How fast is neural winner-take-all when deciding between many options?

Birgit Kriener[1,2*], Rishidev Chaudhuri[1*], and Ila R. Fiete[1,3†]

[1]Center for Learning and Memory and Department of Neuroscience, The University of Texas at Austin, Austin, Texas, USA.
[2]Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway.
[3]Department of Physics, The University of Texas at Austin, Austin, Texas, USA.
[*]B.K. and R.C. contributed equally to this work.
[†]Correspondence: ilafiete@mail.clm.utexas.edu (I.R.F.)

Abstract

Identifying the maximal element (**max**, **argmax**) in a set is a core computational element in inference, decision making, optimization, action selection, consensus, and foraging. We show that running sequentially through a list of N fluctuating items takes $N\log(N)$ time to accurately find the **max**, prohibitively slow for large N. The power of computation in the brain is ascribed to its parallelism, yet it is theoretically unclear whether, even on an elemental task like the **max** operation, leaky and noisy neurons can perform a distributed computation that cuts the required time by a factor of N, a benchmark for parallel computation. We show that conventional winner-take-all circuit models fail to realize the parallelism benchmark and worse, in the presence of noise altogether fail to produce a winner when N is large. If, however, neurons are equipped with a second nonlinearity so that weakly active neurons cannot contribute inhibition to the circuit, the network matches the accuracy of the serial strategy but does so N times faster, partially self-adjusting integration time for task difficulty and number of options and saturating the parallelism benchmark without parameter fine-tuning. Finally, in the regime of few choices (small N), the same circuit predicts Hick's law of decision making; thus Hick's law behavior is a symptom of efficient parallel computation. Our work shows that distributed computation that saturates the parallelism benchmark is possible in networks of noisy and finite-memory neurons.

Introduction

Finding the largest entry in a list of N numbers is a basic and ubiquitous computation. It is invoked in a wide range of tasks including inference, optimization, decision

making, action selection, consensus, and foraging [18, 23, 9, 63]. In inference and decoding, finding the best-supported alternative involves identifying the largest likelihood (**max**), then finding the model corresponding to that likelihood (**argmax**); decision making, action selection and foraging involve determining and selecting the most desirable alternative (option, move, or food source, respectively) according to some metric, again requiring **max**, **argmax** operations.

Because **max**, **argmax** are basic building blocks in these myriad computations, it is important to characterize how long these operations take. The results on the time-complexity of **max**, **argmax** in an optimal serial procedure, as would be carried out a computer, are simple: Finding the largest number or its index in an unsorted list of N elements involves a sequential loop through the list, with one comparison per adjacent pair; the computational complexity is linear in N. We will refer to this as the *serial scaling* of **max**, **argmax**. If each element is observed along with some noise, the computational complexity of solving the task with a desired level of accuracy increases to $N \log(N)$, as we will show (assuming that the gap in the mean value of the top input and the rest remains fixed as N varies).

It is hypothesized that a major source of efficiency of computation in neural systems is the potential for massive parallelism: a computation that would take a long time to perform serially can be distributed across neurons in a circuit containing many (N in the range of tens of thousands) neurons, for a speed-up of a factor N. However, for the benefits of parallelism to hold, the brain must extract the final information it requires from across the N neurons in a time that is independent of, or at most very weakly dependent on, N. In this paper, we will refer to a factor-N speed-up relative to the serial strategy as the *parallelism benchmark*.

There are two distinct regimes in which it is interesting to consider how (fast) the brain computes **max**, **argmax**: The first is finding the most active neuron across thousands of neurons or neuron pools. An example of this large-N **max**, **argmax** computation in the brain is the dynamics that lead to the sparsification of Kenyon cell activity within the fly mushroom bodies [63]. It is possible that many more areas with strong recurrent inhibition and gap-junction coupled interneurons display similar dynamics, including the vertebrate olfactory bulb [60, 52], hippocampal area CA1 [1, 19, 62], and basal ganglia [50, 55]. The second is the problem of explicit decision making across a small number of externally presented alternatives.

