

On the effects of signal acuity in a connectionist model of decision making

Tyler McMillen
Program in Applied and Computational Mathematics,
Princeton University, Princeton, NJ 08544, U.S.A.
E-mail: mcmillen@princeton.edu

June 6, 2006

Abstract

We consider the effects of signal ‘sharpness’ or acuity on the performance of neural models of decision making. In these models, a vector of signals is presented and the subject must decide which of the elements of the vector is the largest. In an earlier paper (McMillen and Holmes, 2006) we derived asymptotically optimal tests under the assumption that the elements of the signal vector were all equal except one, which was a quantity a greater than the rest. In this paper we consider the case of signals spread around a peak. The acuity is a measure of how strongly peaked the signal is. We find that the best test is one in which the detectors are multiplied by a matrix containing the possible signal vectors, and comparisons made on the resulting product. This matrix multiplication is a way to encode knowledge of the shape of the incoming signals. The advantage to this method, in terms of reducing the time to make a decision under error constraints, is correlated to the spread of the signals. The greater the spread of the signals around the peak, the greater is the advantage to performing the matrix multiplication, which can result in orders of magnitude improvement in decision times.

KEYWORDS: leaky accumulator, drift-diffusion model, neural network, signal acuity, multi-hypothesis sequential test, sequential ratio test.

1 Introduction

Time-controlled decision processes often involve a trade-off between speed and accuracy. Generally, accuracy improves as more time is allotted to the decision, but fewer decisions can be made in a fixed duration block of trials. In a setting of successive trials in which a reward is given for each correct decision, the optimal test is the one that maximizes the number of correct decisions per unit time. If the subject is free to make a decision at any time, the optimal test is the one that minimizes the *mean reaction time* (MRT)¹ as a function of the *error rate* (ER). If the subject is constrained to make a decision at a fixed time after the presentation of the stimulus, then the optimal test is the one that minimizes the ER. However, other considerations may make a sub-optimal test preferable. Energy costs associated with the decision process may give greater weight to reducing the MRT; situations where a correct decision is a matter of life-and-death give greater weight to reducing the ER. Decision making processes in organisms have presumably evolved to balance these trade-offs in a manner conducive to survival.

¹In this paper we use RT to denote the time taken by a computational process or dynamical model to render a decision. This is more commonly called *decision time* (DT) in the psychological literature.

In this report we consider tasks in which an observer must identify 1 out of N possible stimuli with an appropriate response. We consider these in the context of the leaky competing accumulator model of Usher and McClelland (2001) in which N accumulators accumulate evidence and inhibit the rest in proportion to their values.

Most studies of multiple choice tasks consider the alternatives as equivalent and equi-distant from each other. This approach goes back to Hick (1952), and is implicitly assumed in models such as those considered by Usher and McClelland (2001). However, this assumption is at odds with the observation that the alternatives “close” to the one presented (the correct alternative) will be chosen more often in incorrect responses than those incorrect responses that are further away from the correct alternative. The obvious difficulty theoretically is to quantify a notion of distance between alternatives. This report is a step toward an understanding of the effect of distance between alternatives. We build on the work in McMillen and Holmes (2006), in which all alternatives are assumed to be equivalent and mutually equidistant. We propose new experiments and possible explanation for the observed limitation in information transmission between stimulus and response (Miller, 1956).

We will investigate the performance of various tests in relation to the optimal tests on the values of the accumulator states. Optimal tests have been derived for tests between two hypotheses. In that case, the Neyman-Pearson lemma (Neyman and Pearson, 1933) provides the best fixed sample size test in that it minimizes error rates, and Wald’s *sequential probability ratio test* (SPRT) is optimal for sequentially accumulating data, in the sense that it minimizes the expectation of decision time among all tests for which the probabilities of error do not exceed pre-determined bounds. (See Wald and Wolfowitz (1948), or Lehmann (1959) for a simpler proof.) For decisions between three or more alternatives, there is no general algorithm for minimizing the MRT for all error rates. However, tests that bound error rates and reaction times and that are optimal in an asymptotic setting have been designed (Lorden, 1977; Golubev and Khas’minskii, 1983; Verden-skaya and Tartakovskii, 1991; Baum and Veeravalli, 1994; Veeravalli and Baum, 1995; Tartakovsky, 1998a,b; Dragalin et al., 1999, 2000). We will consider these “asymptotically optimal” tests and their implementations in a neural model of competing, leaky accumulators, and examine how they might be implemented in the brains of certain organisms.

The outline of this report is as follows. We will first describe signals and motivate a definition of distance that we will employ. In the next section we briefly describe asymptotically optimal tests in decision tasks in which there are more than two alternatives. Then we describe the leaky competing accumulator model for decision making. We consider tests on the accumulators in both the free response and interrogation protocols (that is, tests in which the subject is free to respond at any time after presentation of a stimulus, and those in which the subject is compelled to respond at a given time). We will describe a simple neural network implementation of the asymptotically optimal tests. In the conclusion we will describe possible experiments and propose explanations for various phenomena in the psychological literature.

2 Acuity and distance

Generally, statistical testing involves signals and detectors. The goal is to determine, on the basis of information present in the detectors, which of the signals is the largest. Suppose there are n signals S_1, \dots, S_n . McMillen and Holmes (2006) have explored the case in which all signals are equal save one, which is a quantity a greater than the rest, i.e. for some j (the ‘correct’ signal), $S_j = S_0 + a$, and $S_i = S_0$ for $i \neq j$. Then the problem is to determine which signal is $S = S_j$ from the data present in the detectors. Here we consider the more general case in which the signals are drawn from a distribution. During a trial a vector of signals $S = (S_1, S_2, \dots, S_n)$ is presented, and the goal, as before, is to determine which element of S is the largest. Suppose the signals have a

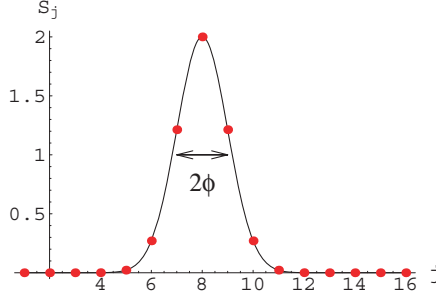


Figure 1: Signals with a Gaussian form. The acuity is $\phi^{-1} = 1$

Gaussian form as in Fig. 1:

$$S_i = S_i^{(j)} = S_0 + a \exp \left[-\frac{(i-j)^2}{2\phi^2} \right], \quad i = 1, \dots, n. \quad (1)$$

Then the largest signal is S_j . We call ϕ^{-1} the *acuity* in the signal. In the limit as the acuity approaches infinity, this reduces to the previous case, $S_i = S_0 + a \delta_{ij}$ (Kronecker delta), but when $\phi^{-1} < \infty$, the signals are “spread” around the maximum. Let

$$S^{(j)} = (S_1^{(j)}, S_2^{(j)}, \dots, S_n^{(j)})$$

be the vector of signals as defined above. $S^{(j)}$ is the vector of signals with maximum in the j th signal. If all possible signals have this form, then the task is to determine which of the $S^{(j)}$ s is equal to S . If it is known *a priori* that the signals have the shape given above then this knowledge can be used in the decision process. That is, if we can reduce the space of signals to a subset of all possible signals, this knowledge can be employed in some of the tests we will consider below.

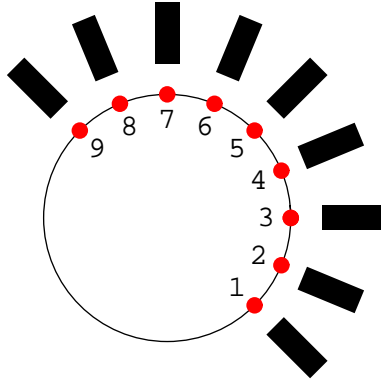


Figure 2: Angles of a bar represented on a circle. Angle 1 is the same as angle 9.

The above type of signal might represent the signals sent to a decision layer in the brain upon viewing a bar on a screen. The angle of the bar excites certain areas of the brain, and one task is to determine the angle the bar makes relative to the x -axis. In the simplest task, the subject must determine if the bar is horizontal or vertical. The problem is the more difficult the more angles are possible. If one only has to discern between a horizontal and a vertical bar, this is much easier than discerning between bars that might lie at angles $0, 10, 20, \dots, 90$. Then the excitation of centers

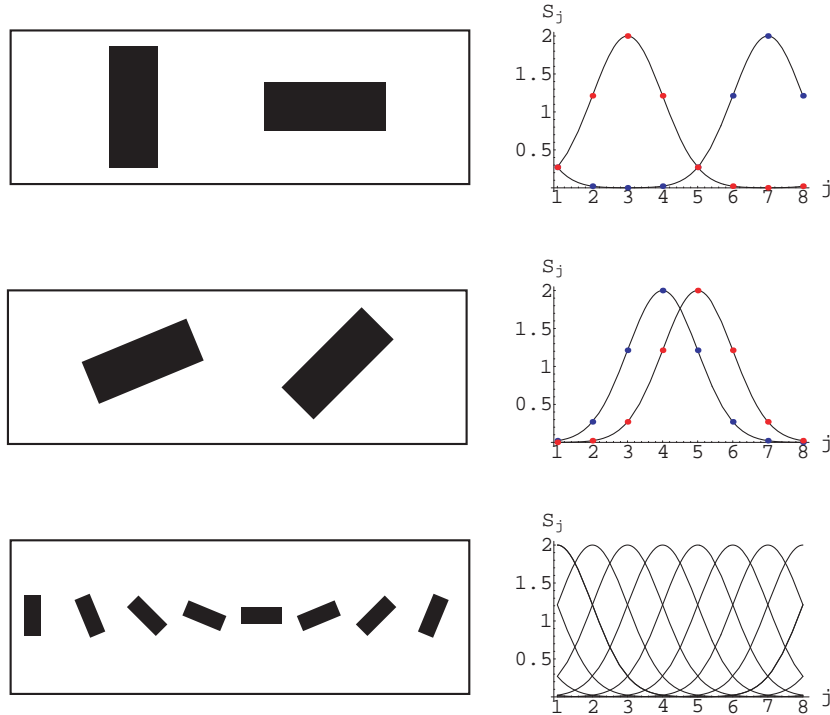


Figure 3: Bar angles (left panels) and corresponding signals (right panels) with acuity $\phi^{-1} = 1$. The top panels show angles 3 and 7; the middle panels show angles 4 and 5; and the bottom panels show all angles 1 through 8.

registering an angle of 10 degrees will also excite those corresponding to 0 and 20, but very little excitation will occur in those registering 90 degrees. The ‘spillover’ of excitation into neighboring angles is related to the acuity. Figures 2 and 3 show the angles for bars in increments of $\pi/8$, and the respective signals S corresponding to these with a finite acuity.

We note that the signals as considered here do not necessarily correspond to the tuning curves of neurons. In this report when we speak of the signals these are the signals inputted into a decision layer. We do not consider how the signals are formed.

The difficulty of the task depends on the number of possible alternatives, and the ‘distance’ between the alternatives. In the angle discrimination task, untrained monkeys and humans can discern between angles 15-20 degrees apart; monkeys can be trained to improve their discernment to 4-5 degrees (Ghose et al., 2002). Improvement is orientation specific: when monkeys learn to discern angles within 5 degrees of a 45 degree angle, they are not able to achieve this level of accuracy for angles around 20 degrees. We will see that there are different ways in which accuracy can be improved. Adjusting the tuning curves in the neurons sensitive to angles is one way to alter the acuity in the signal related to decision making centers in the brain. However, Ghose et al. (2002) note that changes in the tuning curves are not sufficient to explain the improvements in performance observed in monkeys in the angle discrimination task. One might ask why these tuning curves have the shapes that they do, and if there is an optimal shape. We will see that depending on the task, it can be beneficial to have a larger or smaller acuity. Furthermore, we will see how the processing of the signals can be done so as to improve performance.

We will use the angle discrimination task as our canonical example. However, the results are more general, and can be applied to many different decision making tasks, for example, determining the direction of wind, the direction which dots on a screen are tending to move, etc. We also use

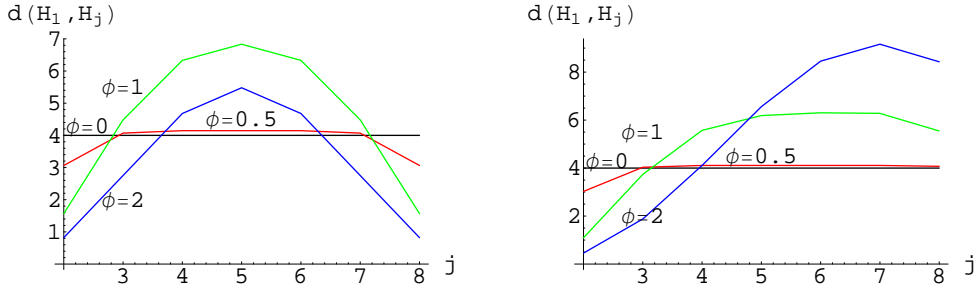


Figure 4: KL-divergence between hypothesis H_1 and H_j for inverse acuity $\phi = 0, 0.5, 1, 2$. The left panel shows distances between signals on a circle (signals given by (4)); the right panel shows distances between signals on an interval (signals given by (1)).

