nature neuroscience

ARTICLES

https://doi.org/10.1038/s41593-019-0453-9

Optimal policy for multi-alternative decisions

Satohiro Tajima[1,4], Jan Drugowitsch[2,4*], Nisheet Patel[1] and Alexandre Pouget[1,3*]

Everyday decisions frequently require choosing among multiple alternatives. Yet the optimal policy for such decisions is unknown. Here we derive the normative policy for general multi-alternative decisions. This strategy requires evidence accumulation to nonlinear, time-dependent bounds that trigger choices. A geometric symmetry in those boundaries allows the optimal strategy to be implemented by a simple neural circuit involving normalization with fixed decision bounds and an urgency signal. The model captures several key features of the response of decision-making neurons as well as the increase in reaction time as a function of the number of alternatives, known as Hick's law. In addition, we show that in the presence of divisive normalization and internal variability, our model can account for several so-called 'irrational' behaviors, such as the similarity effect as well as the violation of both the independence of irrelevant alternatives principle and the regularity principle.

In a natural environment, choosing the best of multiple options is frequently critical for an organism's survival. Such decisions are often value-based, in which case the reward is determined by the chosen item (such as when individuals choose between food items; Fig. 1a), or perceptual, in which case individuals receive a fixed reward if they pick the correct option (Fig. 1b). Compared to binary choice paradigms[1–3], much less is known about the computational principles underlying decisions with more than two options[4]. Some studies have suggested that decisions among 3 or 4 options could be solved with coupled drift diffusion models[4–6], which are optimal for binary choices[7], but, as we are going to show, these become suboptimal once the number of choices grows beyond two. Another option for modeling such choices is to use 'race models'. In race models, the momentary choice preference is encoded by competing evidence accumulators, one per option, which trigger a choice as soon as one of them reaches a decision threshold (Fig. 1c). Such standard race models imply that both races and static decision criteria are independent across individual options. However, in contrast to race models, the nervous system features dynamic neural interactions across races, such as activity normalization[8,9] and a global urgency signal[10]. Whether such coupled races are compatible with optimal decision policies for three or more choices is unknown.

At the behavioral level, individuals choosing between three or more options exhibit several seemingly suboptimal behaviors, such as the similarity effect or violations of both the regularity principle and the independence of irrelevant alternatives (IIA) principle[11]. However, before concluding that such behaviors are suboptimal, it is critical to first derive the optimal policy and check whether they are compatible with this policy.

In this study, we adopt such a normative approach. Unlike previous models motivated by biological implementations, we start by deriving the optimal, reward-maximizing strategy for multi-alternative decision-making, and then ask how this strategy can be implemented by biologically plausible mechanisms. To do so, we first extend a recently developed theory of value-based decision-making with binary options[7] to N alternatives, revealing nonlinear and time-dependent decision boundaries in a high-dimensional belief space. Next, we show that geometric symmetries allow reducing the optimal strategy to a simple neural mechanism. This yields an extension of race models with time-dependent activity normalization controlled by an urgency signal[10].

The model provides an alternative perspective on how normalization and an urgency signal cooperate to implement close-to-optimal decisions for multi-alternative choices. We also demonstrate that the optimal policy is compatible with divisive normalization, which has been widely reported throughout the nervous system[8,9]. Additionally, in the presence of internal variability, our network replicates the similarity effect and violates both the IIA and regularity principles. Thus, our model isolates the functional components required for optimal decision-making and replicates a range of essential physiological and behavioral phenomena observed for multi-alternative decisions.

Results

The optimal policy for multi-alternative decisions. Suppose we have N alternatives to choose from in perceptual or value-based decisions. The decision-maker's aim is to make choices whose outcome depends on a priori unknown variables (for example, true rewards (Fig. 1a), or stimulus contrasts (Fig. 1b)) associated with the individual options, whose values vary across choice trials. We will assume that on a given trial, each short time duration δt yields a piece of noisy momentary evidence about the true values of the hidden variables. For perceptual decision-making, this would correspond to observing new sensory information, while for value-based decision-making, this might be the result of recalling past experiences from memory[12]. Our derivation shows that the optimal way of accumulating such evidence is to simply sum it up over time (Methods). This reduces the process of forming a belief about these variables to a diffusion (or random walk) process, $x(t)$, in an N-dimensional space, as implemented by race models (Fig. 1d).

Next, we derive the optimal stopping strategy: when should the decision-maker stop accumulating evidence and trigger a choice? To do so, and in contrast to experiments where participants wait until the end of the trial to respond, we only consider the more natural scenario where the decision-maker is in control of their decision time. In a standard race model, evidence accumulation stops whenever one of the races reaches a threshold that is constant over time and identical across races. In other words, evidence accumulation stops once the diffusing particle hits any sides of an N-dimensional

[1]Department of Basic Neuroscience, University of Geneva, Geneva, Switzerland. [2]Department of Neurobiology, Harvard Medical School, Boston, MA, USA. [3]Gatsby Computational Neuroscience Unit, University College London, London, UK. [4]These authors contributed equally: Satohiro Tajima, Jan Drugowitsch. *e-mail: jan_drugowitsch@hms.harvard.edu; Alexandre.Pouget@unige.ch

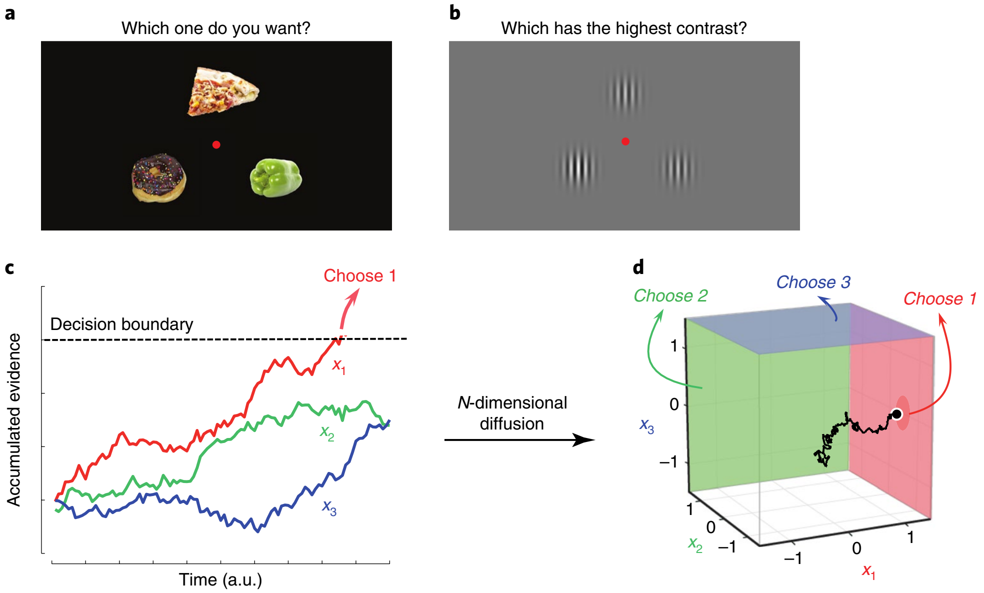


Fig. 1 | Multi-alternative decision tasks and the standard race model. a, An example value-based task in a laboratory setting. In a typical experiment, participants are rewarded with one of the objects they chose (in a randomly selected trial from the whole trial sequence). **b**, An example perceptual task, where participants are required to choose the highest-contrast Gabor patch—in this example, the one on the bottom left. **c**, The race model. The colored traces represent the accumulated evidence for individual options (x_1, x_2 and x_3). In the race model, the accumulation process is terminated when either race reaches a constant decision boundary (a.u., arbitrary units). **d**, An alternative representation for the same race model, where the races of accumulated evidence are shown as an *N*-dimensional diffusion. With this representation, the decision boundary for each option corresponds to a side of an *N*-dimensional cube, reflecting the independence of decision boundaries across options in the race model.

(half-)cube (Fig. 1d). While simple, this stopping policy is not necessarily optimal. To find the optimal policy, we use tools from dynamic programming[7,13]. One such tool is the ‘value function’ $V(t,x)$, which corresponds to the expected reward for being in state x at time t, assuming that the optimal policy is followed from there on. This value function can be computed recursively through a Bellman equation (Methods). For the simple case of a single, isolated choice, the decision-maker aims to maximize the expected reward (or reward per unit time) for this choice minus some cost c for accumulating evidence per unit time. One can imagine several different types of costs, such as, for example, the metabolic cost of accumulating more evidence. Once we embed this single choice within a long sequence of similar choices, an additional cost ρ emerges that reflects missing out on rewards that future choices yield (Methods). Overall, the optimal decision policy results in:

$$V(t,\mathbf{x};\rho)=\max\left\{\underbrace{\max_i\ \hat{r}_i(t,x_i)-\rho t_w}_{\text{deciding immediately}},\ \underbrace{\langle V(t+\delta t,\mathbf{x})\rangle-(c+\rho)\delta t}_{\text{deciding later}}\right\}\qquad(1)$$

This value function compares the value for deciding immediately, yielding the highest of the N expected rewards $\hat{r}_1,\ldots,\hat{r}_N$, with that for accumulating more evidence and deciding later; ρ is the reward rate (see Methods for the formal definition), t_w is the intertrial interval including the nondecision time required for motor movement. The expected reward for each option, $\hat{r}_i(t,x_i)$ is computed by combining the accumulated evidence with the prior knowledge about the reward mean and variance through Bayes’ rule (Methods). As shown by dynamic programming theory, the larger of these two terms yields the optimal value function; their intersection determines the decision boundaries for stopping evidence accumulation and thus the optimal policy. In realistic setups, decision-makers make a sequence of choices, in which case the aim of maximizing the total reward becomes equivalent (assuming a very long sequence of choice) to maximizing their reward rate, which is the expected reward for either choice divided by the expected time between consecutive choices. The value function for this case is the same as that for the single-trial choice, except that both values for deciding immediately and for accumulating more evidence include the opportunity cost of missing out on future rewards (Methods).

We found the optimal policy for this general problem by computing the value function numerically[14] from which we derived the complex, nonlinear decision boundaries (Fig. 2a). Clearly, the structure of the optimal decision boundaries differs substantially from that of standard race models (Fig. 1d). Interestingly, we found that they have an important symmetry. They are parallel to the diagonal, that is, the line connecting (0,0,…,0) and (1,1,…,1) (Supplementary Note 1 shows this formally). This symmetry implies that any diffusion parallel to the diagonal line is irrelevant to the final decision, such that we only need to consider the projection of the diffusion process onto the hyperplane orthogonal to this line (Fig. 2b). The decision boundaries remain nonlinear even in this projection, as depicted by the curvatures of the solid lines in Fig. 2b. Note that for binary choices, our derivation indicates that the projection of the diffusion process onto an $(N-1)$-dimensional subspace becomes a projection onto a line since $N=2$. On this line, the stopping boundaries are just two points and therefore cannot exhibit any nonlinearities. Thus, for $N=2$, the optimal policy corresponds to the well-known drift diffusion model of decision-making[7,13].

Numerical solutions also revealed that the optimal decision boundaries evolve over time; they approach each other as time elapses and finally collapse (Fig. 2b). These nonlinear collapsing boundaries differ from the linear and static ones of previous approximate models, such as multihypothesis sequential probability ratio tests (MSPRTs)[15–17], which are known to be only asymptotically optimal under specific assumptions (Methods).

We show in Supplementary Note 4 that these results generalize to models where the streams of noisy momentary evidence are

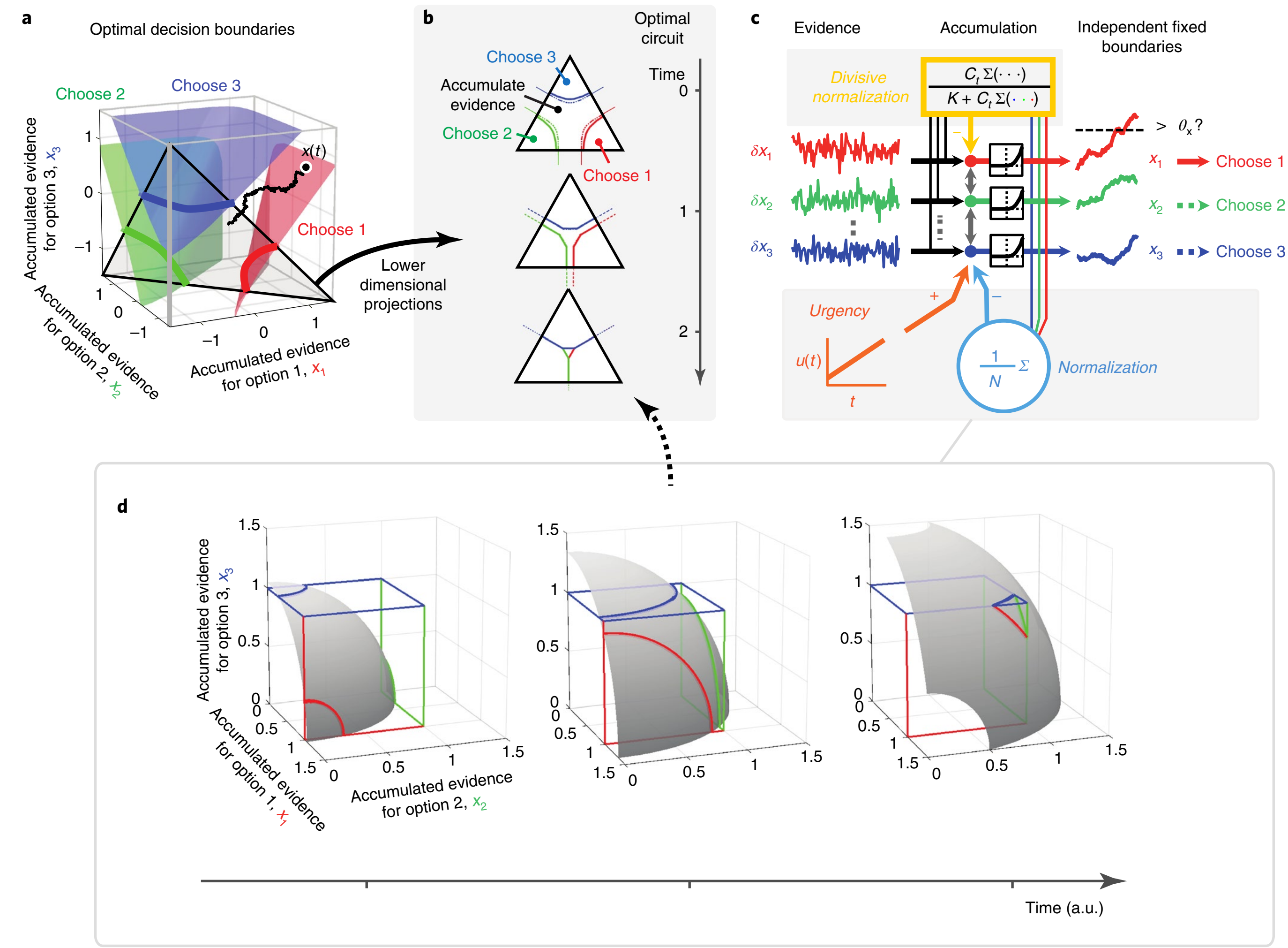

a
Optimal decision boundaries
Choose 2
Choose 3
Choose 1
x(t)
Accumulated evidence for option 3, x_3
Accumulated evidence for option 2, x_2
Accumulated evidence for option 1, x_1
1
0
−1
Lower dimensional projections
b
Optimal circuit
Choose 3
Accumulate evidence
Choose 2
Choose 1
Time
0
1
2
c
Evidence
Accumulation
Independent fixed boundaries
Divisive normalization
$\frac{C_t \Sigma(\cdot\cdot\cdot)}{K + C_t \Sigma(\cdot\cdot\cdot)}$
δx_1
δx_2
δx_3
$> \theta_x$?
x_1 Choose 1
x_2 Choose 2
x_3 Choose 3
Urgency
u(t)
t
+
−
$\frac{1}{N}\Sigma$
Normalization
d
Accumulated evidence for option 3, x_3
Accumulated evidence for option 1, x_1
Accumulated evidence for option 2, x_2
1.5
1
0.5
0
Time (a.u.)

Fig. 2 | The optimal decision policy for three alternative choices. a, The derived optimal decision boundaries in the diffusion space. In contrast to the standard race model's decision boundaries (Fig. 1d), they are nonlinear but symmetric with respect to the diagonal (that is, the vector (1,1,1)). **b**, Lower dimensional projections of decision boundaries at different time points. The solid curves are the optimal decision boundaries projected onto the plane orthogonal to the diagonal (the black triangle in **a**). The dashed curves indicate the effective decision boundaries implemented by the circuit in **c**. **c**, The circuit approximating the optimal policy. Like race models, it features constant decision thresholds that are independently applied to individual options. However, the evidence accumulation process is now modulated by recurrent global inhibition after a nonlinear activation function (the 'normalization' term), a time-dependent global bias input ('urgency signal') and rescaling ('divisive normalization'). **d**, Schematic illustrations of why the circuit in **c** can implement the optimal decision policy. The nonlinear recurrent normalization and urgency signal constrain the neural population states to a time-dependent manifold (the gray areas). Evidence accumulation corresponds to a diffusion process on this nonlinear ($(N-1)$-dimensional) manifold. The stopping bounds are implemented as the intersections (the colored thick curves) of the manifold and the cube (colored thin lines), where the cube represents the independent, constant decision thresholds for the individual choice options. Due to the urgency signal, the manifold moves toward the corner of the cube as time elapses, causing the intersections (that is, the stopping bounds) to collapse onto each other over time.

correlated in time, either with short-range temporal correlations, as is often observed in spikes trains, or with long-range temporal correlations as postulated, for example, in the linear ballistic accumulator model[18,19]. Our results also apply to experiments such as the ones performed by Thura and Cisek[20,21] where the momentary evidence is accumulated directly on the screen, in which case there is no need for latent integration.

