_{w,i} + y^* \frac{\xi_w(\gamma)}{\sqrt{1 + \sum_j \sigma_{w,j}^2 \tilde{x}_j^2}} \sigma_{w,i}^2 \tilde{x}_i, \tag{28}$$

$$\sigma_{w,i}^{2*} = \sigma_{w,i}^2 - \xi_{cov}(\gamma) \frac{\sigma_{w,i}^4 \tilde{x}_i}{\sqrt{1 + \sum_j \sigma_{w,j}^2 \tilde{x}_j^2}} \tag{29}$$

where $\mu_{w,i}$ and $\mu_{w,i}^*$ are the ith element of $\boldsymbol{\mu}_w$ and $\boldsymbol{\mu}_w^*$, respectively, and γ is given by

$$\gamma = y^* \frac{\boldsymbol{\mu}_w^T \tilde{\boldsymbol{x}}}{\sqrt{1 + \sum_j \sigma_{w,j}^2 \tilde{x}_j^2}}. \tag{30}$$

Thus, other than a global divisive normalization and the marginal decision confidence-related term γ, all updates are local.

Approximating the weight posterior by a second-order Taylor series. A simpler alternative to ADF that also yields a normally distributed approximate posterior is to approximate the true log-posterior, $\log p(\boldsymbol{w}|\boldsymbol{x}, t, y^*)$ by a second-order Taylor series in $\boldsymbol{w}$ around $\boldsymbol{w} = \boldsymbol{\mu}_w$. The relevant terms in this log-posterior are

$$\log p(\boldsymbol{w}|\boldsymbol{x}, t, y^*) = \Phi\left(y^* \boldsymbol{w}^T \tilde{\boldsymbol{x}}\right) - \frac{1}{2}\boldsymbol{w}^T \boldsymbol{\Sigma}_w^{-1} \boldsymbol{w} + \boldsymbol{w}^T \boldsymbol{\Sigma}_w^{-1} \boldsymbol{\mu}_w + \text{const.} \quad [31]$$

The required gradient and Hessian are

$$\nabla_w \log p(\boldsymbol{w}|\boldsymbol{x}, t, y^*)\big|_{w=\mu_w} = \xi_w(\gamma) y^* \tilde{\boldsymbol{x}}, \quad [32]$$

$$\nabla\nabla_w \log p(\boldsymbol{w}|\boldsymbol{x}, t, y^*)\big|_{w=\mu_w} = -\boldsymbol{\Sigma}_w^{-1} - \xi_{cov}(\gamma)\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T, \quad [33]$$

where $\xi_w(\cdot)$ and $\xi_{cov}(\cdot)$ are defined as for ADF, but γ changes to $\gamma = y^* \boldsymbol{\mu}_w^T \tilde{\boldsymbol{x}}$. Using the above to find the second-order Taylor series and reading off the resulting posterior moments yields the moment updates

$$\boldsymbol{\mu}_w^* = \boldsymbol{\mu}_w + y^* \xi_w(\gamma) \boldsymbol{\Sigma}_w^* \tilde{\boldsymbol{x}}, \quad [34]$$

$$\boldsymbol{\Sigma}_w^* = \left(\boldsymbol{\Sigma}_w^{-1} + \xi_{cov}(\gamma)\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T\right)^{-1}. \quad [35]$$

These have a similar form as for ADF, Eqs. [25] and [26], with the main difference that they are missing the normalization by $\sqrt{1 + \tilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}}$. Given that this normalization modulates the moment update strength by the weight uncertainty, this implies that the update equations based on the second-order Taylor series will be less influenced by this uncertainty.

Assumed density filtering with noisy feedback. So far we have assumed the feedback y^* to always be correct. We will now consider how ADF changes when the feedback itself is noisy. In particular, we assume that feedback is inverted with probability β, such that the weight likelihood given feedback y^* becomes

$$p(y^*|\boldsymbol{x}, t, \boldsymbol{w}) = \beta\Phi\left(-y^* \boldsymbol{w}^T \tilde{\boldsymbol{x}}\right) + (1-\beta)\Phi\left(y^* \boldsymbol{w}^T \tilde{\boldsymbol{x}}\right) = \beta + (1-2\beta)\Phi\left(y^* \boldsymbol{w}^T \tilde{\boldsymbol{x}}\right). \quad [36]$$

In this case, the marginal decision confidence about feedback y^* becomes

$$p(y^*|\boldsymbol{x}, t, \beta) = \beta + (1-2\beta)\Phi\left(y^* \frac{\boldsymbol{\mu}_w^T \tilde{\boldsymbol{x}}}{\sqrt{1 + \tilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}}}\right). \quad [37]$$

The gradients of the log marginal decision confidence thus become

$$\nabla_{\mu_w} \log p(y^*|\boldsymbol{x}, t, \beta) = \xi_{\beta,w}(\gamma) y^* \frac{\tilde{\boldsymbol{x}}}{\sqrt{1 + \tilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}}}, \quad [38]$$

$$\nabla_{\Sigma_w} \log p(y^*|\boldsymbol{x}, t, \beta) = -\xi_{\beta,w}(\gamma) y^* \frac{\boldsymbol{\mu}_w^T \tilde{\boldsymbol{x}}}{2\left(1 + \tilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}\right)^{3/2}} \tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T \quad [39]$$

with $\xi_{\beta,w}(\gamma)$ given by

$$\xi_{\beta,w}(\gamma) = \frac{\partial}{\partial\gamma} \log\left(\beta + (1-2\beta)\Phi(\gamma)\right) = \frac{(1-2\beta)\mathcal{N}(\gamma|0,1)}{\beta + (1-2\beta)\Phi(\gamma)} \quad [40]$$

and where γ is, as for vanilla ADF, given by Eq. [24]. Using again Eqs. [18] and [19] results in the update equations

$$\boldsymbol{\mu}_w^* = \boldsymbol{\mu}_w + y^* \frac{\xi_{\beta,w}(\gamma)}{\sqrt{1 + \tilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}}} \boldsymbol{\Sigma}_w \tilde{\boldsymbol{x}}, \quad [41]$$

$$\boldsymbol{\Sigma}_w^* = \boldsymbol{\Sigma}_w + \xi_{\beta,cov}(\gamma)\left(\left(\boldsymbol{\Sigma}_w^{-1} + \tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T\right)^{-1} - \boldsymbol{\Sigma}_w\right), \quad [42]$$

with covariance learning rate

$$\xi_{\beta,cov}(\gamma) = \xi_{\beta,w}(\gamma)^2 + \xi_{\beta,w}(\gamma)\gamma. \quad [43]$$

This illustrates that the only impact of noisy feedback is on the update strength modulators, $\xi_{\beta,w}(\cdot)$ and $\xi_{\beta,cov}$. As shown in Fig. S1, these modulators become smaller for larger feedback noise. For high-confidence choices that the feedback flags as incorrect, $\xi_{\beta,cov}$ even becomes negative, indicating that uncertainty in $\boldsymbol{w}$ increases. This increase arises from approximate inference, as additional information in strictly Bayes-optimal inference should not increase uncertainty, even if this information is (knowingly) noisy.