Gaussian functions for the signals, just as a matter of convenience, but the results extend naturally to different kinds of functions like cosines, etc. The main thing we are interested in in this report is the spread of the signal. More generally, the acuity is the reciprocal of the standard deviation of the signal.

2.1 The distance between alternatives

It is more difficult to discern between angles that are close together than between those that are far apart. For instance, if a subject is presented with a bar at an angle of either 45 or 135 degrees and asked which of those two is the correct angle, this is easier than the case in which a bar lies at an angle of either 45 or 50 degrees. The task is more difficult still if the possible angles are 45, 50 and 60. When presented with a single bar, it is more difficult to discern between the two bars in the middle panel of Fig. 3 than between those in the upper panel. The two in the middle panel differ by only $\pi/8$, whereas those in the upper panel differ by $\pi/2$. If the signal when presented is strongly peaked, i.e. has very large acuity, then the difficulty in discerning between the angles in the top panel would be only slightly greater than discerning between those in the middle panel of Fig 3. In the limit of perfect acuity, i.e. when $S_m^{(i)} = S_0 + a \delta_{mi}$, there is no difference between the two cases—the distance between all alternatives is the same. The ‘distance’ between alternatives is therefore dependent not only on the distance between the means, but also on the spread of the densities and their overlaps. Furthermore, the distance between non-overlapping densities should not depend on their means.

A pseudo-metric reflecting this intuition is the Kullback-Leibler divergence (KL-divergence), defined for probability measures P_i as follows:

$$d(P_i, P_j) = E_i \left[\log \left(\frac{dP_i}{dP_j} \right) \right], \quad (2)$$

where $E_i[\cdot]$ is the expectation under the measure P_i . If we consider the signals in noise, $s_i = S_i + c \eta_i$, where η_i is a standard normal random variable, and H_i the hypothesis that the signal vector is $S^{(i)}$, then

$$dP_i(s) = \prod_{m=1}^n \frac{1}{\sqrt{2\pi c^2}} \exp \left[-(s_m - S_m^{(i)})^2 / 2c^2 \right].$$

Therefore, $d(H_i, H_j) = \frac{1}{2c^2} \left| S^{(i)} - S^{(j)} \right|^2.$ (3)

In the case of perfect acuity $S_m^{(i)} = S_0 + a \delta_{mi}$, $d(H_i, H_j) = (a/c)^2$ for all $i \neq j$: all alternatives are the same 'distance' apart.

We will distinguish between two kinds of signals, those that can be represented on a circle and those on an interval. In the case of the signals on an interval, the signal vectors will be given by (1). When the signals are on a circle, as in Fig. 2, we must account for the 'wrap around.' Consider the case of a given signal shape, and the task is to determine where the peak of the signal is. If the signal is on a circle, and has a Gaussian shape, then the signals take the following form:

$$\begin{aligned} \hat{S}_i &= S_0 + a \exp [-(i - \hat{n})^2/2\phi^2] , \\ S_i^{(j)} &= \hat{S}_{\{i-j\}} , \end{aligned} \tag{4}$$

where \hat{n} is the smallest integer larger than $n/2$ and $\{i-j\} = (i-j + \hat{n} + n) \bmod n$. In other words, \hat{S} is the signal with its peak in the middle and the $S^{(j)}$'s are circular shifts of \hat{S} . Figure 4 shows the distance between H_1 and other hypotheses using the signals in (4). We see that, as expected, the distance between a hypothesis and its next nearest neighbor decreases with the acuity. However, the distance between hypotheses that are farther apart can *increase* as the acuity decreases. It should be easier to discern between alternatives that are farther apart than those that are closer together. Thus, we expect that the spread of the signals can be an aid to decision making, if one can use the knowledge of the shape of the signals. But notice that as the acuity decreases, the distance of all the alternatives eventually decreases. Depending on the task and the decision algorithm employed, there can be an optimal value of the acuity that is non-zero.

We will want to examine MRT as the number of alternatives varies. In particular we will be interested to see in what cases we can recover the classical Hick's Law (Hick, 1952), which says that if the accuracy is held at a high level, the MRT increases logarithmically with the number of alternatives N :

$$\text{MRT} = A + B \log N . \tag{5}$$

(See McMillen and Holmes (2006) for a discussion of Hick's Law, esp. regarding the $\log N$ scaling; in some cases a $\log(N-1)$ scaling is a better fit.)

There are several ways to order the alternatives. We will consider the two main ones, which we will call the *inside-out* and the *outside-in* orderings. If we perform an experiment with $n = 2$ alternatives, and increase the number to $n = 3$, how shall we do this? The top and middle panels of Fig. 3 show two cases. In the top panel the alternatives are a maximal distance apart. The two alternatives in the middle panel are much closer. Starting with the alternatives in the top panel and then adding an alternative between them would correspond to the outside-in ordering; starting with angles 4 and 5, as in the middle panel, and then adding angle 6 as an alternative corresponds to the inside-out ordering. If we increase the number of alternatives from $N = 2$ to $N = 8$, then we might do this as proposed in Fig. 5. In the outside-in ordering, we start with the alternatives a maximal distance apart, and then add alternatives so that the distance between the new alternative and the existing alternatives is maximized. In the inside-in ordering, we start with the alternatives close together, and then add alternatives so that the minimal distance of the next additional alternative to the others is fixed. Following the pattern in Fig. 5, if N alternatives are possible, these would be the alternatives $1, \dots, N$, according to the order in the figure, depending on which type of ordering we are following.

3 The MSPRT

We now define some notation and discuss the *multi-sequential probability ratio test* (MSPRT) in an abstract setting. Suppose $X_t = (x_1(t), x_2(t), \dots, x_n(t))$ is a vector-valued random variable, and

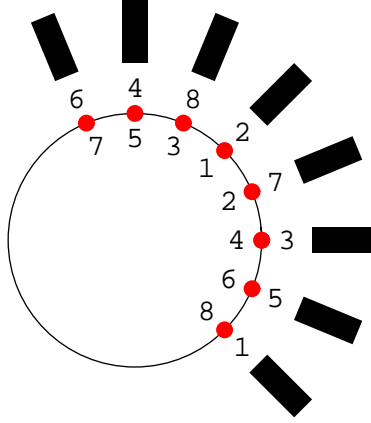


Figure 5: Two orderings of alternatives in Fig. 2. The outside numbers show the order of alternatives as they are presented in the outside-in ordering; the inside numbers correspond to the inside-out ordering.

that one of N hypotheses H_i must be chosen. We denote the prior probability of H_i by $P(H_i) = \pi_i$. Let (Ω, \mathcal{F}) be a measure space on which $\{X_t, t \in \mathbf{R}\}$ is observed, and \mathcal{F}_t the sub- σ -algebra of \mathcal{F} generated by $X^t = \{X_u, 0 \leq u \leq t\}$ observed up to time t . Let P_i be the probability measure under hypothesis H_i , and $P_i^t(\cdot)$ the restriction of the measure P_i to the σ -algebra \mathcal{F}_t . We define the log-likelihood ratios as follows:

$$Z_i(t) = \log \frac{dP_i^t}{dQ^t}(X^t), \quad i = 1, \dots, N, \quad (6)$$

where $dP_i^t/dP_j^t(X^t)$ is the Radon-Nikodym derivative of the measure P_i with respect to a dominating measure Q^t , restricted to \mathcal{F}_t . If $Q^t = P_j^t$, the corresponding process will be denoted $Z_{ij}(t)$, and is the log-likelihood ratio of posterior probabilities given the information in \mathcal{F}_t . We denote a sequential test by $\delta = (\tau, d)$, where τ is a stopping time and $d \in \{1, 2, \dots, N\}$ is the decision.

There are two differently-structured tests which give the same asymptotically optimal results (Dragalin et al., 1999): one determines when the posterior probability of a correct response crosses a threshold; the other test considers the ratios of posterior probabilities. These are called the δ_a and δ_b tests, respectively. Let

$$\begin{aligned} \Pi_i(t) &= \frac{\pi_i \exp[Z_i(t)]}{\sum_{j=1}^N \pi_j \exp[Z_j(t)]} = P(H_i \text{ is true} | X^t), \\ \text{and } L_i(t) &= \frac{\pi_i \exp[Z_i(t)]}{\max_{j \neq i} \pi_j \exp[Z_j(t)]} = \frac{P(H_i \text{ is true} | X^t)}{\max_{j \neq i} P(H_j \text{ is true} | X^t)}, \end{aligned} \quad (7)$$

be the posteriori probability of hypothesis H_i , and the generalized likelihood ratio between H_i and the remaining hypotheses, respectively. For constant thresholds a_i and b_i , the stopping times for

the two tests are as follows:

$$\begin{aligned}
\delta_a : \quad \tau_i &= \inf \{t : \Pi_i(t) \geq \exp(a_i)\} \\
&= \inf \left\{ t : Z_i(t) - \log \left(\sum_{j \neq i} \frac{\pi_j}{\pi_i} \exp [Z_j(t)] \right) \geq a_i \right\}, \\
\delta_b : \quad \nu_i &= \inf \{t : L_i(t) \geq \exp(b_i)\} \\
&= \inf \left\{ t : \min_{j \neq i} \left[Z_{ij}(t) - \log \left(\frac{\pi_j}{\pi_i} \right) \right] \geq b_i \right\}, \tag{8}
\end{aligned}$$

which are the Markov ‘accepting’ times for the hypothesis H_i . $\delta_a = (\tau_a, d_a)$ and $\delta_b = (\nu_b, d_b)$ are then the tests with stopping times and decisions given by

$$\begin{aligned}
\tau_a &= \min_{1 \leq i \leq n} \tau_i, & d_a &= j \text{ if } \tau_a = \tau_j, \\
\nu_b &= \min_{1 \leq i \leq n} \nu_i, & d_b &= j \text{ if } \nu_b = \nu_j. \tag{9}
\end{aligned}$$

Thus the δ_a test accepts hypothesis H_i when the posteriori probability of hypothesis H_i crosses a threshold, and the δ_b test accepts hypothesis H_i when the ratio of the posteriori probability of hypothesis H_i to the posteriori probability of the next most likely hypothesis crosses a threshold. In the case of two hypotheses both tests reduce to the optimal SPRT; in the case of $N > 2$ hypotheses, both tests achieve the same asymptotic optimality in the limit as the error rate approaches zero. In what follows we describe the behavior of the δ_b test. Similar results apply to the δ_a test, with the necessary changes in threshold values.

The δ_b test is *asymptotically optimal* in the following sense. Let R_i be the probability of accepting hypothesis H_i incorrectly. Then, for the δ_a and δ_b tests there exist bounds \bar{R}_i such that among all tests that satisfy similar error bounds, the $\delta_{a,b}$ tests minimize the expectation of reaction (see Dragalin et al., 1999, Theorem 4.2):

$$R_i(\delta_{a,b}) \leq \bar{R}_i, \tag{10}$$

$$E_i[\tau_a, \nu_b] \sim \inf_{\{\delta | R_i \leq \bar{R}_i\}} E_i[\tau], \text{ as } \max_j \bar{R}_j \rightarrow 0, \tag{11}$$

where the notation $x_\gamma \sim y_\gamma$ as $\gamma \rightarrow \gamma_0$ means that $\lim_{\gamma \rightarrow \gamma_0} (x_\gamma/y_\gamma) = 1$. Equation (11) implies that the tests $\delta_{a,b}$ give an expected value of the stopping time that is asymptotically the infimum of the stopping times of all sequential tests satisfying the error bound in (10). This is the multi-dimensional analog of the optimality result for the SPRT, but unlike that case, the result for $n > 2$ only holds in this asymptotic setting. For $n = 2$ we may replace ‘ \sim ’ in equation (11) with ‘=’, and conclude that equality holds for *all* values of $\bar{R}_i > 0$.

