The Diffusion Model of Speeded Choice, from a Rational Perspective

Matt Jones, University of Colorado

August 27, 2017

1 Binary Decision Tasks

This chapter considers tasks in which an experimental subject must make a binary decision, such as a perceptual discrimination or a semantic or mnemonic judgment. For some examples, a character could be presented on a monitor and the subject must decide whether it is red or green, or a letter or a number; or a string of letters is presented and the subject must decide whether it is a word or not; or a word is presented and the subject must decide whether it was part of a previously studied list or not. We are interested in both the probability that the subject will give the correct answer and the response time (RT) of whatever answer the subject gives.

Our aim is to develop formal, mathematical models of this type of task that make predictions for response probability and RT. This chapter will begin by following the historical progression of models developed in the literature, from *signal-detection* models to *random walk* models to *diffusion* models, before presenting some new variations and results regarding the last of these.

All of the models considered here are based on the idea of *evidence sampling*. The assumption is that the subject, in perceiving the stimulus, observes or calculates some sort of information that bears on the correct answer. In a color discrimination task, this information would presumably concern the wavelength of the light coming from the stimulus, originating in the subject's photoreceptors and further processed in visual cortex, for example in red-green opponent-process cells. In a recognition memory task, the information would come from comparing the stimulus to memory, perhaps retrieving an explicit memory of the item from the study phase, or perhaps generating a continuous-valued familiarity signal. For present purposes, we will not be concerned with the specific nature of this information or how it is computed. Instead, the focus will be on how the observations, once obtained, are used to generate a response in the binary decision task.

Under this view, the problem facing the subject is to determine the relative support that the observed information lends to the two responses. The models considered here take a normative approach to this problem, treating it as one of statistical inference. Under this approach, each stimulus category (i.e., correct

response) is associated with a hypothesis. For example, hypothesis H_1 could be the that the stimulus is red (Category 1), and hypothesis H_2 could be that the stimulus is green (Category 2). The subject's goal is to use the observations to infer which hypothesis is probably correct and to select the corresponding response. More specifically, the models here take a Bayesian approach, using the likelihood of the observations under each hypothesis to determine a posterior belief in which hypothesis is correct, which in turn drives decision making.

2 Signal Detection Model

Consider first the simplest version of the model framework outlined above, where the subject observes a single sample and uses it to make a decision. This is referred to as a *signal detection model*, because historically it was developed for psychophysical tasks in which the subject's goal was to detect when a signal (e.g., an auditory tone) was presented, versus just background noise.

Define $\mathcal{X}$ as the space of all possible observation values that could be observed, and $x \in \mathcal{X}$ as the value the subject actually observes. For each hypothesis, there exists a probability distribution over the value of x when that hypothesis is true (i.e., when a stimulus from that category is presented). For example in a recognition memory task, if the observation takes the form of some familiarity signal, then we can define the distribution of familiarity values across all trials on which the stimulus is new (i.e., not on the studied list), as well as the distribution of familiarity values across all trials on which the stimulus was old. Formally, we write these two distributions as

$$P_1(x) = \Pr[x|H_1] \tag{1a}$$

and

$$P_2(x) = \Pr[x|H_2]. \tag{1b}$$

The notation $\Pr[x|H_i]$ indicates conditional probability, meaning the probability that x will be observed given that H_i is true.

Note that we need make no assumptions about the structure of the space $\mathcal{X}$. It could be a one-dimensional continuum (a subset of the real line), as in the case of a recognition familiarity signal or net activation of red-green opponent cells. In richer perceptual tasks the space of observations could be multidimensional (a subset of $\mathbb{R}^n$), and in higher cognitive tasks like lexical decision, it could be some complex structured space of orthographic and semantic representations. The models considered here require only the functions P_1 and P_2.[1]

Considered as a function of the hypotheses, $P_i(x)$ (i.e., $P_1(x)$ vs. $P_2(x)$) is referred to as *likelihood*, and it determines the relative support that the observation lends to the hypotheses. The intuition is that, if x is more probable

[1]The definitions in (1) are written assuming $\mathcal{X}$ is a discrete space. $\mathcal{X}$ could also be taken as a continuous space, with P_1 and P_2 representing probability density (with respect to some measure on $\mathcal{X}$) rather than probability mass. Everything that follows applies equally well to the continuous case as to the discrete case.

under Category 1 than Category 2, then observing x should increase the subject's belief that H_1 is the correct hypothesis. Assuming that the subject knows the functions $P_i(x)$ (or that these are the functions the subject believes, regardless of whether they are objectively accurate), then (s)he can use Bayes' rule to calculate the relative probabilities of the two hypotheses given the observation:

$$\begin{aligned}\frac{\Pr[H_1|x]}{\Pr[H_2|x]} &= \frac{\Pr[H_1]}{\Pr[H_2]} \cdot \frac{\Pr[x|H_1]}{\Pr[x|H_2]} \\ &= \frac{\Pr[H_1]}{\Pr[H_2]} \cdot \frac{P_1(x)}{P_2(x)}. \end{aligned} \tag{2}$$

The expression $\Pr[H_i|x]$ is called the *posterior probability* for H_i given x. This is the probability that the subject should assign to Category i after observing x (assuming inference is done optimally). The relation in (2) shows that the posterior probabilities of the two hypotheses depend on two things: the likelihoods and the *prior probabilities* $\Pr[H_i]$. The prior probabilities reflect the subject's beliefs about which response will be correct prior to observing the stimulus (i.e., before the start of the trial), for example due to learning of base rates or of sequential patterns in the trial sequence.

If the subject's goal is to maximize the probability of choosing the correct answer, then the optimal decision rule is to select response R_1 if $\Pr[H_1|x] > \Pr[H_2|x]$, that is if $\Pr[H_1|x] / \Pr[H_2|x] > 1$, and to select response R_2 otherwise (the choice is arbitrary in case of equality). In the simplest case where the priors are equal, $\Pr[H_1] = \Pr[H_2] = \frac{1}{2}$, this decision rule reduces to comparing $P_1(x)$ and $P_2(x)$ and choosing whichever hypothesis has the greater likelihood. One can imagine dividing the space $\mathcal{X}$ into two regions, according to whether $P_1(x) > P_2(x)$ or vice versa (again, cases of equality are assigned arbitrarily) and associating each region to the corresponding response. This partitioning might be done in advance, so that the decision-making process reduces to determining which region the observation lies in and selecting the associated response.

Figure 1 gives an illustration of this model for a simple case where $\mathcal{X}$ is a unidimensional continuum and P_1 and P_2 are both Gaussian distributions with equal variance. This *equal-variance signal detection model* is the simplest in the family of models to be described in this chapter, and probably the most frequently applied in the psychological literature. In this model, there is a decision criterion that lies midway between the two distributions, at the point where $P_1(x) = P_2(x)$. Optimal decision-making in this model corresponds to selecting a response according to which side of the criterion the observation lies on.

We can generalize this simple model in three ways. First, we can allow arbitrary likelihood functions $P_i(x)$ on an arbitrary space of observations $\mathcal{X}$. Second, we can allow arbitrary values for the prior probability $\Pr[H_1]$ (with $\Pr[H_2] = 1 - \Pr[H_1]$). Third, we can introduce asymmetric reward structures, such that the reward for being correct or the penalty for being wrong is different for the two responses, by writing r_{ij} as the payoff for selecting response R_j when the correct response is R_i (with $r_{11} > r_{12}$ and $r_{22} > r_{21}$). Under this notation,

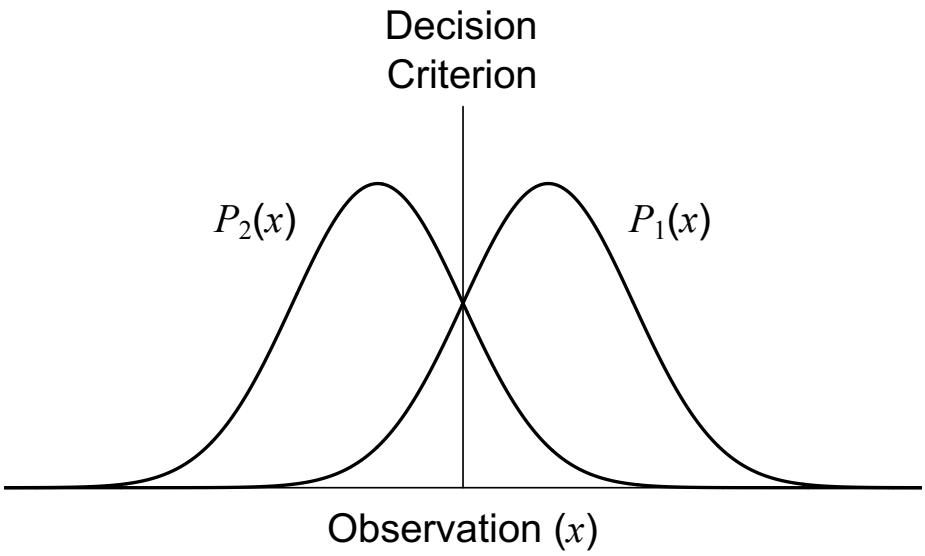

Figure 1: Illustration of the unbiased equal-variance signal detection model. An observation (x) is sampled from a unidimensional continuum represented by the horizontal line at the bottom. The probability distributions for x under the two hypotheses (i.e., stimulus categories) are represented by the curves labeled $P_1(x)$ and $P_2(x)$. Optimal decision making amounts to determining where the observation lies relative to the decision criterion, which is the point where the likelihoods $P_1(x)$ and $P_2(x)$ are equal.

the expected payoff for selecting R_j, conditioned on the observation, is given by

$$E\left[r|x, R_j\right] = r_{1j} \Pr\left[H_1|x\right] + r_{2j} \Pr\left[H_2|x\right]$$

and therefore the criterion for selecting response R_1 is:

$$E\left[r|x, R_1\right] > E\left[r|x, R_2\right] \iff \frac{\Pr\left[H_1|x\right]}{\Pr\left[H_2|x\right]} > \frac{r_{22} - r_{21}}{r_{11} - r_{12}}. \tag{3}$$

Combining this decision rule with (2) implies that R_1 is optimal if and only if

$$\frac{r_{11} - r_{12}}{r_{22} - r_{21}} \cdot \frac{\Pr\left[H_1\right]}{\Pr\left[H_2\right]} \cdot \frac{P_1\left(x\right)}{P_2\left(x\right)} > 1. \tag{4}$$

That is, the optimal response is determined by the net contribution of the prior belief, likelihood ratio, and ratio of reward contingency (i.e., the dependence of the reward on the subject's response, under each category). This is the optimal decision rule for the general signal detection model. In the unbiased case (equal priors and symmetric rewards), the ratios for the prior and reward contingency in (4) equal unity, and we recover the simpler decision rule of selecting R_1 whenever $P_1\left(x\right) > P_2\left(x\right)$.

Finally, because of the multiplicative nature of the optimal decision rule, and because it depends on the observation only through the likelihood ratio, $P_1\left(x\right)/P_2\left(x\right)$, it is mathematically convenient to introduce the *log-likelihood ratio*:

$$L\left(x\right) = \ln \frac{P_1\left(x\right)}{P_2\left(x\right)}. \tag{5}$$

The log-likelihood ratio concisely captures the net evidence that the observation contributes to the two hypotheses. Taking the logarithm of (2) shows that $L\left(x\right)$ determines how much the subject's beliefs should change from prior to posterior, when those beliefs are expressed on a log-odds scale:

$$\ln \frac{\Pr\left[H_1|x\right]}{\Pr\left[H_2|x\right]} = \ln \frac{\Pr\left[H_1\right]}{\Pr\left[H_2\right]} + L\left(x\right). \tag{6}$$

Likewise, the criterion (4) for selecting R_1 can be re-expressed in logarithmic form,

$$\ln \frac{r_{11} - r_{12}}{r_{22} - r_{21}} + \ln \frac{\Pr\left[H_1\right]}{\Pr\left[H_2\right]} + L\left(x\right) > 0, \tag{7}$$

showing that the log reward contingency, prior log-odds, and log-likelihood ratio combine additively to determine the optimal response.

