NOTES ON NORMATIVE SOLUTIONS TO THE SPEED-ACCURACY TRADE-OFF IN PERCEPTUAL DECISION-MAKING

JAN DRUGOWITSCH

Abstract. These notes contain a brief introduction to finding normative solutions to the speed-accuracy trade-off in perceptual decision-making. They were written for the FENS-Hertie Winter School 2015 "The neuroscience of decision making". The presented approach is based on Drugowitsch et al. (2012), Drugowitsch et al. (2014a), and Drugowitsch et al. (2014b). The code for all figures is available on my Github page, `https://github.com/jdrugo/FENS2015`.

Contents

Date: January 2015.

1. Accumulating evidence

Assume we want to decide between two possible states of the world, $z = 1$ and $z = -1$. This might correspond to the opponent's return in tennis going to the right ($z = 1$) or left ($z = -1$), or the coherent motion direction of a random-dot motion (RDM) stimulus being right ($z = 1$) or left ($z = -1$). A-priori, we will assume both options to be equally likely, as expressed by the prior

$$p(z = 1) = p(z = -1) = \frac{1}{2}. \tag{1}$$

1.1. **Discrete chunks of evidence.**

If we could directly observe z, the task would be easy. However, in the real world, we usually only observe some noisy instantiations, $x_1, x_2, \ldots$ of z, which we call the *momentary evidence*. That is, z is a *hidden state*, and we want to infer its value based on some *observations*, x_n. In the RDM task, for example, each observation would correspond to an instantaneous precept of random-dot motion[1].

The way we infer the hidden state z from the observations depends on how these observations are generated by z. For simplicity, we assume these observations to represent the hidden state, perturbed by some Gaussian noise, $x_n = z + \varepsilon_n$, where $\varepsilon_n \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$ is a zero-mean Gaussian with variance σ_ε^2. This leads of the likelihood of z given x_n to be given by

$$p(x_n|z) = \mathcal{N}\left(x_n|z, \sigma_\varepsilon^2\right) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} e^{-\frac{(x_n - z)^2}{2\sigma_\varepsilon^2}}. \tag{2}$$

where the last term is just the probability density function of the Gaussian.

1.1.1. *A single observation.*

Given a single such observation, or a single chunk of evidence, x_n, we can infer z by Bayes' rule,

$$\begin{aligned} p(z|x_n) &\propto_z p(x_n|z)p(z) \\ &\propto_z e^{\frac{x_n z}{\sigma_\varepsilon^2} - \frac{z^2}{2\sigma_\varepsilon^2}} \\ &\propto_z \frac{1}{1 + e^{-2\frac{x_n z}{\sigma_\varepsilon^2}}}, \end{aligned} \tag{3}$$

where, in the second line, we have replaced the likelihood $p(x_n|z)$ by its probability density function, while keeping only the z-related terms, and in the third line, we have added the normalisation constant such that $p(z = 1|x_n) + p(z = -1|x_n) = 1$. For $z = 1$, the above posterior moves monotonically from 0 for $x_n \to -\infty$, over $1/2$ for $x_n = 0$, to $x_n \to \infty$ for $x_n \to \infty$ (see Fig. 1). Therefore, observing $x_n = 0$ is completely uninformative about z, while $x_n < 0$ ($x_n > 0$) slant the evidence towards $z = -1$ ($z = 1$). The likelihood variance σ_ε^2 modulates how informative x_n is about z. A small σ_ε^2 causes the likelihoods for $z = -1$ and $z = 1$ overlap only weakly, such that it is easier to tell apart, which of them generated the observed x_n. This is also reflected in the posterior $p(z|x_n)$, which, for the same z and x_n, moves away from $1/2$ (i.e., increasing certainty) for smaller σ_ε^2's. Thus, σ_ε^2 modulates the difficulty of the task, with a small (large) σ_ε^2 implying an easy (hard) task.

1.1.2. *Multiple observations.*

Usually, we observe multiple chunks of evidence before committing to a decision. For each z, we will assume these chunks, $x_1, x_2, \ldots$, to

[1] As motion extends over time, there is practically no such thing as an instantaneous percept of motion. This concept is only meant as an approximation to reality.

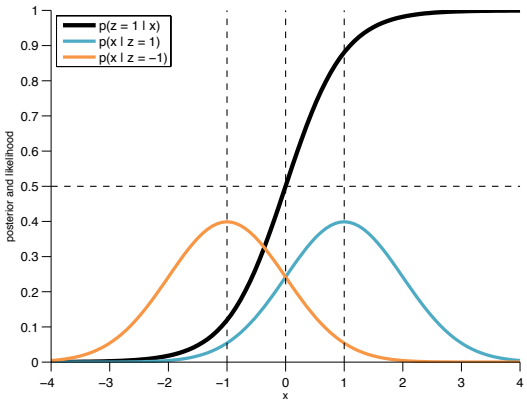


Figure 1. For $\sigma_\varepsilon^2 = 1$, the figure shows the posterior belief $p(z = 1|x)$ (black) upon observing a single x, and the two likelihood functions, $p(x|z = 1)$ (blue) and $p(x|z = -1)$ (orange). The figure was generated by `01evacc/single_obs.m`.

be independent and identically distributed, that is $p(x_1, x_2, \dots|z) = \prod_n p(x_n|z)$. With this property, the poster z given N observations $x_1, \dots, x_N$ is given by

$$\begin{aligned} p(z|x_1, \dots, x_N) &\propto_z p(z) \prod_{n=1}^{N} p(x_n|z)p(z) \\ &\propto_z e^{\frac{z}{\sigma_\varepsilon^2}\sum_{n=1}^{N} x_n - \frac{z^2}{2\sigma_\varepsilon^2}N} \\ &\propto_z \frac{1}{1 + e^{-2\frac{X_N z}{\sigma_\varepsilon^2}}}, \end{aligned} \tag{4}$$

where we have defined $X_N = \sum_{n=1}^{N} x_n$. Thus, the posterior z is the same as for single observations, Eq. (3), only with x_n replaced by X_N. This shows that, for the chosen Gaussian likelihood, evidence can be accumulated over observations by simply summing them up (see Fig. 2a). This sum, X_N, is a *sufficient statistic* of the posterior z.

1.2. **Continuous-time evidence.** In the real world, evidence often does not arrive in chunks, but rather as a continuous stream. To handle such situations, we chop this continuous stream into chunks of size δt (in time), and later let $\delta t \to 0$. Each chunk n is associated with some momentary evidence δx_n with likelihood

$$p(\delta x_n|z) = \mathcal{N}\left(\delta x_n|z\delta t, \sigma_\varepsilon^2 \delta t\right), \tag{5}$$

where the scaling of mean and likelihood of the above Gaussian ensures that with $\delta t \to 0$, the information that δx_n provides about z goes to zero. If this were not the case, then $\delta t \to 0$ would cause z to be immediately known, as every small time step provides more and more information.