Circuit implementation of the optimal policy. In the optimal policy we have derived, evidence accumulation is simple: it involves N accumulators, each summing up their associated momentary evidence independent of the other accumulators. By contrast, the stopping rule is complex: at every time step, the policy requires computing N time-dependent nonlinear functions that form the individual stopping boundaries. This rule is nonlocal because whether an accumulator stops depends not only on its own state but also on that of all the other accumulators. A simpler stopping rule would be one where a decision is made whenever one of the accumulators reaches a particular threshold value, as in independent race models. However, this would require a nonlinear and nonlocal accumulation process to implement the same policy through a proper variable transformation. Nonetheless, such a solution would be appealing from a neural point of view since it could be implemented in a nonlinear recurrent network endowed with a simple winner-takes-all mechanism that selects a choice once the threshold is reached by one of the accumulators.

Armed with this insight, we found that a recurrent network with independent thresholds (Fig. 2c) can indeed approximate the optimal solution very closely. It consists of N neurons (or N groups of identical neurons), one per option, which receive evidence for their

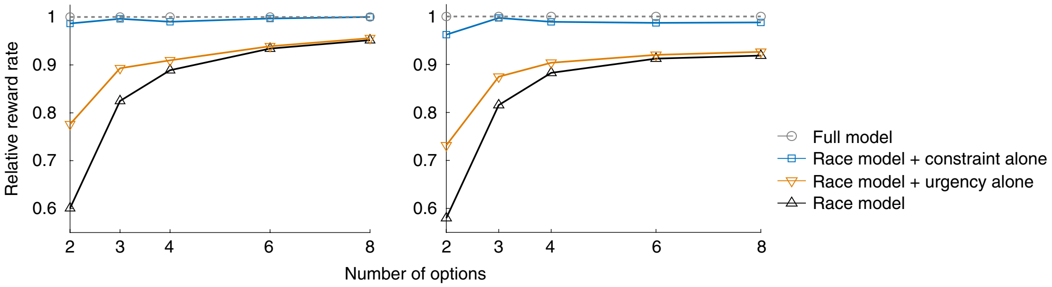


Fig. 3 | Normalization and urgency improve task performance. Relative reward rates in value-based (left) and perceptual tasks (right). To quantify the contribution of each circuit component, we compared the performance of four different circuit models: (1) the standard race model with independent evidence accumulation within each accumulator; (2) a race model with only an urgency signal; (3) a race model with only normalization; and (4) the full model with both urgency signal and normalization. We quantified the reward rates of models 1–3 ('reduced models') relative to that of the full model by $\rho_k^{\mathrm{Rel}} \equiv (\rho_k - \rho^{\mathrm{Rand}})/(\rho^{\mathrm{Full}} - \rho^{\mathrm{Rand}})$, where $\rho_k(k=1,2,3)$ denotes the reward rates of reduced models 1–3; $\rho^{\mathrm{Rand}} = \bar{z}/t_w$ is the baseline reward rate of a decision-maker who makes immediate random choices after trial onset; ρ^{Full} is the reward rate of the full model with both normalization and urgency. The performance differences across models shrink with an increasing number of options because the performance shown is relative to a model making random, immediate choices. Indeed, as the number of options to choose from increases, the absolute reward rates of the full and reduced models increase at similar rates, while the performance of the random model remains the same. Each point represents the mean reward rate across 10^6 simulated trials.

associated option. The network operates at two timescales. On the slower timescale, neurons accumulate momentary evidence independently across options according to:

$$\tilde{\boldsymbol{x}}_t = \delta \boldsymbol{x}_t + \frac{\boldsymbol{x}_{t-1}}{C_{t-1}} \tag{2}$$

$$\boldsymbol{x}_t = C_t \tilde{\boldsymbol{x}}_t \tag{3}$$

where x_t is the vector of accumulated evidence at time t, δx_t is the vector of momentary evidence at time t and C_t is the commonly used divisive normalization, $C_t = K/(\sigma_h + \sum_{n=1}^{N} \tilde{x}_{t,n})$, $\tilde{x}_{t,n}$ denotes the nth component of the vector $\tilde{\boldsymbol{x}}_t$. This form of divisive normalization merely rescales the space of evidence accumulation, leaving the relative distances between accumulators and stopping bounds intact. As a result, it has no impact on the performance of the model if the stopping bounds are adequately rescaled, and no appreciable impact even without this rescaling. It is included for biological realism because this nonlinearity is found throughout the cortex and in particular in the lateral intraparietal (LIP) area[8,22].

On the faster timescale, activity is projected onto a manifold defined by $\frac{1}{N}\sum_i f(x_i) = u(t)$ (gray surface in Fig. 2d), where $u(t)$ is the urgency signal. This operation is implemented by iterating:

$$x_{t,n} \leftarrow x_{t,n} + \gamma \left(u(t) - \frac{1}{N} \sum_i f(x_{t,i}) \right) \tag{4}$$

until convergence; γ is the update rate and f is a rectified polynomial nonlinearity (see Methods and Supplementary Note 2 for details). This process is stopped whenever one of the integrators reaches a preset threshold. The choice of this projection was motivated by two key factors. First, this particular form ensures that the projection is parallel to the diagonal, that is, the line connecting (0,0,…,0) and (1,1,…,1). As we have seen, diffusion along this axis is indeed irrelevant. Second, the use of a nonlinear function f implies that we do not merely project on the hyperplane orthogonal to the diagonal. Instead, we project onto a nonlinear manifold. This step is what allow us to approximate the original complex stopping surfaces with simpler independent bounds on each of the integrators, as illustrated in Fig. 2d (see Supplementary Note 2 for a formal explanation). The time-dependent urgency signal, $u(t)$, implements a collapsing bound, which is also part of the optimal policy (Fig. 2b). Indeed, this urgency signals brings all the neurons closer to their threshold and, as such, is equivalent to the collapse of the stopping bounds over time (Fig. 2d).

Equations (2), (3) and (4) can be turned into a single differential equation (see equation (40) in the Supplementary Note). The iterative difference equations we show in this article are a particular form of the implementation, making it easier to interpret the diffusion process. Importantly, equations (2) and (3) provide a generalization of divisive normalization, which ensures that evidence is still integrated optimally over time.

The model contains three parameters: the power of the nonlinearity, and the starting point and slope of the urgency signal (Methods). When these parameters are optimized to maximize the reward rate, the network approximates very closely the optimal stopping bounds (Fig. 2b). As a result, the reward rate achieved by the network is within 98 and 95% of the optimal reward rate for 3 and 4 options, respectively (across a wide range of prior distributions over rewards; see Methods).

Normalization and urgency improve task performance. Our circuit model comprises independent decision thresholds for individual options, as in standard race models (consistent with recordings in the LIP area[10]), but features time-dependent normalization in addition to an urgency signal. To quantify the contribution of each circuit component, we compared the performance of four different circuit models: (1) the standard race model with independent evidence accumulation within each accumulator; (2) a race model with the urgency signal alone; (3) a race model with normalization alone, where normalization refers to equation (4); and (4) the full model with both urgency signal and normalization. Note that all models included divisive normalization (equations (2) and (3)). This comparison revealed that adding the urgency signal and/or normalization to the standard race model indeed improved the reward rate (Fig. 3). Intriguingly, for both value-based and perceptual decisions, normalization had a much larger impact than the urgency signal, demonstrating the relative importance of normalization in improving the reward rate.

Relation to physiological and behavioral findings. *Urgency signal.* We examined how the neural dynamics and behavior predicted by the proposed circuit relates to previous physiological and behavioral findings. First, we found that the average activity in model neurons rises over time, independently of the sensory evidence, consistent with the urgency signals demonstrated in physiological recordings

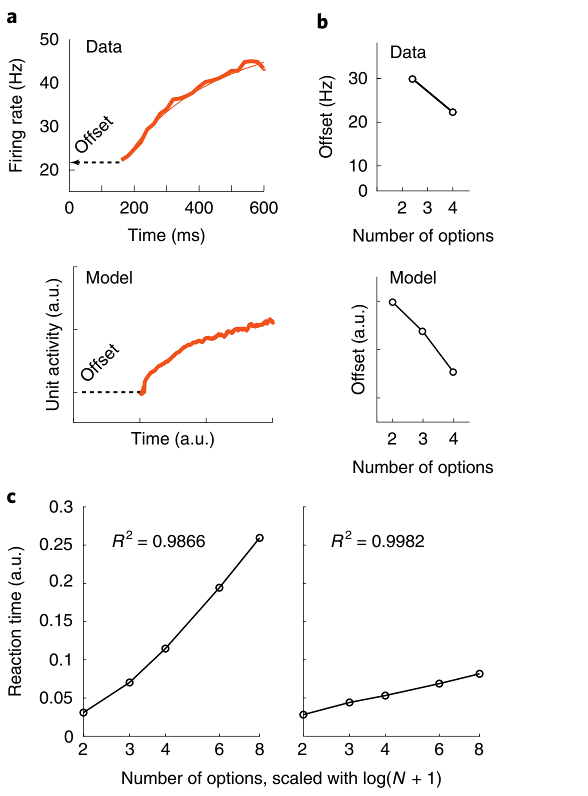


Fig. 4 | The model replicates the neuronal urgency signal and Hick's law in choice reaction times. a, Urgency signals in LIP cortex neurons (top) and in the model (bottom). The data points represent mean values across 10^4 simulated trials. In typical physiological experiments, urgency signals are extracted by averaging over neural activities across the entire recorded population, including different stimulus conditions. The thick trace (top) represents such an average for 0% motion coherence trials, whereas the thin trace is its fit to a hyperbolic function[10]. The rationale behind this procedure is that the urgency signal has been considered as a uniform additional input to all parietal neurons involved in the evidence accumulation process. A signal extracted this way is not exactly the same as the global input signal (function $u(t)$ in Fig. 2c) to the circuit, which includes nonlinear activity normalization through recurrent neural dynamics; thus, it does not trivially relate to the empirically observed urgency signals. Nonetheless, the average activity in model neurons was found to replicate the temporal increase, including the saturating temporal dynamics. a.u., arbitrary unit. **b**, The initial offset activities decrease with an increasing number of options, in both LIP neurons[10] (top) and the model (bottom). The data points represent the mean values across 10^4 simulated trials. **c**, Choice reaction times following Hick's law. The reaction times increase with the number of options (N) in both perceptual (left) and value-based (right) tasks. Note the logarithmic scaling of the horizontal axis. Each circle is the mean value across 10^4 simulated trials. The coefficient of determination (R^2 value) is obtained from linear regression, $\text{RT} = a + b\log(N+1)$ (Methods).

of neurons in the LIP area[10] (Fig. 4a). Interestingly, our model also replicates a gradual decrease in the slope of the average neural activity over time that arises in the model as a consequence of the nonlinear recurrent process.

Decrease in offset activities in multi-alternative tasks. Second, it has been reported that the initial 'offset' (that is, the average neural activity) of evidence accumulation[10,23] decreases as the number of options increases (Fig. 4b), although to our knowledge no normative explanation has been offered for this observation. Our circuit model replicates this property when optimized to maximize the reward rate (Fig. 4b). Indeed, in our model, increasing the number of options while leaving the initial offset unchanged causes a decrease in both accuracy and reaction time, and an associated drop in reward rate. This drop in reward rate can be compensated by lowering the initial offset, which increases both accuracy and reaction time but has a proportionally stronger effect on accuracy such that the reward rate increases.

Hick's law in choice reaction times. Third, the change in the optimal offset also explains the behavioral effects in choice reactions times known as 'Hick's law'[24,25]. Hick's law is one of the most robust properties of choice reaction times in perceptual decision tasks[24,25]. In its classic form, it states that mean reaction time (RT) and the logarithm of the number of options (N) are linearly related via $\text{RT} = a + b\log(N+1)$. Our model replicates this near-logarithmic relationship (Fig. 4c). Interestingly, the reaction time dependency on the number of options tends to be much weaker for value-based than perceptual decisions[26].

Value normalization. Fourth, our model replicates the suppressive effects of neurally encoded values among individual options (Fig. 5a). In particular, the activity of LIP neurons encodes values of targets inside the neuronal receptive fields, but is also affected by values associated with targets displayed outside the receptive fields[8,9,27]. The larger the total target values outside these receptive fields, the lower the neural activity, which is usually described as normalization.

IIA violation. So far, our neural model only has one source of variability, namely the noise corrupting the momentary evidence. However, there are other sources of variability that quite probably exist in the brain. For instance, the decision-maker must learn how to properly adjust the decision bounds to optimize the reward rate, which would result in trial-to-trial variability in the value of the bound. There is experimental evidence suggesting that learning can indeed induce extra variability in decision-making tasks[28]. Variability in bounds and neural responses could also be purposely induced by neural circuits to ensure that the decision-maker does not always choose the option with the highest value but also explores alternatives. Such an exploration behavior is critical in environments where the value of the options varies over time, which is common in real-world situations.

In our neural model, we added such extra variability directly to the accumulator by adding zero-mean Gaussian white noise to the state of the accumulator at each time step t after applying both normalizations (equations (2), (3) and (4)). Despite this extra variability, our neural model continues to outperform the race model (Fig. 5c and Supplementary Fig. 1). Stripping the normalization from the full model results in a large drop in reward rate with a further drop, although less pronounced, when the urgency signal is also removed.

Importantly, this version of the model also replicates apparently 'irrational' behavior in humans and animals that violates the IIA principle[29], an axiomatic property assumed in traditional rational theories of choice[30,31]. Behavioral studies have shown that the choice between two highly valued options depends on the value of a third alternative option[32–36], even if the value of this third option is so low that it is never chosen. One example of such an interaction is shown in Fig. 5b. In this experiment, participants found it increasingly harder to pick among their two top choices as the value of the third option increased. Our noisy neural model exhibits a similar IIA violation (Fig. 5b), which is primarily caused by divisive normalization. Divisive normalization decreases the mean value difference between the two top options as the value of the third option is increased, making these two options harder to distinguish due the presence of internal variability.

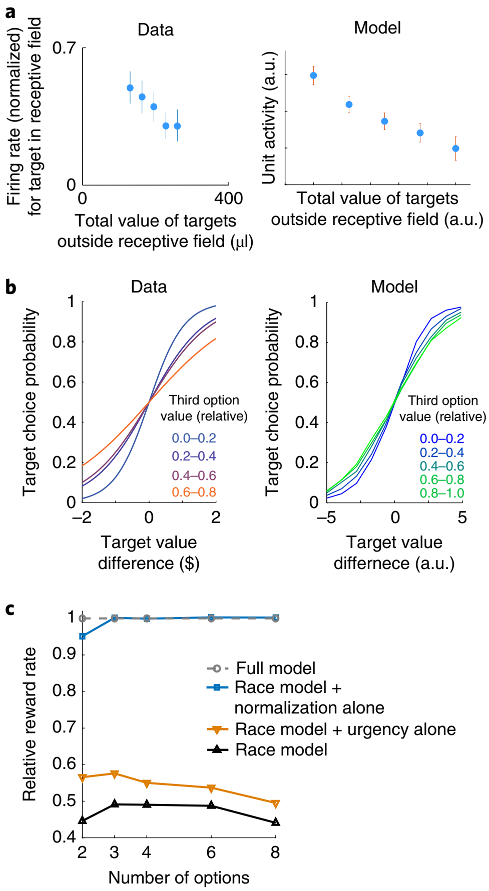


Fig. 5 | Activity normalization and violation of the axiom of IIA independence. a, Neuronal response to a saccadic target associated with a fixed reward as a function of the total amount of reward for all other targets on the screen in the LIP area (left) and in the model (right). Data from ref. [8] were used to create this panel; in the original experiment, subjects were monkeys and the targets were drops of juice. In both LIP and the model, the response of a neuron to a target associated with a fixed amount of reward decreases as the reward to the other targets increases. In the model, this effect is induced by the normalization. The points represent the mean ± s.d. across 10^6 simulated trials. **b**, Left: as the value of a third option is increased, the psychometric curve (for a fixed decision time, as set by the experimenter) corresponding to the choice between options 1 and 2 becomes shallower—a result that violates the IIA axiom. Data from ref. [11] were used to create this panel. Right: the model with added neural noise after activity normalization exhibits the same behavior over a total of 10^6 simulated trials. **c**, In the presence of internal variability, the race model variants without constrained evidence accumulation approximating the optimal policy (second term in equation (2)) perform much worse than our model variants with that constraint (when compared to Fig. 2d). Each point represents mean reward rate across 10^6 simulated trials.

Violation of the regularity principle. In multi-alternative decision-making, individuals not only violate the IIA but also the regularity principle. The regularity principle asserts that adding extra options cannot increase the probability of selecting an existing option. We found that the same model that violates the IIA also violates this regularity principle. At first, this may seem counterintuitive. Introducing a third option into a choice set must decrease the probability of picking either of the first two options. However, consider the probability of picking option 1 when option 2 is more valuable. In the absence of a third option, this probability will tend to be very small. When the third option is introduced and its value is increased, IIA violation implies that the probability of picking option 1 relative to option 2 will increase, as illustrated by the shallower psychometric curves in Fig. 5b. Therefore, two factors are with opposite effects are at play: the presence of a third option implies that choices 1 and 2 are picked less often, but the probability of picking option 1 relative to option 2 increases as a result of IIA violation. Our simulations reveal that the second factor dominates when the value of option 1 is smaller than that of option 2, as illustrated in Fig. 6a.

The similarity effect. Our model also replicates the similarity effect that has been reported in the literature[35,37,38]. This effect refers to the fact that when individuals are given a third option similar to, say, option 1, the probability of choosing option 1 decreases. To model this effect, we postulated that each object is defined by a set of features and that its overall value is a linear combination of the values of its features. As before, we also assumed that the values of the features are not known exactly. Instead, the brain generates noisy samples of these values over time. In this scenario, the similarity between two objects is proportional to the overlap between their features. This overlap implies that the stream of value samples for the two similar options are correlated while being independent for the third, dissimilar option. Accordingly, we simulated a three-way race where the momentary evidence for options 1 and 3 are positively correlated. As illustrated in Fig. 6b, we found that the probability of choosing option 1 decreases relative to option 2 as the value of option 3 increases, thus replicating the similarity effect. As has been observed experimentally[39,40], we found that the similarity effect grows over time during the course of a single trial (Fig. 6c).

Predictions. Our model makes a number of experimental predictions at both the behavioral and neural levels (see Supplementary Note 3 for further details).