3 Random Walk Model

We now extend the signal detection model to assume that the subject observes not just one sample but a series of samples. For example, a subject in a perceptual discrimination task might process the stimulus multiple times in succession, or a subject in a recognition memory task might make multiple queries

to memory. Formally, we model this observation process by assuming a series of observations x_n for $n = 1, 2, 3 \ldots$, all jointly independent conditioned on the stimulus category and all sampled from the same distribution, P_1 or P_2.

Denoting the sequence after n observations as $\mathbf{x}_n = (x_1, \ldots, x_n)$, we can use the conditional independence assumption to write the likelihood as

$$\begin{aligned} \Pr\left[\mathbf{x}_n | H_i\right] &= \prod_{m=1}^{n} \Pr\left[x_n | H_i\right] \\ &= \prod_{m=1}^{n} P_i\left(x\right). \end{aligned} \tag{8}$$

Therefore the log-likelihood ratio between the two hypotheses is equal to

$$\begin{aligned} \ln \frac{\Pr\left[\mathbf{x}_n | H_1\right]}{\Pr\left[\mathbf{x}_n | H_2\right]} &= \sum_{m=1}^{n} \ln \frac{P_1\left(x_m\right)}{P_2\left(x_m\right)} \\ &= \sum_{m=1}^{n} L\left(x_m\right). \end{aligned} \tag{9}$$

That is, the evidence provided by the sequence of observations is equal to the sum of the evidence provided by all of the individual observations. This is another important mathematical convenience of the log-odds representation, in addition to the additive relations in (6) and (7). Henceforth we use the term *evidence* specifically to refer to information or beliefs quantified on the log-odds scale.

Paralleling the derivation for the signal detection model, Bayes' rule gives the posterior odds, conditioned on the first n observations, as

$$\begin{aligned} \frac{\Pr\left[H_1 | \mathbf{x}_n\right]}{\Pr\left[H_2 | \mathbf{x}_n\right]} &= \frac{\Pr\left[H_1\right]}{\Pr\left[H_2\right]} \cdot \frac{\Pr\left[\mathbf{x}_n | H_1\right]}{\Pr\left[\mathbf{x}_n | H_2\right]} \\ &= \frac{\Pr\left[H_1\right]}{\Pr\left[H_2\right]} \cdot \prod_{m=1}^{n} \frac{P_1\left(x_m\right)}{P_2\left(x_m\right)}. \end{aligned} \tag{10}$$

As above, this relationship can also be written in terms of log-odds:

$$\ln \frac{\Pr\left[H_1 | \mathbf{x}_n\right]}{\Pr\left[H_2 | \mathbf{x}_n\right]} = \ln \frac{\Pr\left[H_1\right]}{\Pr\left[H_2\right]} + \sum_{m=1}^{n} L\left(x_m\right). \tag{11}$$

That is, the posterior log-odds equals the prior log-odds plus the sum of the log-likelihood ratios of the observations.

The expression in (11) suggests an intuitive process-level psychological model of decision making. First, define the net evidence E_n as the posterior log-odds after n observations:

$$E_n = \ln \frac{\Pr\left[H_1 | \mathbf{x}_n\right]}{\Pr\left[H_2 | \mathbf{x}_n\right]}. \tag{12}$$

Thus E_n represents the strength of belief that an optimal decision maker will have in H_1 versus H_2 after observing x_1 through x_n. The expression for the posterior log-odds in (11) then implies a recursive relationship for E,

$$E_n = E_{n-1} + L(x_n), \tag{13}$$

for all $n \geq 1$. Under the convention $\mathbf{x}_0 = \emptyset$ (i.e., an empty set of observations after zero trials), the starting point for E is the prior log-odds:

$$E_0 = \ln \frac{\Pr[H_1]}{\Pr[H_2]}. \tag{14}$$

Thus the decision maker starts at an evidence level determined by prior beliefs (E_0), and then uses the evidence provided by each successive observation ($L(x_n)$) to increment the cumulative evidence total (E_n).

Because x_n is a random variable (sampled from P_i when the stimulus comes from Category i), so is $L(x_n)$. This makes E a *stochastic process*, meaning a sequence of jointly distributed random variables. It is in fact a *Markov process*, meaning that the probability distribution for the next member of the sequence is fully determined by the value of the most recent member:

$$\Pr[E_n | \{E_0, \ldots, E_{n-1}\}] = \Pr[E_n | E_{n-1}]. \tag{15}$$

This relationship follows immediately from (13) together with the conditional independence between x_n and x_m for all $m < n$. Intuitively, a Markov process is thought of as memoryless, because its history has no impact on its future once its present state is known. This property makes this model appealing as a psychological model, because it implies that an optimal decision maker needs only to track E_n from one observation to the next, rather than remembering the full sequence of past observations $\mathbf{x}_n$.

Geometrically, E can be conceived as a unidimensional random walk, wherein the posterior log-odds starts at the prior log-odds and moves up or down according to the evidence (i.e., log-likelihood ratio) provided by each successive observation. A (stationary) random walk is a Markov process where the probability distribution for the increment $E_n - E_{n-1}$ is independent of E_{n-1} and is the same for all n. In this case, the distribution for the increment $L(x_n)$ is independent of n because of the assumption that all observations are drawn from the same distribution (P_1 or P_2). It is also useful to consider the expected value of this increment, which determines how rapidly the evidence grows, on average. This is the *drift rate* of the random walk. When the stimulus is from Category 1, the drift rate is equal to

$$\begin{aligned} \mathrm{E}[L(x) | H_1] &= \mathrm{E}_{P_1}\left[\ln \frac{P_1(x)}{P_2(x)}\right] \\ &= D_{\mathrm{KL}}(P_1 \| P_2), \end{aligned} \tag{16a}$$

where E_{P_1} indicates expected value according to the distribution P_1, and D_{KL} denotes *Kullback-Leibler (KL) divergence*, a standard measure of the difference

between two probability distributions ($D_{\mathrm{KL}}(p \| q) = \mathrm{E}_p\left[\ln \frac{p}{q}\right]$). Likewise, the drift rate when the stimulus is from Category 2 is given by

$$\begin{aligned} \mathrm{E}\left[L(x) | H_2\right] &= \mathrm{E}_{P_2}\left[\ln \frac{P_1(x)}{P_2(x)}\right] \\ &= -D_{\mathrm{KL}}(P_2 \| P_1). \end{aligned} \tag{16b}$$

These are sensible results, because they mean that the more different the two hypotheses are (i.e., the greater the divergence between P_1 and P_2 or vice versa), the faster an optimal observer will typically be able to discriminate between them based on a sequence of samples from one or the other.

To take an example, consider the Gaussian equal-variance signal detection model (Figure 1), generalized from one observation to a sequence of observations. Let μ_1 and μ_2 be the respective means of the Gaussian distributions P_1 and P_2, and let σ^2 be their shared variance. For a given observation x, the log-likelihood ratio is equal to

$$\begin{aligned} L(x) &= \ln \frac{\exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2}\right)}{\exp\left(-\frac{(x-\mu_2)^2}{2\sigma^2}\right)} \\ &= \frac{\mu_1 - \mu_2}{\sigma^2}\left(x - \frac{\mu_1 + \mu_2}{2}\right). \end{aligned} \tag{17}$$

Thus the evidence provided by an observation is a linear function of the value of that observation, with slope proportional to the mean difference between the two distributions and with neutral point (i.e., $L(x) = 0$) at the midpoint between the distributions. Because $L(x)$ is a linear function of a Gaussian variable, it has a Gaussian distribution as well. From (17), the mean of $L(x)$ is equal to

$$\begin{aligned} \mathrm{E}\left[L(x) | H_i\right] &= \frac{\mu_1 - \mu_2}{\sigma^2}\left(\mathrm{E}\left[x | H_i\right] - \frac{\mu_1 + \mu_2}{2}\right) \\ &= \pm\frac{1}{2}\left(\frac{\mu_1 - \mu_2}{\sigma}\right)^2 \end{aligned} \tag{18}$$

(with positive sign for $i = 1$ and negative sign for $i = 2$), and its variance is equal to

$$\begin{aligned} \mathrm{var}\left[L(x) | H_i\right] &= \left(\frac{\mu_1 - \mu_2}{\sigma^2}\right)^2 \mathrm{var}\left[x | H_i\right] \\ &= \left(\frac{\mu_1 - \mu_2}{\sigma}\right)^2. \end{aligned} \tag{19}$$

The ratio $(\mu_1 - \mu_2)/\sigma$ is referred to as d' in signal detection theory and is used as a measure of the standardized difference between distributions under the two hypotheses (most frequently when those distributions are assumed to be

Gaussian with equal variance). The derivation in (18) shows that the drift rate of the random walk is proportional to the square of this measure, consistent with the intuition that the mean rate of evidence accumulation is greater with greater separation between the distributions.

The absolute value of the drift rate is equal under the two hypotheses in the equal-variance Gaussian model, but it should be noted that this is not true in general. This is related to the fact that KL divergence is generally asymmetric. For example, in a model with Gaussian sampling distributions of unequal variances (σ_1^2 and σ_2^2), the drift rates turn out to be

$$D_{\mathrm{KL}}\left(P_1 \| P_2\right) = \frac{1}{2}\left(\left(\frac{\mu_1 - \mu_2}{\sigma_2}\right)^2 - 1 + \frac{\sigma_1^2}{\sigma_2^2} - \ln\frac{\sigma_1^2}{\sigma_2^2}\right) \tag{20a}$$

and

$$-D_{\mathrm{KL}}\left(P_2 \| P_1\right) = -\frac{1}{2}\left(\left(\frac{\mu_1 - \mu_2}{\sigma_1}\right)^2 - 1 + \frac{\sigma_2^2}{\sigma_1^2} - \ln\frac{\sigma_2^2}{\sigma_1^2}\right), \tag{20b}$$

which are generally not equal in absolute value. If, say, $\sigma_1 > \sigma_2$, then the drift rate under Category 1 will be greater in magnitude than the drift rate under Category 2. Intuitively this is because is it easier for extreme samples from Category 1 to provide strong evidence against H_2 than vice versa.

The random-walk model can be applied as a psychological model of decision making under two additional assumptions. First, there must be some decision rule that specifies at every possible state of the process—that is, for any value of the pair (n, E_n)—whether the subject terminates the sampling process and responds with R_1 or R_2, or whether the subject defers the decision and continues with another sample. Second, a time constant Δt must be specified, representing the physical time elapsed between successive samples (for simplicity, we assume samples are equally spaced). Under specifications of these assumptions, the model's prediction on any trial comprises the response dictated by the decision rule together with the response time (RT). The predicted RT will equal $n \cdot \Delta t$, where n is the number of samples observed, plus perhaps some nondecision time t_0 to model processes such as sensory encoding and motor execution that occur outside of the decision process itself. Across trials, the response and RT are jointly distributed random variables. Thus the model's predictions constitute a joint distribution over the response and RT, separately for each stimulus category. Equivalently, for each stimulus category, the model yields predictions for the probabilities of both responses together with a conditional distribution of RT for each response (i.e., correct and incorrect).