Assume we now observe momentary evidence $\delta x_{0:t}$ from time 0 (onset of the stimulus / start of the observation) to t. In discrete chunks of size δt, this would correspond to $N \approx t/\delta t$ observations. Based on these observations, the posterior z

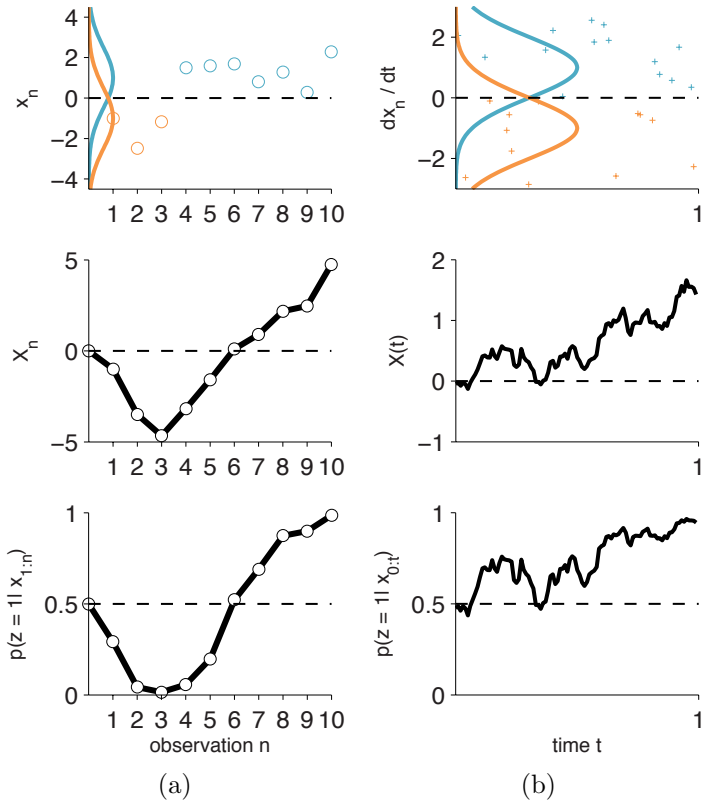


Figure 2. Evidence accumulation with (a) discrete or (b) continuous evidence. In both cases, the top panels show the momentary evidence, either per observation, or in small, finite steps of δt. Here, momentary evidence providing evidence towards $z = 1$ ($z = -1$) is shown in blue (orange). Note that the momentary evidence is always drawn from the blue density. The middle panels show the sufficient statistic that sums up this momentary evidence. The bottom panels shown the posterior belief $p(z = 1|\dots)$ provided the given evidence. The figures were generated with `01evacc/discrete_obs.m` and `01evacc/continuous_obs.m`.

is given by

$$\begin{aligned} p(z|\delta x_{1:t}) &\propto_z p(z) \prod_{n=1}^{N} p(\delta x_n|z) \\ &\propto_z e^{\frac{z}{\sigma_\varepsilon^2}\sum_{n=1}^{N}\delta x_n - \frac{z^2}{2\sigma_\varepsilon^2}\delta t N} \\ &\propto_z \frac{1}{1+e^{-2\frac{X(t)z}{\sigma_\varepsilon^2}}}, \end{aligned} \tag{6}$$

where we have defined $X(t) = \sum_{n=1}^{N} \delta x_n$. If we now take $\delta t \to 0$, this sum turns into the integral

$$X(t) = \int_0^t \delta x(s), \tag{7}$$

where $\delta x(s)$ is the momentary evidence at time s after stimulus onset. This illustrates that the principle is the same as for the discrete case: the sufficient statistic $X(t)$ is simply the sum of the momentary evidences (see Fig. 2b).

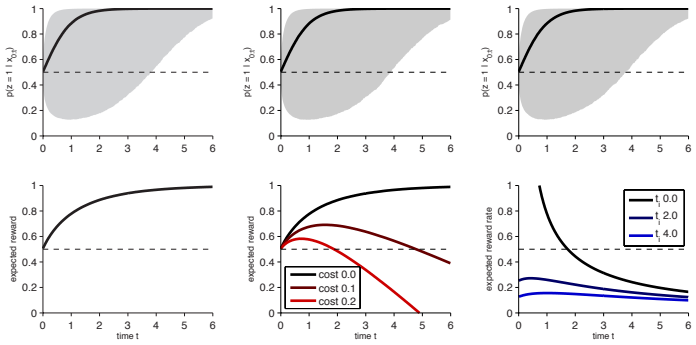


Figure 3. Posterior belief and expected reward/reward rate assuming (a) only rewards and no costs, (b) a cost for accumulating evidence, (c) the reward rate as the relevant measure. In all cases, the top panel shows the median belief (solid line), as well as its 2.5 and 97.5 percentile (shaded area). The bottom panel shows how the expected reward (averaged over 10^5 runs) evolves over time. In (b) this expected reward is shown for different cost per unit time for accumulating evidence. In (c) the inter-trial interval t_i is varied without introducing a cost for accumulating evidence. The figures were generated with `02lossfn/loss_01.m`, `02lossfn/loss_cost.m`, and `02lossfn/loss_rr.m`.

What is interesting in this case is that $X(t)$ describes a diffusion process with drift z and diffusion variance σ_ε^2, such that (in the absence of any bounds)

$$X(t) \sim \mathcal{N}\left(zt, \sigma_\varepsilon^2 t\right). \tag{8}$$

As will be shown later, this property allows us to perform optimal decisions with diffusion models.

2. What do we want to maximise when making decisions?

In order to turn the posterior z into a decision, we need to additionally define a loss function. For each decision, this loss function specifies the loss that occurs for each choice, given that the hidden state z has a certain value. The optimal decision is the choice that minimises the expected loss (i.e., maximising the negative expected loss), where the expectation is taken over the posterior z.

2.1. Maximising decision accuracy.

The simplest loss is the 0-1 loss. It assumes a loss of 0 for correct choices (e.g., choosing "right" if $z = 1$), and a loss of 1 for incorrect choices (e.g. choosing "right" if $z = -1$). In this case, it is easy to show that the optimal choice corresponds to that for which the posterior $p(z|\dots) > 1/2$. For example, if $p(z = 1|\dots) > 1/2$, we would choose "right".

For some given accumulated evidence, this loss tells us which choice is best, but how do we know how much evidence to accumulate? To answer this, we need to consider in which case making a decision with the current evidence leads to a lower loss (i.e., a higher decision accuracy) than accumulating evidence and deciding later. In light of the fact that accumulating more evidence will on average make us more certain about which decision is correct (this is easy to show formally), we would always accumulate more evidence (see Fig. 3a). Thus, according to the 0-1 loss, we would never cease to accumulate more evidence, and thus never make a decision. Clearly, this does not correspond to what humans and other animals are doing. To describe their behavior, we need to use other loss functions.