In the first case, our goal is to understand whether it is possible for neural circuits to achieve the gains of parallelism, in the context of the elemental computations of **max**, **argmax**. In the second case, our goal is to refine our understanding of how circuits in the brain perform multi-choice decision making, by generating predictions about decision time from dynamical and biologically plausible models of neural networks and comparing these with findings from psychophysics experiments. Across both cases, we seek a more unified understanding of neural circuit computation of **max**, **argmax**, whether the options being considered number in the thousands, as in the microscopic states of neurons, or ≤ 10, as in explicit decision-making among externally presented options in psychophysical tasks.

Not surprisingly, because of the importance of **max**, **argmax** operations, they are well-studied in neuroscience in the guise of *winner-take-all* (WTA) neural circuit models and phenomenological *accumulate-to-bound* (AB) models. A WTA network

consists of N neurons, each driven by an external input, each amplifying its own state, and all interacting competitively through global inhibition. Self-amplification and lateral inhibition result, under the right conditions, in a final state in which only the neuron with the largest integrated input (**max**) remains active, while the rest are silenced [24, 11, 41] (i.e., here we do not consider $K-$winners-take-all with $K > 1$ [42]). If the activation level of the winner is moreover proportional to the size of its input [73], the network also solves the **argmax** problem. The final state of the network is the completed output of the computation.

By contrast, AB models [33, 68, 6, 46, 53] consist of individual integrators that sum their inputs and increase their outputs in proportion. In contrast to WTA models, AB models require a separate downstream readout that applies a threshold across the integrators to determine which is largest. Thus they do not, by themselves, output an answer to the **argmax** and **max** problems. More importantly – unlike WTA models, which consist of a network of leaky, interacting neurons – AB models are phenomenological, not neural. For these reasons, our focus is on WTA networks.

Despite the elemental nature of WTA computation in neural circuits, and the extensive literature on the topic, the time-complexity of WTA – how long it takes a system to compute **argmax** and **max** from a set of N inputs, as a function of N – is not well-characterized for recurrent continuous-time neural systems with noisy inputs (see Discussion for background and related work). On the one hand, one might expect that the parallel architecture of neural networks could speed up the computation – trading temporal complexity for space. On the other hand, the neural elements are leaky – hardly ideal parallel processors or integrators – and their nonlinear thresholds combined with noise could discard information relevant to computation. Thus, it is unclear whether parallel processing with such elements can manage the tradeoff efficiently, reducing temporal complexity by an amount in proportion to the increase in spatial complexity and thus achieving the parallelism benchmark.

At the same time, there is an important body of human psychophysics literature on the speed of multi-alternative decision making as a function of the number of options [47, 27, 67], showing that at high accuracy, the duration of human decision-making increases with the number of options as $\log(N)$ [27, 68, 69, 6, 46] – a result known as Hick's law. Theoretical works reproduce Hick's law starting from different frameworks [68, 69, 6, 46], but what is missing is an examination of whether a self-terminating network model with continuous-time dynamics, leaky neurons and noisy inputs, that reports on the results of its own computation, is consistent with Hick's law.

Here, we show that for constant inputs conventional neural WTA networks achieve the parallelism benchmark for strong, but not weak inhibition. However, when the inputs are noisy, conventional WTA networks with strong inhibition altogether fail to exhibit WTA behavior for large N. Making inhibition weak and exquisitely fine-tuning the weights rescues WTA behavior, but yields suboptimal parallelism gains, together with an overly conservative accuracy tending toward zero error and thus a failure to exhibit a speed-accuracy tradeoff.

We introduce a modified form of neural network WTA dynamics, nWTA dynamics, in which inhibition is strong, but only sufficiently active neurons contribute inhibition to the circuit. These nWTA networks can trade time for space efficiently,

fully saturating the parallelism benchmark by producing a factor-N speed-up relative to the serial strategy for both constant and noisy inputs – all without any fine-tuning of network parameters. The independence of decision time T from N means that neural circuits may be able to infer the maximum from a large pool of individually competing neurons in real-time, suggesting that it might indeed be possible for neural circuits to perform and exploit truly parallel computation. We show that these networks are self-adjusting for task difficulty, integrating for longer when the inputs are noisier or when the gap between the top option and the rest is smaller.