An appealing feature of this psychological model is that it corresponds to the classic sequential probability-ratio test (SPRT), a statistical procedure wherein a sequence of observations is used to decide between two hypotheses. A central question in that statistical setting concerns the optimal stopping rule: when should the observer stop sampling and make a decision, versus continuing to draw more samples? A standard result known as the Wald-Wolfowitz theorem

states that the optimal decision rule for the SPRT is to sample until the posterior reaches either an upper threshold, α, or a lower threshold, β (typically $\beta < 0 < \alpha$), and then to choose R_1 or R_2 accordingly. Both thresholds are fixed across time, that is, independent of n. Figure 2 illustrates the random walk model with this decision rule. According to the Wald-Wolfowitz theorem, this model is optimal in the sense that any other decision rule with the same error rates (i.e., probabilities of choosing R_2 when H_1 is true and of choosing R_1 when H_2 is true) will require at least as many samples on average. Thus it is impossible to achieve superior accuracy with a shorter mean RT. In particular, one can set the thresholds as α and $-\alpha$ in cases where it is desirable to equate the two error rates, for example because they incur equal costs (i.e., the reward contingency ratio in (3) equals unity). The choice of α determines the observer's *speed-accuracy tradeoff*, in that smaller α leads to faster responses but larger α leads to fewer errors. We defer analysis of RT predictions until Section 7, but it is easy to see that the log-odds of an error will approximately equal $-\alpha$, or equivalently the error rates will both be approximately $(1 + e^{\alpha})^{-1}$. This is because the observer is terminating the trial when the posterior probability of the chosen response being correct is approximately α (assuming the observer has access to the correct prior probability and likelihood functions, and so can compute the correct posterior). This value for the error rate is nonetheless approximate because, with a discrete sequence of samples, the random walk will generally jump across the threshold rather than landing exactly on it. In the continuous-time model that we consider next, the approximation becomes exact.

4 Continuous-Time Model

We now build on the random walk model to derive a model in which evidence accumulation occurs not in discrete steps but continuously. To achieve this, we consider a sequence of random walk models in which the time between samples (Δt) approaches zero, and we derive a continuous-time evidence process that is the limit of the discrete-time random walks. Under the right assumptions about the means and variances of the evidence increments in the discrete models, the limiting process follows a directed Wiener diffusion (i.e., Brownian motion) process. This limiting model, when viewed as a psychological model of decision making, is known as the *diffusion model*.

Formally, assume there exists a sequence of random walk models, indexed by $k \in \mathbb{N}$, with $\lim_{k\to\infty} \Delta t_k = 0$ and with sampling distributions P_i^k ($i \in \{1, 2\}$) and corresponding log-likelihood functions $L_k(x)$ that satisfy

$$\mathrm{E}\left[L_k(x) \,|H_i\right] = \xi_i \Delta t_k \tag{21}$$

and

$$\mathrm{var}\left[\, L_k(x) |\, H_i\right] = \eta^2 \Delta t_k. \tag{22}$$

That is, both the conditional means and the conditional variance of the evidence increments in each model are proportional to its time step. Thus the mean rate

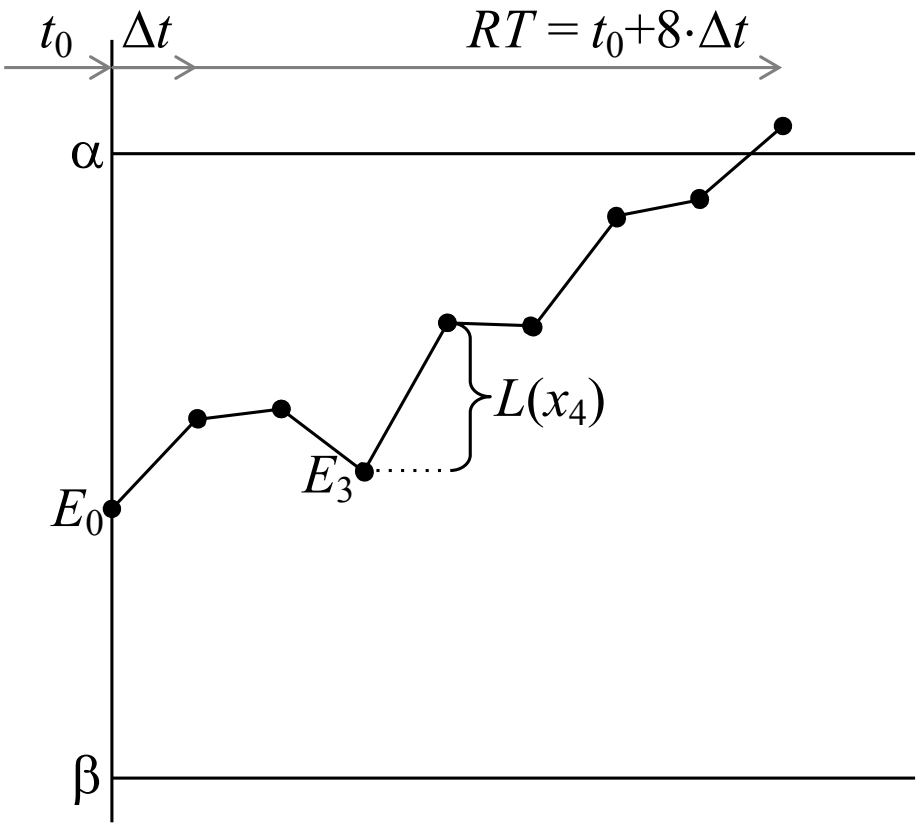

Figure 2: Illustration of the random walk model. Evidence (E) for deciding between two hypotheses is accumulated over time based on a series of observations (x_n). The evidence contributed by each observation is equal to its log-likelihood ratio ($L(x_n)$). The starting point for the evidence process (E_0) is determined by the prior probabilities of the hypotheses, specifically as the prior log-odds. After n observations, the posterior log-odds is given by E_n. Each hypothesis has a threshold, α for H_1 and β for H_2, and crossing this threshold will trigger the sampling process to stop and the corresponding response to be selected. The temporal spacing between successive observations is denoted Δt, and the model's response time (RT) is equal to that spacing times the total number of observations taken, plus a nondecision time (t_0).

of evidence accumulation, $\mathrm{E}\left[L\left(x\right)|H_i\right]/\Delta t$, and the growth rate of the variance, $\mathrm{var}\left[L\left(x\right)|H_i\right]/\Delta t$, are both constant across all models in the sequence. This property is critical for the continuous-time limit to be well-behaved, as will be seen shortly.

It is easy to construct such a sequence of models. For example, a sequence of equal-variance Gaussian sampling models can be defined by setting

$$d'_k = c\sqrt{\Delta t_k} \tag{23}$$

for any constant $c > 0$. (Recall that $d' = \left(\mu_1 - \mu_2\right)/\sigma$; therefore (23) could be achieved in various ways, for example by fixing μ_1 and σ and varying μ_2.) Thus this construction implies that the discriminability between P_1 and P_2 converges to zero in proportion to the square-root of the time step. From (18) and (19), the mean and variance of the evidence increments are

$$\mathrm{E}\left[L_k\left(x\right)|H_i\right] = \pm\frac{c}{2}\Delta t_k \tag{24}$$

and

$$\mathrm{var}\left[L_k\left(x\right)|H_i\right] = c\Delta t_k. \tag{25}$$

Therefore we can take any sequence of Δt_k satisfying $\lim_{k\to\infty}\Delta t_k = 0$, and the random walk models defined by (23) will satisfy the scaling properties in (21) and (22).

Given any sequence of random walk models satisfying (21) and (22), all with the same prior log-odds, define each model's evidence trajectory through time as

$$E_k\left(t\right) = E^k_{\lfloor t/\Delta t_k\rfloor}, \tag{26}$$

where $\lfloor\cdot\rfloor$ is the floor function (i.e., $\lfloor z\rfloor$ is the greatest integer less than or equal to z), and E^k_n is the evidence level (i.e., posterior log-odds) of model k after n observations as in (12). The definition in (26) formalizes the assumption that the model's evidence is updated after each Δt_k interval, by setting $E_k\left(t\right)$ equal to the value of the evidence process after the last update before time t. Note that $E_k\left(0\right)$ is equal to the common prior log-odds, which we denote as $E\left(0\right)$. For each k, the trajectory $E_k\left(t\right)$ is a random function of t, with distribution governed by the updating rule in (13) and by the distribution of increments $L_k\left(x\right)$. The question now is how the distribution of $E_k\left(t\right)$ behaves under the limit $k\to\infty$.

For any single value of t, $E_k\left(t\right)$ is equal to $E\left(0\right)$ plus a sum of loglikelihood increments from $\lfloor t/\Delta t_k\rfloor$ independent observations, with the conditional mean and variance of these increments given by (21) and (22). Because the expressions in (21) and (22) are proportional to Δt_k, the mean and variance of the sum of the increments depends only on t and not on Δt_k, except for contributions of rounding error in $\lfloor t/\Delta t_k\rfloor$. Therefore in the limit, the distribution of $E_k\left(t\right)$ obeys

$$\begin{aligned}\lim_{k\to\infty}\mathrm{E}\left[E_k\left(t\right)|H_i\right] &= \lim_{k\to\infty}\left(E\left(0\right) + \xi_i\Delta t_k\left\lfloor\frac{t}{\Delta t_k}\right\rfloor\right)\\ &= E\left(0\right) + \xi_i t\end{aligned} \tag{27}$$

and

$$\lim_{k\to\infty} \operatorname{var}\left[E_k\left(t\right)|H_i\right] = \lim_{k\to\infty}\left(\eta^2 \Delta t_k \left\lfloor \frac{t}{\Delta t_k} \right\rfloor\right)$$
$$= \eta^2 t. \tag{28}$$

Moreover, the central limit theorem implies that the limiting distribution is Gaussian. The same considerations apply to the difference $E_k\left(t_2\right) - E_k\left(t_1\right)$ for any two time points $t_1 < t_2$, and because of the Markov property of the random walk models these properties hold conditioned on the history of the process up to t_1, denoted $\mathbf{E}_k\left(t_1\right)$:

$$\lim_{k\to\infty} \mathrm{E}\left[E_k\left(t_2\right) - E_k\left(t_1\right)|H_i, \mathbf{E}_k\left(t_1\right)\right] = \xi_i\left(t_2 - t_1\right) \tag{29}$$

and

$$\lim_{k\to\infty} \operatorname{var}\left[E_k\left(t_2\right) - E_k\left(t_1\right)|H_i, \mathbf{E}_k\left(t_1\right)\right] = \eta^2\left(t_2 - t_1\right), \tag{30}$$

with a distribution that is Gaussian in the limit.

The properties in (27) through (30) imply that the processes $E_k\left(t\right)$ (conditioned on each hypothesis H_i) converge in distribution to a Wiener diffusion process, $E\left(t\right)$, with drift rate ξ_i and diffusion rate η^2. This Wiener diffusion process is defined as a stochastic process where the marginal distribution at any point in time is Gaussian with linear growth in the mean and variance,

$$E\left(t\right) \sim \mathcal{N}\left(E\left(0\right) + \xi_i t, \eta^2 t\right), \tag{31}$$

and with increments that are independent of the history (a property that can be seen to imply the Markov property):

$$E\left(t_2\right) - E\left(t_1\right) \perp \mathbf{E}\left(t_1\right)|H_i. \tag{32}$$

More precisely, for any finite set of time points $t_1 < \cdots < t_m$, the values of the process $E\left(t_1\right)$ through $E\left(t_m\right)$ have a multivariate Gaussian distribution, with means and variances given by (31), and with covariances given by

$$\operatorname{cov}\left[E\left(t_i\right), E\left(t_j\right)\right] = \eta^2 \min\left\{t_i, t_j\right\}. \tag{33}$$

Figure 3 illustrates the relationship between the discrete-time random walk model and the continuous-time diffusion model. Under the construction described here, one can think of the time step of the random walk model being repeatedly subdivided until (in the limit) the process evolves continuously. The linear scaling properties in (21) and (22) imply that this subdivision operation preserves the growth rate in both the expected value and the variance of the process, so that these rates are well-defined in the limit. The growth rate of the mean is referred to as the *drift rate*, and the growth rate of the variance the *diffusion rate*.