4 The leaky competing accumulator model

We consider a connectionist model of n leaky, competing units, each of which integrates an incoming signal S_i or stimulus with additive noise, decays at rate k , and is inhibited by all others (Hopfield, 1984; Grossberg, 1988; Usher and McClelland, 2001). The state variable x_i represents the ‘activations’ (e.g. mean soma potential) of the i -th group of neurons and $f(x_j)$ the firing rate of the j -th group of neurons. The function f is typically taken to be sigmoidal, often using the logistic form:

$$f_g(x) = \frac{1}{1 + \exp(-4g(x - b))} = \frac{1}{2} [1 + \tanh(2g(x - b))], \tag{12}$$

where g denotes the maximum slope of f_g , which occurs at $x = b$, although sometimes piecewise linear functions are used (Usher and McClelland, 2001). The lower and upper limits (0 and 1 for (12)) obtained for small and large input levels model the fact that neural activities in general, and firing rates in particular, are bounded above and below. For a single layer of competing units, the resulting nonlinear SDEs take the form:

$$dx_i = \left(-kx_i - w \sum_{j \neq i} f(x_j) + S_i \right) dt + c dW_i, \quad i = 1, \dots, n. \quad (13)$$

Linearization yields the following form in both cases, with different interpretation of parameters (see Brown et al., 2005).

$$dx_i = \left(-kx_i - w \sum_{j \neq i} x_j + S_i \right) dt + c dW_i, \quad i = 1, \dots, n. \quad (14)$$

The above is referred to as the *leaky competing accumulator model* (LCAM). Here the W_i 's are independent, identically distributed (i.i.d.) Wiener processes, with unit variance, representing random fluctuations in the signal, intrinsic accumulator noise and unmodeled inputs, and we assume that the S_i 's are deterministic and constant during each trial. In (14) we have linearized the input or activation function (see Usher and McClelland, 2001); see McMillen and Holmes (2006); Usher and McClelland (2001) for comments on the nonlinear system in which inhibitory inputs and stimuli are passed through a rate-limiting function f . In this report we consider *only* the linearized model (14), but we note that nonlinear effects may (in fact, probably do) play an important role. The linearized system dealt with in this report may be viewed as a first order approximation to the full nonlinear system. The race model, considered below, is obtained by setting leak k and inhibition w both equal to zero. The number of integrators n can be larger than the number of alternatives N .

The subject's problem is to determine which of the signals S_i is the largest. Depending on our knowledge of the possible 'shapes' of the incoming vector of signals, we can improve on performance over simply choosing the option corresponding to the largest signal. Conventionally, each trial starts with initial data $x_i(0) = x_{i0}$ and ends either when the first accumulator passes its preset threshold b_i (for an absolute test under the free response protocol), or at a fixed time $t = T$ (in the interrogation protocol), in which case the accumulator with largest $x_i(T)$ wins. For unbiased trials with equal prior probabilities of H_i , $\pi_i = 1/n$, we conventionally take $x_i(0) = 0$ and set $b_i = \theta$, $\forall i$.

Substantial support for models of this type, and the drift-diffusion systems that emerge from them, come both from fits of human behavioral data (e.g. Ratcliff (1978); Ratcliff et al. (1999)) and direct neural recordings in behaving monkeys (Roitman and Shadlen, 2002; Ratcliff et al., 2003). For a recent review of theoretical work on such SDEs, see Smith (2000).

In the case of two alternatives the theory is fairly complete. A brief summary of results is presented in Appendix F. Here we will deal with an arbitrary number of alternatives N .

5 The free response protocol

Under the free response protocol, subjects may choose at any time after stimulus presentation and must therefore determine when sufficient knowledge has been gathered to make a decision. In the context of the connectionist model (14) there are several different types of response criteria. The simplest is to set a threshold θ , and choose i if the i th accumulator crosses θ before any of the others. We refer to this as an *absolute* test, since it does not account for the magnitudes of other

accumulator states. Tests which account for differences between states are referred to as *relative* tests. One such test compares the i th accumulator with the average of the other accumulators, and chooses i if $\left(x_i - \frac{1}{n-1} \sum_{j \neq i} x_j\right)$ crosses a threshold; we refer to this as the *max-vs-ave* test. Another is the *max-vs-next* test, which chooses i when the difference between x_i and the next greatest x_j reaches threshold. We will also see that the implementation of the MSPRT in the connectionist model requires that we multiply the vector of accumulators by the matrix of possible signal vectors before comparing them.

There are advantages and drawbacks to the different types of tests. Generally, relative tests perform better in that they deliver lower MRTs for given ER and n , but they are more complex to implement. The (asymptotically) optimal tests also require a matrix multiplication, as well as the computation of the matrix itself. In some situations energy costs might dictate a sub-optimal test.

In the following sections we will first see how the application of the MSPRT in the context of competing accumulators leads to relative tests that are optimal in the limit as error rates approach zero. Then we will consider absolute tests and compare their performance with the relative tests.

5.1 Asymptotically optimal tests

We consider the connectionist model (14), with incoming signals taking the form of n possible vectors $S^{(i)}$. We make the following assumption:

$$\sum_{m=1}^n S_m^{(i)2} \quad \text{and} \quad \sum_{m=1}^n S_m^{(i)} \quad \text{do not depend on } i. \quad (15)$$

Under this assumption the signal shape is the same regardless of which alternative is true. Throughout the remainder of this paper we will assume that this (15) holds. The calculations when this assumption does not hold are presented in Appendix B. The task is to determine, from $x_i(t)$, which of the signals is largest, or, equivalently, which of the vectors $S^{(i)}$ is the actual signal. Let H_i be the hypothesis that the signal is $S^{(i)}$.

The derivation of Π_i and L_i in the case when assumption (15) holds is in Appendix A. Let A be the $N \times n$ matrix of alternative signals ($A_{im} = S_m^{(i)}$):

$$A = \begin{pmatrix} S^{(1)} \\ S^{(2)} \\ \dots \\ S^{(N)} \end{pmatrix}, \quad (16)$$

and let

$$\mathbf{y} = \frac{1}{c^2} A \left(\mathbf{x} - \lambda \int_0^t \mathbf{x}(s) ds \right) + \log \boldsymbol{\pi}, \quad (17)$$

where $\log \boldsymbol{\pi}$ is the vector whose i th entry is $\log \pi_i$. Then

$$\log \Pi_i = y_i - \log \sum_{j=1}^n e^{y_j}, \quad (18)$$

$$\log L_i = y_i - \max_{j \neq i} y_j. \quad (19)$$

The asymptotically optimal tests, therefore, depend only on the *difference* λ between inhibition and decay, and not on their absolute values. If $\lambda = 0$ and $S_m^{(i)} = a \delta_{mi}$, as in the case of infinite acuity (the signal is in one channel only), then $y_i = a/c^2 x_i$, and the tests are tests on the accumulators themselves. In that case, the δ_b test is a ‘max-vs-next’ test, which chooses hypothesis H_i when

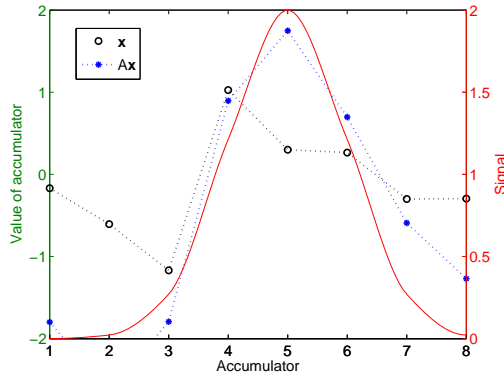


Figure 6: Values of accumulators (circles) and their transformed values (stars), along with the underlying signal (right axis), for a single simulation. In this example, x_4 crosses the threshold before x_5 , but $(A\mathbf{x})_5$ crosses before any of the other elements of $A\mathbf{x}$.

the difference between the i th accumulator and the next largest accumulator crosses a threshold. This case is studied in McMillen and Holmes (2006). There is an advantage to multiplying the accumulators by the matrix A if and only if, on average, $(A\mathbf{x})_i$ is the first to cross the given threshold more often than is x_i when H_i is the correct hypothesis. Figure 6 shows an example of a single simulation in which the multiplication by A results in the correct answer, but the test on the accumulators themselves does not.

5.2 Absolute tests

In the absolute test a decision is rendered once one of the accumulators crosses a pre-determined threshold θ , regardless of the state of any of the other accumulators. In this case, there is no consideration of the difference between accumulators, only their absolute values. However, when there is inhibition and decay, the values of the accumulators effect the values of all the others, and these interactions tend to create a relative effect on the decision via the absolute test.

5.2.1 The absolute test in the race model

By the race model we mean the system of n independent Wiener processes with drift:

$$dx_i = S_i dt + c dW_i, \quad i = 1, \dots, n, \quad (20)$$

which is the limiting process of n independent random walks with i.i.d. increments. This is a special case of the connectionist model (14), in which inhibition and decay are both set to zero ($w = k = 0$), and also represents a bench-mark as being a direct representation of the signals, with additive noise. Since the accumulators are independent in this case, their first passage time densities can be explicitly calculated. For a given threshold θ and drift a , the distribution of threshold crossing times of a single accumulator is (Ricciardi, 1977; Luce, 1986):

$$g(a, \theta, t) = \frac{\theta}{\sqrt{2\pi c^2 t^3}} \exp \left[-\frac{(\theta - at)^2}{2c^2 t} \right]. \quad (21)$$

For a signal $S = (S_1, \dots, S_N)$, where $S_j > S_{i \neq j}$ is the correct signal, the distribution of crossing times for the correct choice is $g(S_j, \theta, t)$. The probability that an accumulator with drift a has not

crossed θ by time t is given by

$$G(a, \theta, t) = \int_t^\infty g(a, \theta, s), \quad (22)$$

which is equal to an error function in the case when the drift $a = 0$. Since the processes are independent the density for a correct response under hypothesis H_i is

$$\rho_i(n, \theta, t) = g(S_i, \theta, t) \prod_{j \neq i} G(S_j, \theta, t). \quad (23)$$

The probability that the correct accumulator crosses θ before any of the others, and the mean time to threshold crossing, are then given by:

$$P(n, \theta) = \int_0^\infty \rho_i(n, \theta, t) dt, \text{ and} \quad (24)$$

$$\text{MRT} = \frac{1}{P(n, \theta)} \int_0^\infty t \rho_i(n, \theta, t) dt. \quad (25)$$

Usher et al. (2002) computed MRTs and ERs in the case in which $S_j = a$ and $S_i = 0$ for $i \neq j$, and found asymptotic expansions for the probability of correct responses and MRT. See also McMillen and Holmes (2006). In that simplified case it is possible to find relatively simple expressions for $P(n, \theta)$ and the MRT, but in the more general case no such simple expressions are possible. Instead, we will evaluate the integrals in (25) numerically.

5.2.2 The absolute test in the LCAM approximates a max-vs-ave test

In the more general case in which inhibition and decay are non-zero ($w, k > 0$), it is not possible to obtain explicit formulas for the ER and MRT. However, McMillen and Holmes (2006) have shown that if w and k are moderately large, the absolute test on the accumulators approximates a max-vs-ave test. That is, for a given threshold z there is a Z such that

$$x_i \geq z \text{ and } x_i - \frac{1}{n-1} \sum_{j \neq i} x_j \geq Z \quad (26)$$

are equivalent, on average, in the limit as $(k + (n-1)w) \rightarrow \infty$. In fact, the two tests are indistinguishable in Monte Carlo simulations for even moderate values such as $w = k = 1, n = 4$. The absolute test on a system with moderately large inhibition and decay occupies a middle ground in performance between the absolute test and asymptotically optimal tests on the race model.

5.3 Comparison of tests

We have seen four different tests that might be applied to the accumulators to decide between the various hypotheses. In addition to these, we will also consider approximations of the asymptotically optimal tests. The ‘max-vs-next’ test—choose the largest accumulator when the difference between the largest and the next largest valued accumulator crosses a threshold—is an approximation of the δ_b test. We also consider an approximation to the δ_a test, which proceeds in the comparisons without the values of a and c . We call this approximate δ_a test the Δ_a test. The max-vs-next and Δ_a tests do not use knowledge of the matrix A . It will also be useful to consider the absolute test on $A\mathbf{x}$, that is, to consider only the effect of pre-multiplying the signal by the matrix of possible signals. The seven tests we consider, then, are as follows. The tests on the balanced model ($\lambda = 0$)

choose hypothesis H_i when the following quantities cross a threshold:

$$\begin{aligned}
& (A\mathbf{x})_i - \max_{j \neq i} (A\mathbf{x})_j && (\delta_b \text{ test}) \\
& x_i - \max_{j \neq i} x_j && (\text{max-vs-next}) \\
& (A\mathbf{x})_i && (\text{absolute in transformed coordinates}) \\
& x_i && \left(\begin{array}{l} \text{absolute: approximates} \\ \text{max-vs-ave when } w, k, n \text{ moderate} \end{array} \right) \\
& x_i - \frac{1}{n-1} \sum_{j \neq i} x_j && (\text{max-vs-ave}) \\
& \frac{1}{c^2} (A\mathbf{x})_i - \log \left[\sum_{j=1}^n \exp \left(\frac{1}{c^2} (A\mathbf{x})_j \right) \right] && (\delta_a \text{ test}) \\
& x_i - \log \left[\sum_{j=1}^n \exp(x_j) \right] && (\Delta_a \text{ test}) \tag{27}
\end{aligned}$$

The thresholds for the different tests will be different for a given accuracy level. The performance of tests is determined by how long it takes, on average, to make a decision at a given accuracy level. Therefore, we are interested in the performance of the tests at a fixed ER, i.e. the MRT for a given ER. To calculate this, we adjust the thresholds for the individual tests until the ER is as desired, and compute the MRT. We will present results for only the first four of the above tests, since results for the δ_a and δ_b tests are indistinguishable, as are the max-vs-next and Δ_a tests. In these, we will set the inhibition and decay equal to $w = k = 1$, unless otherwise noted. Additionally, we will consider the race model, in which $w = k = 0$. There is also a slight abuse of notation here relating to the selection of hypotheses. The hypothesis numbering does not necessarily correspond to the accumulator numbering. In a test on two hypotheses and eight accumulators, for example if we compare angle 3 to angle 7 in Fig. 2, hypothesis H_1 corresponds to angle 3 and H_2 to angle 7. In the third test in the list (27), if $(A\mathbf{x})_1$ crosses the threshold before $(A\mathbf{x})_2$, we choose H_1 (angle 3), and vice versa. However, in the absolute test, we choose H_1 (angle 3) if x_3 crosses before x_7 ; the other accumulators evolve but their values are not considered in making a decision. Comparisons of these tests are seen in Fig. 7. All results are obtained through Monte Carlo simulations (see Appendix E).