Alternative learning heuristics. Let us now discuss alternative heuristics that do not track a belief over $\boldsymbol{w}$, but instead update a point estimate. The first alternative is the delta rule that performs stochastic gradient descent on the sum of squared distances between the chosen decision boundary and the correct decision boundary. For the current choice, this squared distance is $\left(\boldsymbol{w}^T\boldsymbol{x}(t) - y^*\theta(t)\right)^2$, where $\boldsymbol{w}^T\boldsymbol{x}(t) \in \{-\theta(t), \theta(t)\}$ at decision time t equals the chosen bound, and $y^*\theta(t) \in \{-\theta(t), \theta(t)\}$ is the boundary that would have led to the correct choice. Thus, the delta rule update is given by

$$\boldsymbol{w}^* = \boldsymbol{w} + \frac{\alpha}{2\theta(0)}\left(y^*\theta(t) - \boldsymbol{w}^T\boldsymbol{x}\right)\boldsymbol{x}, \tag{44}$$

where we have chosen to normalize the learning α by $2\theta(0)$ to make the update magnitude less dependent of the bound height. The residual in the above is either zero or $\pm 2\theta(t)$, such that the learning rule only makes adjustments to the weight estimate in case of incorrect choices.

The delta rule aims to minimize the probability that incorrect choices are made. In diffusion models this can be achieved by accumulating more evidence before reaching the decision boundary. This, in turn, can be accomplished by reducing the overall magnitude of $\boldsymbol{w}$. In particular for small learning rates, this is exactly what the delta rule does, leading to progressively smaller $\|\boldsymbol{w}\|$, and weight learning that does not converge in expectation. To work around this degeneracy, we introduced the normalized delta rule. This rule performs the update exactly like the standard delta rule, but subsequently adjusts the weight magnitude to match that of the true weights. It therefore needs access to the true weight's magnitude in each trial, making it a rule that has access to an oracle that other rules don't. Thus, it uses strictly more information than other rules, which needs to be kept in mind when comparing its performance to that of other rules.

As a last heuristic we considered performing stochastic gradient ascent on the log-likelihood of the feedback, $\log p\left(y^*|\boldsymbol{x}, \boldsymbol{w}, t\right) = \log \Phi\left(y^*\boldsymbol{w}^T\tilde{\boldsymbol{x}}\right)$. Taking the gradient of this log-likelihood results in the learning rule

$$\boldsymbol{w}^* = \boldsymbol{w} - \alpha y^* \xi_w\left(y^*\boldsymbol{w}^T\tilde{\boldsymbol{x}}\right)\tilde{\boldsymbol{x}}, \tag{45}$$

where $\xi_w(\cdot)$ is defined as for ADF. Due to the inclusion of $\xi_w(\cdot)$, this rule modulates the update strength by decision confidence, unlike the normalized delta rule above. It differs from probabilistic learning rules in that it uses a fixed learning rate α, instead of a learning rate modulation by a current estimate of the certainty about $\boldsymbol{w}$.

Tracking non-stationary combination weights

So far we have assumed the true weights, underlying the generation of the momentary evidences, $\delta\boldsymbol{x}$, to be stationary, allowing us to use a sequence $\boldsymbol{x}$'s, t's, and y^*'s to learn successively better posteriors over $\boldsymbol{w}$. In the ideal case (i.e., if we wouldn't use approximate inference), this would — after enough observations — lead to a very good approximation of the true $\boldsymbol{w}$. We now change this setup to assume that the true weights change slightly across successive trials, and the learner's task is to track these changes as well as possible. This implies that, as the weights are now a moving target, they can never be learned perfectly.

We model the non-stationary of the weights by a first-order autoregressive process. That is, we assume that the true weights $\boldsymbol{w}_n$ in trial n depend on the true weights $\boldsymbol{w}_{n-1}$ in trial n by

$$\boldsymbol{w}_n|\boldsymbol{w}_{n-1} \sim \mathcal{N}\left(\boldsymbol{A}\boldsymbol{w}_{n-1} + \boldsymbol{b}, \boldsymbol{\Sigma}_d\right), \tag{46}$$

where $\boldsymbol{A}$, $\boldsymbol{b}$ and $\boldsymbol{\Sigma}_d$ are parameters of the process.

Let us now consider a probabilistic learner that maintains belief $\boldsymbol{w}_n \sim \mathcal{N}\left(\boldsymbol{\mu}_{w,n}, \boldsymbol{\Sigma}_{w,n}\right)$ before observing $\boldsymbol{x}$, t, and y^* in the nth trial. Despite the successive weight change across trials, the learner would first follow its standard learning rule (discussed above different approximations) to compute posterior parameters $\boldsymbol{\mu}^*_{w,n}$ and $\boldsymbol{\Sigma}^*_{w,n}$. This is followed by taking account of the weight change by updating its parameters according to

$$\boldsymbol{\mu}_{w,n+1} = \boldsymbol{A}\boldsymbol{\mu}^*_{w,n} + \boldsymbol{b}, \qquad \boldsymbol{\Sigma}_{w,n+1} = \boldsymbol{A}\boldsymbol{\Sigma}^*_{w,n}\boldsymbol{A}^T + \boldsymbol{\Sigma}_d. \tag{47}$$

These weights then act as a starting point, i.e., prior, for learning in the next trial. No other changes to the learning rules are required to take the non-stationarity of the combination weights into account.

Sampling the Bayes-optimal posterior

Finding a tractable closed-form expression for the Bayes-optimal posterior over $\boldsymbol{w}$ is unfortunately impossible. However, we can approximate this posterior to almost arbitrary precision by drawing samples from this posterior. We will first discuss such sampling for stationary combination weights, in which case we can use Gibbs sampling.

Gibbs sampling for stationary weights. For Gibbs sampling, we assume prior $\boldsymbol{w} \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)$, and observations $\boldsymbol{x}_n$, t_n, and y^*_n in the nth trial. The aim is to, after N trials, draw samples from $p\left(\boldsymbol{w}|\boldsymbol{x}_{1:N}, t_{1:N}, y^*_{1:N}\right)$. With the per-trial likelihood $p\left(y^*_n|\boldsymbol{x}_n, t_n, \boldsymbol{w}\right) = \Phi\left(y^*_n\boldsymbol{w}^T\tilde{\boldsymbol{x}}_n\right)$, this posterior is given by

$$p\left(\boldsymbol{w}|\boldsymbol{x}_{1:N}, t_{1:N}, y^*_{1:N}\right) \propto \mathcal{N}\left(\boldsymbol{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)\prod_{n=1}^{N}\Phi\left(y^*_n\boldsymbol{w}^T\tilde{\boldsymbol{x}}_n\right). \tag{48}$$