First, during evidence accumulation, the neural population activity should be near an $(N-1)$-dimensional continuous manifold (that is, a nonlinear surface), where N is the number of choices (Fig. 2d). This is a direct consequence of evidence accumulation paired with nonlinear normalization. As the activity of D-neurons is D-dimensional, and since $N \ll D$ in general, our prediction implies that neural activity should be constrained to a small subspace of the neural activity space. This prediction can be tested with standard dimensionality reduction techniques using multielectrode recordings, although this analysis should be done carefully since our model also predicts that the position of this manifold changes over time. Failure to take this time dependency into account could significantly bias the estimate of the dimensionality of the constraining manifold. Our theory makes 11 additional predictions related to the existence and properties of the manifold, which are listed in Supplementary Note 3.

Second, our model correctly predicted the decrease in the initial activity offset (baseline firing rate) value of LIP neurons with the number of choices. Remarkably, this offset decrease results from an economic strategy that maximizes the reward rates by balancing the speed and accuracy in a long sequence of trials under the opportunity cost for future rewards. Thus, the offset should also be modulated by other reward rate manipulations. For example, we predict that increasing the average reward rate by either increasing the reward associated with the choices or decreasing the intertrial interval should raise the offset for a fixed number of choices.

Third, previous studies have considered two types of strategies for multi-alternative decision-making: the 'max versus average' (Fig. 7b) and the 'max versus next' (Fig. 7c) (refs. [6,26,41]). Our theory predicts that individuals should smoothly transition between these two modes depending on the pattern of rewards across choices (Fig. 7a), a prediction that can be tested with standard psychophysical experiments. More specifically, when all choices are equally rewarded, or only one choice is highly rewarded, our model predicts that

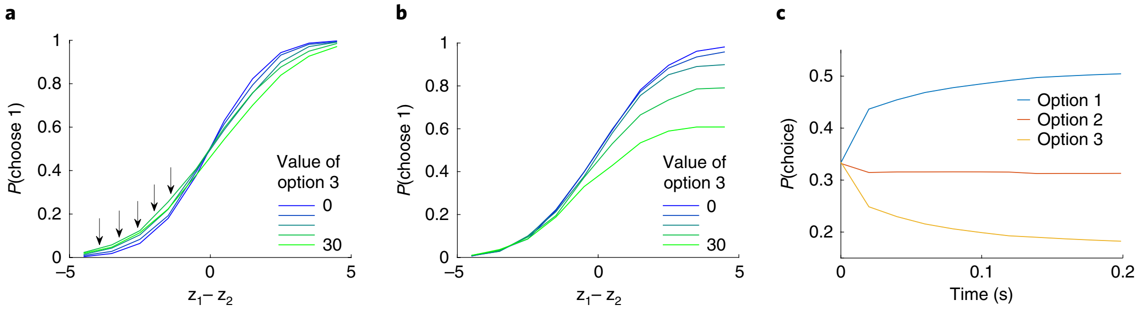


Fig. 6 | Regularity and similarity principles. a, Violation of the regularity principle. When a third choice is introduced, the probability of choosing option 1 increases as the value of option 3 increases. This effect is only observed when option 1 is much less valuable than option 2. **b**, The similarity effect: adding a third option, similar to option 1, reduces the probability of choosing option 1 relative to option 2 as the value of option 3 increases. The inset shows that the probability of picking option 1 also decreases as the value of option 3 increases. For both **a** and **b**, the model was simulated for 10^6 trials and binned into the five relative reward categories shown. Each of the five lines shows the mean of the respective reward category. **c**, The strength of the similarity effect increases with time within the course of a single trial, as shown by the decrease in the probability of choosing option 1 as time elapses.

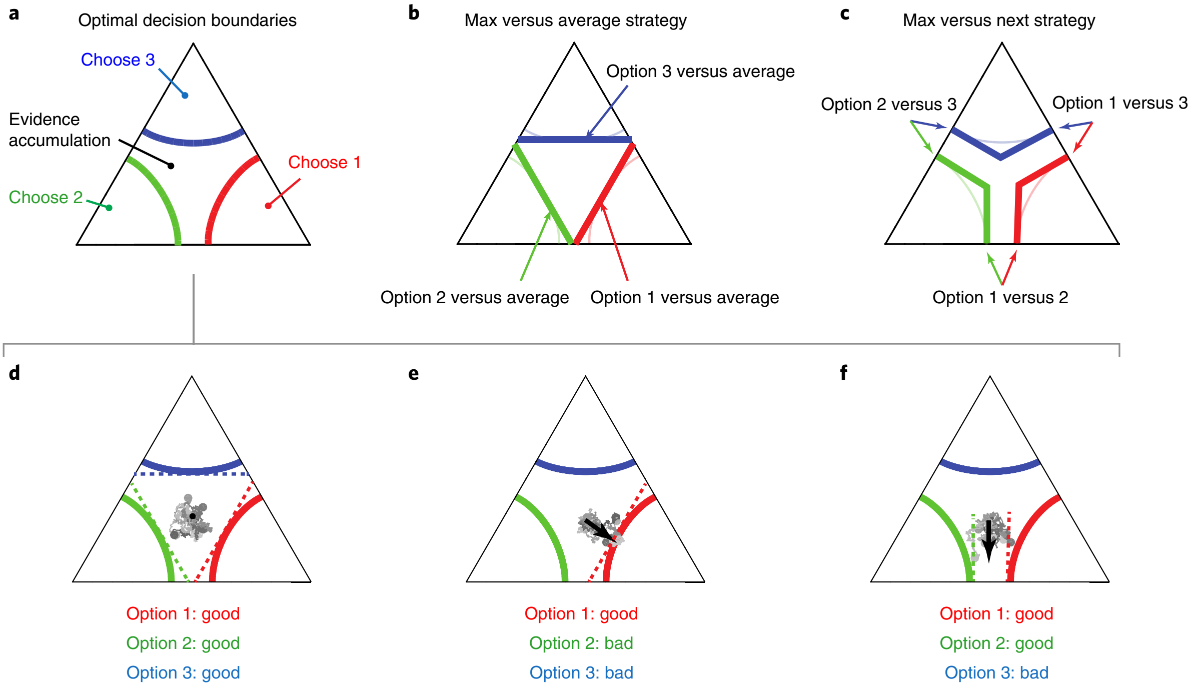


Fig. 7 | The optimal policy predicts a smooth transition between the max versus next and max versus average decision strategies depending on the relative values of the three options. a, The stopping bounds for the optimal policy after projecting the diffusion onto the hyperplane orthogonal to the diagonal. **b**, The stopping bounds corresponding to the max versus average strategy (thick colored lines). In this strategy, the decision-maker computes the difference between each option's value estimate and the average of the remaining options' values and triggers a choice when this difference hits a threshold. The stopping bounds in this case overlap with the optimal bounds from **a** (shown as thin colored lines) in the center but not on the side. **c**, The stopping bounds for the max versus next strategy (thick colored lines). In this strategy, the decision-maker compares the best and second-best value estimates and makes a choice when this difference exceeds a threshold. In a three-alternative choice, this is implemented with three pairs of linear decision boundaries (colored thick lines) corresponding to the three possible combinations of two options. In contrast to the bounds for the max versus average strategy, the bounds for the max versus next strategy overlap with the optimal bounds (thin colored lines) on the edge of the triangle but not in the center. **d**, When all three options are equally good, the diffusion of the particle is isotropic and therefore more likely to hit the stopping bounds in their centers, where they overlap with the max versus average strategy. **e**, When one option is much better than the other two, the diffusion is now biased toward the center of the bound corresponding to the good option, which is once again equivalent to the max versus average strategy. **f**, When two options are equally good, while the third is much worse, the particle will tend to drift toward the part of the triangle corresponding to the two good options (black arrow), where the optimal bound overlaps with the bounds for the max versus next strategy. The blotchy gray curves in **d**–**f** illustrate accumulator trajectories that are typical for the considered scenarios, and the black arrows represent the mean drift direction.

individuals should adopt a max versus average strategy (Fig. 7d,e), whereas when two options are highly rewarded, our model predicts that individuals should adopt a max versus next strategy (Fig. 7f).

Discussion

In this study, we discussed the optimal policy for decisions between more than two valuable options, as well as a possible biological implementation. The resulting policy has nonlinear boundaries and thus differs qualitatively from the simple diffusion models that implement the optimal policy for the two-alternative case[7]. More specifically, this work makes four major contributions. First, we prove analytically that the optimal policy involves a nonlinear projection onto an $(N-1)$-dimensional manifold, which can be closely approximated by neural circuits with nonlinear normalization (equation (4)). Second, apparently 'irrational' choice behaviors, such as IIA violation, are reproduced by our model in the presence of internal variability and divisive normalization. Third, we found that the distance to the threshold must increase with a set size for optimal performance. This has already been observed experimentally[10,23]. To our knowledge, no computational explanation has been offered for this effect until now. Fourth, the model follows Hick's law, that is, it predicts that reaction times in value-based decisions should be proportional to the log of the number of choices plus one, as is commonly observed in behavioral choice data. However, our model does not account for the violation of Hick's law for saccadic eye movement effects[42,43], or the well know pop-out effect reported in visual search, where reaction times are independent of the number of items on the screen[44]. Capturing these effects requires that we specialize our model to the specific context of these experiments; this is beyond the scope of the present article.

Our replication of IIA violation is similar to what Louie et al.[11] have proposed recently, although they did not consider noise in the momentary evidence and did not derive the optimal policy for multi-alternative decision-making. Therefore, our work demonstrates that an optimal policy for multi-alternative decision-making using divisive normalization violates the IIA in the presence of internal noise. Note that our work shows that divisive normalization is not required for optimal performance when the only source of noise is in the sensory evidence, although another form of normalization (equation (4)) is needed. However, preliminary work by Steverson et al.[45] clarified the conditions under which networks with divisive normalization implement the optimal policy for decision-making with regard to internal noise, thus suggesting that divisive normalization might indeed be required for optimal decision-making when all sources of noise are considered. Moreover, recent proof of equivalence between divisive normalization and an information processing model offers another explanation for the role of divisive normalization—to optimally balance the expected value of the chosen option with the entropic cost of reducing uncertainty in the choice[45].

A well-known strategy to decide between multiple options is the MSPRT[15,16]; previous studies have shown that the MSPRT could be implemented/approximated by neural circuits[17,41,46]. However, the MSPRT has not been designed for the problems we considered in this study. First, it assumes that the decision-maker receives a fixed magnitude of reward based on choice accuracy (that is, whether they are correct or incorrect) in each trial, as in conventional perceptual decision tasks. Value-based decisions, where the reward magnitude can vary across trials, clearly violate this assumption. Second, it assumes a constant task difficulty whereas the present study assumes the difficulty of both value-based and perceptual choices to vary across these choices. Third, since the MSPRT is only asymptotically optimal in the limit of infinitely small error rates (that is, when the model's performance is nearly 100% correct), it deviates from the optimal policy when this error rate is not negligible[15,16]. Our present analysis clarifies the properties of the optimal decision policy under multiple options, which differs from the MSPRT by characteristic nonlinear and collapsing decision boundaries. Despite the apparent complexity of those decision boundaries, we found that a symmetry in these boundaries allows the optimal strategies to be approximated by a circuit that features well-known neural mechanisms—race models whose evidence accumulation process is modulated by normalization, an urgency signal and nonlinear activation functions. The model provides a consistent explanation for the functional significance of normalization and urgency signal. They are necessary to implement optimal decision policies for multi-alternative choices where participants control the decision time.

Although we modeled the uncertainty about the true hidden states or values with a single Gaussian process that represents the noisy momentary evidence, in realistic situations the uncertainty could have multiple origins, including both external and internal sources. Potential sources of external noises include the stochastic nature of stimuli, sensory noise and incomplete knowledge about the options (for example, having not yet read the dessert of a particular menu option when choosing among different lunch menus). On the other hand, internal noises could result from learning, exploration, suboptimal computation[47], uncertain memory or ongoing value inference (for example, sequentially contemplating features of a particular menu course over time). We assumed simplified generative models with an unbiased and uncorrelated Gaussian prior; future extensions should consider more complex setups, including asymmetric mean rewards among options.

Note that the present study considers a simplified case where the value of each option is represented with a scalar variable. We have shown that this model is sufficiently complex to replicate basic behavioral properties, such as Hick's law, similarity effect and violation of both the IIA and regularity principle in multi-alternative choices. Future studies should cover more complex situations, including value comparisons based on multiple features (for example, speeds and designs of cars), which can lead to other forms of context-dependent choice behavior[34,35,48]. Decision-making with such a multidimensional feature space requires computing each option's value by appropriately weighting each feature. Some studies suggest that apparently irrational human behavior could be accounted for by heuristic weighting rules for features that integrate feature valences through feedforward[26,37,38] or recurrent[39,40] neural interactions. Interestingly, a recent study reported that a context-dependent feature weighting can increase the robustness of value encoding to neural noise in later processing stages[38,49], whereas another recent study provided a unified adaptive gain-control model that produces context-dependent behavioral biases[50]. However, to our knowledge, the optimal policy for these more complex models where the value function is computed by combining multiple features, presented sequentially, remains unknown. Once this policy is derived, it will be interesting to determine whether all, or part, of the seemingly irrational behaviors that have been reported in the literature are a consequence of using the optimal policies for such decisions or genuine limitations of the human decision-making process.

Finally, the current model provides several interesting predictions on neural population dynamics. Because of normalization, the collective neural activity could be constrained to a low-dimensional manifold during decision-making. The dimensionality of this manifold depends on the number of options ($N-1$ dimensions for N-alternative choices), whereas the position of the manifold should depend on time, reflecting the effect of the urgency signal. These predictions could be tested with neurophysiological population recordings combined with advanced dimensionality reduction techniques.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41593-019-0453-9.

Received: 7 February 2017; Accepted: 19 June 2019;
Published online: 05 August 2019

References

1. Gold, J. I. & Shadlen, M. N. The neural basis of decision making. *Annu. Rev. Neurosci.* **30**, 535–574 (2007).
2. Platt, M. L. & Glimcher, P. W. Neural correlates of decision variables in parietal cortex. *Nature* **400**, 233–238 (1999).
3. Wang, X. J. Decision making in recurrent neuronal circuits. *Neuron* **60**, 215–234 (2008).
4. Churchland, A. K. & Ditterich, J. New advances in understanding decisions among multiple alternatives. *Curr. Opin. Neurobiol.* **22**, 920–926 (2012).
5. Ditterich, J. A comparison between mechanisms of multi-alternative perceptual decision making: ability to explain human behavior, predictions for neurophysiology, and relationship with decision theory. *Front. Neurosci.* **4**, 184 (2010).
6. Krajbich, I. & Rangel, A. Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proc. Natl Acad. Sci. USA* **108**, 13852–13857 (2011).
7. Tajima, S., Drugowitsch, J. & Pouget, A. Optimal policy for value-based decision-making. *Nat. Commun.* **7**, 12400 (2016).
8. Louie, K., Grattan, L. E. & Glimcher, P. W. Reward value-based gain control: divisive normalization in parietal cortex. *J. Neurosci.* **31**, 10627–10639 (2011).
9. Louie, K., LoFaro, T., Webb, R. & Glimcher, P. W. Dynamic divisive normalization predicts time-varying value coding in decision-related circuits. *J. Neurosci.* **34**, 16046–16057 (2014).
10. Churchland, A. K., Kiani, R. & Shadlen, M. N. Decision-making with multiple alternatives. *Nat. Neurosci.* **11**, 693–702 (2008).
11. Louie, K., Khaw, M. W. & Glimcher, P. W. Normalization is a general neural mechanism for context-dependent decision making. *Proc. Natl Acad. Sci. USA* **110**, 6139–6144 (2013).
12. Shadlen, M. N. & Shohamy, D. Decision making and sequential sampling from memory. *Neuron* **90**, 927–939 (2016).
13. Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N. & Pouget, A. The cost of accumulating evidence in perceptual decision making. *J. Neurosci.* **32**, 3612–3628 (2012).
14. Brockwell, A. E. & Kadane, J. B. A gridding method for Bayesian sequential decision problems. *J. Comput. Graph. Stat.* **12**, 566–584 (2003).
15. Baum, C. W. & Veeravalli, V. V. A sequential procedure for multihypothesis testing. *IEEE Trans. Inf. Theory* **40**, 1994–2007 (1994).
16. Dragalin, V. P., Tartakovsky, A. G. & Veeravalli, V. V. Multihypothesis sequential probability ratio tests. II. Accurate asymptotic expansions for the expected sample size. *IEEE Trans. Inf. Theory* **46**, 1366–1383 (2000).
17. Bogacz, R. & Gurney, K. The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural Comput.* **19**, 442–477 (2007).
18. Carpenter, R. H. & Williams, M. L. Neural computation of log likelihood in control of saccadic eye movement. *Nature* **377**, 59–62 (1995).
19. Brown, S. & Heathcote, A. A ballistic model of choice response time. *Psychol. Rev.* **112**, 117–128 (2005).
20. Thura, D. & Cisek, P. Deliberation and commitment in the premotor and primary motor cortex during dynamic decision making. *Neuron* **81**, 1401–1416 (2014).
21. Thura, D. & Cisek, P. Modulation of premotor and primary motor cortical activity during volitional adjustments of speed-accuracy trade-offs. *J. Neurosci.* **36**, 938–956 (2016).
22. Carandini, M. & Heeger, D. J. Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* **13**, 51–62 (2012).
23. Keller, E. L. & McPeek, R. M. Neural discharge in the superior colliculus during target search paradigms. *Ann. N. Y. Acad. Sci.* **956**, 130–142 (2002).
24. Hick, W. E. On the rate of gain of information. Q. *J. Exp. Psychol.* **4**, 11–26 (1952).
25. Hyman, R. Stimulus information as a determinant of reaction time. *J. Exp. Psychol.* **45**, 188–196 (1953).
26. Usher, M. & McClelland, J. L. The time course of perceptual choice: the leaky, competing accumulator model. *Psychol. Rev.* **108**, 550–592 (2001).
27. Pastor-Bernier, A. & Cisek, P. Neural correlates of biased competition in premotor cortex. *J. Neurosci.* **31**, 7083–7088 (2011).
28. Mendonça, A. G. et al. The impact of learning on perceptual decisions and its implication for speed-accuracy tradeoffs. Preprint at *bioRxiv* https://doi.org/10.1101/501858 (2018).
29. Luce, R. D. *Individual Choice Behavior: a Theoretical Analysis* (Wiley, 1959).
30. Samuelson, P. A. *Foundations of Economic Analysis* (Harvard Univ. Press, 1947).
31. Stephens, D. W. & Krebs, J. R. *Foraging Theory* (Princeton Univ. Press, 1986).
32. Shafir, S., Waite, T. A. & Smith, B. H. Context-dependent violations of rational choice in honeybees (*Apis mellifera*) and gray jays (*Perisoreus canadensis*). *Behav. Ecol. Sociobiol.* **51**, 180–187 (2002).
33. Tversky, A. & Simonson, I. Context-dependent preferences. *Manage. Sci.* **39**, 1179–1189 (1993).
34. Huber, J., Payne, J. W. & Puto, C. Adding asymmetrically dominated alternatives: violations of regularity and the similarity hypothesis. *J. Consum. Res.* **9**, 90–98 (1982).
35. Tversky, A. Elimination by aspects: a theory of choice. *Psychol. Rev.* **79**, 281–299 (1972).
36. Gluth, S., Spektor, M. S. & Rieskamp, J. Value-based attentional capture affects multi-alternative decision making. *eLife* **7**, e39659 (2018).
37. Tsetsos, K., Chater, N. & Usher, M. Salience driven value integration explains decision biases and preference reversal. *Proc. Natl Acad. Sci. USA* **109**, 9659–9664 (2012).
38. Tsetsos, K. et al. Economic irrationality is optimal during noisy decision making. *Proc. Natl Acad. Sci. USA* **113**, 3102–3107 (2016).
39. Pettibone, J. C. Testing the effect of time pressure on asymmetric dominance and compromise decoys in choice. *Judgm. Decis. Mak.* **7**, 513–523 (2012).
40. Trueblood, J. S., Brown, S. D. & Heathcote, A. The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychol. Rev.* **121**, 179–205 (2014).
41. McMillen, T. & Holmes, P. The dynamics of choice among multiple alternatives. *J. Math. Psychol.* **50**, 30–57 (2006).
42. Kveraga, K., Boucher, L. & Hughes, H. C. Saccades operate in violation of Hick's law. *Exp. Brain Res.* **146**, 307–314 (2002).
43. Lawrence, B. M., St John, A., Abrams, R. A. & Snyder, L. H. An anti-Hick's effect in monkey and human saccade reaction times. *J. Vis.* **8**, 26.1–7 (2008).
44. Treisman, A. & Souther, J. Search asymmetry: a diagnostic for preattentive processing of separable features. *J. Exp. Psychol. Gen.* **114**, 285–310 (1985).
45. Steverson, K., Brandenburger, A. & Glimcher, P. Choice-theoretic foundations of the divisive normalization model. *J. Econ. Behav. Organ.* **164**, 148–165 (2019).
46. Bogacz, R., Usher, M., Zhang, J. & McClelland, J. L. Extending a biologically inspired model of choice: multi-alternatives, nonlinearity and value-based multidimensional choice. *Philos. Trans. R. Soc. Lond. B* **362**, 1655–1670 (2007).
47. Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E. & Pouget, A. Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron* **74**, 30–39 (2012).
48. Simonson, I. Choice based on reasons: the case of attraction and compromise effects. *J. Consum. Res.* **16**, 158–174 (1989).
49. Howes, A., Warren, P. A., Farmer, G., El-Deredy, W. & Lewis, R. L. Why contextual preference reversals maximize expected value. *Psychol. Rev.* **123**, 368–391 (2016).
50. Li, V., Michael, E., Balaguer, J., Herce Castañón, S. & Summerfield, C. Gain control explains the effect of distraction in human perceptual, cognitive, and economic decision making. *Proc. Natl Acad. Sci. USA* **115**, E8825–E8834 (2018).