The results in Fig. 7 show that the effects of the type of ordering are only seen—as expected—at smaller values of the acuity, but that these differences are very marked when the acuity is small (ϕ is large). The behavior of the MRT as a function of the number of alternatives is qualitatively different depending on the ordering of alternatives when $\phi = 1$. We see that Hick’s Law holds, at least approximately, for both orderings when the acuity is very large, or when the signal is strongly peaked. For small acuity, Hick’s Law does *not* hold when the inside-out ordering is used, and it only holds for small numbers of alternatives when the outside-in ordering is used.

Perhaps the most striking fact about these results is that the MRT declines in the outside-in ordering as the number of alternatives increases, when the acuity is small (ϕ is large), and when the tests that use knowledge of the shape of the signals are used. This is in fact, intuitively correct, since if the spread of the signals is large the difference between the signals is small. For tests that don’t use the shape of the signals, this simply makes the test more difficult. However, if we know the possible shapes, then additional signals gives us more information about the shape, and we can use this to match the signal, in the manner of using a template. This is what the multiplication of the accumulators by the matrix A does.

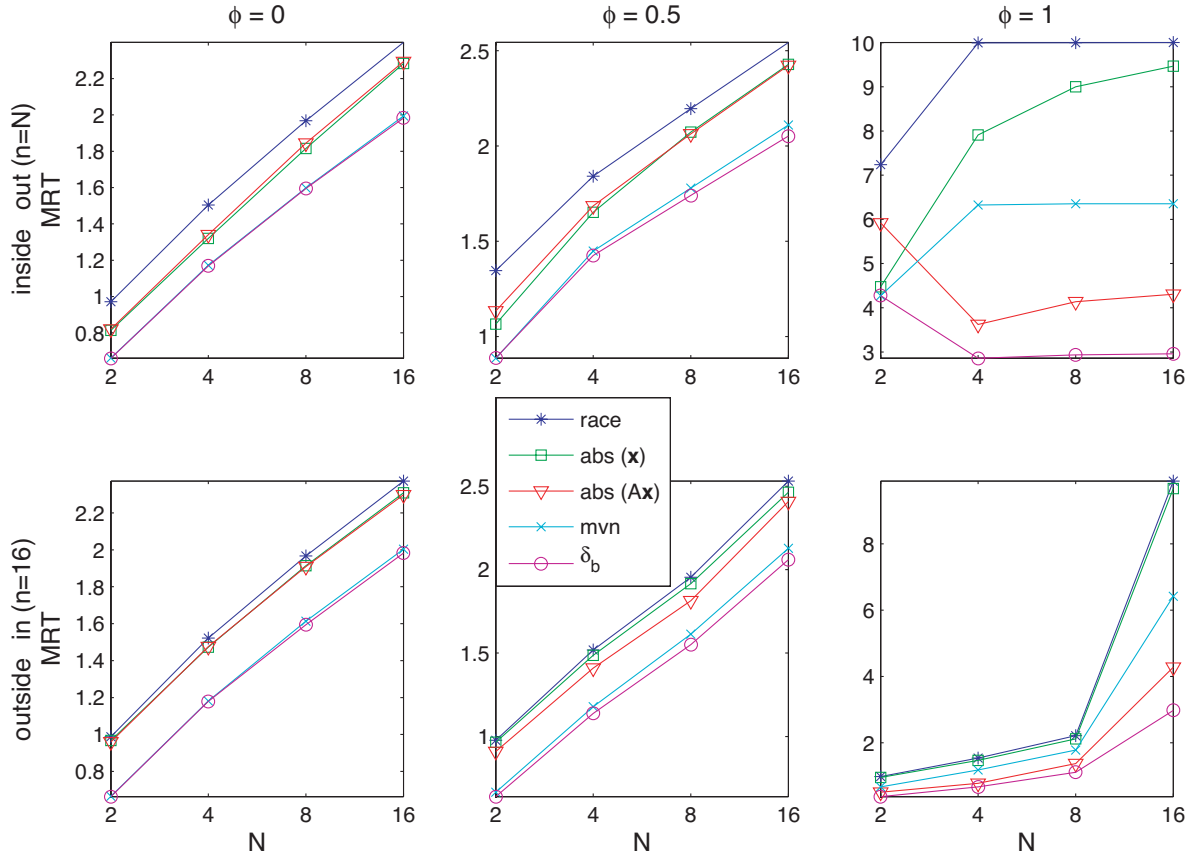


Figure 7: Comparison of MRT, as a function of the number of alternatives N for different values of the acuity. The ER is fixed at 0.05, and $a = 2$, $c = 1$. The top panels show results for the inside-out ordering where the number of accumulators is the same as the number of alternatives ($n = N$); the bottom panels use the outside-in ordering and the number of accumulators is fixed at $n = 16$. When $N < 16$ all of the accumulators evolve but only those corresponding to the alternatives are considered when making a decision.

Δ_a and max-vs-next are the best one can do without knowledge of the signal matrix A . The next best is the absolute test in the connectionist model, which is an approximation to the max-vs-average test, and the worst performing test is the race model. The differences in performance between the test grows with the spread of the signal, or as the acuity declines. For an infinite acuity there is not much difference between the tests, but when the acuity is moderately large, the best tests can achieve the same error rate at 1/3 the time of the worst tests. There is a significant advantage, then, to multiplying the accumulators by A , when the acuity is small. We see, in the outside-in ordering, for small ϕ , the max-vs-next test outperforms the absolute test on the transformed coordinates (i.e. choose i when $(A\mathbf{x})_i$ crosses a threshold). But, as the acuity decreases (ϕ increases), the test on $A\mathbf{x}$ performs better than the max-vs-next test. If the signals have a significant spread, it is much more beneficial to perform the pre-multiplication of signals by A than to implement a relative test.

Figure 7 also allows us to contrast results when the number of accumulators is the same as the number of alternatives ($n = N$), and when the number of accumulators is fixed ($n = 16$), and we

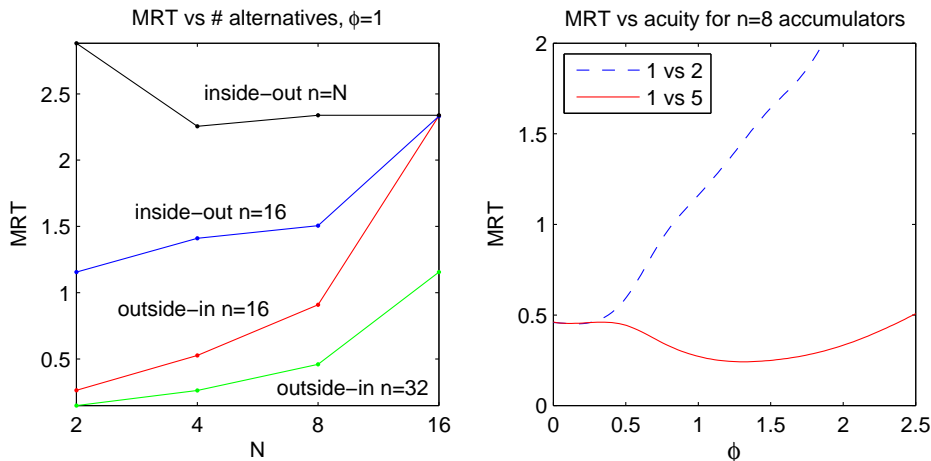


Figure 8: MRT at ER=0.1 as a function of the number of alternatives (left panel) and reciprocal of the acuity (right panel). The optimal δ_a test is used with $a = 2$, $c = 1$, $w = k = 1$.

vary the number of alternatives by allowing all accumulators to evolve but base our decision only on the accumulators corresponding to the alternatives. In the left panels the acuity is infinite, so the only difference is $n = N$ vs $n = 16$. We see that in both cases the MRT is approximately linear in $\log N$: Hicks Law holds in both cases when the acuity is infinite. However, we see that when $\phi = 0$ and $n = 16$, there is only a slight difference between the race model and the connectionist system with $w = k = 1$. We also see that when the number of accumulators is fixed at 16, there is very little improvement in the absolute test on the LCAM and in the race model. This means that the advantage gained through inhibition and decay attenuates as the number of accumulators increases. If an absolute test is used, it is best to use only as many accumulators as alternatives, which could be accomplished by ‘turning off’ those accumulators that don’t correspond to an alternative.

Figure 8 shows the effects of the acuity on different tests. In the left panel we hold the acuity fixed at $\phi = 1$ and vary the number of accumulators and alternatives. In these simulations the optimal δ_b test is used, but the results are qualitatively similar for the absolute test on $A\mathbf{x}$, that is, for tests in which knowledge of the shape of the signals is used. We see that performance increases with the number of accumulators, since more accumulators means that we have more information. In the right panel we compare the MRT to ϕ in tests with 8 accumulators but only two alternatives. When these alternatives are close together (1 vs 2), an increase in ϕ causes an increase in the MRT. However, when there are accumulators ‘between’ the alternatives, when we compare 1 vs 5, an increase in ϕ actually leads to a decline in the MRT. In fact, there is an optimal value of ϕ , which is greater than zero, and at which the MRT is almost half the MRT at $\phi = 0$. In other words, the spread of the signals can be an advantage if there are accumulators between those corresponding to the alternatives we are considering. But note that this can only be an advantage when we employ tests that utilize knowledge of the shape of the signals, as encoded in the matrix A .

6 The interrogation protocol

We now turn to the connectionist implementation of the interrogation protocol in which the subject is given a period of time T after stimulus presentation to arrive at a decision, which must be rendered at T . The problem, then, is to minimize the error rate for a process that ends at time T . This is

a fixed sample size hypothesis test. We ask, what is the best strategy to follow? There are several possible strategies one might follow, depending on energy costs, and availability of computational ability. The posterior probability that H_i is true is given by $\Pi_i(T)$ (eqn 18). Thus, the optimal test is to choose the largest of the Π_i s. However, this requires that we compute not only the value of the accumulators, but \mathbf{y} , requiring a multiplication by the matrix A , and the calculation of $\lambda \int_0^T \mathbf{y} dt$. We will explore strategies apart from this optimal one, that do not involve all of these calculations. For instance, we may not know A , or we may not know λ . There are, then, four possible strategies we might follow, whose decision rules d are given by the following, noting that since the decision time is given no thresholds are involved:

$$(i) \quad d = \left\{ i : x_i(T) > \max_{j \neq i} x_j(T) \right\}, \quad (28)$$

$$(ii) \quad d = \left\{ i : x_i(T) - \lambda \int_0^T x_i(s) ds > \max_{j \neq i} \left[x_j(T) - \lambda \int_0^T x_j(s) ds \right] \right\}, \quad (29)$$

$$(iii) \quad d = \left\{ i : (A\mathbf{x}(T))_i > \max_{j \neq i} (A\mathbf{x}(T))_j \right\}, \quad (30)$$

$$(iv) \quad d = \left\{ i : y_i(T) > \max_{j \neq i} y_j(T) \right\}. \quad (31)$$

The optimal strategy is strategy (iv), because this is merely choosing the alternative with the largest posterior probability; since the $\log \sum \exp(y_j)$ term in the Π_i 's (eqn 18) does not depend on i , $\Pi_i - \Pi_j = y_i - y_j$.

We consider the case in which the prior probabilities are equal, $\pi_i = 1/n$, in which case the appropriate initial conditions are $x_i(0) = 0$ for $i = 1, \dots, n$.