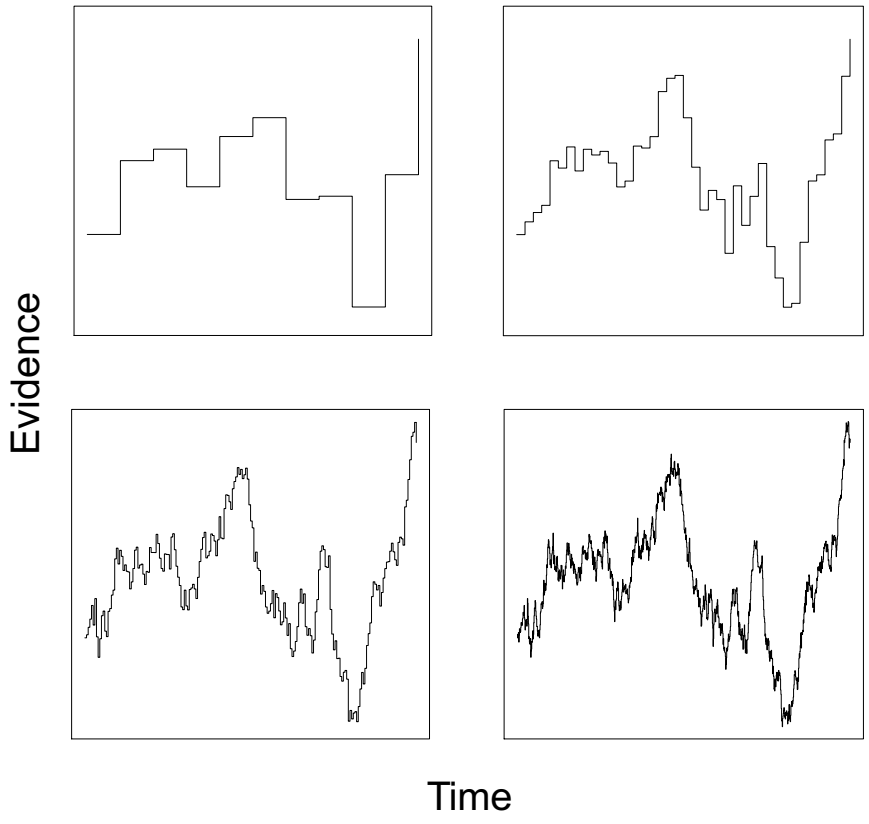

Figure 3: Illustration of the relationship between random walk and diffusion models of evidence accumulation. Each plot shows a sample trajectory from a random walk process. Proceeding from upper-left to lower-right, the time step becomes finer and the evidence increments become smaller. If the mean and variance of the increments are proportional to the time step, then the sequence of random walks converges in distribution to a diffusion process.

The diffusion process can be built into a psychological model of speeded decision making just like the random walk model was. The model embodies the idea that decisions are made based on a continuous stream of evidence, formally described by white noise. Completing the model requires specifying upper and lower thresholds, α and β, for the two responses. Because the evidence process is continuous, these thresholds will exactly determine the log-odds of a correct response, unlike the random walk model where the evidence jumps across the threshold. That is, when the model selects response R_1, the log-odds that Category 1 is correct exactly equals α, and similarly with R_2 and β (again, this assumes the observer knows the correct prior $E(0)$ and log-likelihood ratio function $L(x)$). Also, the diffusion model predicts a dense set of possible RT values, as opposed to the discrete set of possible RTs under a pure random walk model.

5 Bayesian Diffusion Model

Although the historic development of random walk and diffusion models of decision making was founded on the framework of statistical inference in the SPRT, current treatments of the diffusion model in the psychological literature depart from the normative framing, casting the diffusion process at a purely mechanistic level. That is, they simply posit that decision making is based on some cognitive variable that obeys the dynamics of a diffusion process, without interpreting that process as the result of Bayesian inference over some input stream. Here we show how this mechanistic diffusion model can be reformulated in the normative framework of Section 4, and how such a reformulation offers further insights into the properties of the model.

The starting point of the mechanistic diffusion model is to assume a stochastic process, denoted here as $e(t)$ for $t \geq 0$, defined by the stochastic differential equation

$$\mathrm{d}e = \mu_i \mathrm{d}t + \sigma \mathrm{d}B(t). \tag{34}$$

Here i indicates the true stimulus category, and $B(t)$ represents a standard Brownian motion process (i.e., with zero drift and with diffusion rate equal to unity). We assume without loss of generality that the input drift rates satisfy $\mu_1 > \mu_2$, and that $e(0) = 0$. The general version of the mechanistic diffusion model assumes the starting point $e(0)$ can take on arbitrary values, as a free parameter. This assumption is reintroduced in the next section, but it is irrelevant for the Bayesian derivation in the present section because the starting point can simply be subtracted away (e.g., one could define $e'(t) = e(t) - e(0)$ and do inference from e').

The Brownian motion process can be thought of as a limit of random walks with time step approaching zero (just as in Section 4). It has the properties that any increment $B(t_2) - B(t_1)$ is a random variable with Normal distribution, zero mean, and variance equal to $t_2 - t_1$ (for $t_1 \leq t_2$); and that the value of such an increment is independent of the prior history, $(B(\tau))_{\tau \leq t_1}$. Thus $e(t)$ evolves according to a sum of a deterministic, linear process with slope μ_i and a

stochastic process with diffusion rate σ^2. To be clear, we use the term *diffusion rate* to refer to σ^2 (not σ), because it represents the rate at which the variance in e grows over any interval of time:

$$\text{var}\left[e\left(t_2\right) - e\left(t_1\right)\right] = \sigma_i^2\left(t_2 - t_1\right) \tag{35}$$

for $t_1 \leq t_2$. Typically the diffusion rate is assumed to be the same under both stimulus categories, and below we give a justification for that assumption based on the Bayesian interpretation.

Our Bayesian approach is to treat e as comprising the observations that the subject uses to infer the correct hypothesis. This is consistent with the mechanistic diffusion model, except that whereas that model directly applies an arbitrary threshold to $e\left(t\right)$, here we use $e\left(t\right)$ to calculate posterior log-odds and define the decision rule on the log-odds. Thus $e\left(t\right)$ is a continuous analogue of x_n from the random walk model. More precisely, we write $\mathbf{e}_t = \left(e\left(\tau\right)\right)_{\tau \leq t}$ as the full trajectory up to time t, and do Bayesian inference using $\mathbf{e}_t$ in the same way the random walk model does inference using the discrete sequence $\mathbf{x}_n$.

We calculate a posterior from $\mathbf{e}_t$ as follows. For any finite value of Δt, define a discrete sequence of observations x_n as increments of e:

$$x_n = e\left(n\Delta t\right) - e\left(\left(n-1\right)\Delta t\right). \tag{36}$$

The properties of Brownian motion given above imply that these observations are jointly independent for different n and distributed as

$$x_n \sim \mathcal{N}\left(\mu_i \Delta t, \sigma^2 \Delta t\right). \tag{37}$$

The log-likelihood ratio for x_n can be calculated as in (17), as

$$L\left(x_n\right) = \frac{\mu_1 - \mu_2}{\sigma^2}\left(x_n - \frac{\mu_1 + \mu_2}{2}\Delta t\right). \tag{38}$$

Therefore the log-likelihood ratio for the sequence $\mathbf{x}_n = \left(x_1, \ldots, x_n\right)$ equals

$$\begin{aligned} \ln \frac{p\left[\mathbf{x}_n | H_1\right]}{p\left[\mathbf{x}_n | H_2\right]} &= \sum_{m=1}^{n} \frac{\mu_1 - \mu_2}{\sigma^2}\left(x_m - \frac{\mu_1 + \mu_2}{2}\Delta t\right) \\ &= \frac{\mu_1 - \mu_2}{\sigma^2}\left(e\left(n\Delta t\right) - \frac{\mu_1 + \mu_2}{2} n\Delta t\right). \end{aligned} \tag{39}$$

(We use p to indicate probability density, rather than Pr for probability mass, because x and hence $\mathbf{x}$ are now necessarily continuous-valued.) If we fix t and let $\Delta t = t/n$, then this equation becomes

$$\ln \frac{p\left[\mathbf{x}_n | H_1\right]}{p\left[\mathbf{x}_n | H_2\right]} = \frac{\mu_1 - \mu_2}{\sigma^2}\left(e\left(t\right) - \frac{\mu_1 + \mu_2}{2} t\right). \tag{40}$$

Thus we see that inference does not depend on the step size Δt, and by letting $\Delta t \to 0$ (i.e., $n \to \infty$), that $e\left(t\right)$ is a sufficient statistic for inferring the correct

hypothesis from the full trajectory of the process, $\mathbf{e}_t$. That is, the posterior probability depends only on the current value of the diffusion process and not on its history.

Therefore we have a well-defined continuous-time Bayesian evidence-accumulation model, with input defined by the stochastic process $e(t)$. Using the expression for the log-likelihood ratio in (40), and the invariance across sampling densities (i.e., Δt), the posterior log-odds are given by

$$\ln \frac{\Pr[H_1|\mathbf{e}_t]}{\Pr[H_2|\mathbf{e}_t]} = \ln \frac{\Pr[H_1]}{\Pr[H_2]} + \frac{\mu_1 - \mu_2}{\sigma^2} \left(e(t) - \frac{\mu_1 + \mu_2}{2} t \right). \tag{41}$$

Therefore the posterior is a linear transformation of the input, and thus follows a diffusion process itself, with new drift and diffusion rates.

To understand the relationship between the mechanistic diffusion model (where decisions are based directly on the input $e(t)$) and the Bayesian model (where the input is first transformed according to (41)), we reparameterize the latter as follows. First, define a *drift criterion*,

$$\theta = \frac{\mu_1 + \mu_2}{2}, \tag{42}$$

as the midpoint between the input drift rates for the two categories. The drift criterion is a dynamic analogue of the equal-likelihood criterion in signal-detection theory, in that $e(t) > \theta t$ implies a positive log-likelihood ratio (i.e., evidence for H_1), and $e(t) < \theta t$ implies a negative log-likelihood ratio (i.e., evidence for H_2). Second, define a *signal-to-noise ratio*,

$$\phi = \frac{\mu_1 - \mu_2}{\sigma^2}, \tag{43}$$

as the difference in input drift rates between hypotheses divided by the diffusion rate. The signal-to-noise ratio is analogous to the concept of d' in signal detection theory in that it gives a standardized measure of how separated are the samping distributions under the two hypotheses. Another connection between d' and ϕ is that both parameters determine how informative the observations are, by providing scaling factors to convert from the input to log-likelihood ratio. From (17), the log-likelihood ratio in Gaussian equal-variance signal detection and random walk models is $L(x) = d' \cdot (x - \theta)/\sigma$, whereas in the Bayesian diffusion model the log-likelihood ratio is $L(e(t)) = \phi \cdot (e(t) - \theta t)$. Thus larger values of ϕ imply that the posterior moves more rapidly with changes in e.