The covariance of this posterior is given by

$$\boldsymbol{\Sigma}_w = \left(\boldsymbol{\Sigma}_0^{-1} + \sum_{n=1}^{N} \tilde{\boldsymbol{x}}_n \tilde{\boldsymbol{x}}_n^T \right)^{-1} \quad [49]$$

which can be efficiently updated with each successive trial by the Sherman-Morrison update. To sample from the posterior $\boldsymbol{w}$, we introduce the auxiliary variables $a_n \sim \mathcal{N}\left(y_n^* \boldsymbol{w}^T \tilde{\boldsymbol{x}}_n, 1\right)$ for each n, such that $a_n \geq 0$ for a choice consistent with y_n^*. Thus, for a fixed $\boldsymbol{w}$, we can draw a_n according to

$$a_n | \tilde{\boldsymbol{x}}_n, \boldsymbol{w}, y_n^* \sim \mathcal{N}_{\geq 0}\left(y_n^* \boldsymbol{w}^T \tilde{\boldsymbol{x}}_n, 1\right), \quad [50]$$

where $\mathcal{N}_{\geq 0}$ denotes a draw from a truncated normal distribution, guaranteeing $a_n \geq 0$. With these samples, the posterior $\boldsymbol{w}$ is given by

$$\boldsymbol{w} | \boldsymbol{x}_{1:N}, t_{1:N}, y_{1:N}^*, a_{1:N} \sim \mathcal{N}\left(\boldsymbol{\Sigma}_w \left(\Sigma_0^{-1} \boldsymbol{\mu}_0 + \sum_{n=1}^{N} y_n^* \tilde{\boldsymbol{x}}_n a_n \right), \boldsymbol{\Sigma}_w \right). \quad [51]$$

Overall, Gibbs sampling consists in alternating between sampling $a_{1:N}$ and $\boldsymbol{w}$ until a sufficient number of $\boldsymbol{w}$-samples are drawn.

Particle filtering for non-stationary weights. Once the weights become non-stationary, particle filtering turns out to be a more efficient approach to posterior sampling. The aim is to approximate the sequential weight update

$$p\left(\boldsymbol{w}_n | \boldsymbol{x}_{1:n}, \boldsymbol{t}_{1:n}, y_{1:n}^*\right) \propto p\left(y_n^* | \boldsymbol{x}_n, t_n, \boldsymbol{w}_n\right) \int p\left(\boldsymbol{w}_n | \boldsymbol{w}_{n-1}\right) p\left(\boldsymbol{w}_{n-1} | \boldsymbol{x}_{1:n-1}, t_{1:n-1}, y_{1:n-1}^*\right) \mathrm{d}\boldsymbol{w}_{n-1}, \quad [52]$$

by using the particle approximation

$$p\left(\boldsymbol{w}_n | \boldsymbol{x}_{1:n}, \boldsymbol{t}_{1:n}, y_{1:n}^*\right) \approx \frac{1}{K} \sum_{k=1}^{K} \delta_{w_n^{(k)}}, \quad [53]$$

consisting of the K particles $\left\{ \boldsymbol{w}_n^{(1)}, \ldots, \boldsymbol{w}_n^{(K)} \right\}$. With this approximation, the above sequential update becomes

$$p\left(\boldsymbol{w}_n | \boldsymbol{x}_{1:n}, t_{1:n}, t_{1:n}^*\right) \overset{\propto}{\sim} \sum_k p\left(y_n^* | \boldsymbol{x}_n, t_n, \boldsymbol{w}_n\right) p\left(\boldsymbol{w}_n | \boldsymbol{w}_{n-1}^{(k)}\right). \quad [54]$$

We can sample from this posterior by an importance sampling re-sampling scheme in three steps. First, we draw K samples $\tilde{\boldsymbol{w}}_n^{(k)}$ from a Gaussian proposal density

$$\tilde{\boldsymbol{w}}_n^{(k)} \sim \mathcal{N}\left(\boldsymbol{\mu}_w\left(\boldsymbol{w}_{n-1}^{(k)}\right), \boldsymbol{\Sigma}_w\left(\boldsymbol{w}_{n-1}^{(k)}\right) \right). \quad [55]$$

Second, we compute the importance sampling weights,

$$\lambda_n^{(k)} = \frac{p\left(y_n^* | \boldsymbol{x}_n, t_n, \tilde{\boldsymbol{w}}_n^{(k)}\right) p\left(\tilde{\boldsymbol{w}}_n^{(k)} | \boldsymbol{w}_{n-1}^{(k)}\right)}{\mathcal{N}\left(\boldsymbol{\mu}_w\left(\boldsymbol{w}_{n-1}^{(k)}\right), \boldsymbol{\Sigma}_w\left(\boldsymbol{w}_{n-1}^{(k)}\right) \right)}. \quad [56]$$

Third, we re-sample the $\boldsymbol{w}_n^{(k)}$'s from the $\tilde{\boldsymbol{w}}_n^{(k)}$'s with probabilities proportional to their respective weights, $\lambda_n^{(k)}$. To ensure efficiency of the procedure, the proposal density for each weight should be close to $p\left(y_n^* | \boldsymbol{x}_n, t_n, \boldsymbol{w}_n\right) p\left(\boldsymbol{w}_n | \boldsymbol{w}_{n-1}\right)$, appropriately normalized, which we achieve by computing the proposal moments $\boldsymbol{\mu}_w\left(\boldsymbol{w}_{n-1}^{(k)}\right)$ and $\boldsymbol{\Sigma}_w\left(\boldsymbol{w}_{n-1}^{(k)}\right)$ according to the ADF variant that assumes non-stationary combination weights.

Relating learning through inference to learning through optimization

In all of the above we have treated learning as an inference problem, where we want to find the posterior weights given all of the observed evidence. Here, we address the parallels between inference and optimization in two ways. First, we will describe more general decision theoretical principles that highlight these parallels. Second, we will show explicitly how our learning problem can be formulated as an optimization problem.

Decision theoretic perspective. In decision theory, the Bayes-optimal decision rule is the rule that minimizes some expected loss (9). In our case, we have defined the loss as the negative reward rate, which is the negative average number of correct decisions per unit time, across a long sequence of such decisions. Furthermore, we have tuned the diffusion model boundaries such that there exists an optimal set of weights $\boldsymbol{w}^*$ that maximize the reward rate, and thus minimize the loss. Formally, the loss function $L(\boldsymbol{w}^*, \boldsymbol{w})$ returns this loss for a given action $\boldsymbol{w}$ (in our case a particular set of chosen weights) given some unobserved state of nature, $\boldsymbol{w}^*$ (in our case the set of weights that maximize the reward rate).

Given observations X (here, all information gathered from past trials), the Bayes-optimal action is the one minimizing the posterior loss, that is

$$\underset{\boldsymbol{w}}{\operatorname{argmin}} \left\langle L(\tilde{\boldsymbol{w}}, \boldsymbol{w}) \right\rangle_{p(\tilde{\boldsymbol{w}}|X)}, \tag{57}$$

where $p(\tilde{\boldsymbol{w}}|X)$ are the posterior weights given all past information. If we assume the loss to be approximately quadratic around $\tilde{\boldsymbol{w}}$, then it is (approximately) minimized by $\langle \tilde{\boldsymbol{w}} | \boldsymbol{X} \rangle$ (9). This justifies computing the posterior to perform learning through inference, and the use of the posterior mean for decision-making, as used in the main text.