6.1 Strategy (i): Choose the largest accumulator

In the first strategy we simply choose the largest of the x_i 's. Since no knowledge of the possible signals is used, the probability of a correct response depends only on the signal that is presented: $S = (S_1, S_2, \dots, S_n)$. In this case, if H_i is true, then the probability of making a correct response is

$$P_i(n, T) = P(x_i(T) > x_j(T) \forall j \neq i \mid H_i). \quad (32)$$

The derivation of $P_i(n, T)$ using strategy (i) is in Appendix C. We find that

$$P_i(n, T) = \frac{1}{2^{n-1} \sqrt{\pi}} \int_{-\infty}^{\infty} \prod_{j \neq i} \left\{ 1 + \operatorname{erf} \left[y + \frac{S_i - S_j}{c} \sqrt{\frac{\tanh(\lambda T/2)}{\lambda}} \right] \right\} e^{-y^2} dy. \quad (33)$$

We see from (33) that since $\operatorname{erf}(x)$ is a strictly increasing function, the probability of making a correct response is maximized when the arguments in the error function in (33) are maximized. Thus, $P(n, T)$ increases when the differences $S_i > S_{j \neq i}$ between the signals to the correct and incorrect accumulators increase. Furthermore, $\tanh(\lambda T/2)/\lambda$ is an even function of λ and attains its maximum as a function of λ at $\lambda = 0$, i.e. when decay is balanced with inhibition. Therefore, for any number n of choices, the minimal ER in the interrogation protocol occurs when decay is balanced by inhibition, and the ER, and hence the RR, is an even function of λ .

6.2 Strategy (ii): Choose the largest $x_i - \lambda \int_0^T x_i(s) ds$

In this case, the calculations are similar to the previous section, except now we define $V_{ij}(t) \stackrel{\text{def}}{=} x_j(t) - x_i(t) - \lambda \int_0^T [x_j(s) - x_i(s)] ds$. Then, as before, $P_i(n, T) = P\left(\bigcap_{j \neq i} V_{ij}(T) < 0\right)$, and eqn

(61) becomes

$$dV_{ij} = -a_{ij} dt + c(dW_j - dW_i), \quad (34)$$

In other words, when we include the λ term in the decision, it drops out of the probability calculation. In this case, following the calculations of the previous section, but with $\lambda = 0$, we find that the probability of a correct response is

$$P_i(n, T) = \frac{1}{2^{n-1} \sqrt{\pi}} \int_{-\infty}^{\infty} \prod_{j \neq i} \left\{ 1 + \operatorname{erf} \left(y + \frac{S_i - S_j}{c} \sqrt{\frac{T}{2}} \right) \right\} e^{-y^2} dy. \quad (35)$$

The probability of a correct response thus depends only on the differences of the signal strengths, and not on the inhibition and decay. Furthermore, using strategy (ii), the probability of a correct response is the maximum over λ of the probabilities of a correct response using strategy (i).

6.3 Strategy (iii): Choose the largest $(Ax)_i$

In this case we assume that the matrix A of alternative signals (eqn 16) is known, for signal vectors $S^{(i)}$, $i = 1, \dots, n$. We will therefore use this information in making a decision. The probability of a correct response using this strategy is derived in Appendix D.

Let \hat{A} be the $N - 1 \times n - 1$ matrix with elements

$$\hat{A}_{jk} = \frac{S_{[k]}^{([j])} - S_{[k]}^{(i)}}{S_i^{(i)} - S_i^{([j])}}, \quad 1 \leq j \leq N - 1, \quad 1 \leq k \leq n - 1, \quad (36)$$

where $[j] = j$ if $j < i$ and $[j] = j + 1$ if $j \geq i$. Then the probability of a correct response is

$$P_i(n, T) = \frac{1}{\sqrt{\pi^n |\hat{A} \hat{A}^T|}} \int_{-\infty}^{\infty} \left[\int_{\mathbf{z} \leq \mathbf{y} + g(S, c, T, \lambda)} \exp \left(-\mathbf{z}^T \left(\hat{A} \hat{A}^T \right)^{-1} \mathbf{z} \right) d\mathbf{z} \right] e^{-y^2} dy. \quad (37)$$

where g in the integration limits is

$$g(S, c, T, \lambda) = \frac{(S^{(i)}, S^{(i)} - S^{([j]])})}{S_i^{(i)} - S_i^{([j]])} \frac{1}{c} \sqrt{\frac{\tanh(\lambda T/2)}{\lambda}}. \quad (38)$$

The above formula (37) simplifies to (33) when $\hat{A} = \mathbf{I}$, but when \hat{A} is not the identity, the formula does not give much transparent information about the probabilities of correct responses. For quantitative comparisons with strategies (i) and (iii), we will find it simpler to evaluate the SDEs with Monte Carlo simulations, than through the numerical evaluation of (37). However, we may glean important information about the dependence of $P_i(n, T)$ on λ . For one, $P_i(n, T)$ depends on λ only through the expression $g(S, c, T, \lambda)$. Since this is an even function of λ , $P_i(n, T)$ is also an even function of λ . Furthermore, since the integrand of the inner integral of (37) is positive, $P_i(n, T)$ is a strictly increasing function of the expression $g(S, c, T, \lambda)$. If assumption (15) holds then all elements in the first quotient in the expression $g(S, c, T, \lambda)$ are positive, and in this case $g(S, c, T, \lambda)$ achieves a global maximum at $\lambda = 0$, hence $P_i(n, T)$ is maximized when decay is balanced by inhibition $\lambda = 0$, as in strategy (i).

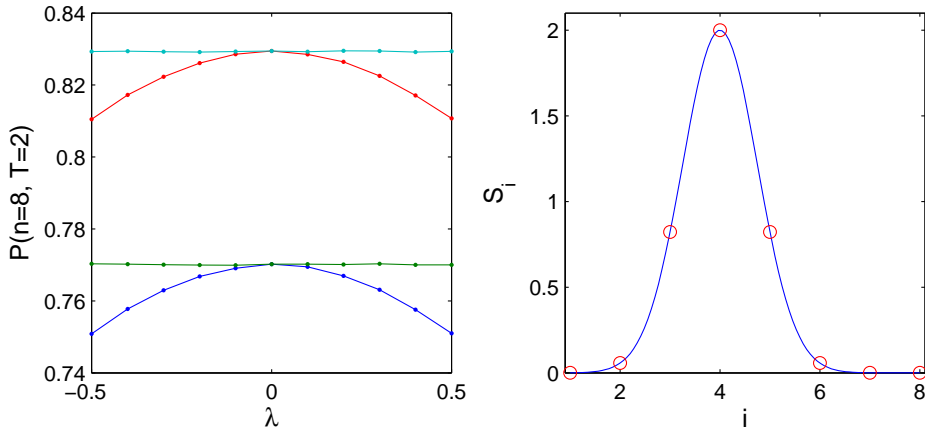


Figure 9: Comparison of different strategies in the interrogation protocol (left panel). They are, from top to bottom, strategies (iv), (iii), (ii) and (i). The acuity is $\phi^{-1} = 4/3$, and the probabilities are computed for $n = 8$ alternatives, at $T = 2$. The signal vector with drift $a = 2$ and acuity $\phi^{-1} = 4/3$ in case H_4 is true ($S = S^{(4)}$) is shown in the right panel.

6.4 Strategy (iv): Choose the hypothesis with the largest posterior probability

In Section 5.1 we computed the posterior probabilities $\Pi_i(t)$ of the hypotheses H_i , given $\mathbf{x}(s)$, $0 \leq s \leq t$, which, for equal prior probabilities, are given by

$$\mathbf{y} = \frac{1}{c^2} A \left(\mathbf{x} - \lambda \int_0^t \mathbf{x}(s) ds \right) \quad (39)$$

$$\log \Pi_i = y_i - \log \sum_{j=1}^n e^{y_j}, \quad (40)$$

Since $\log \sum_{j=1}^n e^{y_j}$ does not depend on i , choosing the most likely hypothesis is equivalent to choosing the largest y_i . Following this strategy, if we set $V_{ij}(t) \stackrel{\text{def}}{=} c^2 [y_j(t) - y_i(t)]$, then eqn (71) holds, with this new definition, and the equation for the V_{ij} 's is the same as in (72), but with $\lambda = 0$. The same calculation as in the previous section therefore follows, but with $\lambda = 0$. Therefore, strategy (iv) achieves the maximal probability of a correct response, as a function of λ , as achieved in strategy (iii).

6.5 Comparison of strategies

If the acuity is infinite ($S_m^{(i)} = S_0 + a \delta_{im}$), strategy (i) is equivalent to strategy (iii), and strategy (ii) is equivalent to strategy (iv). In this case, performance depends on λ in strategies (i) and (iii). If the acuity is large (concentrated signal) then there is little difference between the strategies. But, if the signal has a significant spread, there is a large advantage to using strategies (iii) or (iv). Figure 9 shows the probability of a correct response for 8 alternatives at time $T = 2$, following the four strategies outlined in the previous sections. For a small acuity, there is a significant advantage to multiplying the vector of accumulators by the matrix of possible signals. Figure 10 shows the effect of acuity on the performance of the four strategies. In all four of the strategies, performance improves as acuity increases, if the time to make a decision is fixed. Furthermore, the performance converges for Strategies (i) and (iii), and Strategies (ii) and (iv), at moderate values of the acuity

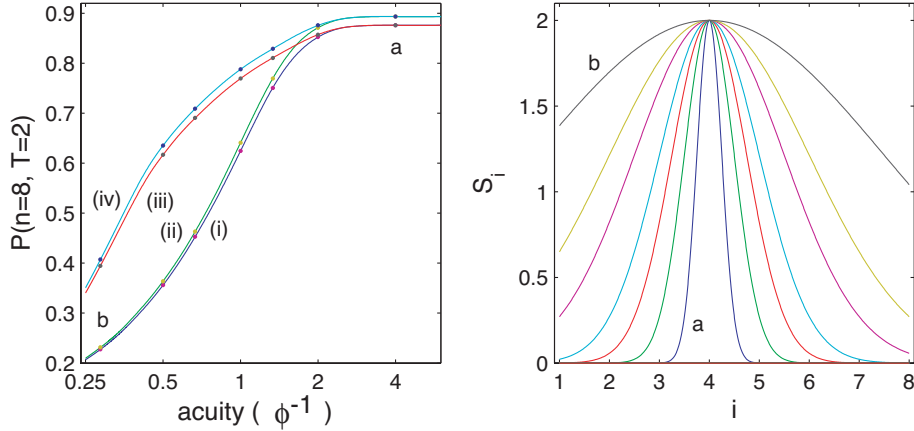


Figure 10: Probability of a correct response as a function of acuity for the four strategies. The right panel shows the signals corresponding to each dot in the left panel. This is the unbalanced case, $\lambda = 0.5$.

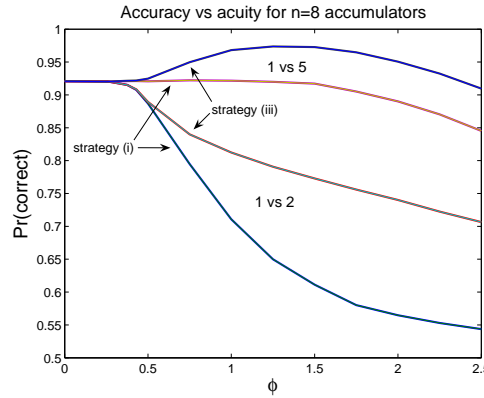


Figure 11: Probability of a correct response as a function of acuity at $T = 1$. The number of accumulators is $n = 8$, but only those corresponding to the given alternatives are considered when making a decision.

(around 2). In other words, for strongly peaked signals there is no advantage to multiplying the accumulators by A .

Figure 11 shows the effects of acuity on decisions between two alternatives when the number of accumulators is $n = 8$. When there are accumulators ‘between’ the alternatives there is a significant advantage to multiplying the accumulators by A , and the optimal value of the acuity is less than infinity. In this case the spread in the signal is an advantage.

7 Implementation in a neural network

As we have seen, there is a significant advantage to be gained by multiplying the accumulators by the matrix A of alternative signals. This can be accomplished in a simple network in which each accumulator x_i , $i = 1, \dots, n$, excites units z_j , $j = 1, \dots, N$, in an output layer with weight matrix W . Figure 12 shows such a network with four accumulators and four output units. In this network the weight of the connection from x_i to z_j is W_{ji} . Then, if we neglect noise in the output layer, the

vector of output units \mathbf{z} is governed by

$$\tau_z \frac{dz_j}{dt} = -z_j + f_j \left(\sum_{i=1}^n W_{ji} x_i \right), \quad (41)$$

therefore,

$$z_j \rightarrow f_j \left(\sum_{i=1}^n W_{ji} x_i \right). \quad (42)$$

When \mathbf{z} has reached its equilibrium, $\mathbf{z} = \mathbf{f}(W\mathbf{x})$. The matrix multiplication $A\mathbf{x}$ is therefore accomplished by tuning the weight matrix so that $W = A$. In other words, the weight matrix W should be the matrix A of alternative signals to perform the required matrix multiplication.