Finally, for psychophysics decision problems with a few ($N < 10$) external options, we show that the networks exhibit a speed-accuracy tradeoff. In particular, we find $T \sim \log(N)$ at fixed accuracy, and hence a reward rate that decreases as $\log(N)$, in accord with achieving the parallelism benchmark, and consistent with Hick's law. Hick's law may thus be viewed as a behavioral signature of a neural computation that saturates the gains of parallel computation. We generate predictions about accuracy at fixed T and remark on the shapes of the responses of single neurons during the evidence integration period. Interestingly, the networks can achieve near-optimal performance across different N with fixed parameters, showing how decision making circuits could solve multi-alternative decision tasks across varying numbers of options, with little or no parameter re-tuning. The 2AFC task, standard in both psychophysics and modeling efforts, is too simple or underconstrained to be fully diagnostic of more general decision making dynamics in a circuit. Our work provides a set of expectations to compare with neural recordings and behavior when generalizing to multi-AFC tasks.

Results

Consider a network of N neurons (or neuron pools), whose states are described by their outgoing synaptic activations $x_i(t)$ or firing rates $r_i(t)$, $i \in \{1, \ldots, N\}$. The neurons receive inputs $b_1 = b_2 + \Delta > b_2 \geq b_3 \ldots \geq b_N$ respectively, and interact through self-excitation (strength α) and mutual inhibition (strength β), Figure 1a:

$$\tau \frac{dx_i}{dt} + x_i = \left[b_i + \alpha x_i - \beta \sum_{j(j \neq i)} x_j \right]_+ \equiv \quad r_i \,. \tag{1}$$

Here, $[\cdot]_+ = \max[0, \cdot]$ is a rectification nonlinearity. For appropriate values of self-excitation, inhibition and inputs, the network exhibits winner-take-all dynamics with a unique winner ([73] and S1.1). These dynamics can be understood as movement downhill on an energy landscape, which drives the network to one of N possible stable states, each corresponding to solo activation of a different neuron (Figure S1a).

Our goal is to understand how WTA dynamics behave as a function of network size N, and in particular, to examine how fast the network can pool information across neurons to arrive at a single winner. We call this duration the decision time T_{WTA} of the network.

In all that follows, we will assume that the gap Δ between the largest and next-largest input is held fixed as N varies. In the *quasi-2D input* case, the remaining

inputs are equal to each other ($\mathbf{b} = (b + \Delta, b, \ldots, b)^\top$ and $b, \Delta > 0$); in the *uniform input* case, the remaining inputs are uniformly distributed ($b_1 = b + \Delta, b_2 = b$, and $b_i = \mathrm{U}[0, b]$ for $i \geq 3$). We will begin by briefly considering the case in which inputs are constant, then move to the more natural setting where inputs fluctuate about their means over time.

(In the SI, we consider the case where the gap shrinks as N grows, as $1/N$ (all inputs are drawn uniformly, $b_i \sim \mathrm{U}[0, 1]$; see Figure S1g,h and S1.4 for deterministic WTA, and Figure S1i–q and S1.8 for noisy WTA).)

Noise-free max

The decision time T_{WTA} is the time taken before the firing rate $r_i(t)$ of the last losing competitor drops to zero (Figure 1b), or in other words, when the activations $\mathbf{x}$ of all neurons but the winner are decaying exponentially to zero while the winner approaches its asymptotic state $x_1^\infty = b_1/(1 - \alpha)$ (Figure 1c and S1.2).

Weak inhibition: Linear growth in T_{WTA} with network size The total inhibition in Equation (1) grows with the number of neurons. A reasonable possibility is thus to scale the inhibitory interaction strengths as $\beta = \beta_0/N$, where β_0 is some constant independent of N. We call this "weak" inhibition.

The strength of self-excitation (α) must then be set to maintain stability and to assure a WTA state. Setting $\alpha < 1$ guarantees that the WTA activation states remain bounded; further, if $1 - \beta < \alpha < 1$, there will be a unique winner ([73], S1.1). The two-sided constraint $1 - \beta_0/N < \alpha < 1$ is a *fine-tuning* condition: excitation must be within $1/N$ of 1, with the allowed range shrinking to zero width as N grows.

For quasi-2D inputs, all $N - 1$ neurons with input b exhibit identical dynamics; hence the name of this input condition. The network converges to the correct solution, where the neuron with input $b + \Delta$ is the winner. The equations can be solved analytically (S1.2):

$$T_{\mathrm{WTA}} \overset{N \gg 1}{=} 2N \log\left[1 + \frac{b}{2\Delta}\right]. \tag{2}$$

Though the problem is effectively two-dimensional, T_{WTA} grows linearly with N, Figure 1d, the same scaling as the serial strategy. Interestingly, in contrast to the serial strategy, T_{WTA} depends on the gap Δ, growing logarithmically as Δ shrinks, Figure 1f and Equation (2).