2.2. **Costly accumulation of evidence.** Why would humans only accumulate limited evidence to make decisions if they could increase their decision accuracy by waiting forever? One possibility is that accumulating evidence (e.g. staring at an RMD stimulus) comes at an implicit (e.g. attention) or explicit (e.g. loss of reward doing something else) cost. Assuming a reward of 1 for correct choices, no reward for incorrect choices, and a cost of c per unit time for accumulating evidence, the decision maker ought to adopt a strategy that maximises

$$ER\,(PC, RT) = PC - cRT, \tag{9}$$

where PC denotes the probability of making a correct choices, RT is the expected evidence accumulation time, and ER is the expected reward as a function of the two previous quantities. Introducing this (positive) cost causes strategies that accumulate evidence forever to become sub-optimal, as at some point, the marginal increase in choice accuracy when accumulating more evidence does not justify the additional cost to pay for doing so (see Fig. 3b). For most of the rest of this note, we will focus on finding the optimal strategy that maximises the above expected reward.

2.3. **Implicit opportunity costs.** Another possible loss function that leads to early choices is to explicitly take into account the loss of future reward in current choices. We can do so by aiming to maximise the reward rate, which is the average reward per unit time. Assuming a time t_i inbetween consecutive similar choices (or trials), and an additional penalty time t_p for incorrect choices, this reward rate is given by

$$RR\,(PC, RT) = \frac{ER(PC, RT)}{RT + t_i + (1 - PC)t_p}, \tag{10}$$

where ER is the single-choice expected reward for the previous section. In this case, even if the accumulation cost c is zero, earlier choices are preferred, as late choices cause an increase in the denominator that reduces the reward rate (see Fig. 3c). We will briefly discuss handing reward rate cases towards the end of this note.

3. Optimal stopping for known evidence reliability

Here we consider the following problem. Assume continuous evidence accumulation, as in Sec. 1.2 where we assume $z \in \{-1, 1\}$ and some known likelihood variance σ_ε^2. What is the decision strategy that maximised the expected reward, Eq. (9)? We find this strategy by Dynamic Programming (Bellman, 1954).

3.1. **Dynamic Programming.** Before applying Dynamic Programming to our problem, let us first describe it in a more general setting. Assume a set of states, $s \in \mathcal{S}$, in each of which we can perform one of multiple actions, $a \in \mathcal{A}$. Choosing action a in state s leads to state s' with probability $p(s'|s, a)$. Furthermore, the transition yields reward $r(s, a)$, which is a function of the current state s and the taken action a. The aim is to find the optimal policy π, which, for each state s, chooses the action $\pi(s) \in \mathcal{A}$ that maximises the overall expected reward. It has been previously shown by Bellman that such a deterministic (rather than stochastic) policy π is the best one can do. To keep the reward bounded, we also assume a terminal state $\tilde{s}$ that will always be reached, loops onto itself, and yields no reward, that is, $p(\tilde{s}|\tilde{s}, a) = 1$ for all a, and $r(\tilde{s}, a) = 0$.

To find the optimal policy π, we first define the value function $V(s)$. For some, not necessarily optimal policy π, this value function $V^\pi(s)$ maps each state into the

total expected reward from this state onwards, following this policy, that is

$$V^{\pi}(s) = \left\langle \sum_{n=0}^{\infty} r(s_n, \pi(s_n)) \right\rangle_{p(s_1,s_2,\ldots|\pi,s_0=s)}, \tag{11}$$

where the expectation is over all state trajectories starting in s and following policy π. The optimal policy is the policy that maximises the value for each state. Thus, the value function corresponding to this optimal policy is given by

$$V(s) = \max_{\pi} \left\langle \sum_{n=0}^{\infty} r(s_n, \pi(s_n)) \right\rangle_{p(s_1,s_2,\ldots|\pi,s_0=s)} \tag{12}$$

The insight in Dynamic Programming is that this value function can be computed in smalls steps by relating the optimal values to each other, leading to Bellman's equation[2],

$$\begin{aligned} V(s) &= \max_{\pi} \left\langle r(s, \pi(s)) + \sum_{n=1}^{\infty} r(s_n, \pi(s_n)) \right\rangle_{p(s_1,s_2,\ldots|\pi,s_0=s)} \\ &= \max_{a} \left[r(s,a) + \left\langle \max_{\pi} \left\langle \sum_{n=1}^{\infty} r(s_n, \pi(s_n)) \right\rangle_{p(s_2,s3,\ldots|\pi,s_1=s')} \right\rangle_{p(s'|s,a)} \right] \\ &= \max_{a} \left[r(s,a) + \langle V(s') \rangle_{p(s'|s,a)} \right], \end{aligned} \tag{13}$$

where, in the second line, we have split the max operator into the current action a and future actions, as defined by π, and have split the expectation into the expectation over the next state s' and that over future states $x_2, x_3, \ldots$. The optimal policy corresponds to choosing in each state the action that satisfies the above maximum.

3.2. **Dynamic Programming applied to perceptual decision-making.** Equipped with this method, we can return to our perceptual decision-making problem. The first question is how the states and actions of dynamic programming map onto our decision-making problem. The state space corresponds to the internal state of the decision maker when performing evidence accumulation. As previously shown this internal state can be summarised by the sufficient statistic $X(t)$ of the posterior $p(z|\ldots)$. As $X(t) \in [-\infty, \infty]$, this choice would make the state space infinite, and thus hard to numerically store the value function over this state space. We avoid this problem by instead using the belief g,

$$g(X) \equiv p(z=1|X) = \frac{1}{1+e^{-2\frac{X}{\sigma_\varepsilon^2}}}, \qquad \text{s.t. } X(g) = \frac{\sigma_\varepsilon^2}{2} \log \frac{g}{1-g}, \tag{14}$$

as another sufficient statistic of the posterior (see Eq. (6)). This belief g is bounded by $g \in [0,1]$ and is a sufficient statistic of the posterior as every X maps uniquely on a particular g. Thus, our value function $V(g)$ will be a function of this belief.

As a next step, we need to find the possible actions. In our case, these correspond to either making a choice (two actions; choosing either "right" or "left") or to continue accumulating more evidence and making a choice later. Thus, we have three possible actions.

What would the left-hand side expression of Bellman's Eq. (13) represent for each of these actions? If we choose "right", we expect a reward of 1 with probability

[2]This derivation of Bellman's equation is only approximate, as it ignores the possibility of returning to the same state s in future transitions. A derivation that takes this into account is slightly more technical, but results in the same final equation.