Using Bayes-optimal decision rules for decision making has several appealing properties. One of particular interest in relation to learning through optimization is that it is an admissible rule (9, Ch. 4, Th. 9). Here, admissability is a concept from the frequentist school of decision theory, and specifies a (not necessarily unique) decision rule $\delta(\cdot)$ whose associated risk function $R(\boldsymbol{w}^*, \delta)$ is smallest among all possible decision rules and all possible states of nature $\boldsymbol{w}^*$. Here, the risk function is the expected loss for a given $\boldsymbol{w}^*$, with the expectation taken over possible observations X given $\boldsymbol{w}^*$, that is $R(\boldsymbol{w}^*, \delta) = \left\langle L(\boldsymbol{w}^*, \delta(X)) \right\rangle_{p(X|\boldsymbol{w}^*)}$. Therefore, the Bayes-optimal decision rule doesn't only minimize the expected loss under the posterior, but also the expected loss across different (frequentist) repetitions of the same "experiment", that is, different observations for the same state of nature $\boldsymbol{w}^*$, and does so across all possible states of nature. As a consequence, finding the posterior $\boldsymbol{w}$ through inference allows us to make decisions that (approximately) minimize the loss in multiple senses, which, in our case, maximizes the reward rate.

Explicit learning through optimization. Here we demonstrate for the stationary-weight case that our inference problem can be formulated as an optimization problem that aims at maximizing performance — here for simplicity measured as the probability of making correct choices. To do so, assume that, in each trial, the decision maker observes some J-dimensional momentary evidence $\delta\boldsymbol{x}$ that relates to the underlying latent state μ by Eq. [4], as before. They accumulate this evidence into $\boldsymbol{x}(t)$, and at some point (e.g., when a decision boundary is reached) decide according to $y = \operatorname{sign}\left(\boldsymbol{w}^T \boldsymbol{x}(t)\right)$, using some decision strategy weight parameters $\boldsymbol{w}$. Their aim is to optimize these weight parameters to maximize their probability of making correct choices.

To find the solution to this maximization problem, let us establish which weight parameters maximize the probability of making correct choices. For this, note that by Eq. [4], the accumulated evidence is distributed as

$$\boldsymbol{x}(t)|\mu^* \sim \mathcal{N}\left((\boldsymbol{a}\mu^* + \boldsymbol{b})\, t, \boldsymbol{\Sigma} t\right), \tag{58}$$

where μ^* is the (unobserved) latent state that determines the correct choice by $y^* = \operatorname{sign}(\mu^*)$. As a consequence, $\boldsymbol{w}^T \boldsymbol{x}(t)/t$ is distributed as

$$\frac{\boldsymbol{w}^T \boldsymbol{x}(t)}{t} | \mu^* \sim \mathcal{N}\left(\boldsymbol{w}^T \boldsymbol{a}\mu^* + \boldsymbol{w}^T \boldsymbol{b}, \frac{1}{t} \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w}\right). \tag{59}$$

Recall that $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{\Sigma}$ in Eq. [4] have been defined to satisfy $\boldsymbol{w}^{*T}\boldsymbol{a} = 1$, $\boldsymbol{w}^{*T}\boldsymbol{b} = 0$, and $\boldsymbol{w}^{*T}\boldsymbol{\Sigma}\boldsymbol{w}^* = 1$ for some particular $\boldsymbol{w}^*$. For these parameters, we thus have $\boldsymbol{w}^{*T}\boldsymbol{x}(t)/t \sim \mathcal{N}\left(\mu^*, t^{-1}\right)$, which provides the best estimate of μ^* (in the mean squared error sense (9)), that can in turn be used as a basis for decision-making.

To find $\boldsymbol{w}^*$ from an observed sequence of $(\boldsymbol{x}_1(t_1), t_1, y_1^*), (\boldsymbol{x}_2(t_2), t_2, y_2^*), \ldots (\boldsymbol{x}_N(t_N), t_N, y_N^*)$, we can use maximum likelihood, which is consistent and asymptotically efficient. For the diffusion model, the likelihood of $\boldsymbol{w}$ for a particular choice y is by Eq. [7] (using $m = 0$) given by $p(y|\boldsymbol{x}, t, \boldsymbol{w}) = \Phi\left(y\boldsymbol{w}^T\tilde{\boldsymbol{x}}\right)$, where $\tilde{\boldsymbol{x}} \equiv \boldsymbol{x}/\sqrt{t + \sigma_\mu^{-2}}$, as previously defined. Therefore, the maximum (log-)likelihood estimate for the observed sequence is given by

$$\hat{\boldsymbol{w}}_{ML} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \sum_{n=1}^{N} \log \Phi\left(y_n^* \boldsymbol{w}^T \tilde{\boldsymbol{x}}_n\right). \tag{60}$$

Finding this estimate is an optimization problem. For a small number of observations N, this optimization problem might be underdetermined. To avoid instabilities, we can additionally add a regularization term that penalizes too large $\|\boldsymbol{w}\|^2$, leading to

$$\hat{\boldsymbol{w}}_{ML,reg} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \left(-\lambda \boldsymbol{w}^T \boldsymbol{w} + \sum_{n=1}^{N} \log \Phi\left(y_n^* \boldsymbol{w}^T \tilde{\boldsymbol{x}}_n\right)\right), \tag{61}$$

where $\lambda > 0$ is some regularization parameter. Overall, this demonstrates how to formulate our learning problem as a an optimization problem. We have used this approach to formulate one of our heuristics, resulting in Eq. [45].

To see how this approach relates to learning through inference, compare the expression for $\hat{\boldsymbol{w}}_{ML,reg}$ to Eq. [48]. As can be seen, with prior parameters $\boldsymbol{\mu}_0 = \boldsymbol{0}$ and $\boldsymbol{\Sigma}_0 = \lambda^{-1}\boldsymbol{I}$, $\hat{\boldsymbol{w}}_{ML,reg}$ finds the maximum of the Bayesian parameter posterior over $\boldsymbol{w}$, given by Eq. [48]. In other words, it equals the maximum a-posteriori estimate. However, the optimization approach does not directly provide an estimate of the uncertainty in $\hat{\boldsymbol{w}}_{ML,reg}$. This makes it hard to form consistent sequential updates, in which uncertain weights should be updated more strongly than certain weights. More generally, formulating the learning problem as

an optimization problem reduces our ability to interpret the resulting expressions. For example, we might not have been able to identify that the learning rate is modulated by decision confidence without the inference formulation. All of these points made us follow the learning-as-inference route instead.

Implementing prior biases

So far we have assumed $P^+ \equiv p(\mu \geq 0) = 1/2$, making both $\mu \geq 0$ and $\mu < 0$ equally likely. Let us now consider how to consistently implement prior biases for which $P^+ \neq 1/2$. To do so, we will restrict our discussion to the one-dimensional momentary evidence δz. The high-dimensional momentary evidence case follows the same principles, and yields the same conclusions, but it notationally more burdensome.

With a *consistent* implementation of a prior bias we mean that we want to be able to choose a pair of arbitrary, potentially time-changing boundaries* $\pm\theta(t)$, each of which triggers a different Bayes-optimal choice. This requirement turns out to become critical.