As above, let $E(t)$ denote the evidence level (i.e., posterior log-odds) at time t:

$$E(t) = \ln \frac{\Pr[H_1|\mathbf{e}_t]}{\Pr[H_2|\mathbf{e}_t]}. \tag{44}$$

Then, using the above definitions, (41) becomes

$$E(t) = E(0) + \phi(e(t) - \theta t). \tag{45}$$

Therefore optimal inference requires knowledge of three parameters: the prior probability ($E(0)$), the drift criterion (θ), and the signal-to-noise ratio (ϕ). These can be considered as properties of the task environment or the psychological processes generating the input signal $e(t)$. Provided the observer knows these three values, then calculating the posterior log-odds for the hypotheses can be accomplished by the linear transformation in (45). The result is a new diffusion process, with drift rates

$$\begin{aligned}\xi_i &= \phi(\mu_i - \theta) \\ &= \pm\frac{(\mu_1 - \mu_2)^2}{2\sigma^2},\end{aligned} \tag{46}$$

(positive for $i = 1$ and negative for $i = 2$) and diffusion rate

$$\begin{aligned}\eta^2 &= \phi^2\sigma^2 \\ &= \frac{(\mu_1 - \mu_2)^2}{\sigma^2}.\end{aligned} \tag{47}$$

If the observer wants to achieve a given accuracy level $\Pr[\text{correct}] = \rho$ (equal for the two categories), then the optimal decision rule is to terminate sampling and choose a response as soon as $|E(t)| \geq \alpha$, using a threshold given by

$$\alpha = \ln\frac{\rho}{1-\rho}. \tag{48}$$

As in the random walk model, the choice of threshold controls the subject's speed-accuracy tradeoff, with larger threshold values yielding greater accuracy but slower RT.

6 Translation between Diffusion Models

Because the evidence $E(t)$ itself follows a diffusion process, in practice one might dispense with e and use E directly as the starting point for modeling. From a mechanistic standpoint, e adds nothing to the model's predictions, barring some physiological theory of the input signal that would allow it to be measured. Indeed, the Bayesian diffusion model can be treated mathematically as simply a special case of the standard, mechanistic model: It is defined by a stochastic evidence process, $E(t)$, with particular values for the drift rates and diffusion rate, as well as a starting point $E(0)$ and thresholds $\pm\alpha$. As seen in (46) and (47), the drift rate of $E(t)$ is necessarily equal to half of its diffusion rate, and otherwise the model's parameters (i.e., starting point, threshold, and drift/diffusion rates) can be independently specified.

Therefore we can think of the Bayesian diffusion model as a reparameterization of the mechanistic one. Under this view, the value of the derivation presented in Section 5 is that it provides a normative Bayesian foundation for the diffusion model. To summarize, under the Bayesian diffusion model,

the observer is assumed to have access to a continuous stream of input information conforming to a Wiener diffusion process with unknown drift rate, or equivalently an integral of a white-noise process with unknown mean. Optimal Bayesian inference is applied to this input to infer the correct hypothesis via its influence on the input drift rate. The evidence process $E(t)$ is obtained from the input by subtracting the drift criterion, which corresponds to neutral input that has equal likelihood under both hypotheses; scaling by the signal-to-noise ratio, which determines how much information the input carries, to transform to units of log-odds; and adding the prior log-odds to reflect prior expectations. The drift and diffusion rates of this evidence process are determined by the difference in input rates between the two hypotheses, together with the noise (diffusion rate) in the input. The starting point, $E(0)$, is determined by the prior probabilities of the two stimulus categories. The decision thresholds correspond to the observer's choice of the log-odds that each response will be correct. Provided the observer knows the input rates μ_1 and μ_2 (or equivalently, θ and ϕ) and the prior probabilities (e.g., the base rate), then the inference process described above can be carried out, and any desired performance level ρ can be achieved optimally—that is, while minimizing the mean RT. Thus the Bayesian diffusion model also raises the question of how an experimental subject might come to know these task parameters, which suggests rich opportunities for integrating diffusion models of decision making with mathematical models of learning.

This normative framing offers answers to two conceptual challenges within the diffusion framework. First, one might ask why the diffusion model assumes the same diffusion rate under both hypotheses. After all, this is not the case under the signal detection or random walk models, where the variance of the input or of the evidence increment can differ between the hypotheses. Nevertheless, the equal-variance assumption is universal in applications of the diffusion model, and the Bayesian framing given here offers a justification. To see this, consider an alternative model in which the input process is defined by

$$\mathrm{d}e = \mu_i \mathrm{d}t + \sigma_i \mathrm{d}B(t), \tag{49}$$

with unequal diffusion rates $\sigma_1^2 \neq \sigma_2^2$. As above, define a sequence of discrete observations by (36). These observations now have a variance that depends on the correct hypothesis:

$$x_n \sim \mathcal{N}\left(\mu_i \Delta t, \sigma_i^2 \Delta t\right). \tag{50}$$

The log-likelihood ratio implied by this unequal-variance model is given by

$$L(x) = -\frac{(x-\mu_1\Delta t)^2}{2\sigma_1^2\Delta t} + \frac{(x-\mu_2\Delta t)^2}{2\sigma_2^2\Delta t} - \ln\frac{\sigma_1}{\sigma_2}, \tag{51}$$

and the expected value of this quantity under each stimulus category equals

$$\mathrm{E}\left[L(x)\,|H_1\right] = \frac{(\mu_1-\mu_2)^2}{2\sigma_2^2}\Delta t + \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} - 1 - \ln\frac{\sigma_1^2}{\sigma_2^2}\right) \tag{52a}$$

and

$$\mathrm{E}\left[L\left(x\right)|H_2\right] = -\frac{\left(\mu_1-\mu_2\right)^2}{2\sigma_1^2}\Delta t - \frac{1}{2}\left(\frac{\sigma_2^2}{\sigma_1^2} - 1 - \ln\frac{\sigma_2^2}{\sigma_1^2}\right), \quad (52b)$$

paralleling the result for the unequal-variance discrete models in (20). Now fix t and let $\Delta t = t/n$. The expected posterior log-odds after time t is given by

$$\begin{aligned}\mathrm{E}\left[E\left(t\right)|H_1\right] &= \ln\frac{\Pr\left[H_1\right]}{\Pr\left[H_2\right]} + \sum_{m=1}^{n}\mathrm{E}\left[L\left(x_m\right)|H_1\right] \\ &= \ln\frac{\Pr\left[H_1\right]}{\Pr\left[H_2\right]} + \frac{\left(\mu_1-\mu_2\right)^2}{2\sigma_2^2}t + \frac{n}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} - 1 - \ln\frac{\sigma_1^2}{\sigma_2^2}\right). \quad (53a)\end{aligned}$$

and

$$\mathrm{E}\left[E\left(t\right)|H_1\right] = \ln\frac{\Pr\left[H_1\right]}{\Pr\left[H_2\right]} + \frac{\left(\mu_1-\mu_2\right)^2}{2\sigma_1^2}t - \frac{n}{2}\left(\frac{\sigma_2^2}{\sigma_1^2} - 1 - \ln\frac{\sigma_2^2}{\sigma_1^2}\right). \quad (53b)$$

If the variances are equal then the last main summand in both (53a) and (53b) equals zero, and the posterior log-odds follows a diffusion process with drift rate $\pm\left(\mu_1-\mu_2\right)^2/\left(2\sigma^2\right)$ as already found above. However, if the variances differ then the last term will be strictly positive in (53a) and strictly negative in (53b). As $\Delta t \to 0$, meaning $n \to \infty$, this term approaches infinity, implying the observer approaches perfect certainty of the correct category. This is a property of statistical inference with diffusion processes: Observation on an arbitrarily fine timescale provides no advantage in inferring the drift rate (because the final value of the process is a sufficient statistic for that parameter), but it provides perfect information for inferring the diffusion rate. Therefore any diffusion model of decision making must assume equal diffusion rates for the two categories in order to be sensible from a Bayesian perspective. Otherwise, an ideal observer would be able to respond with perfect accuracy, using an arbitrarily short amount of time. This is also why we stated above that the observer needs only to know θ, ϕ, and the prior probabilities of the categories: σ does not need to be known because it can be exactly inferred from the input.

Second, the Bayesian diffusion model offers an elegant solution to a particular redundancy among the parameters in the mechanistic diffusion model. Specifically, the mechanistic diffusion model suffers a *scaling degeneracy* due to the fact that the scale of $e\left(t\right)$ has no impact on the model's predictions. If the drift rate, square-root of the diffusion rate (i.e., σ), starting point, and thresholds for the mechanistic model were all multiplied by any positive number, the only consequence would be a change in the internal scaling of $e\left(t\right)$, and the model's predictions for the joint distribution of response and RT would be unchanged. In practice, many modelers have adopted a convention to fix σ at an arbitrary value (usually 0.1), or in related models to fix the sum of the drift rates $\mu_1+\mu_2$, so as to remove this degeneracy when estimating parameter values from data. In the Bayesian diffusion model, this indeterminacy never arises, because

the evidence process has a uniqely determined scale, defined by log-odds (and log-likelihood ratio) of the two hypotheses. This unique scaling manifests in the constraint noted above for the model's parameters, that the diffusion rate is exactly twice the drift rate. Thus the Bayesian diffusion model offers a more principled solution to the scaling degeneracy than the arbitrary solutions just mentioned, specifically by constraining the drift and diffusion rates to obey a 1:2 ratio. It is easy to see that, given any parameterization of the mechanistic model, it can always be rescaled to fit this form by multiplying all parameters by ϕ. This implies that the 1:2 constraint is not a predictive constraint at all; that is, it does not make the Bayesian model any more restricted than the mechanistic model. Moreover, the derivation of (45) shows that this rescaling always puts the evidence process into units of log-odds, and therefore has the benefit of making all model parameters more interpretable. In particular, under the parameterization entailed by the 1:2 constraint, the model's starting point can be directly interpreted as the observer's subjective prior log-odds, and the thresholds can be directly interpreted as values chosen by the observer for the log-odds that each response will be correct.

Despite the close correspondence between the Bayesian and mechanistic diffusion models, there are some important differences. First, the mechanistic model is often applied to experimental settings involving multiple stimulus subtypes within each category, such as perceptual stimuli of varying salience or lexical stimuli of varying corpus frequency. These cases are usually modeled by assuming a different drift rate for each subtype. The Bayesian model can easily incorporate this assumption, by allowing different drift rates in the input process. It does, however, raise the question of how this additional complexity in the input relates to the hypotheses held by the observer. One possibility is that the input has multiple drift rates, μ_{ij}, indexed by both stimulus category (i) and subtype (j), but that the observer's hypotheses assume a single drift rate per category, $\hat{\mu}_i$, perhaps equal to the frequency-weighted mean of the true drift rates across all subtypes. Another possibility is that the hypotheses incorporate the subtyping of stimuli, such that H_i specifies a mixture distribution over diffusion processes with different drift rates, μ_{ij} for all possible j. The latter possibility would make the calculation of the posterior more complex than the result in (45). Either way, the 1:2 condition offered above for resolving the mechanistic model's scaling degeneracy is too simple to be applicable to experiment designs with multiple stimulus subtypes, but there should be extensions of this idea that would achieve similar results.

A second set of differences lie in constraints on the parameters of the two models. On one hand, the mechanistic model is more general than the Bayesian model in that it can assume arbitrary drift rates for the two stimulus categories, μ_1 and μ_2, whereas in the Bayesian model the drift rates must satisfy $\xi_2 = -\xi_1$. However, if the mechanistic model assumes asymmetric drift rates ($\mu_2 \neq -\mu_1$) then its decision rule is non-optimal in the sense of the Wald-Wolfowitz theorem. That is, its thresholds (which are time-invariant as criteria on $e(t)$) become time-varying when translated to criteria on $E(t)$ (i.e., on posterior log-odds). This follows from (45) and the fact that $\theta \neq 0$ when the mechanistic model's drift

rates are asymmetric. This issue is well known in the literature on the diffusion model, and in practice it is often assumed that the drift rates are symmetric ($\mu_2 = -\mu_1$), under the rationale that the subject knows the correct drift criterion and has already adjusted for it (i.e., by transforming $e(t)$ to $e(t) - \theta t$). On the other hand, the Bayesian model is more general than the mechanistic one because it has two forms of response bias, one in the prior ($E(0)$) and the other in the thresholds (α and β, perhaps reflecting asymmetries in the reward structure). These two types of bias are not separately interpretable in the mechanistic model, where they both manifest as differences in the distances from the starting point to the two thresholds. The mechanistic model appears to have three free parameters for the starting point and thresholds (thus giving it two degrees of freedom to capture response bias, in addition to the overall difference between the two thresholds for capturing speed-accuracy tradeoff), but as with the scaling invariance discussed above, there is a translation invariance whereby adding a constant to $e(t)$ (and hence to the starting point and both thresholds) has no impact on model predictions. In other words, the evidence process in the mechanistic model is defined only on an interval scale, whereas in the Bayesian model it has a well-defined zero point.