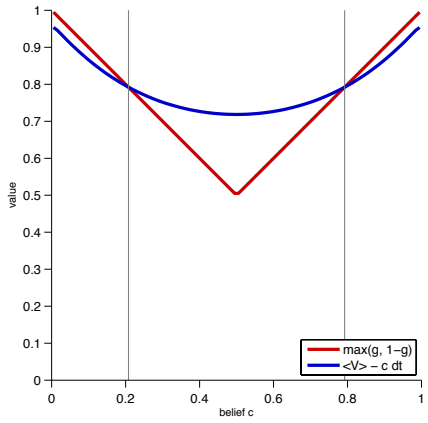


Figure 4. Illustration of finding the optimal decision bound by Dynamic Programming. The figure illustrates the value $\max[g, 1-g]$ of making immediate decisions (red), and that, $\langle V(g')\rangle_{p(g'|g)} - c\delta t$, of accumulating more evidence and deciding later (blue). The optimal bound where to stop accumulating evidence is where these two value functions intersect (grey lines). The plot was generated by `03knownreliab/plot_valueintersect.m`.

$p(z = 1|X) = g$, and no reward otherwise. Therefore, the expected reward in this case will be g. Furthermore, we will not receive any further reward in the future, such that $\langle V(s')\rangle$ will be zero. Choosing "left" instead, we expect reward 1 with probability $p(z = -1|X) = 1 - g$, which results in an expected reward of $1 - g$, and again $\langle V(s')\rangle = 0$. If we instead continue to accumulate more evidence for some more time δt, we will not receive any reward, but instead incur a cost $c\delta t$ (recall that the cost c is the cost per unit time). Furthermore, the expected next value will be $\langle V(g')\rangle_{p(g'|g)}$, which is the expected future value when accumulating more evidence, resulting in belief g' (see next section for how to find $p(g'|g)$).

Putting all the pieces together results in Bellman's equation

$$V(g) = \max\left[g, 1-g, \langle V(g')\rangle_{p(g'|g)} - c\delta t\right]. \tag{15}$$

The optimal action at any point in time corresponds to choosing the largest right-hand side term within the squared bracket. This already gives us a hint about how the optimal policy looks like (see Fig. 4). As long as the last term dominates either g or $1 - g$, we will continue to accumulate more evidence. As soon as the other two terms, either g or $1 - g$, dominate, the corresponding decision will be triggered. This policy will thus yield two boundaries in the space of beliefs within which evidence accumulation is the preferred course of action. As soon as either boundary is reached, the corresponding decision is made.

3.3. **Computing the belief transition $p(g'|g)$.** In order to computed the expected future value, we need to know the probability $p(g'|g)$ of holding belief g' after having observed another δt time units worth of evidence, and before having held belief g. Due to the monotonic relation between X and g, we can first find this density in X and then map it onto g.

To find the density $p(X'|X)$, note that $z = 1$ with probability g, and $z - 1$ with probability $1 - g$. Therefore, the next small chunk of momentary evidence will either be drawn from $\mathcal{N}(\delta t, \sigma_\varepsilon^2 \delta t)$ or $\mathcal{N}(-\delta t, \sigma_\varepsilon^2 \delta t)$. As X' is just this chunk of

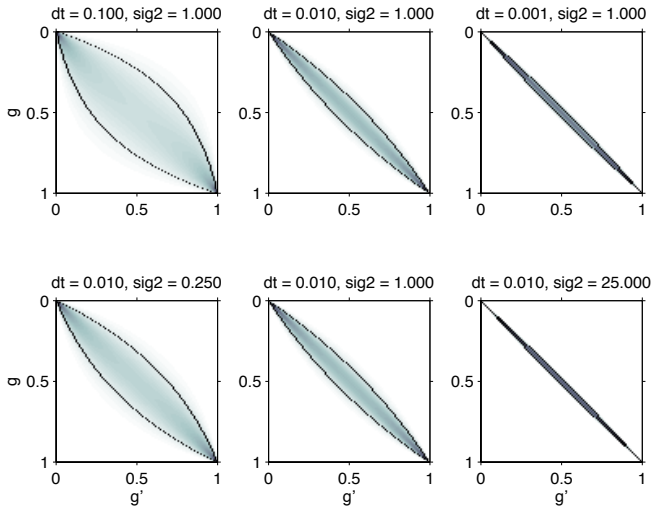


Figure 5. Examples of belief transition densities $p(g'|g)$. In each panel, each row corresponds to a density over g' for a different g. In each of these rows, the black dots indicate the 2.5 and 97.5 percentiles of the distribution. The top row of panels illustrates how a smaller δt causes the transition densities to sharpen towards a Dirac delta. The bottom row illustrates how increasingly harder tasks (from left to right) also introduce such a sharpening, as, within the same time-span, less evidence is expected to be collected. The plots were generated by `03knownreliab/plot_belieftrans.m`.

evidence added to X, the density of X' is given by the mixture of Gaussians,

$$p(X'|X) = g(X)\mathcal{N}(X'|X + \delta t, \sigma_\varepsilon^2 \delta t) + (1 - g(X))\mathcal{N}(X'|X - \delta t, \sigma_\varepsilon^2 \delta t) \tag{16}$$

This density does not take into account any boundary crossings that lead to decisions, and is therefore only valid for very small δt, in which these crossings are unlikely to occur.

To map X' back onto g', we use the relation

$$p(g'|g) \left| \frac{\mathrm{d}g'}{\mathrm{d}X'} \right| = p(X'|X) \tag{17}$$

with the derivative

$$\begin{aligned} \frac{\mathrm{d}g'}{\mathrm{d}X'} &= \frac{2e^{\frac{2}{\sigma_\varepsilon^2}X}}{\sigma^2 \left(1 + e^{\frac{2}{\sigma_\varepsilon^2}X}\right)^2} \\ &= \frac{2e^{\frac{2}{\sigma_\varepsilon^2}\frac{\sigma_\varepsilon^2}{2}\log\frac{g'}{1-g'}}}{\sigma^2 \left(1 + e^{\frac{2}{\sigma_\varepsilon^2}\frac{\sigma_\varepsilon^2}{2}\log\frac{g'}{1-g'}}\right)^2} \\ &= \frac{2}{\sigma_\varepsilon^2} g'(1 - g'), \end{aligned} \tag{18}$$

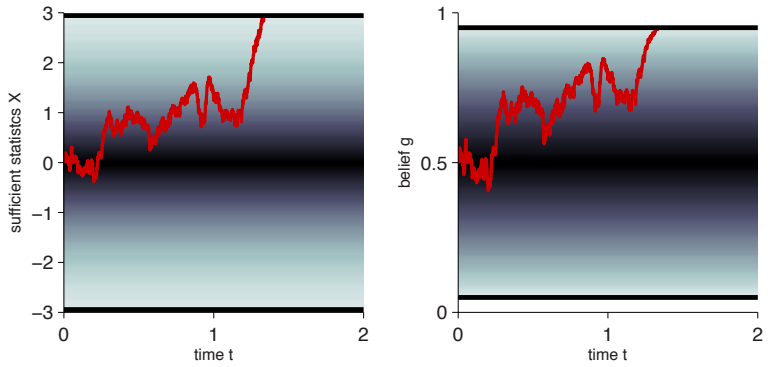


Figure 6. Implementation of the optimal policy by a diffusion model. The figure shows an example decision (red trajectory) implemented by a diffusion model (left panel) or by directly updating the belief (right panel). The mapping between diffusion and belief space is mostly linear, except for close to the boundaries. This is also apparent in the grey-shaded background, whose colour marks the decision maker's decision certainty at different points within the diffusion and belief space (black = completely uncertain, white = completely certain). The plots were generated by `03knownreliab/plot_diffusion_example.m`.