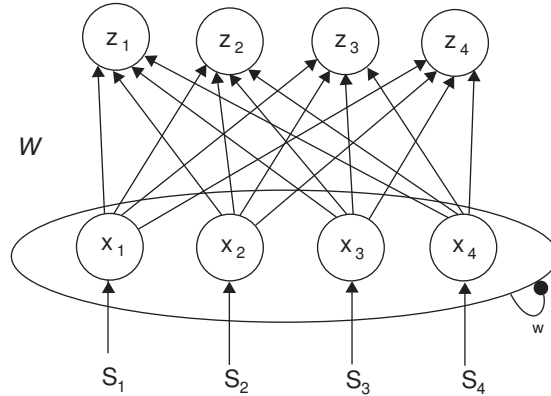


Figure 12: Neural network implementation of matrix multiplication. The closed circles denote excitatory connections and open circles denote inhibitory connections. The weight matrix W denotes the weights of the connections between the x_i 's and z_j 's. The weight of the connection between x_i and z_j is W_{ji} .

8 Comments

Much of the analysis in this paper has proceeded under the assumption that the sum of squares of the signals is the same, regardless of which signal vector is presented (eqn 15). This is a reasonable assumption in many different situations. It is the case when we know the shape of the signal, and the task is to determine where the peak of the signal is. However, it obviously cannot account for effects in unidimensional identification tasks such as the “bow effect,” in which performance is significantly better when presented with stimuli at the ends of the range of possible stimuli (see, e.g., Luce et al. (1982), Lacouture and Marley (1995, 2004), and references therein). Such unidimensional tasks are not the subject of this report. However, we note that preliminary simulations on the LCAM using the signals as proposed by Lacouture and Marley (1995) indicate that subjects *do not* pre-multiply the accumulators by the matrix of possible signals. Such multiplication (i.e., choosing i when $(A\mathbf{x})_i$ crosses a threshold, as opposed to when x_i crosses a threshold results in much improved MRT, but it does not reproduce the bow effect. Further research is needed to determine if some pre-multiplication is done, perhaps on a different kind of signal set.

Given a decision task, such as deciding what angle a bar lies at on a screen, subjects improve their performance over time. The accumulator model, in which accumulators are preferentially

sensitive to various alternatives, suggests several ways in which improvement can occur. For alternatives that are close together, in the sense of there being no or few accumulators ‘between’ them, the way to improve is to increase the acuity in the signal. To discern between angles that are close together, for example, this would mean tightening the tuning curves of the receptors. However, if the task is to decide quickly between angles that are a moderate distance apart, this can be facilitated by adjusting the acuity in the signal. There is an optimal value of the acuity that is non-zero. The spread of the signal can be an aid to decision making.

Significant improvements in performance can be achieved if the incoming signals are pre-multiplied by the matrix A of possible signals. This performance improvement is such that it becomes beneficial to increase the spread of the signals in some cases. This may perhaps go some way to explaining why tuning curves are not always at their maximal sharpness. Why is it that they can be adjusted at all? We showed that optimal performance occurs at positive values of ϕ only if there are accumulators ‘between’ the alternatives under consideration and only if the signal is multiplied by A . This suggests an explanation for the improvement in angle discrimination tasks in monkeys. Presumably, when the test is first taken, the subject uses a criteria without any (or, with little) knowledge of the spread of the signals. After a few trials, the subject learns the shape of the incoming signals, and can use this to her benefit. One possible mechanism for learning, then, is a shift from a test that does not rely on a knowledge of the shape of the signals, to one that does. For example, the subject could initially begin with a max-vs-next test, and then shift to a δ_b test, once the matrix A has been learned.

Additionally, the subject can learn by adjusting the spread of the signals ϕ , by altering the tuning curves in receptor neurons. This will be an advantage if the alternatives are separated by accumulators. The effects of these adjustments on improvement are limited by the number of available accumulators. For the angle discrimination task, for example, if there are no accumulators sensitive to angles between 22 and 26 degrees, then the subject will gain an advantage in tests on these angles by increasing the acuity of signals as much as possible. But in tests between 15 and 30 degrees, it will be an advantage to decrease the acuity. If this model is an accurate description of decision making, then this explains why the acuity can be adjusted. For the most common discrimination tasks it is an advantage to have a moderately small acuity. Training a subject in a particular task can result in a sharpening of the tuning curves. But, the improvements in performance are limited by the number of accumulators. Moreover, performance is limited by the *range* of the signals. Even if all the information in the signals is used, the range of the signal strength limits the amount of information that can be carried. Since the signals are presumably limited by the maximal firing rate of neurons, this limits the amount of information that can be conveyed to a decision process. The ratio of noise inherent in the signals and decision process to the range of signals thus imposes an absolute limit on the amount of information that can be processed. This could perhaps be a clue to understanding the ‘magic number seven’ limit in signal-response information capacity in humans.

Lastly, we note that we know of no experiments that test the difference between performance in tests in which the number of alternatives is increased via the two orderings we have suggested—the ‘inside-out’ and ‘outside-in’ orderings. Such experiments would be interesting for determining whether or not humans perform the multiplication by the matrix of possible signals.

Appendix: Mathematical details

A Derivation of posteriori probabilities under assumption (15)

The log-likelihood ratios $Z_{ij}(t)$ are determined by the infinitesimal increments:

$$dZ_i(t) = \log p_i(d\mathbf{x}), \quad (43)$$

where p_i is the probability density function of the increments $d\mathbf{x}$ which are Gaussian distributed,

$$p_i(d\mathbf{x}) = \prod_{m=1}^n \frac{1}{\sqrt{2\pi\sigma_m^{(i)2}}} \exp \left[-\frac{\left(dx_m - \mu_m^{(i)}\right)^2}{2\sigma_m^{(i)2}} \right], \quad (44)$$

where $\mu_m^{(i)}$ and $\sigma_m^{(i)2}$ are the mean and variance of the m th component of the increment of \mathbf{x} , given that hypothesis H_i is true. Hence

$$\begin{aligned} \mu_m^{(i)} &= \left(\lambda x_m - w \sum_{j=1}^n x_j + S_m^{(i)} \right) dt, \\ \text{and } \sigma_m^{(i)2} &= c^2 dt, \end{aligned} \quad (45)$$

where $\lambda = w - k$. Letting $\gamma_m = \lambda x_m - w \sum_{j=1}^n x_j$, and substituting (45) into (43), we obtain:

$$\begin{aligned} dZ_i &= \log \left(\frac{1}{\sqrt{2\pi c^2 dt}} \right) \\ &\quad - \frac{1}{2c^2 dt} \sum_{m=1}^n \left[dx_m^2 - 2\gamma_m dx_m dt + \left(\gamma_m^2 - 2w \left(\sum_{j=1}^n x_j \right) S_m^{(i)} + S_m^{(i)2} \right) dt^2 \right] \\ &\quad + \frac{1}{c^2} \sum_{m=1}^n \left[S_m^{(i)} dx_m - \lambda S_m^{(i)} x_m dt \right]. \end{aligned} \quad (46)$$

We use the assumption (15), and the fact that all terms that do not depend on i will cancel when we subtract terms. The only terms in (46) that do depend on i are those in the third row of (46). Therefore, after integrating (46), we have

$$Z_i(t) = C + \frac{1}{c^2} \sum_{m=1}^n S_m^{(i)} \left(x_m - \lambda \int_0^t x_m(s) ds \right), \quad (47)$$

where C is a function of time but does not depend on i . We form the $N \times n$ matrix of alternative signals:

$$A = \begin{pmatrix} S^{(1)} \\ S^{(2)} \\ \dots \\ S^{(N)} \end{pmatrix}. \quad (48)$$

Then the logs of the posterior probability and generalized likelihood ratio are

$$\log \Pi_i = \frac{1}{c^2} \left((A\mathbf{x})_i - \lambda \int_0^t (A\mathbf{x})_i dt \right) - \log \sum_{j=1}^n \frac{\pi_j}{\pi_i} \exp \left[\frac{1}{c^2} \left((A\mathbf{x})_j - \lambda \int_0^t (A\mathbf{x})_j dt \right) \right], \quad (49)$$

$$\log L_i = \frac{1}{c^2} \left((A\mathbf{x})_i - \lambda \int_0^t (A\mathbf{x})_i dt \right) - \max_{j \neq i} \left\{ \frac{1}{c^2} \left((A\mathbf{x})_j - \lambda \int_0^t (A\mathbf{x})_j dt \right) + \log \frac{\pi_j}{\pi_i} \right\}. \quad (50)$$

B Relaxation of assumption (15)

We write eqn (46) as

$$dZ_i = \log \left(\frac{1}{\sqrt{2\pi c^2 dt}} \right) - \frac{1}{2c^2 dt} \sum_{m=1}^n [dx_m^2 - 2\gamma_m dx_m dt + \gamma_m^2 dt^2] \\ + \frac{1}{c^2} \sum_{m=1}^n [S_m^{(i)} dx_m - \lambda S_m^{(i)} x_m dt] + \frac{dt}{2c^2} \sum_{m=1}^n \left[S_m^{(i)} \left(S_m^{(i)} - 2w \sum_{j=1}^n x_j \right) \right]. \quad (51)$$

Therefore,

$$Z_i(t) = C(t) + \frac{1}{c^2} \sum_{m=1}^n S_m^{(i)} \left(x_m - \lambda \int_0^t x_m(s) ds \right) \\ - \frac{t}{2c^2} \left(\sum_{m=1}^n S_m^{(i)2} \right) + \frac{w}{c^2} \left(\sum_{m=1}^n S_m^{(i)} \right) \int_0^t \left(\sum_{j=1}^n x_j(s) \right) ds, \quad (52)$$

where $C(t)$ does not depend on i . Let

$$\mathbf{y} = \frac{1}{c^2} A \left(\mathbf{x} - \lambda \int_0^t \mathbf{x}(s) ds \right) + \log \boldsymbol{\pi}, \quad (53)$$

$$\text{and } u_i(t) = -\frac{t}{2c^2} \left(\sum_{m=1}^n S_m^{(i)2} \right) + \frac{w}{c^2} \left(\sum_{m=1}^n S_m^{(i)} \right) \left(\sum_{j=1}^n \int_0^t x_j(s) ds \right). \quad (54)$$

Then

$$\log \Pi_i = y_i + u_i - \log \sum_{j=1}^n e^{y_j + u_j}, \quad (55)$$

$$\log L_i = y_i + u_i - \max_{j \neq i} (y_j + u_j). \quad (56)$$

We see that under assumption (15), the u_i 's drop out of the above formulas. When this assumption does not hold, the decision is biased toward the hypothesis corresponding to the largest sum of squares of the elements of the signal vector. Interestingly, the u_i terms depend on the inhibition w , but not on the decay.

We can make a further simplification by subtracting from both sides of (55-56) those parts of the sums of the signals that do not depend on i . Let

$$\sum_{m=1}^n S_m^{(i)2} = q_0 + q_i, \quad (57)$$

$$\sum_{m=1}^n S_m^{(i)} = r_0 + r_i, \quad (58)$$

where q_0 and r_0 are the minimums of the corresponding sums. Since the terms involving q_0 and r_0 do not depend on i , we can let

$$u_i(t) = -\frac{1}{c^2} \left(2t q_i - w r_i \sum_{j=1}^n \int_0^t x_j(s) ds \right). \quad (59)$$

Then (55-56) holds with this new definition of u_i . We see that the posterior probability depends on the variation in the values of the sums and sums of squares of the possible signals.