Let us discuss two ways to implement biased priors in turn. The first corresponds to a shift in the mean of $p(\mu)$, while the second modulates the mass of $\mu \geq 0$ while keeping the shape of $p(\mu)$ otherwise unchanged (as in (10)).

Shifting the prior mean. If we assume prior $\mu \sim \mathcal{N}(m, \sigma_\mu)$, then $P^+ = \Phi(m/\sigma_\mu)$, such that $P^+ \neq 1/2$ if and only if $m \neq 0$. This is the case we have discussed further above (see Sec. "One-dimensional momentary evidence"), where we have found the posterior

$$p(\mu > 0|z,t) = \Phi\left(\frac{\sigma_\mu^{-2}m + z}{\sqrt{\sigma_\mu^{-2} + t}}\right). \quad [62]$$

Thus, the posterior is $p(\mu \geq 0|z,t) \geq 1/2$ if and only if $\sigma_\mu^{-2}m + z \geq 0$. This implies that Bayes-optimal decisions are determined by the sign of $\sigma_\mu^{-2}m + z$. As a result, we cannot simply bound the accumulated evidence z, as this might not guarantee a unique association between boundaries and Bayes-optimal choices. For example, consider some negative $m < 0$ and a positive $z < \sigma_\mu^{-2}|m|$ that has just reached the upper boundary $z = \theta(t)$. At this point we would intuitively make choice $y = 1$, corresponding to $\mu \geq 0$. However, as $\sigma_\mu^{-2}m + z < 0$, our expression for the posterior shows that $p(\mu \geq 0|z,t) < 1/2$, such that $y = -1$ would be the Bayes-optimal choice. This shows that bounding z directly can in some cases violate the boundary - choice correspondence.

We can regain this correspondence by instead bounding $\tilde{z}(t) = \sigma_\mu^{-2}m + z(t)$, which, by definition, starts at $\tilde{z}(0) = \sigma_\mu^{-2}m$. For this new accumulation variable it is easy to see that $p(\mu \geq 0|\tilde{z},t) \geq 1/2$ if and only if $\tilde{z} \geq 0$, thus restoring the boundary - choice correspondence.

Directly modulating $p(\mu \geq 0)$. An alternative approach to introducing a biased prior, which was taken in (10), is to boost one half of $p(\mu)$, while modulating down the other half,

$$p(\mu) = 2\mathcal{N}\left(\mu|0,\sigma_\mu^2\right)\begin{cases} P^+ & \text{if } \mu \geq 0, \\ 1-P^+ & \text{otherwise,} \end{cases} \quad [63]$$

ensuring again $p(\mu \geq 0) = P^+$. This prior, and the corresponding solution, has previously been investigated by (10).

This choice of prior results in the posterior over μ,

$$p(\mu|z,t) \propto \mathcal{N}\left(\mu|\frac{z}{\sigma_\mu^{-2}+t}, \frac{1}{\sigma_\mu^{-2}+t}\right)\begin{cases} P^+ & \text{if } \mu \geq 0, \\ 1-P^+ & \text{otherwise .} \end{cases} \quad [64]$$

Adding the normalization constant and integrating the above over all $\mu \geq 0$ results in the posterior

$$p(\mu \geq 0|z,t) = \frac{P^+\Phi\left(\frac{z}{\sqrt{\sigma_\mu^{-2}+t}}\right)}{P^+\Phi\left(\frac{z}{\sqrt{\sigma_\mu^{-2}+t}}\right) + (1-P^+)\left(1-\Phi\left(\frac{z}{\sqrt{\sigma_\mu^{-2}+t}}\right)\right)}. \quad [65]$$

This posterior is $p(\mu \geq 0|z,t) \geq 1/2$, and thus promotes choice $y = 1$, if

$$\log\frac{\Phi\left(\frac{z}{\sqrt{\sigma_\mu^{-2}+t}}\right)}{1-\Phi\left(\frac{z}{\sqrt{\sigma_\mu^{-2}+t}}\right)} \geq \log\frac{1-P^+}{P^+}, \quad [66]$$

that is, if the log-odds provided by the accumulated evidence exceeds that of the prior log-odds for $\mu < 0$. For the same accumulator value z, the evidence log-odds drops to zero over time. As a result, it might be that the Bayes-optimal choice at the same boundary changes over time, thus violating the boundary - decision correspondence.

*They might even follow different time-courses, without changing any of the discussed concepts. To keep notation simple, we won't consider this case.

When compared to the previous section, the way the prior impacts the posterior is more complex. This makes recovering the boundary - decision correspondence more complex. The aim is to find a $C_2(P^+,t)$ such that $\tilde{z}(t) = z(t) + C_2(P^+,t)$ determines Bayes-optimal decisions by its sign alone. This can be achieved by

$$C_2(P^+,t) = \sqrt{\sigma_\mu^{-2}+t}\Phi^{-1}\left(\frac{P^+\Phi\left(\frac{z}{\sqrt{\sigma_\mu^{-2}+t}}\right)}{P^+\Phi\left(\frac{z}{\sqrt{\sigma_\mu^{-2}+t}}\right)+(1-P^+)\Phi\left(\frac{-z}{\sqrt{\sigma_\mu^{-2}+t}}\right)}\right)-z(t), \quad [67]$$

which unfortunately doesn't yield a closed-form expression. To gain further insight, we approximate the cumulative Gaussian function by the logistic sigmoid $\Phi(z) \approx (1+\exp(-C_\sigma z))^{-1}$ with $C_\sigma = \pi^2/6$ to have matching slope at $z=0$. After some algebra, this results in

$$C_2(P^+,t) \approx \sqrt{\sigma_\mu^{-2}+t}\frac{1}{C_\sigma}\log\frac{P^+}{1-P^+}, \quad [68]$$

showing that it becomes insufficient to use a shift of the accumulation starting point, as for the previous prior. Instead, we require both a shifted starting point, as well as an additional shift in the accumulated evidence that varies over time.

The relation between decision confidence and choice accuracy for biased priors. For either choice of the prior, the solutions that regains the boundary - decision correspondence result in a decision confidence that is the same at both boundaries, as long as these boundaries are symmetric around zero. For example, for the posterior Eq. [62] this is easy to see by replacing $\sigma_\mu^{-2}m + z$ by $\tilde{z}$, and, at the decision by $\tilde{z} = \pm\theta(t)$, depending on which choice has been made. This is seemingly at odds with expecting a different choice accuracy at either boundary, imposed by the biased prior.

To show that no overall inconsistency between choice accuracy and choice confidence exists, let us consider the simpler case of a prior with a single "difficulty" μ_0, which is given by

$$p(\mu) = \frac{P^+}{2}\delta(\mu-\mu_0) + \frac{1-P^+}{2}\delta(\mu+\mu_0), \quad [69]$$

where $\delta(\cdot)$ is the Dirac delta function. That is, $\mu = \mu_0$ with probability P^+, and $\mu = -\mu_0$ with probability $1-P^+$. With this prior, it is easy to show that the posterior becomes

$$p(\mu=\mu_0|z,t) = p(\mu \geq 0|z,t) = \frac{1}{1+e^{-2\mu_0\left(z-\frac{1}{2\mu_0}\log\frac{P^+}{1-P^+}\right)}}. \quad [70]$$

For symmetric boundaries at $\pm\theta$, rather than shifting the accumulation starting point, we can equivalently shift the boundaries by the same amount to

$$\theta^+ = \theta - \frac{1}{2\mu_0}\log\frac{P^+}{1-P^+}, \qquad \theta^- = -\theta - \frac{1}{2\mu_0}\log\frac{P^+}{1-P^+}, \quad [71]$$

again leading to a constant decision confidence $(1+\exp(-2\mu_0\theta))^{-1}$ at either boundary.