where the second line is based on substituting $X(g)$ for the X, and the third line on simplifying the expression. Combining all the above results in

$$\begin{aligned} p(g'|g) &= \frac{\sigma_\varepsilon^2}{2g'(1-g')}\left(g\mathcal{N}(X'|X+\delta t,\sigma_\varepsilon^2\delta t)+(1-g)\mathcal{N}(X'|X-\delta t,\sigma_\varepsilon^2\delta t)\right) \\ &= \frac{\sigma_\varepsilon^2 e^{-\frac{\delta t}{2\sigma_\varepsilon^2}}\mathcal{N}\left(X(g')|X(g),\sigma_\varepsilon^2\delta t\right)}{2g'(1-g')}\left(ge^{\frac{X(g')-X(g)}{\sigma_\varepsilon^2}}+(1-g)e^{-\frac{X(g')-X(g)}{\sigma_\varepsilon^2}}\right), \end{aligned} \tag{19}$$

where $X(g)$ and $X(g')$ are the summary statistics corresponding to belief g and g' (see Eq. (14)), and the second line results from expanding the Gaussians in the bracket and collecting common terms into the outer Gaussian. See Fig. 5 for examples of transition densities.

3.4. **Optimal decision-making with diffusion models.** As previously derived, the optimal policy corresponds to two boundaries in belief. This belief g maps monotonically into the sufficient statistics X. Therefore, the boundaries on the belief can be transformed into boundaries on X, using the same mapping. As a result, we can perform optimal decision-making without ever representing, and even computing the belief explicitly.

As shown before, $X(t)$ follows a diffusion process. By the optimal policy, this diffusion process is bounded by two constant (i.e., time invariant) boundaries. Thus, this policy corresponds to diffusion models, in which either decision is triggered once either boundary is reached. Therefore, diffusion model perform optimal decision-making in this setting, as long as the boundaries are set to the correct height, which is imposed by the optimal policy (see Fig. 6).

3.5. **Computing the solution by belief discretisation.** So far we have shown to theoretically derive the optimal policy. In this section we describe how the bounds due to the optimal policy can be found numerically.

As the belief g can take an infinite number of possible values, and there is no known functional form of $V(g)$, we need to represent $V(g)$ in some approximate

form. The approach that makes the least assumptions is to discretise g in equally-sized steps of size Δ_g, and represent $v(g)$ only for these discretised values of g. To discretise g, we will skip its extremes $\{0, 1\}$, as at these points the value function is known, and the belief transition $p(g'|g)$ is a Dirac delta at $\delta(g' - g)$. We denote the remaining discretised g by g^k with $k = 1, \ldots, K$. The corresponding value function is denoted $V^k = V(g^k)$. This allows us to express Bellman's equation as

$$V^k = \max \left[g^k, 1 - g^k, \sum_j p(g^j|g^k) V^k - c\delta t \right], \tag{20}$$

where $p(g^j|g^k)$ is normalised, such that $\sum_j p(g^j|g_k) = 1$.

To find the V^k that solves this equation, we use a Dynamic Programming technique called *value iteration*. This technique starts with some arbitrary value function $V^{k,0}$, and then uses this value function in the right-hand side of Eq. (20) while assigning the left-hand side the new value function $V^{k,1}$. This is iterated by

$$V^{k,n} = \max \left[g^k, 1 - g^k, \sum_j p(g^j|g^k) V^{k,n-1} - c\delta t \right], \tag{21}$$

until the value function does not change significantly between two consecutive iterations. At this point, the intersection between the third and the first two terms reveals the boundary in belief that corresponds to the optimal policy.

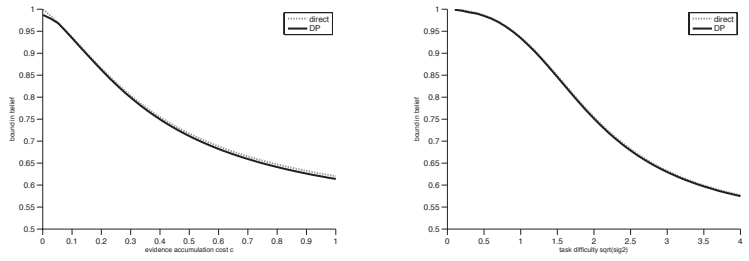
A remaining question is what δt to use when computing the belief transition $p(g'|g)$. This transition does not consider possible boundary crossings, such that δt needs to be small to minimise the possibility of crossing this boundary. A too small δt, however, corresponds to almost no additional accumulated evidence, such that the belief remains almost unchanged. In other words, for small δt, $p(g'|g)$ approaches the Dirac delta $\delta(g' - g)$. This causes numerical issues, as almost no posterior mass of $p(g^j|g^k)$ falls on any $j \neq k$. Thus, a smaller δt requires a finer discretisation of the belief, which causes the speed of the method to decrease. Overall, the best δt is a trade-off between an increased error due to ignoring boundary crossing and numerical inaccuracies due to too little mass being assigned to g^j's other than g^k.

See Fig. 7 for example boundaries computed by the described method. This figure illustrates how the optimal boundary depend, on one hand, on the cost of accumulating evidence, and, on the other hand, on the task difficulty as controlled by σ_ε^2.

3.6. **Finding the optimal bounds by direct optimisation.** In this simple setting, we can take an alternative (i.e., non-Dynamic Programming) route to find the optimal boundary location. We already know that these boundaries will not change with time, and that the optimal decision making model can be implemented by a diffusion model. This allows us to use the known expressions for expected first-passage time and bound-hitting probability to directly find the bound heights that maximise the expected reward. For boundaries at $\{-\theta, \theta\}$ (in X-space, not belief space), These expressions are given by (Cox and Miller, 1965; Palmer et al., 2005)

$$PC(\theta) = \frac{1}{1 + e^{-2\frac{\theta}{\sigma_\varepsilon^2}}}, \quad \text{and} \quad RT(\theta) = \theta \tanh\left(\frac{\theta}{\sigma^2}\right) \tag{22}$$

Substituting these into Eq. (9) allows us to use numerical maximisation techniques to find the bounds that maximise the expected reward. This maximisation is easy, as the expected reward has a unique maximum with respect to the bound height



(a) (b)

Figure 7. Dependence of optimal decision bounds on (a) cost for evidence accumulation, and (b) task difficulty. Both panels show the optimal bound in belief (only top bound; bottom bound symmetric around 1/2) on one hand computed by Dynamic Programming (black line; $\Delta_g = 1/500$, $\delta t = 0.0005$), and on the other hand by direct optimization (grey, dotted line). In (a), the bound decreases with the cost for accumulating evidence, as it does not pay to accumulate a lot of evidence if it comes at a high cost. In (b), the bound decreases with task difficulty (high σ_ε^2 = difficult task), as little is gained from accumulating more uninformative evidence. The plots were generated by `03knownreliab/plot_bound_with_cost.m` and `03knownreliab/plot_bound_with_sig2.m`.