Acknowledgements

A.P. was supported by the Swiss National Foundation (grant no. 31003A_143707) and a grant from the Simons Foundation (no. 325057). J.D. was supported by a Scholar Award in Understanding Human Cognition by the James S. McDonnell Foundation (grant no. 220020462). We dedicate this paper to the memory of S. Tajima, who tragically passed away in August 2017.

Author contributions

S.T., J.D. and A.P. conceived the study. S.T. and J.D. developed the theoretical framework. S.T., J.D. and N.P. performed the simulations and conducted the mathematical analysis. S.T., J.D., N.P. and A.P. interpreted the results and wrote the paper.

Competing interest

The authors declare no competing interest.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41593-019-0453-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.D. or A.P.

Peer review information: *Nature Neuroscience* thanks Jennifer Trueblood and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Task structure and generative models. We consider N-alternative value-based or perceptual decisions where decision-makers respond as soon as they commit to a choice. Value-based and perceptual decisions differ in how choices are associated with reward: in value-based decisions, the decision-maker reaps the reward associated with the chosen item (for example, a food item), whereas in perceptual paradigms the amount of reward depends only on whether the choice is 'correct' in the context of the current task. In contrast to previous models motivated by biological implementations[51–54], we start by deriving the optimal, reward-maximizing strategy for multi-alternative decision-making tasks without assuming specific biological implementations, and then ask how this strategy can be implemented by biologically plausible mechanisms. The following formulation applies to both perceptual and value-based tasks.

Let $\mathbf{z} \equiv (z_1,...,z_N)$ denote hidden variables (for example, reward magnitudes for value-based tasks, or stimulus contrasts for perceptual tasks) associated with N choice options. These true hidden variables vary across trials and are never observed directly and as such unknown to the decision-maker. Instead, the decision-maker observes some noisy momentary evidence with mean $\mathbf{z}\delta t$,

$$\delta \boldsymbol{x}_n | \boldsymbol{z} \sim \mathcal{N}(z\delta t, \boldsymbol{\Sigma}_x \delta t) \tag{5}$$

for each option $i \in \{1,...,N\}$, in every small time step n of duration δt. Σ_x here denotes the covariance matrix of the momentary evidence. Before observing any evidence, the decision-maker is assumed to hold a normally distributed prior belief,

$$\boldsymbol{z} \sim \mathcal{N}(\overline{z}, \boldsymbol{\Sigma}_z) \tag{6}$$

with mean $\overline{z}$ and covariance Σ_z reflecting the statistics of the true prior distribution, $p(\mathbf{z})$. For simplicity, we define the correct option in a perceptual task as the option associated with the largest hidden variable, $i_{\text{correct}} = \text{argmax}_i z_i$, which, for example, can be interpreted as the highest contrast in a contrast discrimination task.

In both value-based and perceptual tasks, we assume that the decision-maker tries to maximize the expected reward under a time constraint. Specifically, we focus on reaction time tasks where the decision-maker is free to choose at any time within each trial and proceeds through a long sequence of trials within a fixed time period. The total number of trials, and thus the total reward throughout the entire trial sequence, depends on how rapidly the decision-maker chooses in each trial: faster decisions allow for more of them in the same amount of time. However, due to noisy evidence, collecting more of this kind of evidence in each trial yields better choices, resulting in a trade-off between speed and accuracy.

Optimal decision policy. We assume that the decision-maker's aim is to maximize the total expected reward obtained in this task. The optimal decision policy comprises two key components: (1) optimal online inference of the hidden variables by accumulating the evidence about them; and (2) optimal rules for stopping the evidence accumulation to make a choice.

Optimal evidence accumulation. We provide a general formulation that includes correlations among options in the generative models. After some time $t = n\delta t$, the decision-maker's posterior belief about the true hidden variables $p(\mathbf{z}|\delta\mathbf{x}_1,...,\delta\mathbf{x}_n)$ is found using Bayes' rule, $p(\boldsymbol{z}|\delta\boldsymbol{x}_1,\ldots,\delta\boldsymbol{x}_n) \propto p(\boldsymbol{z}) \prod_{n'=1}^{n} p(\delta\boldsymbol{x}_{\boldsymbol{n}'}|\boldsymbol{z})$, using the fact that $\delta\mathbf{x}_{n'}$ $(n' = 1,...,n)$ is independent and identically distributed across time. This results in:

$$\boldsymbol{z}|\delta\boldsymbol{x}_1,\ldots,\delta\boldsymbol{x}_n \sim \mathcal{N}(\boldsymbol{\Sigma}(t)(\boldsymbol{\Sigma}_z^{-1}\overline{\boldsymbol{z}} + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{x}(t)), \boldsymbol{\Sigma}(t)) \tag{7}$$

where we have defined $\boldsymbol{x}(t) \equiv \sum_{n'=1}^{n} \delta\boldsymbol{x}_{n'}$ as the sum of all momentary evidence up to time t, and $\boldsymbol{\Sigma}(t) = (\boldsymbol{\Sigma}_z^{-1} + t\ \boldsymbol{\Sigma}_x^{-1})^{-1}$ as the posterior covariance. The temporally accumulated evidence $x(t)$ and the time t provide the sufficient statistics for $\mathbf{z}$ and thus for the rewards $\boldsymbol{r} \equiv (r_1,\ldots,r_N)^{\top}$ associated with individual options. For value-based decision-making, the reward r equals the true hidden variable $\mathbf{z}$, that is $\mathbf{r} = \mathbf{z}$, such that the expected option reward $\hat{r}_i(t, x_i(t)) = \langle z_i | t, x_i(t) \rangle$ is the mean of the posterior. For perceptual decision-making, the rewards associated with individual options are expressed as a vector r such that $r_i = r_{\text{correct}}$ when i is the correct option and $r_i = r_{\text{incorrect}}$ otherwise. Thus, the expected reward for option i is $r_i(t,\mathbf{x}(t)) = r_{\text{correct}}\, p(i = i_{\text{correct}} \mid t,\mathbf{x}(t)) + r_{\text{incorrect}}\, p(i \neq i_{\text{correct}} \mid t,\mathbf{x}(t))$. Because $\delta\boldsymbol{x}_{n'}$ is independent and identically distributed in time, $\mathbf{x}(t)$ is a random walk in an N-dimensional space (the thick black trace in Fig. 2a). The next question is when to stop accumulating evidence and which option to choose at that point.

Optimal stopping rules. To find the optimal policy, we use tools from dynamic programming[7,13,55]. One such tool is the 'value function' $V(\cdot)$, which can be defined recursively through Bellman's equation[56]. This value function returns for each state of the accumulation process (identified by the sufficient statistics) the total reward (including accumulation cost) the decision-maker expects to receive from this state onward when following the optimal policy.

Let us first consider this value function for the case of a single choice, where the aim is to maximize the expected reward for this choice minus some cost c per unit time for accumulating evidence (if there were no such cost, no decisions would ever be made). At any point in time t, the decision-maker can either decide to make a choice, yielding the highest of the N expected rewards, or accumulate more evidence for some small time δt, resulting in cost $-c\delta t$, and expected future reward given by the value function at time $t + \delta t$. According to Bellman's principle of optimality, the best action corresponds to the one yielding the highest expected reward, resulting in Bellman's equation

$$V(t, \boldsymbol{x}) = \max\left\{\max_i r_i(t, \boldsymbol{x}), \left\langle V(t + \delta t, \boldsymbol{x}(t + \delta t)) \right\rangle - c\delta t\right\} \tag{8}$$

where the expected rewards $r_i(t,x)$ differ between perceptual and value-based choices (see previous section; in both cases, they are functions of x and t), and the expectation in the second term is across expected changes of the accumulated evidence, $p(\mathbf{x}(t + \delta t)|\mathbf{x}(t),t)$. The intersection between the two terms within $\{\cdot,\cdot\}$ determines the decision boundaries for stopping the evidence accumulation and thus the optimal policy.

In more realistic setups, decision-makers make a sequence of choices within a limited time period, where the aim of maximizing the total reward becomes equivalent (assuming long time periods) to maximizing their reward rate ρ, which is the expected reward for either choice divided by the expected time between consecutive choices. This reward rate is thus given by $\rho = (\langle r_j | \tilde{\boldsymbol{z}}_j(0:T) \rangle - c\langle T \rangle)/(t_w + \langle T \rangle)$, where T is the evidence accumulation time, t_w is the waiting time after choices (including possible delays in motor responses) before the onset of evidence for the next choice, and the expectation is across choices j. The value function associated with the reward rate maximizing policy differs by introducing an additional opportunity cost ρ per unit time. For immediate choices, this introduces the cost $-\rho t_w$ that the decision-maker has to wait until the next trial (assuming $V(0, \overline{z}; \rho) = 0$). For accumulating more evidence, the associated cost increases from $-c\delta t$ to $-(c + \rho)\delta t$. Overall, this leads to Bellman's equation (equations (1), (8)) as given in the main text. If we set $\rho = 0$, we recover Bellman's equation for single, isolated choices.

To find the optimal policy for the aforementioned cases numerically, we computed the value function by backward induction[14] using Bellman's equation. Bellman's equation expresses the value function at time t as a function of the value function at time $t + \delta t$. Therefore, if we know the value function at some time T, we can compute it at time $T - \delta t$, then $T - 2\delta t$, and so on, until time $t = 0$. To find the reward rate, which is required to compute the value function, we initially set it to $\rho = 0$, computed the full value function, and then update it iteratively by root finding until $V(0, \overline{z}; \rho) = 0$, recomputing the full value function in each root-finding step (see Drugowitsch et al.[57] for the rationale behind this procedure).

Unless otherwise mentioned, we used $T = 10$ s and $\delta t = 0.005$ s for all simulations. That is, we assumed $V(T = 10,\mathbf{x};\rho)$ to be given by the value for immediate choices, and then moved backward in time in steps of 0.005s to find the value function by backward induction until $t = 0$. Furthermore, we set the prior parameters of the true, latent variables $\mathbf{z}$ to $\overline{z} = 1$. The waiting time was fixed to $t_w = 0.5$ s, and the accumulation cost to $c = 0$ (that is, the opportunity cost ρ was the only cost). The results did not change qualitatively when changing the values of these parameters. Supplementary Fig. 2 shows the dependence of stopping boundaries on the task parameters.

Boundary structure analysis. Interestingly, we found that the decision boundaries in value-based tasks generally have a remarkable symmetry that reduces the optimal policy to a simple neural computation. All the decision boundaries are parallel to the diagonal—the line connecting (0,0,...,0) and (1,1,...,1).

In value-based tasks, this symmetry emerges from the fact that the state transition probability $p(\mathbf{x}(t)|\mathbf{x}(t + \delta t))$ is invariant to translational shifts in $\mathbf{x}$. We can prove that the value function increases linearly along the diagonal, $V(t,\mathbf{x} + 1C) = V(t,\mathbf{x}) + C$ and $\nabla_x V(t,\mathbf{x}) \geq 0$, where C is an arbitrary scalar. From these properties of the value function, we can prove that the decision boundaries are 'parallel' to the diagonal: for all i, $B(t,x_i + C) = B(t,x_i) + 1C$, where $B(t,x_i)$ is a set of points that define, for a fixed x_i, the boundary in $x_{j \neq i}$ at which point a decision ought to be made. The formal proofs are provided in Supplementary Note 1.

We can demonstrate the same symmetry in the perceptual tasks, even though it arises from a different mechanism. In perceptual tasks, by construction, the value function is determined by the probability of each option being the correct answer. Because this probability is already normalized such that the sum of all the probabilities across options is 1, the resulting value function is constant along the diagonal (in contrast to the value-based case where the value function increases linearly along the diagonal). This yields the symmetry of decision boundaries along the diagonal.

Circuit implementation of the optimal policy. It may seem difficult for biological systems to implement the optimal decision boundaries since these boundaries are, in general, represented by N time-dependent nonlinear functions $F_i(t,\mathbf{x}(t)) = 0$ corresponding to the individual options, $i = 1,\ldots,N$, that depends on N and other task contingencies. Fortunately, however, because of the symmetry of these boundaries (see main text), the decision policy effectively reduces to a lower dimensional representation ($N - 1$ dimensions for an N-alternative choice), which supports a simpler implementation of these boundaries. The key idea is as follows. The original decision policy representation assumes evidence accumulation by a simple random walk (diffusion) process in a linear space, which is terminated by a set of complex decision boundaries as a stopping rule. However, if we nonlinearly constrain the evidence accumulation space, we can vastly simplify these boundaries and instead can use constant decision thresholds that are independent across options.

More specifically, there exists a variable transformation, $\phi_t : \boldsymbol{x}(t) \mapsto \boldsymbol{x}^\star(t) \equiv \boldsymbol{x} + \Delta_x 1$ with a scalar Δ_x, under which the optimal policy becomes equivalent to comparing each element $x_i^\star(t)$ to a constant threshold θ_x satisfying $\hat{r}_i(t, \theta_x) = \theta$. This variable transformation projects the states $\mathbf{x}$ onto an $(N-1)$-dimensional manifold M_θ that is differentiable everywhere and asymptotically approaches the plane $\{\boldsymbol{x} | x_i = \theta_x + (\sigma^2/\sigma_z^2 + t) c\delta t\}$ in the limit of $\forall j \neq i : x_j \to -\infty$ for each i, where σ^2 and σ_z^2 are the variances of likelihood and prior, respectively. The intersection of M_θ and the constant thresholds $x_i = \theta_x (\forall i)$ implements effectively the same decision policy as the original one (see Supplementary Note 2).

Moreover, for some fixed time t, this manifold M_θ is well approximated by the parameterized surface $\widetilde{M}_\theta = \left\{ \boldsymbol{x} | \frac{1}{N} \sum_i f(x_i) = u(t) \right\}$, where $f(x)$ is an arbitrary increasing, differentiable function that asymptotically approaches zero in the limit of $x_i \to -\infty$ and $u(t)$ is a scalar parameter. The variable transformation $\widetilde{\phi}_t : \boldsymbol{x}(t) \mapsto \widetilde{\boldsymbol{x}}^\star(t) \in \widetilde{M}_\theta$ is achieved by a recurrent neural process shown in Fig. 2c, which implements the following updated rule:

$$\boldsymbol{x} \leftarrow \boldsymbol{x} + 1\Delta_{\widetilde{x}} \tag{9}$$

$$\Delta_{\widetilde{x}} \leftarrow \Delta_{\widetilde{x}} + \gamma \left(u(t) - \frac{1}{N} \sum_i f(x_i + \Delta_{\widetilde{x}}) \right) \tag{10}$$

where γ is the update rate. The second equation finds the appropriate $\Delta_{\widetilde{x}}$, whereas the first equation performs the projection. This circuit comprises a nonlinear normalization of neural activities, $x_i^\star(t)$, controlled by an 'urgency signal', $u(t)$. Further, the circuit performs divisive normalization at a slower timescale (see equation (3)).