To show that this decision confidence equals the probability of making the correct choice on average, we find this probability for each possible latent state value, using known expression for boundary hitting probabilities for diffusion models with asymmetric boundaries, as given in (11, 12). For $\mu = \mu_0$, the upper boundary θ^+ leads to the correct choice. This boundary is reached with probability

$$p\left(z=\theta^+|z\in\left\{\theta^+,\theta^-\right\},\mu=\mu_0\right) = \frac{e^{2\mu_0\theta}-\frac{1-P^+}{P^+}}{e^{2\mu_0\theta}-e^{-2\mu_0\theta}}, \quad [72]$$

which is 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 trials 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

$$p\left(z=\theta^-|z\in\left\{\theta^+,\theta^-\right\},\mu=-\mu_0\right) = \frac{e^{2\mu_0\theta}-\frac{P^+}{1-P^+}}{e^{2\mu_0\theta}-e^{-2\mu_0\theta}}, \quad [73]$$

where the only difference to the previous expression 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, which, after some algebra, results in

$$p(\text{correct}) = p\left(z=\theta^+|z\in\left\{\theta^+,\theta^-\right\},\mu=\mu_0\right)p(\mu=\mu_0) + p\left(z=\theta^-|z\in\left\{\theta^+,\theta^-\right\},\mu=-\mu_0\right)p(\mu=-\mu_0) = \frac{1}{1+e^{-2\mu_0\theta}}, \quad [74]$$

where we have used $p(\mu=\mu_0) = P^+$ and $p(\mu=-\mu_0) = 1-P^+$. 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 their decision confidence. Once this latent state is marginalized out (by averaging over it in Eq. [74]), consistency with the decision confidence is restored (13). 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.

Generating correlated momentary evidence

Recall that, for a given latent state μ, the mutli-dimensional momentary evidence is drawn according to

$$\delta\boldsymbol{x}|\mu \sim \mathcal{N}\left((\boldsymbol{a}\mu + \boldsymbol{b})\,\delta t, \boldsymbol{\Sigma}\delta t\right), \tag{75}$$

where the parameters $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{\Sigma}$ satisfy $\boldsymbol{a}^T\boldsymbol{w} = 1$, $\boldsymbol{b}^T\boldsymbol{w} = 0$, and $\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} = 1$ (see Sec. "High-dimensional momentary evidence").

We satisfy the requirement on $\boldsymbol{a}$ and $\boldsymbol{b}$ by choosing

$$\boldsymbol{a} = \frac{\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{w}}, \qquad \boldsymbol{b} = f_0\left(\mathbf{1} - \frac{\mathbf{1}^T\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{w}}\boldsymbol{w}\right), \tag{76}$$

where f_0 is a parameter. The expression for $\boldsymbol{b}$ minimizes $\|\boldsymbol{b} - f_0\mathbf{1}\|$ under the constraint $\boldsymbol{b}^T\boldsymbol{w} = 0$, effectively introducing an approximate baseline at f_0.

For our choice for the covariance we were guided by observations that the noise covariance spectrum in neural population recordings has few dominant components, and otherwise rapidly drops towards small values. We achieve this while satisfying $\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} = 1$ by designing a $\boldsymbol{\Sigma}$ that has one eigenvector $\boldsymbol{w}/\|\boldsymbol{w}\|$ with associated eigenvalue $1/\boldsymbol{w}^T\boldsymbol{w}$, and otherwise the desired eigenspectrum. To do so, we fill a $J \times J$ matrix $\boldsymbol{B}$ (J is the size of $\delta\boldsymbol{x}$) with zero mean unit variance Gaussian random numbers, except for the first row, which we set to $\boldsymbol{w}$. This is followed by Gram-Schmidt orthonormalization of $\boldsymbol{B}$, such that the first row becomes $\boldsymbol{w}/\|\boldsymbol{w}\|$, while all other rows unit vectors, orthogonal to $\boldsymbol{w}$. We then choose a diagonal $\boldsymbol{D}$ with the first diagonal element $d_{11} = 1/\boldsymbol{w}^T\boldsymbol{w}$, and all other diagonal elements $d_{jj} = \max\left\{\sigma_x^2 e^{-j+1}, \sigma_0^2\right\}/\boldsymbol{w}^T\boldsymbol{w}$, with parameters σ_x^2 and σ_0^2. The final covariance matrix is then given by $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{D}\boldsymbol{B}^T$.

If the weights change across consecutive trials $n-1$ and n, the momentary evidence needs to satisfy $\boldsymbol{a}_n^T\boldsymbol{w}_n = 1$, $\boldsymbol{b}_n^T\boldsymbol{w}_n = 0$, and $\boldsymbol{w}_n^T\boldsymbol{\Sigma}_n\boldsymbol{w}_n$ in each trial. For $\boldsymbol{a}_n$ and $\boldsymbol{b}_n$ this is easily achieved by re-computing them in each trial according to the above expressions.

The generation of $\boldsymbol{\Sigma}_n$ relies on a stochastic process, such that re-generating a new $\boldsymbol{\Sigma}_n$ in each trial might lead $\boldsymbol{\Sigma}_.$ to change significantly across trials despite only small changes in $\boldsymbol{w}_.$. To avoid this, we instead modify $\boldsymbol{\Sigma}_{n-1}$ by finding the smallest rotation $\boldsymbol{U}$ of $\boldsymbol{\Sigma}_{n-1}$ that satisfies $\boldsymbol{w}_n^T\boldsymbol{\Sigma}_n\boldsymbol{w}_n = 1$. To do so, we aim at finding $\boldsymbol{U}$ that satisfies $\boldsymbol{w}_n \propto \boldsymbol{U}\boldsymbol{w}_{n-1}$. This leaves $\boldsymbol{U}$ underconstraint. To introduce additional constraints, we would like to restrict the rotation imposed by $\boldsymbol{U}$ to the $(\boldsymbol{w}_{n-1}, \boldsymbol{w}_n)$ plane. We express this by using $\boldsymbol{\psi}_3, \ldots, \boldsymbol{\psi}_J$ that are orthonormal unit vectors that are also orthogonal to $\boldsymbol{w}_{n-1}$ and $\boldsymbol{w}_n$, which we can find by Gram-Schmidt orthonormalization. For those vectors, we desire $\boldsymbol{\psi}_n = \boldsymbol{U}\boldsymbol{\psi}_n$. Overall, this leads to the linear equation