Considering these differences in parameter constraints or redundancy in the two models, full equivalance between the Bayesian and mechanistic diffusion models is possible only under special cases of each. Specifically, we consider now the mechanistic model with symmetric drift rates ($\mu_2 = -\mu_1$), and the Bayesian model with symmetric thresholds ($\beta = -\alpha$). Effectively, we remove the mechanistic model's ability to implement suboptimal decision rules, and we limit the Bayesian model to response bias in the prior (and not in reward structure). Under these restrictions, we can derive exact translations between the parameters of the two models. The mechanistic parameterization we use is the one that is most common in the literature, with drift rates $\pm\mu$, diffusion rate σ^2, starting point z, and thresholds of 0 and a (with $a > 0$). These four parameters confer three degrees of freedom because of the scaling degeneracy. For the Bayesian parameterization, we denote the drift rates as $\pm\xi$, the diffusion rate 2ξ, the starting point $E(0)$, and the thresholds $\pm\alpha$.

The derivation of (45) assumed for simplicity a starting point of $e(0) = 0$ but, as noted above, in general calculation of the posterior requires subtracting $e(0)$ from $e(t)$. Also, the assumption of symmetric drift rates in the mechanistic model implies $\theta = 0$. Therefore (45) becomes

$$E(t) = E(0) + \frac{2\mu}{\sigma^2}\left(e(t) - z\right) \tag{54}$$

(where we have substituted the definition of ϕ from (43) to write everything in terms of the basic parameters of the mechanistic model). Thus we see as in (46) and (47) that $E(t)$ is a diffusion process with drift rate (for Category 1) equal to

$$\xi = \frac{2\mu^2}{\sigma^2} \tag{55}$$

and diffusion rate equal to twice this quantity. From (54), requiring the thresh-

olds of the two models to agree implies

$$\alpha = E(0) + \frac{2\mu}{\sigma^2}(a - z) \tag{56a}$$

and

$$-\alpha = E(0) + \frac{2\mu}{\sigma^2}(-z). \tag{56b}$$

Subtracting these two equations provides the translation between the models' threshold parameters,

$$\alpha = \frac{\mu a}{\sigma^2}, \tag{57}$$

and then substituting back into (56) gives the translation between starting points:

$$E(0) = \frac{2\mu}{\sigma^2}\left(z - \frac{a}{2}\right). \tag{58}$$

Finally, we can substitute back into (54) to obtain a translation between the two diffusion processes in terms of only the mechanistic model's parameters:

$$E(t) = \frac{2\mu}{\sigma^2}\left(e(t) - \frac{a}{2}\right). \tag{59}$$

This result is sensible, because $a/2$ is the midpoint between thresholds and thus represents a neutral point corresponding to $E(t) = 0$, and because we have already seen that $\phi = 2\mu/\sigma^2$ is the scaling factor that converts the units of $e(t)$ into units of log-odds. Table 1 summarizes the correspondences between the two models. Under this translation, the mechanistic and Bayesian diffusion models make identical predictions and differ only in theoretical interpretation.

7 Predictions

We now consider the diffusion model's predictions for response probability and RT. We do this for both the mechanistic and the Bayesian parameterizations of the model (see Table 1). According to the operation of the model, on each decision trial a diffusion process begins at the specified starting point and evolves stochastically until it reaches one threshold or the other. The model's response is determined by which threshold is crossed first, and its RT by the time it takes for that crossing to occur (plus perhaps some nondecision time). The model's prediction for the joint distribution of response and RT is determined by aggregating across the ensemble of possible trajectories for the diffusion process, as illustrated in Figure 4.

The model's response probabilities, $\Pr[R_j|H_i]$ for $i, j \in \{1, 2\}$, have been presented many times in the literature, but here we show that the Bayesian formulation provides an intuitive method for deriving these quantities. The key insight is that, because $E(t)$ represents the true posterior log-odds at any time t, the value of E at any time determines the objective probability that either response will be correct. Thus, if the model selects response R_1 because

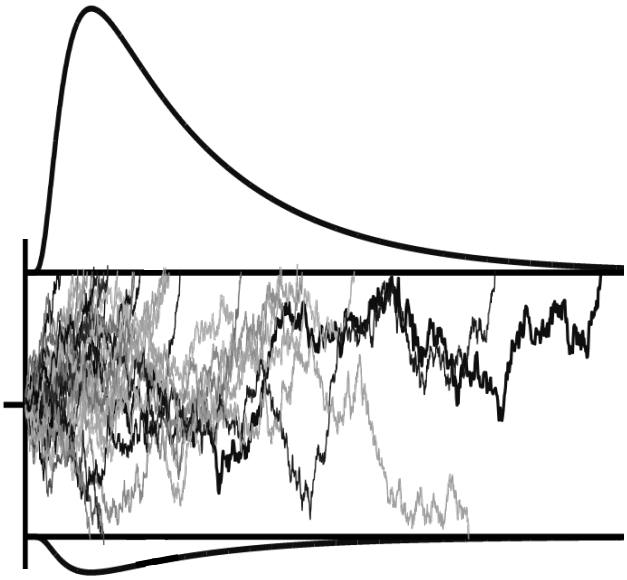

Figure 4: Illustration of the operation of the diffusion model. At the start of each decision trial (indicated by the vertical line, with time flowing to the right), a diffusion process begins at a value determined by the model's starting point parameter (horizontal tick). The process evolves stochastically until it reaches one of the two thresholds (horizontal lines). Greyscale curves show sample trajectories. The model's response and response time (RT) on each trial are determined by which threshold is crossed first and when the crossing occurs. The predicted joint distribution of response and RT is represented by the curves at top and bottom, which indicate the probability density of RT for each response. The area subsumed by each curve represents the marginal probability of that response.

Table 1: Translation between mechanistic and Bayesian versions of the diffusion model

Property	Mechanistic Model	Bayesian Model	Translation
Diffusion process	$e(t)$	$E(t)$	$E(t) = \frac{2\mu}{\sigma^2}\left(e(t) - \frac{a}{2}\right)$
Drift rate	$\pm\mu$	$\pm\xi$	$\xi = \frac{2\mu^2}{\sigma^2}$
Diffusion rate	σ^2	2ξ	$2\xi = \frac{4\mu^2}{\sigma^2}$
Starting point	z	$E(0)$	$E(0) = \frac{2\mu}{\sigma^2}\left(z - \frac{a}{2}\right)$
Thresholds	$\{0, a\}$	$\pm\alpha$	$\alpha = \frac{\mu a}{\sigma^2}$

Note: Models are restricted to assume homogeneous stimuli within each category (no stimulus subtypes), to use symmetric drift rates (i.e., mechanistic model assumes the observer has adjusted for the drift criterion), and symmetric thresholds for the Bayesian model (i.e., response bias lies only in the starting point). Translation column gives the unique translation from the mechanistic parameterization to the Bayesian one. The reverse translation is unique only up to multiplicative scaling of the mechanistic model's parameters, because of the scaling degeneracy in that model.

the process terminates at $E = \alpha$, then the log-odds that this response is correct equals α, and if the model selects R_2 because the process terminates at $E = -\alpha$, then the log-odds that that response is correct equals α and the log-odds that R_1 would have been correct equals $-\alpha$. Writing these dependencies in terms of probabilities, we have

$$\Pr[H_1|R_1] = \frac{1}{1+e^{-\alpha}} \tag{60a}$$

and

$$\Pr[H_1|R_2] = \frac{1}{1+e^{\alpha}}. \tag{60b}$$

From the definition of $E(0)$ as the prior log-odds, the marginal probabilities for the two hypotheses are given by

$$\Pr[H_1] = \frac{1}{1+e^{-E(0)}} \tag{61a}$$

and

$$\Pr[H_2] = \frac{1}{1+e^{E(0)}}. \tag{61b}$$

Combining the above four relations enables us to derive the conditional probability for the response given each hypothesis. We use the following basic identity for binary random variables, which can be derived from the definitions of joint and conditional probability:

$$\Pr[R_1] = \frac{\Pr[H_1] - \Pr[H_1|R_2]}{\Pr[H_1|R_1] - \Pr[H_1|R_2]}. \tag{62}$$

The conditional response probability under Category 1 can then be derived as

$$\begin{aligned}\Pr[R_1|H_1] &= \frac{\Pr[R_1]\Pr[H_1|R_1]}{\Pr[H_1]} \\ &= \frac{e^{\alpha} - e^{-E(0)}}{e^{\alpha} - e^{-\alpha}},\end{aligned} \tag{63a}$$

and by analogous calculations, the response probability under Category 2 is

$$\Pr[R_1|H_2] = \frac{e^{E(0)} - e^{-\alpha}}{e^{\alpha} - e^{-\alpha}}. \tag{63b}$$

Notice that these results do not depend on the drift or diffusion rate of the Bayesian evidence process (ξ), which is a consequence of the 1:2 property of the Bayesian model, and which contrasts with the results for the mechanistic model given below. In the special case of a neutral prior ($E(0) = 0$, corresponding to $\Pr[H_1] = \Pr[H_2] = \frac{1}{2}$), the response probabilities reduce to

$$\Pr[R_1|H_1] = \frac{1}{1 + e^{-\alpha}} \tag{64a}$$

and

$$\Pr[R_1|H_2] = \frac{1}{1 + e^{\alpha}}, \tag{64b}$$

meaning that the log-odds of a correct response under either stimulus category equal α.

To derive the response probabilities for the mechanistic model, we substitute the parameter translations in Table 1 into (63), yielding

$$\Pr[R_1|H_1] = \frac{1 - e^{-\frac{2\mu z}{\sigma^2}}}{1 - e^{-\frac{2\mu a}{\sigma^2}}} \tag{65a}$$

and

$$\Pr[R_1|H_2] = \frac{e^{\frac{2\mu z}{\sigma^2}} - 1}{e^{\frac{2\mu a}{\sigma^2}} - 1}. \tag{65b}$$

These expression match those found elsewhere in the literature for the mechanistic diffusion model's response probabilities. Importantly, although (61) assumes $E(0)$ corresponds to the true (objective) prior log-odds, changing the prior on H has no effect on $\Pr[R|H]$. Therefore (63), and hence (65), hold even if the starting point does not correspond to the objective prior. This fact should be obvious for the mechanistic model, because prior probabilities play no role in that model. Most psychological applications of the diffusion model treat the starting point as a free parameter, not determined by the Bayesian analysis presented here.