(Bogacz et al., 2006). The resulting bounds using this approach are shown in Fig. 7, and match well those found by Dynamic Programming.

4. Optimal stopping for unknown evidence reliability

In what has been discussed so far, we have assumed that the decision-maker has full knowledge of the likelihood functions, $p(x|z = 1)$ and $p(x|z = -1)$. This corresponds to fully knowing the reliability of the evidence for each made decision. However, in the real work, and also in many experimental settings in the laboratory, the difficulty varies across decisions. How does the optimal decision-making strategy change in such a setting?

To address this, we first modify the process with which the evidence is generated to allow its reliability to vary across decisions. Second, we describe how to effects the accumulation of evidence. Third, we show how to use Dynamic Programming to find the optimal policy in such a setup. And, fourth, we describe how to implement the outlined approach numerically.

4.1. Varying evidence reliability across decisions.

In Sec. 1.2, we have assumed the momentary evidence to be generated by $\delta x \sim \mathcal{N}\left(z\delta t, \sigma_\varepsilon^2 \delta t\right)$. The reliability of this evidence was controlled the generative variance σ_ε^2. An alternative to control this reliability is to leave the variance fixed (and set to 1, for convenience), and instead change the magnitude of the mean of δx (above set to $z\delta t$, with constant magnitude $|z\delta t| = \delta t$). In this case, a large magnitude of this mean (irrespective of its sign) corresponds to an easy task, and a small magnitude to a hard one. Thus, this mean can be used to encode both the hidden state z, and the difficulty of the current decision.

To use this feature, we assume the momentary evidence to be generated by $\delta x \sim \mathcal{N}\left(\mu\delta t, \delta t\right)$, where $\mu = \alpha z$, $z \in \{-1, 1\}$ is again the hidden state that the decision maker wants to identify, and $\alpha \geq 0$ is the difficulty of the current decision. We have already defined a prior over z from which this hidden state is drawn before

each trial. We could now also define a prior over α to draw a per-trial difficulty. However, instead, we define a prior over μ directly, given by

$$\mu \sim \mathcal{N}(0, \sigma_\alpha^2). \tag{23}$$

This prior specifies, on one hand, the hidden state z by the sign of μ ($\mu \geq 0$ corresponds to $z = 1$). As $p(\mu < 0) = p(\mu \geq 0) = 1/2$ with the above prior, we have $p(z = 1) = p(z = -1) = 1/2$, as before. The prior over α, on the other hand, is defined my the magnitude of μ, that is, $\alpha = |\mu|$. Therefore, this prior is proportional to a Gaussian over non-negative values only. In other words, the above prior puts more weight on hard trials (small α) than easy ones (large α). The σ_α^2 controls how hard the trials are on average, with a small σ_α^2 corresponding to an overall hard task.

4.2. **Evidence accumulation with unknown evidence reliability.** Before each trial, μ is drawn according to its prior. Based on this drawn μ, a stream of evidence, $\delta x_1, \delta x_2, \ldots$ is generated. The decision maker observes this evidence, and needs to identify the sign of μ (as this sign identifies the hidden state z). We will tackle this problem in two steps: first, we find the posterior over μ, and, second, the posterior over z.

The posterior μ, based on some evidence $\delta x_{0:t}$ from time 0 to t is again found by Bayes' rule, and results in

$$\begin{aligned} p(\mu|\delta x_{0:t}) &\propto_\mu p(\mu) \prod_{n=1}^{N} p(\delta x_n|z) \\ &\propto_\mu e^{-\frac{\mu^2}{2\sigma_\alpha^2}} e^{-\sum_{n=1}^{N} \frac{(\delta x_n - \mu\delta t)^2}{2\delta t}} \\ &\propto_\mu e^{-\left(\frac{1}{\sigma_\alpha^2}+t\right)\frac{\mu^2}{2}+X(t)\mu} \\ &\propto_\mu \mathcal{N}\left(\mu|\frac{X(t)}{\sigma_\alpha^{-2}+t}, \frac{1}{\sigma_\alpha^{-2}+t}\right), \end{aligned} \tag{24}$$

where we have used $N \approx t/\delta t$ such that $t \approx \sum_{n=1}^{N} \delta t$, and $X(t) = \sum_{n=1}^{N} \delta x_n$. Based on this posterior over μ, we find the posterior over z by

$$p(z = 1|\delta x_{0:t}) = p(\mu \geq 0|\delta x_{0:t}) = \int_0^\infty p(\mu|\delta x_{0:t}) \mathrm{d}\mu = \Phi\left(\frac{X(t)}{\sqrt{\sigma_\alpha^{-2}+t}}\right), \tag{25}$$

where $\Phi(\cdot)$ is the cumulative density function of the standard Gaussian.

The cumulative $\Phi(a)$ extends, just like the logistic sigmoid in Eq. (6), from $\Phi(a) = 0$ for $a \to -\infty$, over $\Phi(a) = 1/2$ for $a = 0$, to $\Phi(a) = 1$ for $a \to \infty$. Therefore, as for the constant-reliability case, the sign of $X(t)$ again determines if $z = 1$ or $z = -1$ is more likely. However, in contrast to before, the posterior z now depends on both $X(t)$ and passed time t, rather than only $X(t)$. This implies that we have to track two sufficient statistics $(X(t), t)$ when accumulating evidence, rather than only $X(t)$.

The role of time in the above posterior has that of monitoring the flow of evidence. If the decision maker does not perceive sufficient evidence in some time period, which would be reflected in $X(t)$ not growing in magnitude, that the posterior would drop closer to 1/2. This is because not increasing $X(t)$ in magnitude with time is associated with a difficult trial, and therefore induces a larger degree of uncertainty. The converse applies for a rapidly growing $X(t)$. In this case, the posterior confidence (i.e. proximity to 0 or 1) increases more rapidly over time, as before.