For subsequent simulations, we use the following sequence of discretized steps for each time step of incoming momentary evidence: (1) accumulate evidence according to equation (2); (2) project the newly accumulated evidence onto a nonlinear manifold by iterating equation (4) (or equations (9) and (10)) five times; (3) perform divisive normalization as in equation (3); and (4) add independent noise ξ_i on the individual output units (only for simulations corresponding to Figs. 5 and 6). We follow this sequence because we assume that the projection happens at a much faster timescale than divisive normalization (see main text). However, as we show in Supplementary Note 5, this particular order of time-discretized steps is inconsequential.

We found that a linear urgency signal, $u(t) = \beta t + u_0$, approximates well the collapse of the optimal decision boundaries. In this instance, β and u_0 are the slope and offset of the function, respectively, which we optimized in the subsequent simulations to maximize the reward rates. For the nonlinear function f, we used a rectified power function $f(x_i) = \lfloor x_i \rfloor^\alpha$, with the exponent fixed to $\alpha = 1.5$ (see Supplementary Fig. 3 for the dependence of optimal urgency signal on the nonlinearity). The update rate of the projection in equations (4) and (10) was fixed to $\gamma = 0.4$. We also fixed the gain of the divisive normalization term, K, to the mean reward across all trials and options, whereas σ_h was optimized. We ran the simulation for $T = 10$ s with time steps of $\delta t = 0.005$ s. We identified the optimal parameters (that is, the parameters that maximize the reward rate) with an exhaustive search followed by a simplex optimization[58]. For $N = 3$ and $N = 4$, the circuit was confirmed to yield near-optimal reward rates for a reasonably wide range of the mean reward (from $\bar{z} = 0$ to 5).

IIA violation, similarity effect and violation of the regularity principle. To simulate the third option effect that violates the IIA and regularity principles, and to reproduce the similarity effect, we perform simulations to reoptimize our optimal neural circuit for N choice options with independent variability added to each accumulator at every time step. We simulate the model for a fixed duration of $T = 200$ ms as in Louie et al.[11] with time steps of $\delta t = 1$ ms and pick the option with the highest accumulator value at the end of the trial. The rewards for the three options were chosen uniformly from $z_1 \in [25,35]$, $z_2 = 30$ and $z_3 \in [0,30]$. The momentary evidence was uncorrelated for the IIA and regularity principles with $\Sigma_x = \sigma 1$; for the similarity effect, the momentary evidence for two of the choice options was positively correlated with the correlation coefficient 0.1.

Statistics. Most figures are based on simulating our model using a sufficiently large number of trials (mentioned in the corresponding figure legends); this made the use of statistical testing unnecessary.

For Fig. 4c, we performed linear regression ($\mathrm{RT} = \beta_0 + \beta_1 \log(N+1)$) to predict the reaction time based on a logarithmic function of the number of choices, $\log(N+1)$, where N is the number of choices. We found a significant relation between reaction time and N for both value-based decisions ($P = 5.2 \times 10^{-4}$) with $R^2 = 0.9866$ (non-adjusted) and perceptual decisions ($P = 3.9 \times 10^{-5}$) with $R^2 = 0.9982$ (non-adjusted). Additional information can be found in the accompanying Life Sciences Reporting Summary.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data sharing is not applicable to this article since no datasets were generated or analyzed during the current study.

Code availability

These results of this article were generated using code written in MATLAB. The code is available at https://github.com/DrugowitschLab/MultiAlternativeDecisions.

References

51. Roe, R. M., Busemeyer, J. R. & Townsend, J. T. Multialternative decision field theory: a dynamic connectionist model of decision making. *Psychol. Rev.* **108**, 370–392 (2001).
52. Furman, M. & Wang, X. J. Similarity effect and optimal control of multiple-choice decision making. *Neuron* **60**, 1153–1168 (2008).
53. Albantakis, L. & Deco, G. The encoding of alternatives in multiple-choice decision making. *Proc. Natl Acad. Sci. USA* **106**, 10308–10313 (2009).
54. Teodorescu, A. R. & Usher, M. Disentangling decision models: from independence to competition. *Psychol. Rev.* **120**, 1–38 (2013).
55. Mahadevan, S. Average reward reinforcement learning: foundations, algorithms, and empirical results. *Mach. Learn.* **22**, 159–196 (1996).
56. BellmanR. E. *Dynamic Programming.* (Princeton Univ. Press, 1957).
57. Drugowitsch, J., Moreno-Bote, R. & Pouget, A. Optimal decision bounds for probabilistic population codes and time varying evidence. Preprint at *Nature Precedings* http://precedings.nature.com/documents/5821/version/1/files/npre20115821-1.pdf (2011).
58. Acerbi, L.& Ma, W. J. Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. *Adv. Neural Inf. Process. Syst.* **2017**, 1837–1847 (2017).

nature research

Corresponding author(s): Alexandre Pouget, Jan Drugowitsch

Last updated by author(s): May 31, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No data was collected during this study.
Data analysis	MATLAB R2018a. Code is available at https://github.com/DrugowitschLab/MultiAlternativeDecisions.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes (typically number of trials simulated) have been mentioned in the respective figure legends.
Data exclusions	This study did not involve data collection.
Replication	This study did not perform any experiments that require replication. The code is available as mentioned in the code availability statement.
Randomization	This study did not conduct any experiments that required randomization.
Blinding	This study did not conduct any experiments that required blinding.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
☒	□	Human research participants
☒	□	Clinical data

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

nature neuroscience

SUPPLEMENTARY INFORMATION

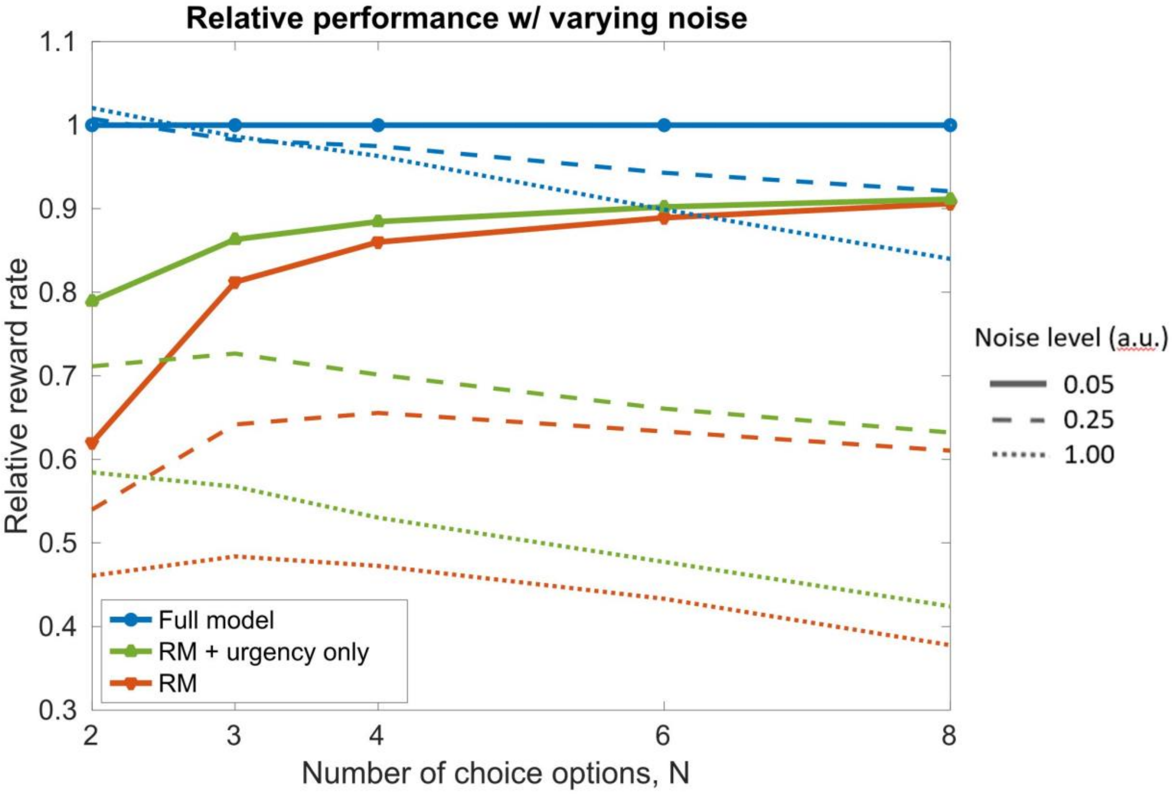
https://doi.org/10.1038/s41593-019-0453-9

In the format provided by the authors and unedited.

Optimal policy for multi-alternative decisions

Satohiro Tajima [1,4], Jan Drugowitsch [2,4]*, Nisheet Patel [1] and Alexandre Pouget [1,3]*

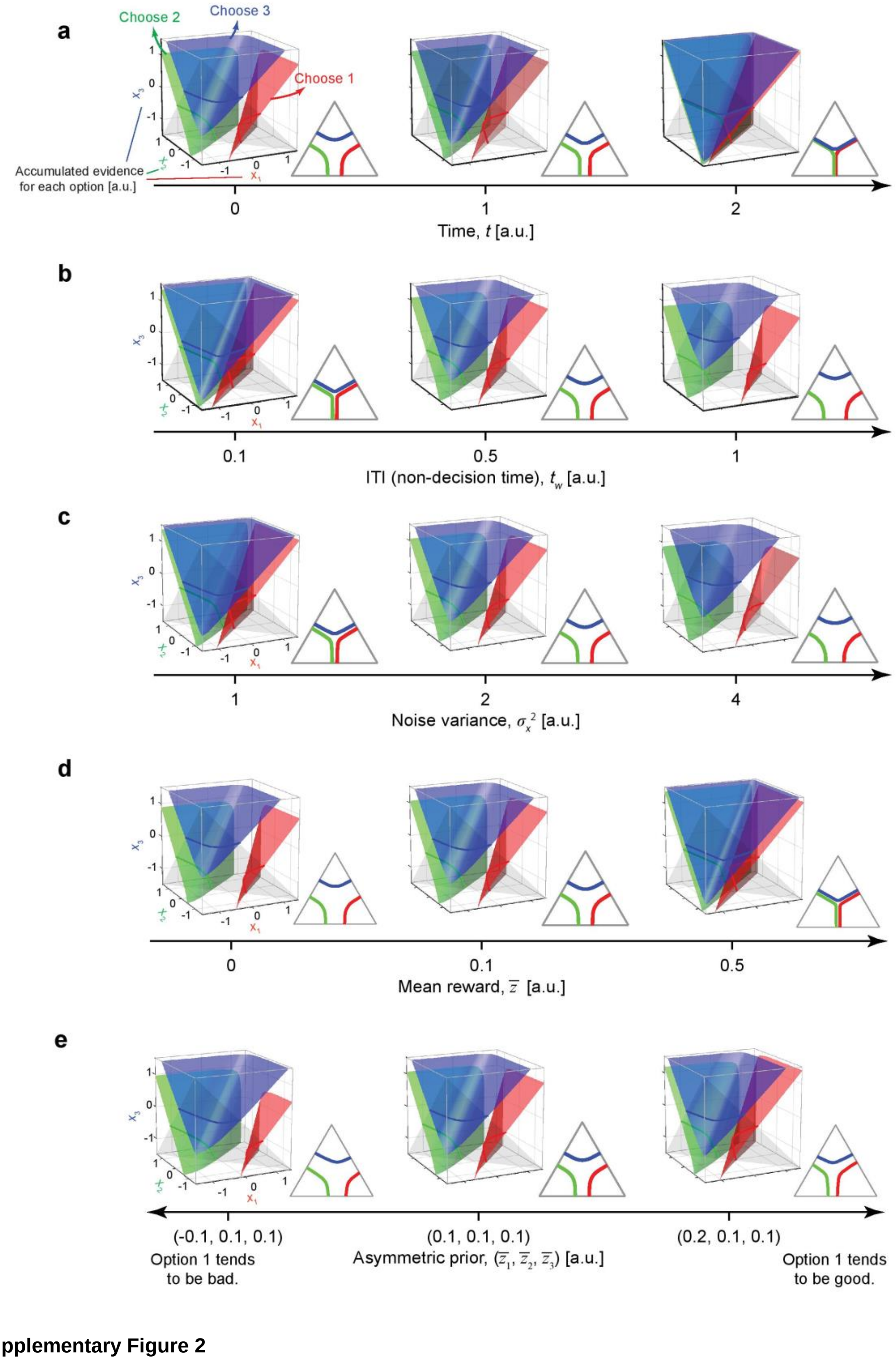
[1]Department of Basic Neuroscience, University of Geneva, Geneva, Switzerland. [2]Department of Neurobiology, Harvard Medical School, Boston, MA, USA. [3]Gatsby Computational Neuroscience Unit, University College London, London, UK. [4]These authors contributed equally: Satohiro Tajima, Jan Drugowitsch. *e-mail: jan_drugowitsch@hms.harvard.edu; Alexandre.Pouget@unige.ch



Supplementary Figure 1

Addition of variability to the accumulator affects models' relative performance

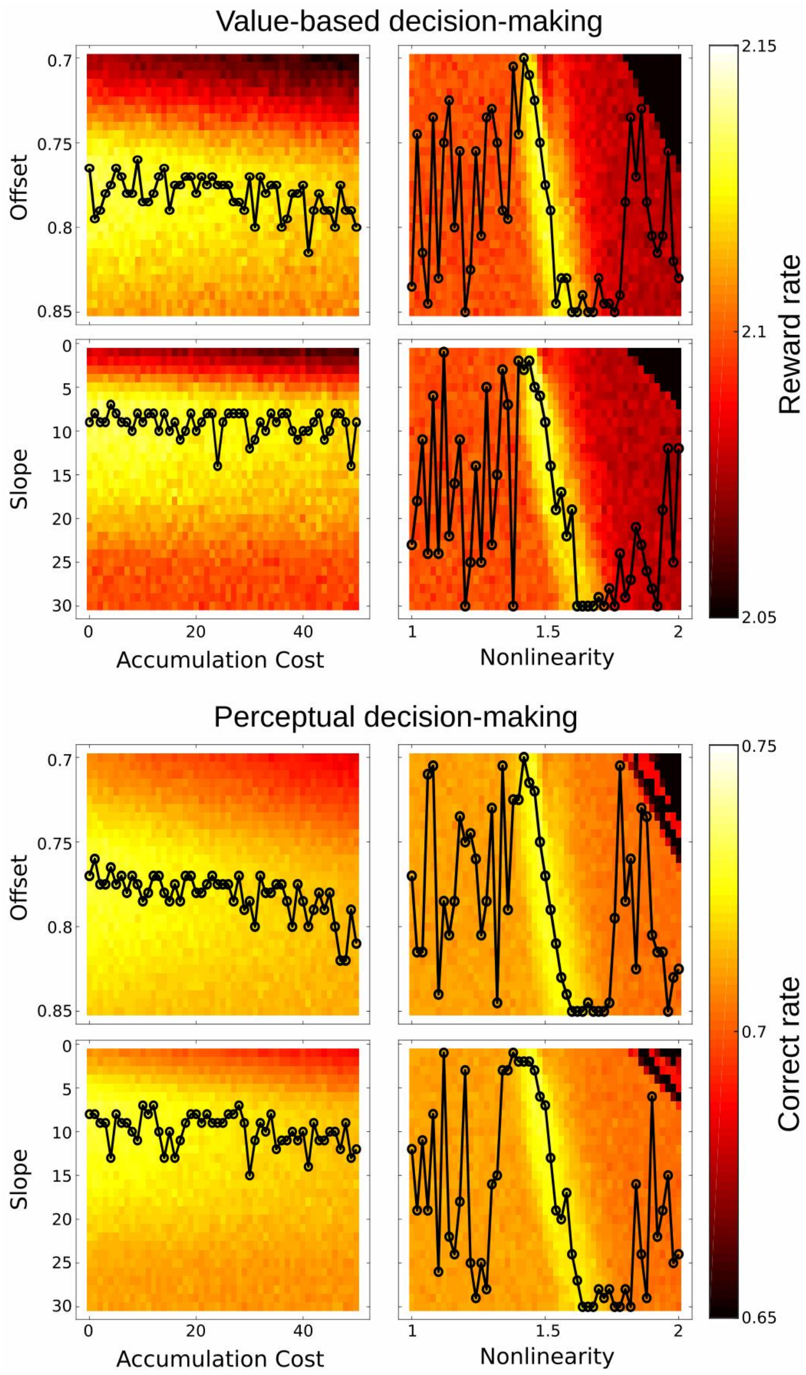
The race model variants without constrained evidence accumulation approximating the optimal policy perform much worse than our model's variants with that constraint, a result that is demonstrated in **Figure 5c**. Here, we show that reducing the amount of variability in the decision bounds brings the models' relative performances closer to each other as was the case in **Figure 3**. As in **Figures 3** and **5c**, this figure shows the reward rate of the race model with (green) and without (orange) the urgency signal relative to our full model with urgency and constrained evidence accumulation (blue). Each point represents the mean reward rate across 10^6 simulated trials.



Supplementary Figure 2

Dependencies of the stopping boundaries on task parameters.

We show how the decision boundaries change as a function of time (**a**), inter-trial interval (**b**), noise variance (**c**), and with symmetric (**d**) and asymmetric (**e**) prior mean of reward. (**a**) Dynamics of decision boundaries over time, t. The decision boundaries approach each other over time. Here, we used the following parameters: reward prior, $(\bar{z}_1, \bar{z}_2, \bar{z}_3) = (\bar{z}, \bar{z}, \bar{z}) = (0.1,0.1,0.1)$; inter trial interval (ITI, including non-decision time), $t_w = 0.5$; noise variance, $\sigma_x^2 = 2$. In (**b**)-(**e**) we varied a single parameter, while keeping all other parameters constant. The shown boundaries are the initial ones, at time $t = 0$. (**b**) Effect of inter trial interval (ITI), t_w. The boundaries start further apart for longer ITIs. $t_w = 0.5$ corresponds to the leftmost plot in panel **a**. (**c**) Effect of the evidence noise variance, σ_x^2. The boundaries start further apart for larger noise. $\sigma_x^2 = 2$ corresponds to the leftmost plot in panel **a**. (**d**) Effect of the reward prior mean, $\bar{z}$. The boundaries start closer to each other for larger mean rewards. $\bar{z} = 0.1$ corresponds to the leftmost plot in panel **a**. (**e**) Effect of the asymmetric reward prior, $(\bar{z}_1, \bar{z}_2, \bar{z}_3)$, where $\bar{z}_1$, $\bar{z}_2$, and $\bar{z}_3$ can be different from each other. The boundaries remain parallel to the cube diagonal but the asymmetric priors cause a shift of the boundary positions when projected on the triangle orthogonal to the diagonal, such that the boundaries corresponding to the most rewarded options start closer to the center of the triangle. $(\bar{z}_1, \bar{z}_2, \bar{z}_3) = (0.1,0.1,0.1)$ is identical to the leftmost plot in panel **a**. We have not been able to derive analytical approximations to the stopping bounds but note that the neural network provides a close approximation to the optimal bound with only three parameters. Given the shape and time dependence of the bounds, it is unlikely that it is possible to obtain an analytical solution with fewer parameters.