In addition to response probability, the diffusion model predicts the joint distribution of response and RT (denoted here by the random variable T). These

predictions are derived in numerous sources, and we simply repeat them here, first for the standard mechanistic parameterization:

$$\begin{aligned}\Pr\left[R_1, T \leq t | H_1\right] &= \Pr\left[R_1|H_1\right] - \frac{\pi\sigma^2}{a^2} e^{\frac{\mu(a-z)}{\sigma^2}} \times \\ &\sum_{k=1}^{\infty} \frac{2ka^2\sigma^2}{k^2\pi^2\sigma^4 + a^2\mu^2} \sin\left(\frac{k\pi\left(a-z\right)}{a}\right) e^{-\frac{k^2\pi^2\sigma^4+a^2\mu^2}{2a^2\sigma^2}t},\end{aligned} \tag{66a}$$

$$\begin{aligned}\Pr\left[R_2, T \leq t | H_1\right] &= \Pr\left[R_2|H_1\right] - \frac{\pi\sigma^2}{a^2} e^{-\frac{\mu z}{\sigma^2}} \times \\ &\sum_{k=1}^{\infty} \frac{2ka^2\sigma^2}{k^2\pi^2\sigma^4 + a^2\mu^2} \sin\left(\frac{k\pi z}{a}\right) e^{-\frac{k^2\pi^2\sigma^4+a^2\mu^2}{2a^2\sigma^2}t},\end{aligned} \tag{66b}$$

$$\begin{aligned}\Pr\left[R_1, T \leq t | H_2\right] &= \Pr\left[R_1|H_2\right] - \frac{\pi\sigma^2}{a^2} e^{-\frac{\mu(a-z)}{\sigma^2}} \times \\ &\sum_{k=1}^{\infty} \frac{2ka^2\sigma^2}{k^2\pi^2\sigma^4 + a^2\mu^2} \sin\left(\frac{k\pi\left(a-z\right)}{a}\right) e^{-\frac{k^2\pi^2\sigma^4+a^2\mu^2}{2a^2\sigma^2}t},\end{aligned} \tag{66c}$$

$$\begin{aligned}\Pr\left[R_2, T \leq t | H_2\right] &= \Pr\left[R_2|H_2\right] - \frac{\pi\sigma^2}{a^2} e^{\frac{\mu z}{\sigma^2}} \times \\ &\sum_{k=1}^{\infty} \frac{2ka^2\sigma^2}{k^2\pi^2\sigma^4 + a^2\mu^2} \sin\left(\frac{k\pi z}{a}\right) e^{-\frac{k^2\pi^2\sigma^4+a^2\mu^2}{2a^2\sigma^2}t}.\end{aligned} \tag{66d}$$

Translating these equations into the parameters of the Bayesian model yields

$$\begin{aligned}\Pr\left[R_1, T \leq t | H_1\right] &= \Pr\left[R_1|H_1\right] - \frac{\pi\xi}{2\alpha^2} e^{\frac{\alpha-E(0)}{2}} \times \\ &\sum_{k=1}^{\infty} \frac{4k\alpha^2}{k^2\pi^2\xi + \alpha^2\xi} \sin\left(\frac{k\pi\left(\alpha-E\left(0\right)\right)}{2\alpha}\right) e^{-\frac{k^2\pi^2\xi+\alpha^2\xi}{4\alpha^2}t},\end{aligned} \tag{67a}$$

$$\begin{aligned}\Pr\left[R_2, T \leq t | H_1\right] &= \Pr\left[R_2|H_1\right] - \frac{\pi\xi}{2\alpha^2} e^{-\frac{\alpha+E(0)}{2}} \times \\ &\sum_{k=1}^{\infty} \frac{4k\alpha^2}{k^2\pi^2\xi + \alpha^2\xi} \sin\left(\frac{k\pi\left(\alpha+E\left(0\right)\right)}{2\alpha}\right) e^{-\frac{k^2\pi^2\xi+\alpha^2\xi}{4\alpha^2}t},\end{aligned} \tag{67b}$$

$$\begin{aligned}\Pr\left[R_1, T \leq t | H_2\right] &= \Pr\left[R_1|H_2\right] - \frac{\pi\xi}{2\alpha^2} e^{-\frac{\alpha-E(0)}{2}} \times \\ &\sum_{k=1}^{\infty} \frac{4k\alpha^2}{k^2\pi^2\xi + \alpha^2\xi} \sin\left(\frac{k\pi\left(\alpha-E\left(0\right)\right)}{2\alpha}\right) e^{-\frac{k^2\pi^2\xi+\alpha^2\xi}{4\alpha^2}t},\end{aligned} \tag{67c}$$

$$\Pr\left[R_2, T \leq t | H_2\right] = \Pr\left[R_2|H_2\right] - \frac{\pi\xi}{2\alpha^2} e^{\frac{\alpha+E(0)}{2}} \times \sum_{k=1}^{\infty} \frac{4k\alpha^2}{k^2\pi^2\xi + \alpha^2\xi} \sin\left(\frac{k\pi\left(\alpha + E\left(0\right)\right)}{2\alpha}\right) e^{-\frac{k^2\pi^2\xi+\alpha^2\xi}{4\alpha^2}t}. \quad (67d)$$

For the remainder of this section we use the mechanistic parameterization, and derive some further properties of the model's predictions. Starting with the (cumulative) distribution functions in (66), we can take the derivative with respect to time to get the RT density functions,

$$p_i^j\left(t\right) = \frac{\mathrm{d}}{\mathrm{d}t} \Pr\left[R_j, T \leq t | H_i\right]. \quad (68)$$

These density functions are equal to

$$p_1^1\left(t\right) = \frac{\pi\sigma^2}{a^2} e^{\frac{\mu(a-z)}{\sigma^2}} \sum_{k=1}^{\infty} k \sin\left(\frac{k\pi\left(a-z\right)}{a}\right) e^{-\frac{k^2\pi^2\sigma^4+a^2\mu^2}{2a^2\sigma^2}t}, \quad (69a)$$

$$p_1^2\left(t\right) = \frac{\pi\sigma^2}{a^2} e^{-\frac{\mu z}{\sigma^2}} \sum_{k=1}^{\infty} k \sin\left(\frac{k\pi z}{a}\right) e^{-\frac{k^2\pi^2\sigma^4+a^2\mu^2}{2a^2\sigma^2}t}, \quad (69b)$$

$$p_2^1\left(t\right) = \frac{\pi\sigma^2}{a^2} e^{-\frac{\mu(a-z)}{\sigma^2}} \sum_{k=1}^{\infty} k \sin\left(\frac{k\pi\left(a-z\right)}{a}\right) e^{-\frac{k^2\pi^2\sigma^4+a^2\mu^2}{2a^2\sigma^2}t}, \quad (69c)$$

$$p_2^2\left(t\right) = \frac{\pi\sigma^2}{a^2} e^{\frac{\mu z}{\sigma^2}} \sum_{k=1}^{\infty} k \sin\left(\frac{k\pi z}{a}\right) e^{-\frac{k^2\pi^2\sigma^4+a^2\mu^2}{2a^2\sigma^2}t}. \quad (69d)$$

The integral of each density function equals the corresponding total response probability:

$$\int_0^{\infty} p_i^j\left(t\right) \mathrm{d}t = \Pr\left[R_j|H_i\right]. \quad (70)$$

Dividing by the response probability gives the conditional RT distribution for each response under each category,

$$q_i^j\left(t\right) = \frac{\mathrm{d}}{\mathrm{d}t} \Pr\left[T \leq t | H_i, R_j\right].$$

These conditional RT distributions are equal to

$$q_1^1\left(t\right) = \frac{\pi\sigma^2}{a^2} \cdot \frac{e^{\frac{\mu a}{\sigma^2}} - e^{-\frac{\mu a}{\sigma^2}}}{e^{\frac{\mu z}{\sigma^2}} - e^{-\frac{\mu z}{\sigma^2}}} \sum_{k=1}^{\infty} k \sin\left(\frac{k\pi\left(a-z\right)}{a}\right) e^{-\frac{k^2\pi^2\sigma^4+a^2\mu^2}{2a^2\sigma^2}t}, \quad (71a)$$

$$q_1^2\left(t\right) = \frac{\pi\sigma^2}{a^2} \cdot \frac{e^{\frac{\mu a}{\sigma^2}} - e^{-\frac{\mu a}{\sigma^2}}}{e^{\frac{\mu(a-z)}{\sigma^2}} - e^{-\frac{\mu(a-z)}{\sigma^2}}} \sum_{k=1}^{\infty} k \sin\left(\frac{k\pi z}{a}\right) e^{-\frac{k^2\pi^2\sigma^4+a^2\mu^2}{2a^2\sigma^2}t}, \quad (71b)$$

$$q_2^1(t) = \frac{\pi\sigma^2}{a^2} \cdot \frac{e^{\frac{\mu a}{\sigma^2}} - e^{-\frac{\mu a}{\sigma^2}}}{e^{\frac{\mu z}{\sigma^2}} - e^{-\frac{\mu z}{\sigma^2}}} \sum_{k=1}^{\infty} k \sin\left(\frac{k\pi(a-z)}{a}\right) e^{-\frac{k^2\pi^2\sigma^4 + a^2\mu^2}{2a^2\sigma^2}t}, \quad (71c)$$

$$q_2^2(t) = \frac{\pi\sigma^2}{a^2} \cdot \frac{e^{\frac{\mu a}{\sigma^2}} - e^{-\frac{\mu a}{\sigma^2}}}{e^{\frac{\mu(a-z)}{\sigma^2}} - e^{-\frac{\mu(a-z)}{\sigma^2}}} \sum_{k=1}^{\infty} k \sin\left(\frac{k\pi z}{a}\right) e^{-\frac{k^2\pi^2\sigma^4 + a^2\mu^2}{2a^2\sigma^2}t}. \quad (71d)$$

Thus we see that $q_1^j(t) = q_2^j(t)$, meaning the distribution of RTs for response j is the same regardless of whether that response is correct (Category $i = j$) or incorrect (Category $i \neq j$). All that differs between categories is the overall probability of that response. Moreover, if $z = a/2$ (i.e., no response bias, or equal priors for the two categories in the Bayesian model), then $q_i^1(t) = q_i^2(t)$. That is, for either stimulus category, the RTs for responses R_1 and R_2 have the same distribution (again, all that differs is the overall probability of the two responses). If the starting point is not midway between the thresholds, the model can predict faster RTs for one response than for the other. However, the direction of the effect will be the same for both categories: If RTs are faster for R_1 than for R_2 following Category 1 stimuli, they will also be faster for R_1 than R_2 following Category 2 stimuli.

8 Intertrial Variability and Unfalsifiability

What the diffusion model as presented thus far cannot predict (under either the mechanistic or Bayesian formulation) is a difference between conditional RT distributions for the two responses that depends on which response is correct. That is, it cannot predict a pattern wherein correct RTs are faster than error RTs (i.e., response R_1 faster than R_2 under Category 1, and vice versa under Category 2), and likewise it cannot predict a pattern of error RTs being faster than correct RTs. As it turns out, numerous experiments show these patterns, presenting an empirical challenge to the model.

One solution to this challenge, which has been widely adopted in the literature, is to extend the plain diffusion model as presented thus far by introducing intertrial variability in the model parameters. For example, one could assume that μ is a random variable, taking on different values on different trials of the same stimulus category. When μ is larger, the evidence rates for the two categories are better separated, and consequently the model responds more quickly and more accurately (for a fixed value of the threshold). Indeed, it is readily seen from (63) that increasing μ increases the probabilities of correct responses, $\Pr[R_1|H_1]$ and $\Pr[R_2|H_2]$, and calculations using (71) show that RTs become shorter as well. Likewise, decreasing μ increases the probabilities of error responses and produces longer RTs. When different values of μ are mixed across trials, the result is a variant of Simpson's paradox: Even though errors and correct responses have the same RT distribution for any given value of μ, errors are more likely on trials when μ is smaller, which is also when RT tends to be

longer, and therefore errors are overall slower than correct responses once μ is integrated out.

Intertrial variability in the starting point can also break the symmetry between correct and error RT distributions, but in the opposite direction. Greater values of $E(0)$ or z produce faster RTs for R_1, as can be calculated from (71a) and (71c), and they increase the probability of that response under either category, as can be seen from (63). However, the increase in response probability is greater under Category 2 (i.e., when R_1 is an error). That is, R_1 is more likely with starting points that lead to a short RT for that response, and this effect is more pronounced when the response is incorrect. The same conclusion holds for R_2. Therefore intertrial variability in the starting point produces error RT distributions that are faster than correct RT distributions.