$$\boldsymbol{U}\begin{pmatrix}\frac{\boldsymbol{w}_{n-1}}{\|\boldsymbol{w}_{n-1}\|} & \frac{\boldsymbol{w}_n}{\|\boldsymbol{w}_n\|} & \boldsymbol{\psi}_3 & \ldots & \boldsymbol{\psi}_J\end{pmatrix} = \begin{pmatrix}\frac{\boldsymbol{w}_n}{\|\boldsymbol{w}_n\|} & \frac{\tilde{\boldsymbol{w}}}{\|\tilde{\boldsymbol{w}}\|} & \boldsymbol{\psi}_3 & \ldots & \boldsymbol{\psi}_J\end{pmatrix}, \tag{77}$$

where $\tilde{\boldsymbol{w}}$ is given by

$$\tilde{\boldsymbol{w}} = 2\frac{\boldsymbol{w}_{n-1}^T\boldsymbol{w}_n}{\boldsymbol{w}_n^T\boldsymbol{w}_n}\boldsymbol{w}_n - \boldsymbol{w}_{n-1}, \tag{78}$$

and which we can easily solve for[†] $\boldsymbol{U}$. With this rotation matrix, $\boldsymbol{\Sigma}_n$ is given by

$$\boldsymbol{\Sigma}_n = \frac{\boldsymbol{w}_{n-1}^T\boldsymbol{w}_{n-1}}{\boldsymbol{w}_n^T\boldsymbol{w}_n}\boldsymbol{U}\boldsymbol{\Sigma}_{n-1}\boldsymbol{U}^T, \tag{79}$$

where the re-scaling by the fraction ensures the correct scaling of the eigenvalues.

Simulation details

We used parameters $\sigma_0^2 = 0.001$, $\sigma_x^2 = 2$ and $f_0 = 0$ to generate the momentary evidence $\delta\boldsymbol{x}|\mu$, as described in the previous section. At the beginning of each trial sequence we drew the true weights according to $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{m}_w, \boldsymbol{S}_w)$, with unit mean $\boldsymbol{m}_w = \mathbf{1}$ and identity covariance $\boldsymbol{S}_w = \boldsymbol{I}$. For that sequence, the diffusion model bounds $\pm\theta$ were time-invariant, and tuned to maximize the reward rate if the true weights were used to combined the inputs. We used $\sigma_\mu^2 = 3^2$ to draw μ in each trial. This μ determined the correct choice by $y^* = 1$ if $\mu \geq 0$, and $y^* = -1$ otherwise. The reward rate was given by

$$RR = \frac{p(\text{correct}) - c_{accum}\,\langle t\rangle}{\langle t\rangle + t_{iti}}, \tag{80}$$

where the average was across trials, and we set the evidence accumulation cost to $c_{accum} = 0.01$ and the inter-trial interval to $t_{iti} = 2s$. For non-stationary weights, we re-drew the weights after each choice according to Eq. [46], with $\boldsymbol{A} = \lambda\boldsymbol{I}$,

[†] Most likely there exists a closed-form expression for $\boldsymbol{U}$. We found it by solving the above expression numerically in each trial.

$\boldsymbol{b} = (1-\lambda)\boldsymbol{m}_w$, and $\boldsymbol{\Sigma}_d = \left(1-\lambda^2\right)\boldsymbol{S}_w$, and set the decay factor to $\lambda = 1 - 0.01$. This yields a weight diffusion that follows a first-order autoregressive process with steady-state mean $\boldsymbol{m}_w$ and covariance $\boldsymbol{S}_m$.

To compare the weight learning performance of ADF to alternative models, we simulated 1,000 learning trials 5,000 times, and reported the reward rate per trial averaged across these 5,000 repetitions. To assess steady-state performance, we performed the same procedure with non-stationary weights, and reported reward rate averaged over the last 100 trials, and over 5,000 repetitions. The sequential choice dependencies in Fig. 4A/B where also computed from these last 100 trials. The learning rate in Fig. 1D in the main text shows the pre-factor to $\boldsymbol{\Sigma}_w\tilde{\boldsymbol{x}}$ in Eq. [41] over decision confidence for a subsample of the last 10,000 trials of a single 15,000 trial simulation with non-stationary weights. For the Gibbs sampler, we drew 10 burn-in samples, followed by 200 samples in each trial. For the particle filter we simulated 1,000 particles.

We sped up the diffusion model 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 (14). The final $\boldsymbol{x}_n\left(t_n\right)$ was recovered by drawing it from

$$\boldsymbol{x}_n\left(t_n\right) \sim \mathcal{N}\left(\frac{\mu_n t_n}{\boldsymbol{w}^{*T}\boldsymbol{w}^*}\boldsymbol{w}^*, \frac{t_n}{\boldsymbol{w}^{*T}\boldsymbol{w}^*}\boldsymbol{I}\right), \qquad [81]$$

subject to the constraint $\boldsymbol{w}^T\boldsymbol{x}_n(t_n) = y_n\theta$, and where $\boldsymbol{w}^*$ and $\boldsymbol{w}$ denote the true weights, and the weights used for evidence accumulation, respectively.

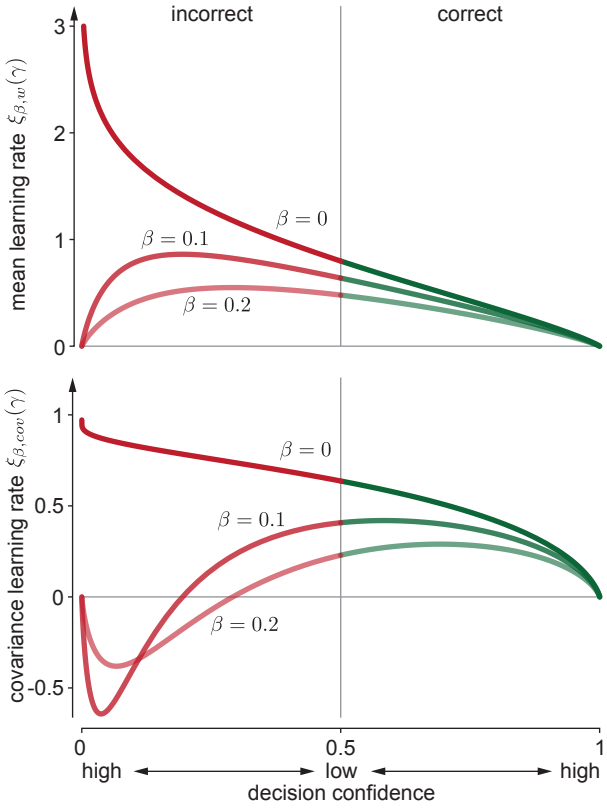

Fig. S1. Assumed density filtering learning rate modulators for noise-free and noisy feedback y^*. The top panel shows the learning rate modulator $\xi_{\beta,w}(\gamma)$ of the mean update for different levels of feedback noise, β. The bottom panel shows the same for the learning rate modulator $\xi_{\beta,cov}(\gamma)$ of the covariance update. In both cases, the marginal decision confidence associated with the feedback $p(y^*|\boldsymbol{x}, t) = \Phi(\gamma)$ is varied along the horizontal axis. This marginal decision confidence is $> 1/2$ for correct (green), and $< 1/2$ for incorrect (red) choices. $\beta = 0$ corresponds to the noise-free case, for which $\xi_w(\gamma) = \xi_{\beta,w}(\gamma)$ and $\xi_{cov}(\gamma) = \xi_{\beta,cov}(\gamma)$.