Supplementary Figure 3

The optimal urgency signal is only weakly dependent on accumulation cost and nonlinearity.

Each panel shows combinations of urgency signal parameters (vertical axis; offset or slope) and cost (left panels) or nonlinearity (right panels) setting the reward rate (value-based decisions; top) or correct rate (perceptual decision; bottom) as a color gradient. For each parameter combination, reward and correct rate were found by simulating 500,000 trials. The black line in each panel indicates for each cost or nonlinearity setting the value of the urgency signal parameter that maximizes the reward/correct rate. This line is noisy due to the simulation-based stochastic evaluation of the reward/correct rates. In general, both optimal slope and offset only weekly depend on the accumulation cost. The same applies to the nonlinearity, except for a narrow band around 1.5, where it is best to decrease both slope and offset for an increase in this nonlinearity.

Optimal policy for multi-alternative decisions

Supplementary Mathematical Note

Supplementary Mathematical Note

1 Structure of the value function and the optimal decision boundaries

The value function

In this section, we provide an analytic characterization of the decision boundary structure. To do so, we focus on the value function in the single-choice value-based decision tasks; the result for the reward rate case is not shown, but follows a similar analysis. Assume that $\boldsymbol{X}(t)$ is the stochastic process (or "decision variable") that describes the expected reward in N-dimensional space. Furthermore, assume that $\boldsymbol{X}(t)$ is shift-invariant, that is $\boldsymbol{X}(\tau) \mid (\boldsymbol{X}(t) + \boldsymbol{C}) = \left(\boldsymbol{X}(\tau) \mid \boldsymbol{X}(t)\right) + \boldsymbol{C}$, where $\tau \geq t$. For simple (even correlated) setups, this will hold. In particular, it holds for all cases discussed in the main text.

In this context, the value function is non-recursively given by

$$V(t, \boldsymbol{x}) = \max_{\tau \geq t} \left\langle \max_i X_i(\tau) - c(\tau - t) \middle| \boldsymbol{X}(t) = \boldsymbol{x} \right\rangle, \tag{1}$$

where the expectation is over the time-evolution of $\boldsymbol{X}$.

Below we show the value function to have the following properties:

1. $V(t, \boldsymbol{x} + \boldsymbol{1}\, C) = V(t, \boldsymbol{x}) + C$.
2. $V(t, \boldsymbol{x})$ is increasing in each element of $\boldsymbol{x}$.
3. $V(t, \boldsymbol{x}) \leq V(t, \boldsymbol{x} + \boldsymbol{e}_i\, C) \leq V(t, \boldsymbol{x}) + C$, where $\boldsymbol{e}_i$ is the ith basis vector of a Cartesian basis.
4. $V(t, \boldsymbol{x}) + \min_i C_i \leq V(t, \boldsymbol{x} + \boldsymbol{C}) \leq V(t, \boldsymbol{x}) + \max_i C_i$, where C_i is the i th element of $\boldsymbol{C}$.

Property 2 implies that $V(t, \boldsymbol{x})$ is continuous and differentiable. Thus, this property can be expressed as $\boldsymbol{\nabla}_x V(t, \boldsymbol{x}) \geq 0$, where the inequality is on each element of the gradient separately. As C in property 3 can be arbitrarily small, it is a generalization of property 2, such that we only need to show property 3. Property 1 is a special case of property 4 in which $\boldsymbol{C} = \boldsymbol{1}\, C$, such that $\min_\mathrm{i} C_i = \max_i C_i = C$.

Property 1

Fix some stopping times $\tau_1, \dots, \tau_N$. Then, the value function at time t is given by

$$\left\langle \sum_i 1_{\tau_i < \min_{j \neq i} \tau_j} X_i(\tau_i) - c(\min_i \tau_i - t) \,\middle|\, \boldsymbol{X}(t) = \boldsymbol{x} \right\rangle, \tag{2}$$

where the indicator function 1_a is 1 if a is true, and 0 otherwise. Thus, if we set the starting point to $\boldsymbol{x} + \boldsymbol{1}\, C$, we find

$$\left\langle \sum_i 1_{\tau_i<\min_{j\neq i}\tau_j} X_i(\tau_i) - c\left(\min_i \tau_i - t\right) \Big| \boldsymbol{X}(t) = \boldsymbol{x} + \boldsymbol{1}C \right\rangle$$

$$= \left\langle \sum_i 1_{\tau_i<\min_{j\neq i}\tau_j} (X_i(\tau_i) + C) - c\left(\min_i \tau_i - t\right) \Big| \boldsymbol{X}(t) = \boldsymbol{x} \right\rangle$$

$$= \left\langle \sum_i 1_{\tau_i<\min_{j\neq i}\tau_j} X_i(\tau_i) - c\left(\min_i \tau_i - t\right) \Big| \boldsymbol{X}(t) = \boldsymbol{x} \right\rangle + C, \qquad (3)$$

where the last line follows because the indicator function is only 1 for a single n. This is true for all choices of stopping times, and so also for the maximum over stopping times and choices.

Properties 2 and 3

Fix some integer k and stopping times $\tau_1, \ldots, \tau_N$. For starting point $\boldsymbol{x} + \boldsymbol{e}_k C$ we get

$$\left\langle \sum_i 1_{\tau_i < \min_{j\neq i}\tau_j} X_i(\tau_i) - c\left(\min_i \tau_i - t\right) \Big| \boldsymbol{X}(t) = \boldsymbol{x} + \boldsymbol{e}_k C \right\rangle$$

$$= \left\langle \sum_{i\neq k} 1_{\tau_i < \min_{j\neq i}\tau_j} \left(X_i(\tau_i)\right) + 1_{\tau_k < \min_{j\neq k}\tau_j} (X_k(\tau_k) + C) - c\left(\min_i \tau_i - t\right) \Big| \boldsymbol{X}(t) = \boldsymbol{x} \right\rangle$$

$$= \left\langle \sum_i 1_{\tau_i < \min_{j\neq i}\tau_j} X_i(\tau_i) - c\left(\min_i \tau_i - t\right) \Big| \boldsymbol{X}(t) = \boldsymbol{x} \right\rangle + 1_{\tau_k<\min_{j\neq k}\tau_j} C,$$

$$(4)$$

Note that, for the last term of the last line, $0 \leq 1_{\tau_k < \min_{j\neq k}\tau_j} C \leq C$, which upper-bounds the increase by C. The above again holds for an arbitrary set of stopping times, such that it also holds for the maximum over stopping times and choices.

Property 4

Following the same argument as in the preceding sections, we find for initial state $\boldsymbol{x} + \boldsymbol{C}$ and fixed stopping times that the value function is given by

$$\left\langle \sum_i 1_{\tau_i < \min_{j\neq i}\tau_j} X_i(\tau_i) - c\left(\min_i \tau_i - t\right) \Big| \boldsymbol{X}(t) = \boldsymbol{x} \right\rangle + \sum_i 1_{\tau_i<\min_{j\neq i}\tau_j} C_i, (5)$$

The last term is bounded by $\min_i C_i \leq \sum_i 1_{\tau_i<min_{j\neq i}\tau_j} C_i \leq \max_i C_i$, such that the result follows.

Characterizing the optimal decision boundaries

In this section we derive a few properties of the optimal decision boundaries, based on the above value function properties.

The expression for the optimal decision boundaries

Note that $V(t,\boldsymbol{x}) \geq \max_i x_i$. Furthermore, the decision maker ought to accumulate more evidence as long as $V(t,\boldsymbol{x}) > \max_i x_i$ and decide as soon as $V(t,\boldsymbol{x}) = \max_i x_i$. Let us assume that $x_1 > \max_{j>1} x_j$, such that, in case of a choice, option 1 ought to be chosen. The argument that follows is valid for all options, but we focus on option 1 for notational convenience. In this case, we have $x_j < x_1$ for all $j > 1$, and $V(t,\boldsymbol{x}) \geq x_1$. Furthermore, we accumulate evidence as long as $V(t,\boldsymbol{x}) > x_1$, and choose option 1 as soon as $V(t,\boldsymbol{x}) = x_1$. Note that $V(t,\boldsymbol{x})$ is increasing in $x_{2:N} \equiv x_2, \dots, x_N$, such that we will have $V(t,\boldsymbol{x}) > x_1$ for large $x_{2:N}$. Lowering $x_{2:N}$ will cause $V(t,\boldsymbol{x})$ to reduce until it reaches its lower bound, $V(t,\boldsymbol{x}) = x_1$, which is the point at which a decision ought to be made. Thus, the decision boundary is the "largest" $x_{2:N}$ (assuming natural vector ordering) at which $V(t,\boldsymbol{x}) = x_1$, or

$$B_1(t,x_1) \equiv \max \{x_{2:N} < x_1 \mid V(t,\boldsymbol{x}) = x_1\}, \tag{6}$$

where $x_{2:N} < x_1$ here denotes $x_j < x_1$ for all $j > 1$. This $B_1(t,x_1)$ is a set of points that, for a fixed x_1, define the boundary in $x_{2:N}$ at which a decision ought to be made. The above argument and resulting expression is valid for the decision boundaries associated with all options.

The decision boundaries are continuous and decreasing

To show that the decision boundaries are continuous, fix again x_1 such that $x_1 > \max_{j>1} x_j$. Furthermore, pick some x_2 and $x_2 + \delta$ that are both part of the vector elements of $B_1(t,x_1)$ (this restriction is necessary, as we cannot arbitrarily increase x_2 and still guarantee it to be part of the decision boundary). As the decision boundary is determined by the largest $x_{2:N}$ such that $V(t,\boldsymbol{x}) = x_1$, increasing x_2 while leaving all other elements constant will cause $V(t,\boldsymbol{x} + \boldsymbol{e}_2\,\delta) > x_1$. Therefore, we need to reduce another element of $x_{2:N}$ such that $V(t,\boldsymbol{x}) = x_1$ is again satisfied. As δ is arbitrarily small and $V(t,\boldsymbol{x})$ is increasing in all elements of $\boldsymbol{x}$, the decision boundary is continuous. Furthermore, as increasing one element of $x_{2:N}$ causes a decrease in other elements, the decision boundary as function of x_2 is decreasing in $x_{3:N}$.

The decision boundaries are "parallel" to the diagonal

Let us add a constant vector $\mathbf{1}C$ to all elements in $B_1(t,x_1)$. Defining $\boldsymbol{x}' = \boldsymbol{x} + \mathbf{1}C$, this results in

$$\begin{aligned} B_1(t,x_1) + \mathbf{1}C &= \max \{x_{2:N} < x_1 \mid V(t,\boldsymbol{x}) = x_1\} + \mathbf{1}C \\ &= \max \{x'_{2:N} < x'_1 \mid V(t,\boldsymbol{x}' - \mathbf{1}C) = x'_1 - C\} \\ &= \max \{x'_{2:N} < x'_1 \mid V(t,\boldsymbol{x}') = x'_1\} \\ &= B(t,x'_1) \\ &= B(t,x_1 + C). \end{aligned} \tag{7}$$

Thus, $B_1(t,\mathrm{x}_1 + C) = B_1(t,\mathrm{x}_1) + \mathbf{1}C$ which implies that the decision boundaries are parallel to the diagonal.

This implies that, for decision-making, only the accumulation space orthogonal to the direction given by $\mathbf{1}$ matters. Mapping onto this space could be achieved by $y_i = x_i - (N - 1)^{-1} \sum_{j \neq i} x_j$, or other arbitrary projections on $N - 1$ dimensional manifolds, which maps the accumulation into an $N - 1$ dimensional subspace.

2 Neural circuit implementation of the decision policy

In this section, we describe step-by-step the reason why the proposed recurrent neural circuit can approximate the optimal decision policy for N-alternative value-based decisions. Again, here we focus on the single-choice value-based decision tasks; the same arguments hold for the reward rate cases.

Decision boundaries as a set of manifold intersections

The optimal decision boundaries are determined by Bellman's equation,

$$V(t, \boldsymbol{x}) = \max\left\{\max_i \hat{r}_i(t, x_i), \langle V(t + \delta t, \boldsymbol{x})\rangle - c\ \delta t\right\}, \tag{8}$$

In the curled bracket, the first term corresponds to the value for deciding and choosing something right now, whereas the second term corresponds to the value for waiting (postponing the decision) to accumulate more evidence. Let us fix some time t. For this fixed time, the boundaries B between deciding and waiting are defined as a set of states where those two value functions equal to each other, i.e.,

$$B \equiv \left\{\boldsymbol{x} \,\middle|\, \max_i \hat{r}_i(t, x_i) = \langle V(t + \delta t, \boldsymbol{x})\rangle - c\ \delta t\ \right\}, \tag{9}$$

which is described as a set of intersections between the following two $N - 1$ dimensional manifolds,

$$L_\theta \equiv \left\{\boldsymbol{x} \,\middle|\, \max_i \hat{r}_i(t, x_i) = \theta\ \right\},$$

$$M_\theta \equiv \{\boldsymbol{x} \mid \langle V(t + \delta t, \boldsymbol{x})\rangle - \ c\ \delta t = \theta\ \}, \tag{10}$$

with a scalar reward parameter θ varied from $-\infty$ to ∞. L_θ and M_θ represent the level sets of value functions for choosing either option right now and for waiting to accumulating more evidence, respectively. Just as B, L_θ and M_θ are defined for some fixed time t. As shown in **Supplementary Math Note Figure 1a** and **S1b**, L_θ represents one corner of an N-dimensional hypercube (i.e., an orthant, as described later) that is intersected by M_θ. The point of this intersection corresponds to the part of the decision boundary that promises reward θ. Therefore, the complete set of decision boundaries B can be expressed as a "chain" of intersections between the two manifolds L_θ and M_θ, ordered by the reward parameter θ (**Supplementary Math Note Figure 1c**):

$$B = \{B_\theta \mid -\infty < \theta < \infty\} \tag{11}$$

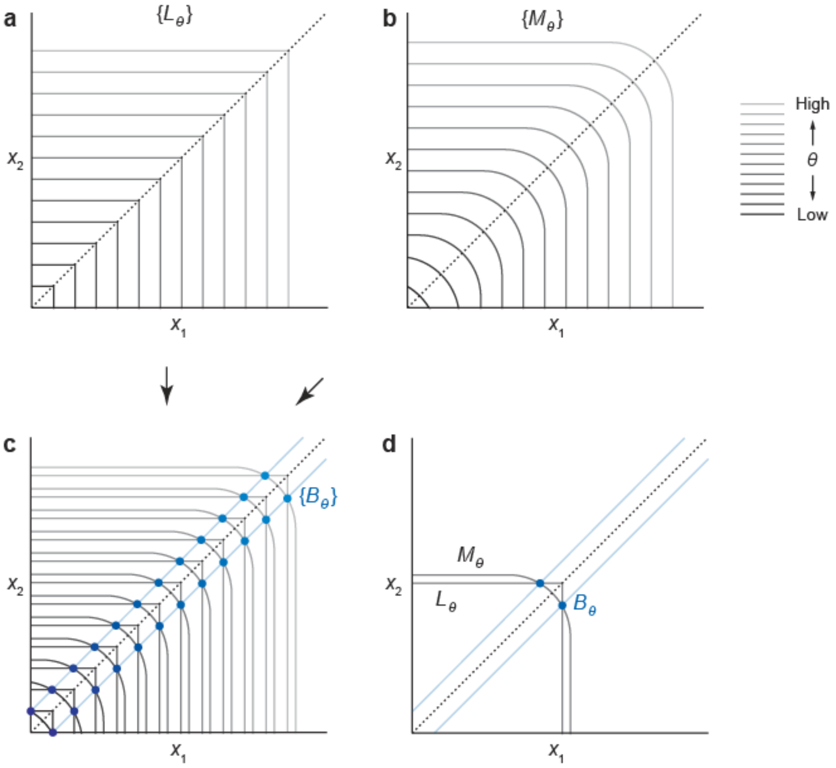
$$B_\theta \equiv L_\theta \cap M_\theta. \tag{12}$$

For each θ the dimensionality of B_θ is $N - 2$ because it is an intersection of two $N - 1$ dimensional manifolds, which makes the full decision boundary B an $N - 1$ dimensional manifold. Recall that all the value functions are shift-invariant in the dimension parallel to the diagonal, and that the set of decision boundaries, B, is "parallel" to the diagonal. Using this fact, B can be expressed in a different way as follows:

$$B = \left\{B_{\theta+\Delta_\theta} \,\middle|\, -\infty < \Delta_\theta < \infty\right\}$$

$$= \{B_\theta + \mathbf{1}\Delta_\theta \mid -\infty < \Delta_\theta < \infty\}, \tag{13}$$

where $+\mathbf{1}\Delta_\theta$ represents a translational shift of the set along the diagonal vector, $\mathbf{1}$, with a distance Δ_θ. Thus, rather than defining the set of all boundaries by the intersection B_θ between L_θ and M_θ for all reward levels θ (first line), we can define it as one such intersection B_θ for some arbitrary fixed θ, translated in directions of the diagonal $\mathbf{1}$ (second line).



Supplementary Math Note Figure 1. Manifold intersections define decision boundaries.

Schematic illustrations of the decision boundaries defined by manifold intersections. A two-alternative case is shown for the visualization purpose although the same argument applies to arbitrary N-alternative problems. (a) The manifold set $\{L_\theta\}$, which describes the value function for "deciding right now." (b) The manifold set $\{M_\theta\}$, which describes the value function for "waiting to accumulate more evidence." (c) The set of decision boundaries, $B \equiv \{B_\theta | -\infty < \theta < \infty\}$, is defined as a set of intersections of L_θ and M_θ. (d) Because the decision boundaries defined by different θ are all symmetric along the diagonal (the dashed line), we can consider a lower-dimensional projection by fixing θ.