Combining intertrial variability in the drift rate and starting point produces a more complex pattern. When the thresholds are relatively large, the impact of starting-point variability is reduced, and the drift-rate variability dominates to produce a pattern of slow errors. When the thresholds are relatively small, starting-point variability dominates to produce a pattern of fast errors. Experimental manipulations that are assumed to influence subjects' speed-accuracy tradeoff, and that are modeled by changes in threshold (i.e., smaller thresholds under conditions encouraging speed, larger threshods under conditions encouraging accuracy), have been found to yield this pattern. For example, when an instructional manipulation is used, errors tend to be faster than correct responses when subjects are told to emphasize speed over accuracy, and when subjects are told to emphasize accuracy this relationship tends to reverse.

The proposal of intertrial variability thus appears to be empirically successful, and to have resolved what is otherwise an important predictive failure of the diffusion model. However, this proposal suffers two problems. First, it is theoretically unmotivated. The general idea that sensory input is variable has a long history in psychophysics, but the diffusion model already incorporates this idea as within-trial variability. The proposal of two separate timescales of variability (within trials and between trials) seems to have been introduced solely to fit the data. The second problem, which is somewhat a consequence of the first, is that without some theory to constrain the form of the intertrial distributions, the model becomes excessively flexible. In fact, if the drift-rate distribution is entirely free, then the model becomes fully unfalsifiable. That is, for any joint distribution over the response and RT, there exist drift-rate distributions under which the model exactly reproduces that joint distribution.

To state this unfalsifiability result more formally: Let $G_i^j(t)$ for $i, j \in \{1, 2\}$ be any set of nondecreasing right-continuous functions with $G_i^j(t) = 0$ and $\lim_{t\to\infty} G_i^1(t) + G_i^2(t) \leq 1$. (Allowing for inequality in the latter constraint allows a nonzero probability that no response is given.) Let a be any value for the upper threshold (with the lower threshold equal to 0), and let z be any value for the starting point (fixed across trials), with $0 < z < a$. Then there exist a value of σ and intertrial drift-rate distributions under the two categories such that the diffusion model exactly predicts $\Pr[R_j, T \leq t | H_i] = G_i^j(t)$ for all

$t \in \mathbb{R}^+$ and $i, j \in \{1, 2\}$.

To prove this statement, let $\sigma = 0$ and define the drift-rate distribution under each category by

$$\Pr\left[\mu \leq x | H_i\right] = \begin{cases} G_i^2\left(-\frac{z}{x}\right) & x < 0 \\ \lim_{t\to\infty} G_i^2(t) & x = 0 \\ 1 - \sup_{t<\frac{a-z}{x}} G_i^1(t) & x > 0. \end{cases} \tag{72}$$

Under the special case of no diffusion ($\sigma = 0$), the response is always R_1 if $\mu > 0$ and R_2 if $\mu < 0$, and the RT is given by

$$T = \begin{cases} -\frac{z}{\mu} & \mu < 0 \\ \infty & \mu = 0 \\ \frac{a-z}{\mu} & \mu > 0. \end{cases} \tag{73}$$

Therefore for any $t > 0$ we have

$$\begin{aligned} \Pr\left[R_2, T \leq t | H_i\right] &= \Pr\left[\mu \leq -\frac{z}{t} \middle| H_i\right] \\ &= G_i^2(t). \end{aligned} \tag{74a}$$

Likewise,

$$\begin{aligned} \Pr\left[R_1, T \leq t | H_i\right] &= \Pr\left[\mu \geq \frac{a-z}{t} \middle| H_i\right] \\ &= 1 - \sup_{m<\frac{a-z}{t}} \left(1 - \sup_{\tau<\frac{a-z}{m}} G_i^1(\tau)\right) \\ &= \inf_{\frac{a-z}{m}>t} \left(\sup_{\tau<\frac{a-z}{m}} G_i^1(\tau)\right) \\ &= G_i^1(t) \end{aligned} \tag{74b}$$

with the last equality due to right-continuity of G_i^1.

Although this proof relies on allowing diffusion to be absent from the model, one can also choose $\sigma > 0$ and obtain a model with predictions arbitrarily close to a given G_i^j. That is, for any $\epsilon < 0$, there exist a diffusion rate and drift-rate distributions such that the model's predictions satisfy

$$\left|\Pr\left[R_j, T \leq t | H_i\right] - G_i^j(t)\right| < \epsilon \tag{75}$$

for all $t > 0$ and $i, j \in \{1, 2\}$. This follows from the fact that (66) is continuous with respect to σ at $\sigma = 0$ and that probability functions are monotonic with compact range (thus ensuring uniform convergence in (75)). Because the model's predictions are invariant under any transformation of its parameters $(z, \mu, \sigma, a) \to (\gamma z, \gamma\mu, \gamma\sigma, \gamma a)$ for $\gamma > 0$, one can then pick a transformation that

results in any desired value of σ. In other words, if the diffusion rate is fixed in advance, one can still obtain a model with predictions arbitrarily close to a given G_i^j by appropriate choice of the drift-rate distributions, starting point, and thresholds.

In practical applications of the diffusion model, the intertrial distributions of drift rate and starting point are not fully unconstrained as they are in the proof just given. Instead, drift rate is typically assumed to vary according to a Gaussian distribution, and starting point according to a uniform one. This more restricted model is not unfalsifiable, and indeed it makes constrained predictions that have been well-supported empirically. However, the choices of Gaussian and uniform distributions are made purely for mathematical convenience; they are considered implementation assumptions rather than theoretical commitments. Therefore we are left in an unusual situation, where a formal model makes constrained and successful predictions, but the theory this model is meant to embody (diffusion process, time-invariant boundaries, and intertrial variability in starting point and drift rate) is unfalsifiable. Clearly the model is capturing regularities in human decision-making behavior, in a way that gives it remarkable predictive power, but at present the reasons for this empirical success are poorly understood.

9 Further Reading

For readers not familiar with Bayesian models of cognition, Griffiths, Kemp, and Tenenbaum (2009) provide a tutorial introduction.

The signal detection model was originated by Tanner and Swets (1954) and Green and Swets (1966), with important later elaborations by Ashby and Townsend (1986).

Reviews of evidence sampling models of speeded choice, including the random walk and diffusion models, can be found in Ratcliff and Smith (2004), Luce (1986), Townsend and Ashby (1983), and Vickers (1979).

The original formulation of the random walk model and its grounding in the SPRT are due to Stone (1960). Later developments can be found in Laming (1968), Link (1975), and Link and Heath (1975).

The Wiener diffusion model was originated by Ratcliff (1978). The Ornstein-Uhlenbeck (OU) model, which is closely related to the Wiener diffusion model but includes a decay component in the dynamics of the evidence process, was developed by Busemeyer and Townsend (1993).

Bogacz et al. (2006) present an optimality analysis of several evidence-accumulation models, including the diffusion and OU models, that is complementary to the Bayesian derivation presented here.

Response and RT predictions for the diffusion model can be found in Ratcliff (1978), with elaboration in Smith (2000). Predictions for the random walk model and relationships to the diffusion model's predictions can be found in Smith (1990).

Intertrial variability was first introduced by Laming (1968), who assumed a

variable starting point for the random walk model. Ratcliff (1978) introduced a variable drift rate for the diffusion model. Ratcliff and Rouder (1998) assumed variability in both starting point and drift rate in the diffusion model and showed that these assumptions together can predict a crossover pattern of fast errors under speed instructions and slow errors under accuracy instructions. Data showing this crossover pattern can be found, for example, in Ratcliff, Van Zandt, and McKoon (1999) and Wagenmakers, Ratcliff, Gomé z, and McKoon (2008).

The unfalsifiability property of the diffusion model with unconstrained drift-rate distribution was proven by Jones and Dzhafarov (2014a). Further discussion of this theorem appears in Heathcote, Wagenmakers, and Brown (2014), Smith, Ratcliff, and McKoon (2014), and Jones and Dzhafarov (2014b).

10 Conclusions

We have shown here that the diffusion model that has been influential in psychological studies of speeded decision-making has a normative basis in Bayesian inference from a continuous evidence stream. The version of the model that results from this rational analysis is formally equivalent to the standard, mechanistic diffusion model and offers new insights on psychological interpretations of its parameters.

A general challenge to Bayesian models of cognition is that exact Bayesian inference becomes intractable as the task environment becomes more complex. For example, the rational analysis presented here does not cover cases where there are different stimulus subtypes within each category, or where the true drift rates of the input process (μ_i) are unknown, or where there are sequential dependencies across trials. Although extensions of the present model to these cases are possible, it seems that the brain must eventually give up exact Bayesian inference in favor of approximate methods. Therefore, as a cognitive theory, Bayesian optimality is better viewed as a guiding principle that is likely to be more accurate in simpler situations. In light of these considerations, we suggest the value of an analysis like the one presented here is that it offers a link between mechanistic and rational levels of explanation. Understanding the normative underpinnings of a mechanistic model, such as the diffusion model, may provide guidance in extending it to cover more complex tasks or phenomena.

One example of how a normative grounding might inform extensions of the diffusion model concerns the relationships between RT distributions for correct and error responses. We have reviewed here how the plain diffusion model makes strong predictions about these relationships that do not hold up empirically. The model can be extended to enable violations of these predictions by incorporating random intertrial variability in its drift rate and starting point, but this extension is atheoretical and comes at the cost of making the underlying theory unfalsifiable. In order to obtain a model that makes a genuine explanatory contribution—that is, that makes theoretically driven, constrained predictions that match the data—it seems that what is needed is a theory of intertrial variability itself, one that implies constraints on the forms that vari-

ability can take. Because the Bayesian treatment presented here offers a rational interpretation of the starting point and drift rate, it may suggest a principled theory of how and why they vary. In particular, one hope is that such a theory could emerge from trial-by-trial learning of the task parameters (the prior probabilities of the categories, the drift criterion, and the signal-to-noise ratio) that the observer must know in order to carry out optimal inference.

References

[1] Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93*, 154-179.

[2] Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & 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*, 700-765.

[3] Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100*, 432–459.

[4] Green, D. M., Swets J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

[5] Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In Ron Sun (ed.), *The Cambridge handbook of computational cognitive modeling*. Cambridge University Press.

[6] Heathcote, A., Wagenmakers, E.-J., & Brown, S. D. (2014). The falsifiability of actual decision-making models. *Psychological Review, 121*, 676–678.

[7] Jones, M., & Dzhafarov, E. N. (2014a). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review, 121*, 1–32.

[8] Jones, M., & Dzhafarov, E. N. (2014b). Analyzability, ad hoc restrictions, and excessive flexibility of evidence-accumulation models: Reply to two critical commentaries. *Psychological Review, 121*, 689-695.

[9] Laming, D. R. J. (1968). *Information theory of choice reaction time*. Wiley.

[10] Link, S. W. (1975). The relative judgement theory of two choice response time. *Journal of Mathematical Psychology, 12*, 114–135.

[11] Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika, 40*, 77–105.

[12] Luce, R. D. (1986). *Response times*. Oxford University Press.

[13] Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59-108.

[14] Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9*, 347–356.

[15] Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111*, 333– 367.

[16] Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106*, 261–300.

[17] Smith, P. L. (1990). A note on the distribution of response time for a random walk with Gaussian increments. *Journal of Mathematical Psychology, 34*, 445–459.

[18] Smith, P. L. (2000). Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology, 44*, 408–463.

[19] Smith, P. L., Ratcliff, R., & McKoon, G. (2014). The diffusion model is not a deterministic growth model: Comment on Jones and Dzhafarov (2014). *Psychological Review, 121*, 679–688.

[20] Stone, M. (1960). Models for choice reaction time. *Psychometrika, 25*, 251–260.

[21] Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61*, 401-409.

[22] Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. Cambridge University Press.

[23] Vickers, D. (1979). *Decision processes in visual perception*. Academic Press.

[24] Wagenmakers, E.-J., Ratcliff, R., Goméz, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language, 58*, 140–159.