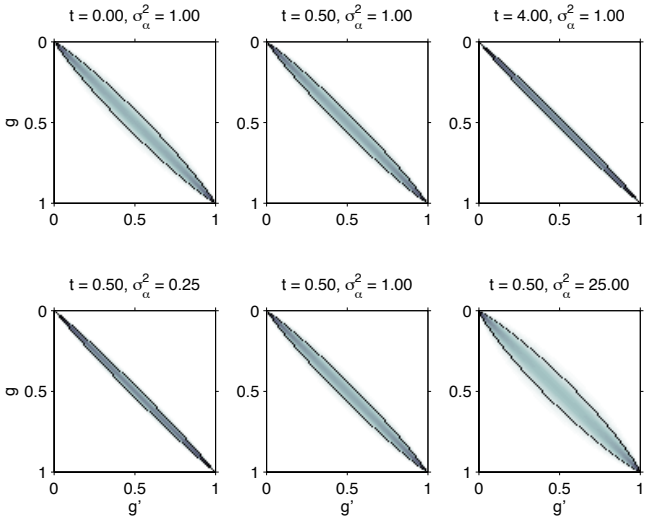


Figure 8. Examples of belief transition densities $p(g'|g,t)$. In each panel, each row corresponds to a density over g' for a different g. In each of these rows, the black dots indicate the 2.5 and 97.5 percentiles of the distribution. The top row shows that, the further time increases, the less likely it is to observe highly reliable evidence, as indicated by sharpening of the transition densities. The bottom row illustrates how increasingly harder tasks (from right to left) also introduce such a sharpening, as, within the same time-span, less evidence is expected to be collected. All shown densities are computed for time-steps of size $\delta t = 0.01$. The plots were generated by `04unknownreliab/plot_belieftrans.m`.

4.3. **Optimal decision-making by Dynamic Programming.** We can use the same principles of Dynamic Programming to find when to best stop accumulating evidence even for the case of an evidence reliability that varies across decisions. The possible actions, choose "right"/"left" or continue to accumulate more evidence, remain the same, but the state space is now determined by X and t, which are the sufficient statistics of the posterior z. For some fixed t, we again map this X into the belief g by

$$g(X) \equiv p(z=1|X,t) = \Phi\left(\frac{X}{\sqrt{\sigma_\alpha^{-2}+t}}\right), \qquad \text{s.t. } X(g) = \sqrt{\sigma_\alpha^{-2}+t}\,\Phi^{-1}(X), \tag{26}$$

where $\Phi^{-1}(\cdot)$ is the inverse of the cumulative density function of a standard Gaussian. Facilitating this mapping, we define the value function $V(g,t)$ over (g,t) rather than over (X,t). This leads to Bellam's equation

$$V(g,t) = \max\left[g, 1-g, \langle V(g',t+\delta t)\rangle_{p(g'|g,t)} - c\delta t\right], \tag{27}$$

where, in contrast to Eq. (15), the last term additionally takes into account the passage of time.

To compute the expected future value $\langle V(g',t+\delta t)\rangle$, we again need to find the belief transition density $p(g'|g,t)$. Its derivation follows the same arguments as in

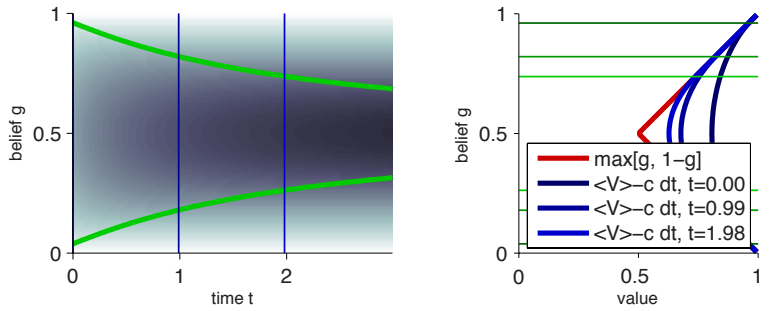


Figure 9. An example value function and resulting bound. The left panel shows the value function over belief g and time t, with values randing from 0.5 (black) to 1 (white). The bounds (green) are the locations where the values for immediate decisions equals the value of continuing to accumulate more evidence. The right panel shows some example value function, with the values for immediate decisions shown in red, the the values for accumulating more evidence shown in different shades of blue. The latter corresponds to different time-points, corresponding to the blue vertical lines in the left panel. The green lines in the right panel indicate the locations of the inferred bounds. The plots were generated by `04unknownreliab/plot_valuefn.m`.

Sec. 3.3, and results in the following expression

$$p(g'|g,t) = \frac{1}{\sqrt{\delta t_{eff}}}\left(\frac{\left[\Phi^{-1}(g')\right]^2}{2} - \frac{\left[\Phi^{-1}(g') - \sqrt{1+\delta t_{eff}}\Phi^{-1}(g)\right]^2}{2\delta t_{eff}}\right), \qquad (28)$$

where we have defined $\delta t_{eff} = \delta t/(t + \sigma_\alpha^{-2})$ (see Drugowitsch et al. (2012) for the derivation; the paper has a typo in δt_{eff}). In contrast to before, this transition density now depends on both the current belief g and time t (see Fig. 8).

The optimal policy is again determined by Bellman's equation, Eq. (27). For some fixed time t, it is best to accumulate more evidence as long as the last term in this equation dominates the one, $\max[g, 1-g]$, that describe the value of immediate decisions. As soon as either of the first two terms dominate, a decision ought to be triggered. This results in two bounds in the decision maker's belief at which decisions are triggered (see Fig. (9)). These bounds are now time-dependent, as will generally, approach $1/2$ over time.

4.4. **Optimal decision-making with diffusion models.** As before, we can map the boundary on belief into the boundary on a diffusing particle, $X(t)$, using Eq. (26). This boundary will now vary with time, for two reasons. First, the optimal boundary in belief already varies with time. Second, the mapping between belief and $X(t)$ is time-dependent, such that, even a time-invariant bound in belief would result in a time-variant bound on $X(t)$. Therefore, optimal decision-making is achieved by a diffusion model with time-varying boundaries. For an example of such a model, see Fig. 10.

4.5. **Computing the solution by belief/time discretisation.** How do we compute the optimal bounds for a given task difficulty σ_α^2 and evidence accumulation cost c? To do so, we can again discretise out state space, which now spans both belief g and time t. Clearly, we cannot compute the value for all time, as this time is not bounded from above. A useful strategy is to assume some large time T at which the decision maker is guaranteed to decide, in which case $V(g,T) = \max[g, 1-g]$. All other $V(g,t)$ can then be computed by backwards induction with Eq. (27),

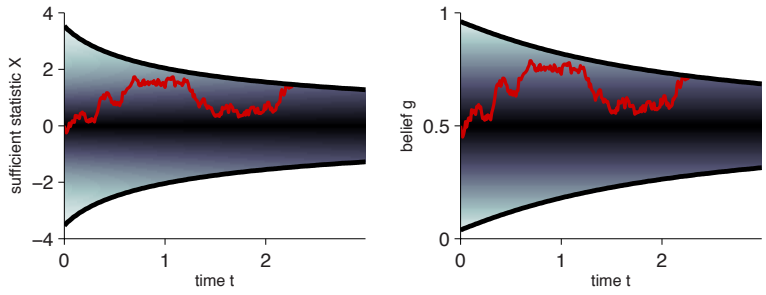


Figure 10. Implementation of the optimal policy for evidence reliability that varies across decisions by a diffusion model. The figure shows an example decision (red trajectory) implemented by a diffusion model (left panel) or by directly updating the belief (right panel). The mapping between diffusion and belief space is non-linear and time-dependent. This is apparent in the grey-shaded background, whose colour marks the decision maker's decision certainty at different points within the diffusion and belief space (black = completely uncertain, white = completely certain). For the diffusion model, the same level of X maps to different levels of certainty at different points in time. The plots were generated by `04unknownreliab/plot_diffusion_example.m`.

starting with $V(g, T - \delta t)$, then $V(g, T - 2\delta t)$, and so on. In practise, setting T to around five times the time-frame of interest has worked well (Drugowitsch et al., 2012). That is, if we are interested in 2s worth of decision boundary, we would set $V(g, 10) = \max[g, 1 - g]$ and work backwards from there.