C Derivation of probability of a correct response in the interrogation protocol, strategy (i)

We define the random variables $V_{ij}(t) \stackrel{\text{def}}{=} x_j(t) - x_i(t)$, and the constants $a_{ij} \stackrel{\text{def}}{=} S_i - S_j$ for $j = 2, \dots, n$ so that (32) becomes:

$$P_i(n, T) = P \left(\bigcap_{j \neq i} (V_{ij}(T) < 0) \right). \quad (60)$$

From (14) we see that the random variables V_{ij} satisfy the equations

$$dV_{ij} = (\lambda V_{ij} - a_{ij}) dt + c (dW_j - dW_i), \quad (61)$$

for independent Wiener processes dW_j , and we may integrate to obtain:

$$V_{ij}(t) = \frac{a_{ij}}{\lambda} (1 - e^{\lambda t}) + c \int_0^t e^{\lambda(t-s)} [dW_j(s) - dW_i(s)]. \quad (62)$$

We further define random variables $W_0(t)$ and $U_{ij}(t)$, for $j \neq i$ as follows:

$$W_0(t) \stackrel{\text{def}}{=} c \int_0^t e^{\lambda(t-s)} dW_i(s), \quad (63)$$

$$\begin{aligned} U_{ij}(t) &\stackrel{\text{def}}{=} V_{ij}(t) + W_0(t) \\ &= \frac{a_{ij}}{\lambda} (1 - e^{\lambda t}) + c \int_0^t e^{\lambda(t-s)} dW_j(s), \end{aligned} \quad (64)$$

so that W_0 and the U_{ij} 's are independent processes. The U_{ij} 's have means $\mu_{ij}(t)$ and variances $\sigma^2(t)$, and W_0 has mean 0 and variance $\sigma^2(t)$, where

$$\mu_{ij}(t) = \frac{a_{ij}}{\lambda} (1 - e^{\lambda t}) \quad \text{and} \quad \sigma^2(t) = \frac{c^2}{2\lambda} (e^{2\lambda t} - 1). \quad (65)$$

Since $V_{ij} = U_{ij} - W_0$, the probability of making a correct response can now be written as

$$P_i(n, T) = P \left(\bigcap_{j \neq i} [U_{ij}(T) < W_0(T)] \right). \quad (66)$$

The random variable W_0 has the density $f(w)$ of a normal random variable with mean 0 and variance σ^2 , and U_{ij} has the distribution $F_{U_{ij}}$ of a normal random variable with mean μ_{ij} . Since the U_{ij} 's and W_0 are independent,

$$\begin{aligned} P_i(n, T) &= \int_{-\infty}^{\infty} \prod_{j \neq i} F_{U_{ij}}(w) f(w) dw \\ &= \int_{-\infty}^{\infty} \prod_{j \neq i} \left\{ \frac{1}{2} \left(1 + \operatorname{erf} \left[\frac{w - \mu_{ij}}{\sqrt{2\sigma^2}} \right] \right) \right\} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{w^2}{2\sigma^2} \right] dw. \end{aligned} \quad (67)$$

Making the change of variables $y = w/\sqrt{2\sigma^2}$, we obtain:

$$P_i(n, T) = \frac{1}{2^{n-1}\sqrt{\pi}} \int_{-\infty}^{\infty} \prod_{j \neq i} \{1 + \operatorname{erf}[y + g(a_{ij}, c, \lambda, T)]\} e^{-y^2} dy, \quad (68)$$

where g in the argument of the error function takes the form

$$g(a_{ij}, c, \lambda, T) = -\frac{\mu_{ij}(T)}{\sqrt{2\sigma^2(T)}} = \frac{a_{ij}}{c} \sqrt{\frac{\tanh(\lambda T/2)}{\lambda}} \quad (69)$$

where we have used the definitions of $\mu(T)$ and $\sigma^2(T)$ in (65). Therefore,

$$P_i(n, T) = \frac{1}{2^{n-1}\sqrt{\pi}} \int_{-\infty}^{\infty} \prod_{j \neq i} \left\{ 1 + \operatorname{erf} \left[y + \frac{S_i - S_j}{c} \sqrt{\frac{\tanh(\lambda T/2)}{\lambda}} \right] \right\} e^{-y^2} dy, \quad (70)$$

D Derivation of probability of a correct response in the interrogation protocol, strategy (iii)

We define $V_{ij}(t) \stackrel{\text{def}}{=} (\mathbf{Ax}(t))_j - (\mathbf{Ax}(t))_i$. Then the probability of choosing correctly, given that signal $S^{(i)}$ is presented is

$$P_i(n, T) = P \left(\bigcap_{j \neq i} V_{ij}(T) < 0 \mid S = S^{(i)} \right). \quad (71)$$

If $S = S^{(i)}$,

$$dV_{ij} = \left[\lambda V_{ij} + (S^{(j)} - S^{(i)}, S^{(i)}) \right] dt + c (S^{(j)} - S^{(i)}, d\mathbf{W}), \quad (72)$$

where (\cdot, \cdot) is the dot product and \mathbf{W} is a vector of Wiener processes. Let \hat{A} be the $(N-1) \times (n-1)$ matrix formed by removing the i th column from the matrix of rows of $S^{(j)} - S^{(i)}$, and $-\tilde{S}^{(i)}$ the column that we remove. That is,

$$\begin{pmatrix} S^{(1)} - S^{(i)} \\ \dots \\ S^{(i-1)} - S^{(i)} \\ S^{(i+1)} - S^{(i)} \\ \dots \\ S^{(N)} - S^{(i)} \end{pmatrix} = \begin{pmatrix} \underbrace{\hat{A}_1}_{\text{1st column}} & \underbrace{\hat{A}_2}_{\text{2nd column}} & \dots & \underbrace{\hat{A}_{i-1}} & \underbrace{-\tilde{S}^{(i)}}_{\textit{ith column}} & \hat{A}_i & \dots & \underbrace{\hat{A}_{n-1}}_{\text{nth column}} \end{pmatrix}, \quad (73)$$

where \hat{A}_k is the k th column of \hat{A} . Likewise, let V be the vector of V_{ij} 's and \hat{S} the vector of $(S^{(j)} - S^{(i)}, S^{(i)})$, with the i th elements (of 0) removed. Then eqn (72) can be written in vector form as

$$dV = \left(\lambda V + \hat{S} \right) dt + c \left(\hat{A} d\hat{\mathbf{W}} - \tilde{S}^{(i)} dW_i \right), \quad (74)$$

where $\hat{\mathbf{W}} = (W_1, \dots, W_{i-1}, W_{i+1}, \dots, W_n)^T$. (Note that $\tilde{S}^{(i)}$ is a vector and W_i is a scalar Wiener process.) Making the further transformations $V_j \rightarrow V_j/\tilde{S}_j^{(i)}$, $\hat{S}_j \rightarrow \hat{S}_j/\tilde{S}_j^{(i)}$, $\hat{A}_{jk} \rightarrow \hat{A}_{jk}/\tilde{S}_j^{(i)}$, the equation becomes

$$dV = \left(\lambda V + \hat{S} \right) dt + c \left(\hat{A} d\hat{\mathbf{W}} - \mathbf{1} dW_i \right), \quad (75)$$

where $\mathbf{1}$ is the vector of 1's. The difference, then between the above equation (75), and eqn (61), is the matrix \hat{A} , which in (61) is simply the identity. To summarize, the quantities appearing in eqn (75) are

$$\hat{S}_j = \frac{(S^{([j])} - S^{(i)}, S^{(i)})}{S_i^{(i)} - S_i^{([j])}}, \quad \hat{A}_{jk} = \frac{S_{[k]}^{([j])} - S_{[k]}^{(i)}}{S_i^{(i)} - S_i^{([j])}}, \quad 1 \leq j \leq N-1, \quad 1 \leq k \leq n-1, \quad (76)$$

where $[j] = j$ if $j < i$ and $[j] = j + 1$ if $j \geq i$.

Following the strategy of Appendix C, we define $U = V + \mathbf{1}W_0$, where $W_0(t)$ is defined by eqn (63), and U satisfies

$$dU = \left(\lambda U + \hat{S} \right) dt + c \hat{A} d\hat{\mathbf{W}}. \quad (77)$$

The elements of U are not correlated with W_0 , but they *are* correlated with each other, and the correlation of elements of U is related to the acuity in the signal. The probability of a correct response is therefore

$$P_i(n, T) = \int_{-\infty}^{\infty} F_U(\mathbf{1}w) f(w) dw, \quad (78)$$

where F_U is the joint distribution of U , and $f(w)$ is the density of a normal random variable with mean 0 and variance σ^2 . The joint distribution F_U is calculated from the joint density:

$$f_U(\mathbf{u}) = \frac{1}{\sqrt{(2\pi)^{n-1} |\boldsymbol{\sigma}|}} \exp \left(-\frac{1}{2} (\mathbf{u} - \boldsymbol{\mu})^T \boldsymbol{\sigma}^{-1} (\mathbf{u} - \boldsymbol{\mu}) \right), \quad (79)$$

$$\text{where } \boldsymbol{\mu} = \frac{\hat{S}}{\lambda} (e^{\lambda t} - 1), \quad (80)$$

$$\text{and } \boldsymbol{\sigma} = \frac{c^2}{2\lambda} (e^{2\lambda t} - 1) \hat{A} \hat{A}^T \quad (81)$$

is the covariance matrix. Note that $\boldsymbol{\sigma} = \sigma^2 \hat{A} \hat{A}^T$, where σ^2 is the same as in eqn (65). Hence, the joint distribution of U is

$$F_U(\mathbf{w}) = \int_{\mathbf{y} \leq \mathbf{w}} f_U(\mathbf{y}) d\mathbf{y}. \quad (82)$$

Making the change of variables $\mathbf{z} = (\mathbf{y} - \boldsymbol{\mu}) / \sqrt{2\sigma^2}$, in the integral for F_U ,

$$F_U(\mathbf{w}) = \frac{1}{\sqrt{\pi^{n-1} |\hat{A} \hat{A}^T|}} \int_{\mathbf{z} \leq \frac{\mathbf{w} - \boldsymbol{\mu}}{\sqrt{2\sigma^2}}} \exp \left(-\mathbf{z}^T (\hat{A} \hat{A}^T)^{-1} \mathbf{z} \right) d\mathbf{z}. \quad (83)$$

Substituting this into the expression eqn (78), and making the further change of variables $y = w / \sqrt{2\sigma^2}$, we have

$$P_i(n, T) = \frac{1}{\sqrt{\pi^n |\hat{A} \hat{A}^T|}} \int_{-\infty}^{\infty} \left[\int_{\mathbf{z} \leq y - \frac{\boldsymbol{\mu}}{\sqrt{2\sigma^2}}} \exp \left(-\mathbf{z}^T (\hat{A} \hat{A}^T)^{-1} \mathbf{z} \right) d\mathbf{z} \right] e^{-y^2} dy. \quad (84)$$

E Fast Monte Carlo simulations using Matlab

Matlab's optimized array computations may be utilized to vastly improve the performance of Monte Carlo simulations. The key is to write all of the calculations in terms of arrays. The SDEs considered in this paper are Ornstein-Uhlenbeck (OU) processes:

$$dX = (AX + S) dt + c dW. \quad (85)$$

In the connectionist system (14), $A = \lambda \mathbf{I} - w \mathbf{1}$, where \mathbf{I} is the identity matrix and $\mathbf{1}$ is the full matrix of 1's. Sample paths of the OU process (85) are approximated by

$$X_{i+1} = BX_i + \Delta S + c\sqrt{\Delta} N_{i+1}, \quad (86)$$

where $X_i \approx X(i \Delta)$, (notice that X_i is a *vector*) N_i is a vector of standard normal random variable, Δ is the time step, and $B = \mathbf{I} + \Delta A$. Naively, we could evaluate (87) in a ‘for’ or ‘while’ loop. However, we can compute all of the X_i ’s for $i = 1, \dots, m$ in a single matrix operation. We can directly sum (86) to obtain

$$X_i = B^i X_0 + \Delta \left(\sum_{j=0}^{i-1} B^j \right) S + c\sqrt{\Delta} \sum_{j=0}^{i-1} B^j N_{i-j}. \quad (87)$$

If Δ is sufficiently small, B can be decomposed into $B = UDU^T$, where D is diagonal and U is an orthogonal matrix. We use the fact that since U is orthogonal, $U^T N_{i-j}$ is also a vector of independent standard normal random variables. Therefore,

$$X_i = UD^i U^T X_0 + \Delta U \left(\sum_{j=1}^i D^{j-1} \right) U^T S + c\sqrt{\Delta} U \sum_{j=1}^i D^{j-1} N_{i-j}. \quad (88)$$

We can compute an $n \times npts$ matrix where the i th column is the vector X_i by the following Matlab operations:

```
% N - number of alternatives
% npts - number of iterations of X_i to calculate
% DT - time step
% A - linear part of SDE
% y0 - initial condition

B = eye(N) + DT * A;
[U, D, V] = svd(B);
dd = diag(D);
ds = dd(diag(1:N) * ones(N, npts)).^ (ones(N, npts) * diag( 0:(npts-1) ));
ds2 = dd(diag(1:N) * ones(N, npts)).^ (ones(N, npts) * diag( (npts-1):-1:0 ));
drift = DT * U * diag(U'*S) * cumsum(ds, 2);

y = U * diag(V*y0) * D * ds + drift + ...
    c * sqrt(DT) * U * (cumsum(ds2.* randn(N, npts),2)./ ds2);
```

y is then an $N \times npts$ matrix whose i th column is X_i , i.e. $X_i = y(:, i)$. The key point is that the computation of a sample path is contained in only the last line of the above. This line can be placed inside a loop. If the initial condition is the same for all sample paths, the first term can be absorbed into “drift.” If only the last element of the sample path is required, “cumsum” can be replaced with “sum” to slightly increase performance. To calculate $\int x_i(s) ds$, however, these terms will be needed. The above algorithm is generally about ten times faster than a simple ‘for’ loop. Also, note that if one is calculating a first passage time problem, these can be done in ‘blocks’, and a search through the block made, which is still about five times faster than a ‘while’ loop.