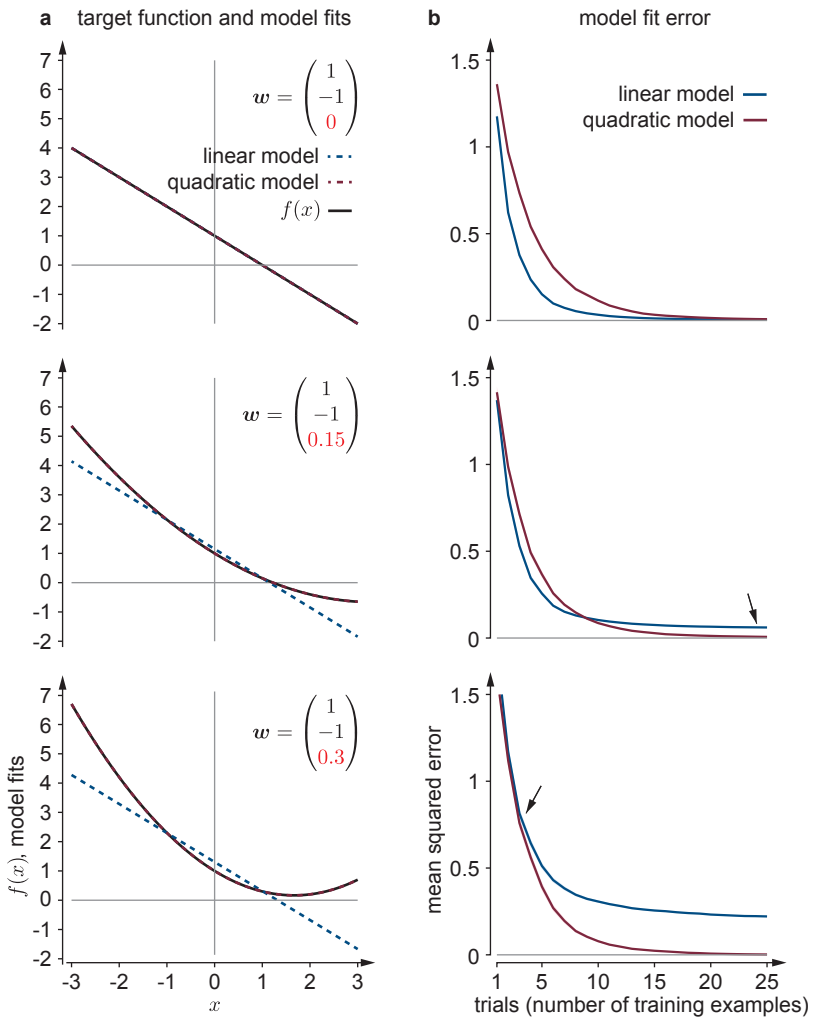

Fig. S2. Simpler models can learn more rapidly than more complex models, even if they are unable to provide perfect fits. We show model fits and model error for a linear (blue) and quadratic (red) model when fitting the target function $f(x) = w_1 + w_2 x + w_3 x^2$ for different w_3's (red number in the $\boldsymbol{w}$'s in (a)) for the top, middle, and bottom row. The quadratic model has the same functional form as $f(x)$ and learns all $\boldsymbol{w}$. The linear model fixes $w_3 = 0$, and only learns w_1 and w_2. Both models are fitted to training data consisting of $(x_n, f(x_n))$-pairs, by finding the model weights that minimize the mean squared error between model predictions and given $f(x_n)$'s across all observed x_n's. (a) With 10^4 training examples, both the linear and the quadratic model can fit a linear function (top; model fits and target function plotted on top of each other). As soon as the target function becomes quadratic (middle & bottom), the linear model fails to perfectly fit this function. (b) The mean squared error, here shown as an average across 500 repetitions across different training sets, drops more rapidly for the linear model than for the quadratic model if the target function is linear (top). This is because the linear model needs to learn fewer parameters for the same training set size. The error of both models goes to zero once the training set size increases. Even if the target function becomes quadratic (middle), the linear model can still learn more rapidly than the quadratic model (blue initially drops faster than red), even if it can't reduce its error to zero (arrow). This is only possible if the target function is still close-to-linear over the range of interest. Once it becomes too non-linear (bottom), the linear model learns slower than the quadratic model (arrow), and features a significantly worse asymptotic error. In (b), the mean squared error was in each repetition and for each training set size computed over 1000 new x's that were not part of the training set. For all simulations, the x's were drawn from $x \sim \mathcal{N}(0, 1)$. All learning was performed through optimization, by minimizing the mean squared error. We could have equally used learning by inference (using Bayesian linear regression with sufficiently uninformative priors), without affecting the results. Therefore, the shown effects are independent of the chosen learning formalism.

References

1. Pouget A, Drugowitsch J, Kepecs A (2016) Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience* 19(3):366–374.
2. Drugowitsch J, Moreno-Bote R, Churchland AK, Shadlen MN, Pouget A (2012) The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience* 32(11):3612–3628.
3. Opper M (1998) A Bayesian approach to on-line learning in *On-line Learning in Neural Networks*, ed. Saad D. (Cambridge University Press, New York, NY, USA), pp. 363–378.
4. Minka TP (2001) Expectation propagation for approximate Bayesian inference in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01. (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA), pp. 362–369.
5. Graepel T, Candela JQ, Borchert T, Herbrich R (2010) Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, eds. Fürnkranz J, Joachims T. (Omnipress, USA), pp. 13–20.
6. Chu W, Zinkevich M, Li L, Thomas A, Tseng B (2011) Unbiased online active learning in data streams in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, eds. Apte C, Gosh J, Smyth P. (ACM, New York, NY, USA), pp. 195–203.
7. Bishop CM (2006) *Pattern Recognition and Machine Learning*. (Springer-Verlag).
8. Murphy KP (2012) *Machine Learning: A Probabilistic Perspective*, Adaptive Computation and Machine Learning Series. (The MIT Press).
9. Berger JO (1993) *Statistical Decision Theory and Bayesian Analysis*, Springer Series in Statistics. (Springer-Verlag), 2nd edition.
10. Hanks TD, Mazurek ME, Kiani R, Hopp E, Shadlen MN (2011) Elapsed decision time affects the weighting of prior probability in a perceptual decision task. *The Journal of Neuroscience* 31(17):6339—-6352.
11. Cox DR, Miller HD (1965) *The Theory of Stochastic Processes*. (Chapman and Hall).
12. Palmer J, Huk AC, Shadlen MN (2005) The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision* 5:376–404.
13. Drugowitsch J, Moreno-Bote R, Pouget A (2014) Relation between belief and performance in perceptual decision making. *PLoS ONE* 9(5):e96511.
14. Drugowitsch J (2016) Fast and accurate Monte Carlo sampling of first-passage times from Wiener diffusion models. *Scientific Reports* 6:20490.