Constrained states

As a next step we demonstrate that, if we restrict our evidence accumulation process state $\boldsymbol{x}$ to its projection $\boldsymbol{x}^*$ parallel to the diagonal $\mathbf{1}$ on the manifold M_θ, then we can make optimal choices as soon as this projected state reaches the manifold L_θ, which implies reaching $B_\theta = L_\theta \cap M_\theta$ (see also **Supplementary Math Note Figure 1d**). For now we assume some arbitrary fixed θ, but will later discuss that the argument is valid for any θ. More formally, fix some arbitrary θ (and some time t, as in the previous section) and consider the following map:

$$\phi_\theta : \mathbb{R}^N \to M_\theta, \tag{14}$$

$$\phi_\theta : \boldsymbol{x} \mapsto \boldsymbol{x}^* \equiv \boldsymbol{x} + \mathbf{1}\Delta_x. \tag{15}$$

which projects each state along the diagonal onto manifold M_θ. We call $\boldsymbol{x}^*$ the "constrained state." For a particular state $\boldsymbol{x} \in M_{\theta+\Delta_\theta}$ the extent Δ_x of this projection corresponds to $\Delta_x = -\Delta_\theta$, which yields the set of states that project into B_θ to be given by

$$\begin{aligned} \{\boldsymbol{x} \mid \exists \Delta_\theta : \boldsymbol{x} \in B_\theta + \mathbf{1}\,\Delta_x\} &= \{\boldsymbol{x} \mid \boldsymbol{x}^* \in B_\theta\} \\ &= \{\boldsymbol{x} \mid \boldsymbol{x}^* \in L_\theta\} \\ &= \left\{\boldsymbol{x} \,\middle|\, \max_i \hat{r}_i\,(t, x_i^*) = \theta\right\}, \end{aligned} \tag{16}$$

where the first equality follows from the definition of the projection, the second from the definition of B_θ as the intersection of M_θ and L_θ (recall that $\boldsymbol{x}^*$ is in M_θ by definition), and the third from the definition of L_θ. Note that by Equation (13) the states that project into B_θ form the set of all decision boundaries B, such that we can re-express the above as

$$\{\boldsymbol{x} \mid \boldsymbol{x} \in B\} = \left\{\boldsymbol{x} \,\middle|\, \max_i \hat{r}_i(t, x_i^*) = \theta\right\}, \tag{17}$$

showing that, as long as the accumulation process is constrained to states in M_θ, the decision boundary is formed by points on L_θ.

In the value-based case, as $\hat{r}_i(t, x_i)$ is an increasing function of x_i for each i, there exists a unique scalar $\theta_x \in \mathbb{R}$ such that $\hat{r}_i(t, \theta_x) = \theta$, with which the decision boundaries are described as

$$\{\boldsymbol{x} \mid \boldsymbol{x} \in B\} = \left\{\boldsymbol{x} \,\middle|\, \max_i x_i^* = \theta_x\right\}. \tag{18}$$

This equation shows that evaluating whether the state $\boldsymbol{x}$ hits a decision boundary or not is equivalent to evaluating whether the largest component of the constrained state $\boldsymbol{x}^*$ equals θ_x or not. Since the choice of θ is arbitrary, we can choose any θ that makes θ_x constant over time. With such a time-invariant θ_x, the decision policy is implemented simply by evaluating whether the largest component of $\boldsymbol{x}^*$ exceeds a fixed threshold. Note also that, because the value function is decreasing for each element x_i as described previously (**Supplementary Math Note 1**), the manifold M_θ is also a decreasing function for each element, thus the projection of the states $\boldsymbol{x}$ to M_θ is generally described as a mutual inhibition among the elements x_i corresponding to the individual options.

The structure of L_θ and M_θ

As a next step, we investigate the structure of the two manifolds in order to find functional forms that capture the symmetry of those manifolds.

As already described further above L_θ is a set of $N-1$ dimensional half-planes, $\{\boldsymbol{x} | x_i = \theta_x, x_{j\neq i} \leq \theta_x\}$ $(i = 1,\dots,N)$. These half-planes collectively form the sides of an N-dimensional orthant whose origin is at $\theta_x\mathbf{1} = (\theta_x, \theta_x,\dots,\theta_x)$ (**Supplementary Math Note Figure 1d**). Due to this straight-forward form, L_θ does not need to be approximated.

On the other hand, M_θ is a surface of a "smoothed orthant," which we define here as a differentiable $N-1$ dimensional manifold that asymptotically approaches $L_{\theta'}$ $(\exists\theta')$ in the limit of $\forall j \neq i : x_j \to -\infty$ for each i (**Supplementary Math Note Figure 1d**); this is because when all the options except for option i have infinitely low values, the decision-maker should choose option i, which makes the value for waiting equal the value for choosing option i minus the cost of time. In particular, from the Bellman equation and the Bayes rule applied to our setup, θ' could be defined explicitly as

$$\theta' = \theta_x + \left(\frac{\sigma^2}{\sigma_z^2} + t\right) c\,\delta t, \tag{19}$$

where σ^2 and σ_z^2 are the variances of the evidence noise and the prior, respectively. Using the symmetry along the diagonal line,

$$L_{\theta'} = L_{\theta_x} + \left(\frac{\sigma^2}{\sigma_z^2} + t\right) c\,\delta t\,\mathbf{1}, \tag{20}$$

This equation implies that $L_{\theta'}$ and thus M_θ move along the diagonal as time elapses.

Due to the symmetry, M_θ is invariant to permutations of the coordinates, $1,\dots,N$. Thus, at the point $\boldsymbol{x} \in M_\theta$ such that $\forall\, i,j : x_i = x_j$, M_θ is orthogonal to the N-dimensional diagonal line, $\{\boldsymbol{x} | \forall\, i,j : x_i = x_j\}$. Note that M_θ has only one intersection with the diagonal line due to the decreasing property as we have described further above. Furthermore, as understood intuitively, if option j's value is very low, the problem becomes effectively a comparison among the remaining $N-1$ options, $\{1,\dots,N\} \setminus$ j. Because we have the same symmetry as before but now among those $N-1$ options; i.e., in the limit of $\forall\, j : x_j \to -\infty$, M_θ is orthogonal to the vector $\mathbf{1}^{\backslash j}$ that is defined by $\mathbf{1}_i^{\backslash j} = 1 - \delta_{ij}$ with Kronecker's delta, when $\boldsymbol{x} \in M_\theta$ satisfies $\forall\, i, i' \neq j : x_i = x_{i'}$. Similarly, if two options j and j' have infinitely low values, the effective problem becomes to compare the remaining $N-2$ options, $\{1,\dots,N\} \setminus \{j,j'\}$, then the manifold is orthogonal to $\mathbf{1}^{\backslash\{j,j'\}}$ (where $\mathbf{1}_i^{\backslash\{j,j'\}} = 1 - \delta_{ij}\delta_{ij'}$) when $\boldsymbol{x} \in M_\theta$ satisfies $\forall\, i, i' \notin \{j,j'\} : x_i = x_{i'}$. Repeating the same argument reveals the whole hierarchy of symmetries in the manifold M_θ. Note that when $\forall\, j \neq i : x_j \to -\infty$, the problem reduces to choosing from only one option i; at this limit, M_θ is orthogonal to $\mathbf{1}^{\backslash(\{1,\dots,N\}\backslash i)} = \boldsymbol{e}_i$, agreeing with the aforementioned property that M_θ asymptotically approaches $L_{\theta'}(\exists\theta')$.

These properties of M_θ are well-captured by a manifold defined as follows:

$$\widetilde{M}_\theta = \left\{ \boldsymbol{x} \,\middle|\, \frac{1}{N}\textstyle\sum_i f(x_i) = u \right\}, \tag{21}$$

where $f(x_i)$ is an arbitrary increasing, differentiable function that asymptotically approaches zero in the limit of $x_i \to -\infty$. Here, u is a scalar value, which generally increases with elapsed time to capture the time-dependent property of M_θ. Moreover, by varying the functional form of f and the value of parameter u, we can make $\widetilde{M}_\theta$ have the same position and curvature as M_θ at its intersection with the diagonal line, $\left\{ \boldsymbol{x} \,\middle|\, \boldsymbol{x} \in \widetilde{M}_\theta,\ \forall i,j : x_i = x_j \right\}$. This indicates that $\widetilde{M}_\theta$ can be a good approximation of M_θ around its intersection with the diagonal line, and thus $\tilde{B}_\theta \equiv L_\theta \cap \widetilde{M}_\theta$ approximates B_θ well at points close to the diagonal line. The approximation around the diagonal line is particularly important because the assumed unbiased prior over rewards requires the initially expected rewards (at $t = 0$, before accumulating any evidence) to be symmetric across options, such that the decision variable $\boldsymbol{x}(t)$ fluctuates around the diagonal line.

A recurrent circuit that approximates the constraining manifold

We design a neural mechanism that constrains the neural population activity that encodes evidence accumulation to the manifold $\widetilde{M}_\theta$. Consider a map that projects each state along the diagonal onto the manifold $\widetilde{M}_\theta$, as follows:

$$\tilde{\phi}_\theta : \mathbb{R}^N \to \widetilde{M}_\theta, \tag{22}$$

$$\tilde{\phi}_\theta : \boldsymbol{x} \mapsto \tilde{\boldsymbol{x}}^* \equiv \boldsymbol{x} + \mathbf{1}\Delta_{\tilde{x}}, \tag{23}$$

Based on the arguments in the previous sections (Supplementary Note 1), the decision boundary $\{\boldsymbol{x} \mid \boldsymbol{x} \in B\}$ is approximated by $\left\{ \boldsymbol{x} \,\middle|\, \max_i \tilde{x}_i^* = \theta_x \right\}$. That is, evaluating whether the state x hits a decision boundary or not is equivalent to evaluating whether the largest component of the constrained state $\tilde{x}^*$ equals θ_x or not. Again, $\widetilde{M}_\theta$ also depends on time, thus so does the map $\tilde{\phi}_\theta$. As shown in the previous section, the map $\tilde{\phi}_\theta$ can be implemented by a circuit that computes $\Delta_{\tilde{x}}$ satisfying the following property:

$$\tilde{\boldsymbol{x}}^* \in \widetilde{M}_\theta \Leftrightarrow \frac{1}{N}\textstyle\sum_i f(\tilde{\boldsymbol{x}}_i) = u, \tag{24}$$

$$\Leftrightarrow u - \frac{1}{N}\textstyle\sum_i f(x_i + \Delta_{\tilde{x}}) = 0. \tag{25}$$

If we define $E \equiv \frac{1}{2}\left(u - \frac{1}{N}\sum_i f(x_i + \Delta_{\tilde{x}})\right)^2$ its gradient is given by

$$-\frac{\partial E}{\partial \Delta_{\tilde{x}}} = \left(\frac{1}{N}\textstyle\sum_i f'(x_i + \Delta_{\tilde{x}})\right)\left(u - \frac{1}{N}\sum_i f(x_i + \Delta_{\tilde{x}})\right). \tag{26}$$

Note that the first term on the left hand side of the equation is always zero or positive because f is an increasing function as described in the previous section. Therefore, the following update rule is able to find $\Delta_{\tilde{x}}$ by approximate gradient descent:

$$\Delta_{\tilde{x}} \longleftarrow \Delta_{\tilde{x}} + \gamma \left(u - \frac{1}{N} \sum_i f(x_i + \Delta_{\tilde{x}}) \right), \tag{27}$$

where γ is a small positive scalar that determines the update rate. The update process terminates when the term in the parenthesis becomes zero. This update rule is implemented by the recurrent neural circuit with activity normalization (implementing the projection) and urgency signal (realizing the time-variant nature of M_θ) as mentioned in the main text.

Corresponding to the update of $\Delta_{\tilde{x}}$, each neuron's output is updated as follows:

$$f(\tilde{x}_i) \longleftarrow f(\tilde{x}_i + \Delta_{\tilde{x}}). \tag{28}$$

Given that f is invertible in $\tilde{x}_i \geq 0$, this update is the same as $\tilde{x}_i \leftarrow \tilde{x}_i + \Delta_{\tilde{x}}$. In our implementation, the projection was performed by applying Eqs. (27) and (28) 5 times for each evidence input at time t, assuming that the relaxation of neural activity is faster compared to the time scale of the evidence dynamics. Within every time step t, $f(\tilde{x}_i)$ was after the projection compared to a constant threshold $\theta \equiv f(\theta_x)$, which, for the monotonically increasing f, is equivalent to comparing $\tilde{x}_i$ with θ_x.

3 Experimental predictions

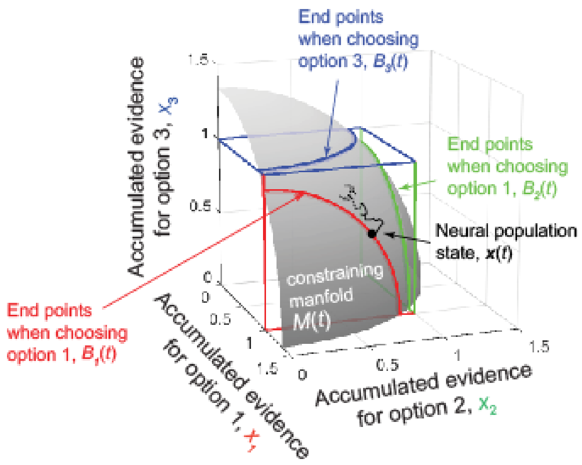
Here we provide a list of detailed predictions derived from our theoretical results. All of them can be tested with neurophysiological or behavioral experiments. We consider human or animal subjects performing a standard N-alternative value-based decision-making task with a reaction-time paradigm, in which the subjects choose one of N options at their own pace, while trying to maximize the total reward within a session of fixed duration (i.e., they can make more choices if each of them is faster).

Suppose that we record the activity of a decision-related neuronal population (serially or simultaneously) during the task. We denote the entire population state by $\boldsymbol{x}(t) = \left(x_1(t), \dots, x_D(t)\right)$, where D is the number of recorded neurons.

Physiological predictions

Here we provide a list of detailed predictions derived from our theoretical results. All of them can be tested with neurophysiological or behavioral experiments. We consider human or animal subjects performing a standard N-alternative value-based decision-making task with a reaction-time paradigm, in which the subject choose one of N options at their own pace, while trying to maximize the total reward within a session of fixed duration (i.e., they can make more choices if each of them is faster).

Suppose that we record the activity of a decision-related neuronal population (serially or simultaneously) during the task. We denote the entire population state by $x(t) = \left(x_1(t), \dots, x_{\mathrm{D}}(t)\right)$, where D is the number of recorded neurons.



Supplementary Math Note Figure 2. The manifold constraining neural population state.

Predictions about the dynamics of evidence accumulation

1. The neural population activity is constrained on a low-dimensional manifold: at each time point t, the neural population state $\boldsymbol{x}(t)$ is constrained on an $N-1$ dimensional manifold, $\mathcal{M}(t)$ (the 'constraining manifold', the gray surface in **Supplementary Math Note Figure 2**).

2. The $N-1$ dimensional manifold is nonlinear: although the topological dimensionality (i.e., a locally defined dimensionality) of the constraining manifold $\mathcal{M}(t)$ is $N-1$, $\mathcal{M}(t)$ is curved and embedded within an N-dimensional space.

3. The $N-1$ dimensional manifold evolves over time: the constraining manifold $\mathcal{M}(t)$ varies over time, meaning that the $N-1$ dimensional manifold can only be observed for a fixed time. Otherwise, we would only observe an N-dimensional structure, resulting from the $N-1$ dimensional manifold being smeared out over time.

4. The effects of prior belief: the position and the shape of the constraining manifold $\mathcal{M}(t)$ depend on the prior knowledge about the value distribution (e.g., mean value over trials), but the dimensionality is always $N-1$.

5. The stability against the trial-to-trial option contingency: the constraining manifold $\mathcal{M}(t)$ is invariant to the option values within each trial. That is, if we compare a trial set with N high-valued options against another trial set in which low and high values are mixed, the neural state trajectories can differ between those trial sets, but all the trajectories are constrained on the same $N-1$ dimensional manifold $\mathcal{M}(t)$.

6. The constraining manifold $\mathcal{M}(t)$ has a hierarchical symmetry as follows: on $\mathcal{M}(t)$, $x_i(t)$ is a decreasing function of $x_j(t)$ for all $i \neq j$. In particular, $\mathcal{M}(t)$ is orthogonal to the vector $(1,1,\ldots,1)$ when the neural state $\boldsymbol{x}(t)$ is nearly proportional to

$(1,1,\dots,1)$—which happens when all the options have equal values. $\mathcal{M}(t)$ is orthogonal to the vector $(1,1,\dots,1,0)$ when $\boldsymbol{x}(t)$ is nearly proportional to $(1,1,\dots,1,0)$ —which happens when we have $N-1$ equally high-valued options and a low-valued option

7. The uniqueness of the manifold over time: for two different time points t and t', the constraining manifolds $\mathcal{M}(t)$ and $\mathcal{M}(t')$ do not intersect with each other.

8. The diffusion process can be recovered by a renormalization: when we renormalize the neural activity $\boldsymbol{x}(t)$ by projecting it onto an $N-1$ dimensional hyperplane (the triangle in Fig. 2a) orthogonal to diagonal vector $(1,1,\dots,1)$, the temporal evolution of population state on this plane is a standard $N-1$ dimensional diffusion process. Namely, the variance of temporal derivative of neural population activity is uniform over time in those dimensions.

9. The effect of opportunity cost: the position and speed of the constraining manifold $\mathcal{M}(t)$ depend not only on the number of options but also on the reward rate. This means that the offset activity of neurons should depend on the average reward size over trials or inter-trial interval, not only on the number of options.

Predictions about the neural states at the termination of evidence accumulation

Let $\boldsymbol{x}(t|\text{choose } i)$ denote the neural population state at the end of evidence accumulation, right before choosing option i.

10. The low-dimensional structure of the neural activity at the end of evidence accumulation: the neural activity when the stopping boundary is hit, $\boldsymbol{x}(t|\text{choose } i)$, is constrained on the $N-2$ dimensional manifold $B_i(t)$ (the 'end-point manifold') that is defined as the intersection of the constraining manifold $\mathcal{M}(t)$ and a hyper-plane, $x_i(t|\text{choose } i) = \theta_i$, where $x_i(t|\text{choose } i)$ is the ith compornent of $\boldsymbol{x}(t|\text{choose } i)$, and θ_i is a constant which is invariant to option sets.