F The two alternative case

The main theoretical results on the two-alternative case are collected in Bogacz et al. (2004). We briefly sketch them here. For $n = 2$ eqn (14) reduces to:

$$\begin{aligned} dx_1 &= (-kx_1 - wx_2 + S_1) dt + c dW_1, \\ dx_2 &= (-kx_2 - wx_1 + S_2) dt + c dW_2, \end{aligned} \quad (89)$$

and the hypotheses H_1 and H_2 become:

$$H_1 = \{S_1 = S_0 + a, S_2 = S_0\}, \text{ and } H_2 = \{S_1 = S_0, S_2 = S_0 + a\}.$$

Making the orthonormal change of variables $y_1 = (x_1 + x_2)/\sqrt{2}$, $y_2 = (x_2 - x_1)/\sqrt{2}$ to an eigenvector basis, equations (89) decouple and become:

$$dy_1 = \left(-(w+k)y_1 + \frac{2S_0+a}{\sqrt{2}} \right) dt + c d\tilde{W}_1, \quad (90)$$

$$dy_2 = \left(\lambda y_2 + \tilde{S} \right) dt + c d\tilde{W}_2, \quad (91)$$

where $\lambda = w - k$, $\tilde{S} = (S_2 - S_1)/\sqrt{2}$, and $d\tilde{W}_1, d\tilde{W}_2$ are i.i.d. Wiener processes that result from the orthogonal transformation. The hypotheses H_1 and H_2 are

$$H_1 = \left\{ \tilde{S} = -\frac{a}{\sqrt{2}} \right\} \text{ and } H_2 = \left\{ \tilde{S} = \frac{a}{\sqrt{2}} \right\}.$$

Since $w+k > 0$, (90) is a stable Ornstein-Uhlenbeck (OU) process whose sample paths approach a normal distribution with mean $(2S_0+a)/\sqrt{2}(w+k)$ and variance

$$\sigma^2 = \frac{c^2}{2(w+k)} \quad (92)$$

(Gardiner, 2004), so if $(w+k)$ is large most sample paths will converge to the attracting line $y_1 = (2S_0+a)/\sqrt{2}(w+k)$ before crossing a threshold. Hence we may reduce to the scalar OU process (91) with drift $\pm a/\sqrt{2}$, depending on which signal is presented, and the modified threshold (Bogacz et al., 2004):

$$\bar{\theta} = \frac{2\theta(w+k) - (2S_0+a)}{\sqrt{2}(w+k)}. \quad (93)$$

When $\lambda = w - k = 0$ we say that (89) is *balanced*, and in this case (91) becomes a Wiener process with drift. The original system of coupled OU processes (14) is therefore transformed into an OU process and a Wiener process with drift. Under the assumption that the stationary distribution of the OU component concentrates inside the absorbing boundaries, the first passage dynamics are determined primarily by its Wiener component (91), which is the continuum limit of the SPRT. Thus, for sufficiently large $w = k$ the network with absolute thresholds is an optimal decision-maker (Bogacz et al., 2004). Figure 13 shows typical sample paths of (89), and the transformation to the coordinates y_1, y_2 . Quantitative comparisons showing good agreement among the full nonlinear, linearised, and reduced systems are given in Brown et al. (2005).

Bogacz et al. (2004) consider the system (89) under both free-response and interrogation protocols. In the latter, performance is characterized solely by the error rate: the probability of making an incorrect choice; in the former it is characterized by the *reward rate* (RR) or average number of correct responses per unit time, as suggested by Gold and Shadlen (2002):

$$\text{RR} = \frac{1 - \text{ER}}{\text{MRT} + D + D_p \text{ER}}. \quad (94)$$

Here D is a fixed delay between response and the next stimulus presentation and D_p is an additional ‘penalty’ delay imposed for incorrect responses. (This RR presumes that no other penalty is imposed for incorrect responses; clearly other objective functions can be proposed that include such effects, or are weighted for accuracy, etc.) Recognizing that when $\lambda = 0$ (91) is a simple drift-diffusion equation, which is also the continuum limit of Wald’s SPRT, it is shown in Bogacz et al.

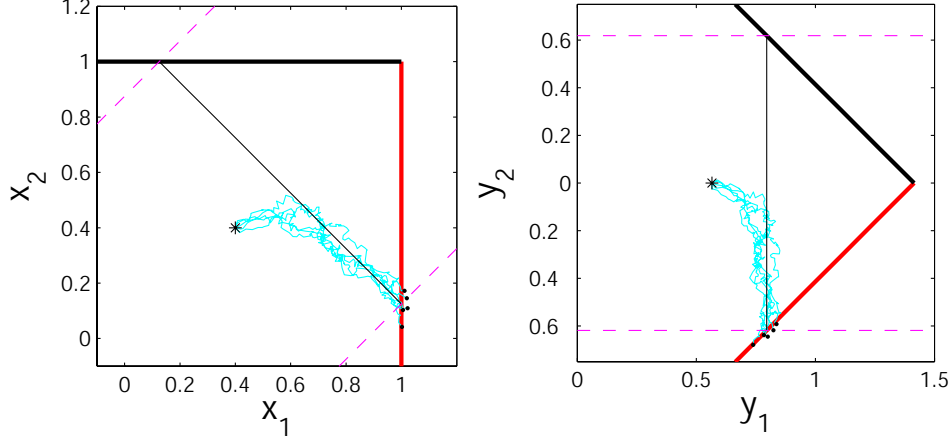


Figure 13: Five typical sample paths in the two-choice connectionist model showing the original and decoupled coordinate systems and illustrating convergence to an attracting line. The thick lines are the thresholds $x_i = \theta$; the thin solid line is the attracting line; the dashed lines are the thresholds for y_2 ; the star is the initial condition, and dots are the final states following threshold crossing. Parameters are $S_0 = 4$, $\theta = 1$, $a = 1$, $c = 0.1$, $w = 4$, $k = 4$.

(2004) that RR has a unique maximum at $\lambda = 0$ and at an optimal threshold $\bar{\theta} = \bar{\theta}_c(a, c; D + D_p)$. Specifically, assuming equiprobable alternatives (in which case it is optimal to start each trial at $y_2(0) = 0$) and using the classical results for mean first passage time and probability of crossing the incorrect threshold (Gardiner, 2004, Eqns 5.2.158, 5.2.189), we have:

$$\text{MRT} = \frac{\sqrt{2}\bar{\theta}}{a} \tanh\left(\frac{a\bar{\theta}}{\sqrt{2}c^2}\right), \quad \text{ER} = \frac{1}{1 + \exp\left(\frac{\sqrt{2}a\bar{\theta}}{c^2}\right)}. \quad (95)$$

Substituting these in (94) and differentiating we obtain the condition that implicitly defines $\theta_c(a, c; D + D_p)$:

$$\exp\left(\frac{\sqrt{2}a\bar{\theta}_c}{c^2}\right) - 1 = \left(\frac{a}{c}\right)^2 \left(D + D_p - \frac{\sqrt{2}\bar{\theta}_c}{a}\right). \quad (96)$$

Behavioral studies of human ability to achieve this optimum appear in (Bogacz et al., 2004).

The parameters a, c and $\bar{\theta}$ appear in (95-96) only in the groups (a/c) and $(\bar{\theta}/a)$. The latter can be eliminated from (95) and MRT written in terms of signal-to-noise ratio a/c (SNR) and ER:

$$\text{MRT} = \left(\frac{c}{a}\right)^2 (1 - 2\text{ER}) \log\left(\frac{1 - \text{ER}}{\text{ER}}\right). \quad (97)$$

References

- Baum, C. and Veeravalli, V. (1994). A sequential procedure for multihypothesis testing. *IEEE Trans. Inform. Theory*, 40:1994–2007.
- Bogacz, R., Brown, E., Moehlis, J., Hu, P., Holmes, P., and Cohen, J. (2004). The physics of optimal decision making: A formal analysis of performance in two-alternative forced choice tasks. *Preprint*.

- Brown, E., Gao, J., Holmes, P., Bogacz, R., Gilzenrat, M., and Cohen, J. (In press, 2005). Simple neural networks that optimize decisions. *Int. J. Bifurcation and Chaos*.
- Dragalin, V., Tartakovsky, A., and Veeravalli, V. (1999). Multihypothesis sequential probability ratio tests, part I: Asymptotic optimality. *IEEE Trans. Inform. Theory*, 45:2448–2461.
- Dragalin, V., Tartakovsky, A., and Veeravalli, V. (2000). Multihypothesis sequential probability ratio tests, part II: Accurate asymptotic expansions for the expected sample size. *IEEE Trans. Inform. Theory*, 46:1366–1383.
- Gardiner, C. (2004). *Handbook of Stochastic Methods*. Berlin: Springer-Verlag, 3rd edition.
- Ghose, G., Yang, T., and Maunsell, J. (2002). Physiological correlates of perceptual learning in monkey v1 and v2. *J. Neurophysiol.*, 87:1867–1888.
- Gold, J. and Shadlen, M. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36:299–308.
- Golubev, G. and Khas'minskii, R. (1983). Sequential testing for several signals in gaussian white noise. *Theory Prob. Appl.*, 28:573–584.
- Grossberg, S. (1988). Nonlinear neural networks: principles, mechanisms, and architectures. *Neural Networks*, 1:17–61.
- Hick, W. (1952). On the rate of gain of information. *Quart. J. Exp. Psych.*, 4:11–26.
- Hopfield, J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Natl. Acad. Sci. USA*, 82:3088–3092.
- Lacouture, Y. and Marley, A. (1995). A mapping model of bow effects in absolute identification. *J. Math. Psych.*, 39:383–395.
- Lacouture, Y. and Marley, A. (2004). Choice and response time processes in the identification and categorization of unidimensional stimuli. *Perception & Psychophysics*, 66(7):1206–1226.
- Lehmann, E. (1959). *Testing Statistical Hypotheses*. New York: Wiley.
- Lorden, G. (1977). Nearly-optimal sequential tests for finitely many parameter values. *Ann. Stat.*, 5(1):1–21.
- Luce, R. (1986). *Response times*. New York: Oxford University Press.
- Luce, R., Nosofsky, R., Green, D., and Smith, A. (1982). The bow and sequential effects in absolute identification. *Perception & Psychophysics*, 32:397–408.
- McMillen, T. and Holmes, P. (2006). The dynamics of choice among multiple alternatives. *J. Math. Psych.*, 50(1):30–57.
- Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psych. Rev.*, 63:81–97.
- Neyman, J. and Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Roy. Soc. London Ser. A.*, 231:289–337.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psych. Rev.*, 85:59–108.

- Ratcliff, R., Segraves, M., and Cherian, A. (2003). A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of simple two-choice decisions. *J. Neurophysiol.*, 90:1392–1407.
- Ratcliff, R., Van Zandt, T., and McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psych. Rev.*, 106:261–300.
- Ricciardi, L. (1977). *Diffusion processes and related topics in biology*. New York: Springer-Verlag.
- Roitman, J. and Shadlen, M. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J. Neurosci*, 22 (1):9475–9489.
- Smith, P. (2000). Stochastic dynamic models of response time and accuracy: A foundational primer. *J. Math. Psych.*, 44:408–463.
- Tartakovsky, A. (1998a). Asymptotically optimal sequential tests for nonhomogeneous processes. *Sequential Anal.*, 17(1):33–61.
- Tartakovsky, A. (1998b). Asymptotically optimality of certain multihypothesis sequential tests: Non-i.i.d case. *Statist. Infer. Stochastic Processes*, 1(3):265–295.
- Usher, M. and McClelland, J. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psych. Rev.*, 108:550–592.
- Usher, M., Olami, Z., and McClelland, J. (2002). Hick’s law in a stochastic race model with speed-accuracy tradeoff. *J. Math. Psych.*, 46:704–715.
- Veeravalli, V. and Baum, C. (1995). Asymptotic efficiency of a sequential multihypothesis test. *IEEE Trans. Inf. Theory*, 41:1994–1997.
- Verdenskaya, N. and Tartakovskii, A. (1991). Asymptotically optimal sequential testing of multiple hypotheses for nonhomogeneous Gaussian processes in an asymmetric situation. *Theory Prob. Appl.*, 36:536–547.
- Wald, A. and Wolfowitz, J. (1948). Optimal character of the sequential probability ratio test. *Ann. Math. Statist.*, 19:326–339.