To apply the discretisation, we choose Δ_g as the step size between two consecutive beliefs, g^k and g^{k+1}, and δt as the discretisation of time. Then, the value function can be solved by backwards induction in time from $t = T - \delta t$ to $t = 0$ by

$$V(g^k, t) = \max \left[g^k, 1 - g^k, \sum_{j=1}^{K} p(g^j | g^k, t) V(g^j, t + \delta t) - c\delta t \right], \tag{29}$$

where $p(g^j | g^k, t)$ is normalised such that $\sum_{j=1}^{K} p(g^j | g^k, t) = 1$. For each point in time, the optimal decision boundaries in g are found where the last term equals either of the first two terms in this expression. As time is now an explicit factor in the value function, value iteration is not anymore required to find its solution.

5. Extensions

The above introduces the general idea of how Dynamic Programming and related methods can be applied to find the optimal policy for decisions under the pressure of time. In this section we discuss several variants and extensions to this idea, to handle a different loss function, generalise the problem domain, and find solutions more efficiently.

5.1. Maximising reward rate rather than expected reward.

The approach outlined so far finds policies that maximise the expected reward of individual decisions, where the only pressure on deciding early is introduced by a cost for the accumulation of evidence. A more realistic assumption is that early decisions are induced to avoid a loss of potential future reward. This assumption can be formulated by maximising the reward rate rather than the expected reward.

Maximising the reward rate is equivalent to maximising the overall reward in an infinite sequence of structurally equivalent trials/decisions. The problem in

handling such an infinite sequence by Dynamic Programming is that the value before the first decision will be infinite, as an infinite number of reward-promising choices follow. The trick to still be able to find the optimal policy is to use an average-adjusted value that penalises the passage of δt time by a cost of $\rho\delta t$, where ρ is the reward rate (reward per unit time). This average-adjusted value thus expresses how much better or worse particular states are when compared to the average. It also allows us to handle all trials equivalently as the same trial.

To illustrate this concept, assume that we want to maximise the reward rate, Eq. (10), while assuming that the evidence reliability is know. In this case, choosing the option corresponding to $z = 1$ promises a reward of g, but causes an expected waiting time of $t_i + (1 - g)t_p$ (incurring a penalty time t_p being incorrect with probability $1 - g$) until the start of the next trial. Thus, the overall expected reward for this choice would be $g + t_i + (1 - g)t_p + V(1/2)$, where $V(1/2)$ is the value at the start of the next trial (as the initial belief will be $g = 1/2$) For choosing the option $z = -1$, the expected reward is $1 - g + t_i + gt_p + V(1/2)$. Accumulating more evidence causes a cost $(c + \rho)\delta t$. Overall, this leads to Bellman's equation

$$V(g) = \max \left[\begin{array}{c} g + t_i + (1 - g)t_p + V(1/2), 1 - g + t_i + gt_p + V(1/2), \\ \langle V(g') \rangle_{p(g'|g)} - (c + \rho)\delta t \end{array} \right]. \quad (30)$$

Adding a constant to all values does not change the resulting policy. Thus, we can choose one of the values freely, for example $V(1/2) = 0$. We can use this property to find the reward rate ρ, as $V(1/2) = 0$ is only guaranteed to hold if this reward rate is set to the correct value. This allows us to find the value function and the reward rate simultaneously. For more information on this approach, see Drugowitsch et al. (2012, 2014b).

5.2. **A cost for accumulating evidence that varies over time.** So far we have assumed the cost for accumulating evidence to remain constant throughout the evidence accumulation period. This does not need to be the case, and, in fact, humans and monkeys feature behavior that is compatible with the idea of a cost that rises over time (Drugowitsch et al., 2012). Such a rising cost is easily included in the Dynamic Programming formulation by making the cost time-dependent. For the case of a reliability that changes across trials this won't change much, as it already features a time-dependent value function. For a constant evidence reliability, in contrast, the value function will become time-dependent. For more information on this approach, see Drugowitsch et al. (2012).

5.3. **Reliability of evidence that fluctuates within individual decisions.** In our above formulation we have allowed the evidence reliability to vary across decisions, but have assumed it to remain constant within individual choices. This is unrealistic, as in real world choices, this reliability fluctuates all the time. Returning to the initial tennis example, assuming a constant reliability corresponding to assuming that we receive as much information about the ball's landing point when it has just left the opponent's racket as when it has already passed the net.

An approach to tackle this is to formulate a process in which the evidence reliability varies according some reliability process, and formulate the evidence accumulation process as an inference of both the hidden state and the momentary reliability. This approach then leads to a value function over both belief and reliability estimate, and consequently to a decision boundary that also depends on this reliability estimate. Details to this approach can be found in Drugowitsch et al. (2014b).

5.4. **Improving the speed of finding the optimal policy.** In particular when moving to higher-dimensional value functions, finding the expected future value can

be time-consuming. The naïve implementations scales quadratically in the size of the state space, and has problems with singularities for small δt. We have recently developed an approach that scales linearly with the state space, and gets around the singularity problem. It is based on finding the expected future value by defining a continuous-time flow from current to future expected value as a stochastic differential equation. This equation can then be solved by standard partial differential equation solvers that feature linear scaling and high robustness. Specificities about this approach are described in Drugowitsch et al. (2014b).

References

Bellman, R. (1954). The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–516.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113(4):700–765.

Cox, D. R. and Miller, H. D. (1965). *The theory of stochastic processes*. Wiley.

Drugowitsch, J., DeAngelis, G. C., Klier, E. M., Angelaki, D. E., and Pouget, A. (2014a). Optimal multisensory decision-making in a reaction-time task. *eLife*. 10.7554/eLife.03005.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., and Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *The Journal of Neuroscience*, 32(11):3612–3628.

Drugowitsch, J., Moreno-Bote, R., and Pouget, A. (2014b). Optimal decision-making with time-varying reliability. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 27*, pages 748–756. Curran Associates, Inc.

Palmer, J., Huk, A. C., and Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, 5(5):376–404.

University of Geneva
E-mail address: jdrugo@gmail.com