11. The symmetry in the end-point manifolds: for different options i and j, two end-point manifolds $B_i(t)$ and $B_j(t)$ do not intersect with each other. Moreover, on each $B_i(t)$, $x_j(t|\text{choose } i) < \theta_j$ for all $j \neq i$, and the distance between $B_1(t)$ and $B_2(t)$ is almost constant when the state $\boldsymbol{x}(t)$ is proportional to $(1,1,0,\dots,0,0)$.

Behavioral predictions

12. The choice accuracy depends on time, the option set size, and the reward: the choice accuracy (the frequency of choosing the best options) decreases with reaction time. The choice accuracy also depends on the number of options as well as the reward rate.

13. The transitions of behavior between the 'max-vs.-next' and the 'max-vs.-average' strategy within a same task: in trials with N almost equally-valued options, or with one high-

valued option and $N-1$ low-valued options, the subject behavior (e.g., the reaction time dependency on choice contexts) is similar to what is predicted by the ‘max-vs.-average’ strategy (i.e., the strategy such that the decision is driven by the difference between the best option and the average of all the options.). In trials with two high-valued options and $N-2$ low-valued options, the subject behavior is similar to what is predicted by the ‘max-vs.-next’ strategy (i.e., the strategy such that the decision is driven by the difference between the best and the second-best options.). In other trials, the results differ from either of ‘max-vs.-average’ and ‘max-vs.-next’ strategies.

4 Evidence with short- and long-range temporal correlations

Here we consider the modifications required to the optimal policy if the evidence features temporal short- and long-range correlations. In our discussion, we focus on positive temporal correlations. Similar arguments can be made for negative correlations. Both short- and long-range correlations reduce the amount of information available to the decision maker per unit time, but in different ways.

Short-range correlations

In the main text we have assumed the momentary evidence to be drawn i.i.d. according to $\delta \boldsymbol{x}_n | \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}\delta t, \boldsymbol{\Sigma}_x \delta t)$, resulting in $\mathrm{cov}(\delta \boldsymbol{x}_n, \delta \boldsymbol{x}_m) = \delta_{mn} \boldsymbol{\Sigma}_x \delta t$, where $\delta_{mn} = 1$ if $m = n$, and $\delta_{mn} = 0$ otherwise. That is, the momentary evidence provides independent information about $\boldsymbol{z}$ within each small time-bin. This makes summing up this momentary evidence the optimal thing to do.

Let us now consider what happens if the momentary evidence becomes correlated across time. We first focus on short-range correlations, which could arise if the momentary evidence with white noise is passed through a circuit that low-pass filters this evidence. Then, we have $\mathrm{cov}(\delta x_n, \delta x_{n+m}) > 0$ for sufficiently small time-differences $m\delta t$, and an autocorrelation that drops to zero with increasing $m\delta t$. What is the impact of these correlations on evidence accumulation and the optimal decision policy?

One effect of such correlations is that the amount of independent information about $\boldsymbol{z}$ per unit time is reduced. This is because consecutive pieces of momentary evidence are correlated, such that their associated noise does not average out when summing them. However, as long as the momentary evidence's auto-correlation structure is known and sufficiently well-behaved, we can apply a linear filter to the incoming momentary evidence to whiten it (e.g., Papoulis & Pillai, 2002[1]). This will result in another stream of momentary evidence that has the overall same amount of information about $\boldsymbol{z}$, but whose individual pieces of evidence are independent across time. Thus, the re-formatted momentary evidence satisfies the assumptions underlying the model developed in the main text, such that its conclusions still apply.

What are the limitations of this approach? First, temporal whitening of the momentary evidence requires knowledge of its auto-correlation structure. Knowing the statistics of the incoming evidence in a general pre-requisite to finding the optimal stopping boundaries, as finding them involves computing an expectation over potential future values of the accumulated evidence. Without knowing these statistics, we would not be able to find optimal stopping boundaries. This also applies to the auto-correlation structure. Second, despite temporal correlations, the amount of information per unit time needs to remain constant. For the original i.i.d. momentary evidence, this was satisfied by a likelihood $p(\delta \boldsymbol{x}_n | \boldsymbol{z})$ that had a fixed covariance structure. For correlated momentary evidence, this remains satisfied as long as its auto-correlation structure does not vary across time. Similar approaches to identifying optimal stopping boundaries can also be applied to scenarios in which the informativeness of momentary evidence fluctuates across time, but the resulting policies will become significantly more complex (e.g., Drugowitsch et al. , 2014[2]).

Long-range correlations

In the context of long-range correlations, we consider the case where the momentary evidence associated with each option is offset by a random, unknown, amount that is fixed within each trial. Let us denote this amount y_j for option j and assume it to be drawn independently in each trial from a zero-mean Gaussian with variance σ_y^2, that is $y_j \sim \mathcal{N}(0, \sigma_y^2)$. The momentary evidence in this trial is then drawn according to $\delta x_{j,n} | y_j, z_j \sim N(z_j + y_j, \sigma_x^2)$.

One extreme of this scenario are ballistic accumulator models[3,4], in which the only stochastic element is y_j, whereas the momentary evidence is noise-free, that is $\sigma_x^2 = 0$. In this case it becomes superfluous to accumulate evidence, as evidence samples beyond the first do not yield any additional information. Thus, the optimal policy would be to await this first sample and decide immediately after that.

For noisy momentary evidence, when $\sigma_x^2 > 0$, it remains optimal to accumulate momentary evidence. In this case, the only impact of an unknown y_j is that the prior over the mean of the momentary evidence becomes less certain. Specifically, it grows in variance by σ_y^2. As a consequence, we can handle this case by ignoring the y_j's, while at the time widening the prior over z by σ_z^2. Therefore, it reduces to the case discussed in the main text, and results in the same optimal policy.

In summary, both short- and long-term temporal correlations in the momentary evidence reduce the amount of evidence we have about the true values of the underlying latent states $\boldsymbol{z}$. Short-term correlations do so by reducing the information in each piece of evidence, effectively making the likelihood less certain. Long-term correlations, in contrast, make the underlying mean less certain, effectively making the prior less certain. Both cases thus impact particular parameters of the model, while leaving the general structure unchanged. Therefore, while they impact the optimal decision boundaries quantitatively, they don't change our conclusions qualitatively.

5 Discrete implementation of circuit with divisive normalization

In this section, we derive equations for our circuit model that approximates the optimal policy. Our network operates at two time-scales. On the slower time-scale, neurons accumulate (noisy) momentary evidence independently across options according to:

$$\boldsymbol{x_t} \; = \; C_t \, \delta \boldsymbol{x}_t + \frac{C_t}{C_{t-1}} \boldsymbol{x}_{t-1} \tag{30}$$

where $\boldsymbol{x_t}$ is the vector of accumulated evidence at time t, $\delta \boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{z}_t dt, \Sigma_t dt)$ is the vector of momentary evidence at time t, where $\boldsymbol{z}_t$ is the vector of "true" rewards, Σ_t is a covariance matrix that makes the momentary evidence noisy, and C_t is the commonly used divisive normalization term[5,6]:

$$C_t = \frac{K}{\sigma_h \; + \; \sum_{n=1}^{N} x_n(t)} \tag{31}$$

On the faster time scale, activity is projected onto a manifold defined by $\frac{1}{N}\sum_i f(x_i) = u(t)$, (shown as a gray surface in **Supplementary Math Note Figure 2**) where $u(t)$ is the urgency signals. This operation is implemented by iterating:

$$x_i \leftarrow x_i + \gamma \left(u(t) - \frac{1}{N} \sum_i f(x_i) \right) \tag{32}$$

until convergence, where γ is the update rate and f is a rectified polynomial non-linearity. Ignoring the fast dynamics, Equation (30) can be rearranged to get:

$$\boldsymbol{x_t} - \boldsymbol{x}_{t-1} \; = \; C_t \, \delta \boldsymbol{x}_t \; + \frac{C_t - C_{t-1}}{C_{t-1}} \boldsymbol{x}_{t-1} \tag{33}$$

$$d\boldsymbol{x}_t = C_t \left(\boldsymbol{z}_t dt + \Sigma_x^{\frac{1}{2}} dW_t \right) + \frac{dC_t}{C_t} \boldsymbol{x}_t \tag{34}$$

$$\frac{dx_j(t)}{dt} = C(t) \delta \boldsymbol{x}_t + \frac{1}{C(t)} \frac{dC(t)}{dt} x_j(t) \tag{35}$$

From Equation (31), it can be shown that

$$\frac{dC(t)}{dt} = \; - \frac{C^2(t)}{K} \sum_{n=1}^{N} \frac{dx_n(t)}{dt} \tag{36}$$

Combining Equations (35) and (36), we get

$$\frac{dx_j(t)}{dt} = C(t) \delta \boldsymbol{x}(t) - \frac{C(t)}{K} \, x_j(t) \sum_{n=1}^{N} \frac{dx_n(t)}{dt} \tag{37}$$

Equation (37) captures the slow dynamics of our circuit model. On a faster time-scale $\tau \ll dt$, this activity is further projected on to the manifold $\frac{1}{N}\sum_i f\big(x_i(t)\big) = u(t)$, which can be expressed as:

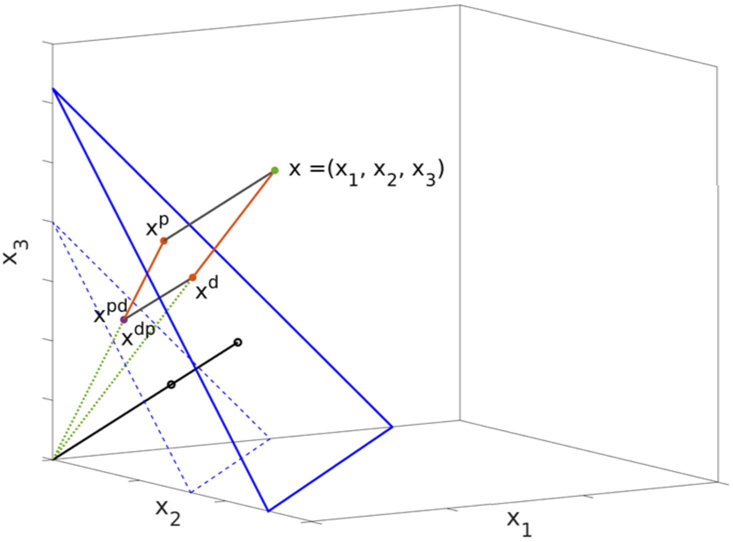
$$\frac{dx_j(t)}{dt} = C(t)\left(\delta\boldsymbol{x}(t) + \frac{1}{\tau}\left(u(t) - \frac{1}{N}\sum_{n=1}^{N} f\left(\frac{x_n(t)}{C(t)}\right)\right) - \frac{1}{K}\,x_j(t)\sum_{n=1}^{N}\frac{dx_n(t)}{dt}\right) \qquad (38)$$

This can be written in vector form as:

$$\frac{1}{C(t)}\frac{d\boldsymbol{x}(t)}{dt} = \delta\boldsymbol{x}(t) + \frac{1}{\tau}\left(u(t) - \frac{1}{N}\left(\mathbb{1}\cdot f\left(\frac{\boldsymbol{x}(t)}{C(t)}\right)\right)\right) - \boldsymbol{x}(t)\left(\mathbb{1}\cdot\frac{d\boldsymbol{x}(t)}{dt}\right) \qquad (39)$$

$$= \delta\boldsymbol{x}(t) + \frac{1}{\tau}\left(u(t) - \frac{1}{N}\left(\mathbb{1}\cdot f\left(\frac{\boldsymbol{x}(t)}{C(t)}\right)\right)\right) - \boldsymbol{x}(t)\left(\frac{C(t)\big(\mathbb{1}\cdot\delta\boldsymbol{x}(\mathrm{t})\big)}{K + C(t)\big(\mathbb{1}\cdot\boldsymbol{x}(t)\big)}\right) \qquad (40)$$

where the explicit Equation (40) follows by summing Equation (39) and substituting the last term on its right-hand side.



Supplementary Math Note Figure 3. Geometry of the projection and divisive normalization.

Geometric depiction of the projection due to the linear constraint ($u(t) - N^{-1}\sum_n^N x_n(t) = 0$) and divisive normalization. Note that the nonlinearity in ($u(t) - N^{-1}\sum_n^N f(x_n(t)) = 0$) makes the blue triangle (plane) a curved surface, thereby disallowing a closed-form solution. However, the geometric intuition for projection and divisive normalization remains the same. $\boldsymbol{x}$ is an arbitrary initial point. $\boldsymbol{x}^p$ is obtained by projecting $\boldsymbol{x}$ according to the linear constraint mentioned above. Further applying divisive normalization to $\boldsymbol{x}^p$ gives $\boldsymbol{x}^{pd}$, *i.e.* $\boldsymbol{x}^{pd} = C\boldsymbol{x}^p$, where C is defined in Eq. 31. On the other hand, $\boldsymbol{x}^d$ is obtained by applying divisive normalization to $\boldsymbol{x}$, *i.e.* $\boldsymbol{x}^d = C\boldsymbol{x}$, and further projecting $\boldsymbol{x}^d$ according to the linear constraint mentioned above gives $\boldsymbol{x}^{dp}$. In the following section, we analytically show that $\boldsymbol{x}^{dp} = \boldsymbol{x}^{pd}$, or that the diffusion is unaffected if one were to apply the constraint first followed by divisive normalization or *vice versa*.

Optimal evidence accumulation

Our full model approximates the optimal policy by accumulating evidence and then projecting it on a non-linear manifold at each time-step. If we rescale the entire evidence accumulation space after this process at each time-step, then the relative distances between the accumulators and decision threshold are preserved, leaving the choices optimal. Mathematically, divisive normalization rescales the evidence accumulation space. As we just argued, doing so after the projection preserves optimality.

However, different instances of the discrete implementation may reverse this order – one may perform the rescaling before projecting at each time-step. Fortunately, it is possible to show analytically that, at least in the linear case, i.e. when $f(x) = ax + b$, the order is irrelevant. For simplicity, but without loss of generality, we will show this for $f(x) = x$.

To formally analyze this problem, we make use of some geometric intuition (see **Supplementary Math Note Figure 3)**. In 3-dimensional evidence accumulation space, consider an arbitrary initial point, $\boldsymbol{x} = (x_1, x_2, x_3)$, that has not hit any decision boundary. If we were to apply divisive normalization to this point, we would get $\boldsymbol{x}^d = (x_1^d, x_2^d, x_3^d)$, and further applying the projection would yield $\boldsymbol{x}^{dp} = (x_1^{dp}, x_2^{dp}, x_3^{dp})$, where the subscripts d and p are defined respectively by the form of divisive normalization and projection on a linear or a non-linear manifold (along the diagonal) as noted in Equations 36-37. On the other hand, if we were to project the initial point $\boldsymbol{x}$ first, that would give us $\boldsymbol{x}^p = (x_1^p, x_2^p, x_3^p)$, and then implementing divisive normalization would yield $\boldsymbol{x}^{pd} = (x_1^{pd}, x_2^{pd}, x_3^{pd})$. Our goal is to show that $\boldsymbol{x}^{dp} = \boldsymbol{x}^{pd}$.

In the linear case,

$$x_i^p = x_i + u - \frac{1}{N}\sum_{n=1}^{N} x_n \tag{36}$$

$$\begin{aligned} x_i^d &= Cx_i \\ &= \frac{K\, x_i}{\sigma_h + \sum_{n=1}^{N} x_n} \end{aligned} \tag{37}$$

Using these, we can calculate x_i^{pd} as

$$\begin{aligned} x_i^{pd} &= \frac{K\, x_i^p}{\sigma_h + \sum_{n=1}^{N} x_n^p} \\ &= \frac{K\, x_i + Ku - \frac{K}{N}\sum_{n=1}^{N} x_n}{\sigma_h + \sum_{n=1}^{N} x_n + Nu - \sum_{n=1}^{N} x_n} \end{aligned}$$

$$= \frac{K\,x_i + Ku - \frac{K}{N}\sum_{n=1}^{N} x_n}{\sigma_h + Nu}$$

and x_i^{dp} as

$$x_i^{dp} = x_i^{d} + u^{d} - \frac{1}{N}\sum_{n=1}^{N} x_n^{d}$$

$$= \frac{K\,x_i}{\sigma_h + \sum_{n=1}^{N} x_n} + \frac{K\,u}{\sigma_h + \sum_{n=1}^{N} x_n} - \frac{\frac{K}{N}\sum_{n=1}^{N} x_n}{\sigma_h + \sum_{n=1}^{N} x_n}$$

$$= \frac{K\,x_i + Ku - \frac{K}{N}\sum_{n=1}^{N} x_n}{\sigma_h + \sum_{n=1}^{N} x_n}$$

$$= x_i^{pd} \qquad (\because \sum_{n=1}^{N} x_n = Nu \text{ after projection})$$

Thus, the order does not matter in the linear case.

Adding the nonlinearity does not allow a closed form solution. The projection now takes place iteratively as $\boldsymbol{x}(t) \leftarrow \boldsymbol{x}(t) + \alpha[u(t) - \frac{1}{N}\Sigma_n x_n(t)]$ until convergence. It is important to note that though the projection is on a non-linear manifold, the direction of the projection is parallel to the diagonal; only the magnitude of the distance is determined iteratively.

References

1. Papoulis, A. & Pillai, S. Probability, random variables, and stochastic processes. (2002).

2. Drugowitsch, J., DeAngelis, G., Klier, E., Elife, D. A.- & 2014, undefined. Optimal multisensory decision-making in a reaction-time task. *cdn.elifesciences.org*

3. Carpenter, R. & Williams, M. Neural computation of log likelihood in control of saccadic eye movement. *Nature* **377**, 59–62 (1995).

4. Brown, S. & Heathcote, A. A Ballistic Model of Choice Response Time. *Psychol. Rev.* **112**, 117–128 (2005).

5. Louie, K., Grattan, L. E. & Glimcher, P. W. Reward value-based gain control: divisive normalization in parietal cortex. *J. Neurosci.* **31**, 10627–10639 (2011).

6. Carandini, M. & Heeger, D. J. Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* **13**, 51–62 (2012).