If inputs are drawn uniformly after holding the top gap fixed, the results remain unchanged in their N-scaling, Figure 1d. In fact, the scaling of T_{WTA} is practically insensitive to the statistics of the inputs beyond the top gap (S1.3, Figure S1e,f).

The decision time T_{WTA} grows linearly with N because the initial total inhibition at each neuron is $O(1)$ and roughly cancelled by the excitatory drive. The eventual winner and losers are thus somewhat isolated from each other, individually integrating their input drives with the slow network time-constant $\sim \tau/(1 - (\alpha + \beta)) \sim N\tau$, until the losers and eventual winner finally separate enough that the nonlinear portion of WTA dynamics pushes them to their steady-state activations in a time given by Equation (2) (see S1.2).

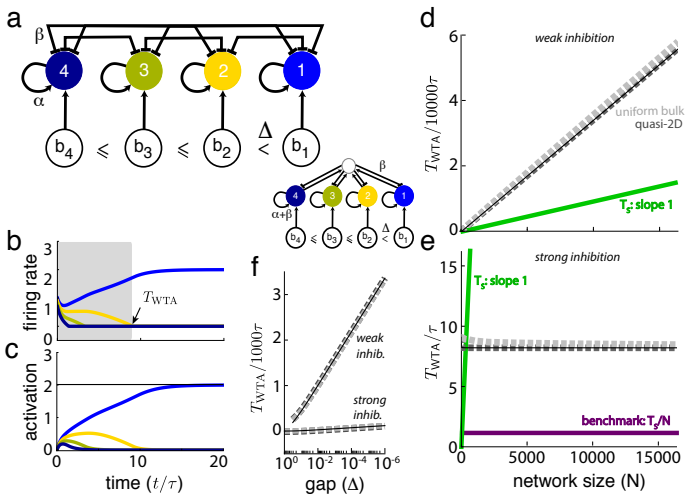

Figure 1: **WTA with non-noisy inputs: weak and strong inhibition.** a) Network schematic: Each neuron (pool), ordered by the size of its external inputs (with gap $\Delta \equiv b_1 - b_2$), is inhibited by the others and excited by itself. Inset: Mathematically equivalent network with a global inhibitory neuron, requiring only N synapses compared to the N^2 synapses for mutual inhibition as in the main schematic. b-c) Neural firing rates and activation (coloring as in a)): The competitive phase (gray area) is over when the firing rate of the second-most-active neuron drops to zero (arrow). The duration of the competitive phase is the decision time T_{WTA}. d) Decision time T_{WTA} versus N for weak inhibition (quasi-2D and uniform inputs: dashed dark and light gray curves), with analytical prediction of Equation (3) (black line). Decision time for the serial strategy (solid green) is also linear in N. The time axis is normalized by 10000τ, where τ refers (here and in subsequent figures) to the single-neuron time-constant for WTA curves and to the duration of a single time-step taken to read an entry in a list for the serial curves. [$b = 0.9$; $\Delta = 0.1$; $\alpha = 0.5$; $\beta = 0.6$] e) T_{WTA} versus N for strong inhibition (same color codes as above). The parallelism benchmark, given by the serial time T_{S} divided by N is shown in purple. Parameters as in d). f) T_{WTA} grows logarithmically as the gap Δ shrinks (upper/lower set of curves: weak/strong inhibition, same color codes as in d)).

In summary, the existence of WTA in a weak-inhibition circuit with constant inputs requires exquisite fine-tuning of excitation. The decision time T_{WTA} increases linearly with the number N of inputs and neurons, independent of the statistics of the input beyond the top gap (and cannot be adjusted for a speed-accuracy tradeoff), exhibiting no gains in speed from parallelism.

Strong inhibition: T_{WTA} can be independent of N An alternate choice is to hold β, the strength of inhibition contributed by each neuron, fixed as N is varied. Total inhibition then grows with N. A unique WTA solution exists for any choice of α in the interval $(1 - \beta, 1]$. Unlike in the weak inhibition case, α need not be fine-tuned because the interval of permissible values does not shrink with N.

For quasi-2D inputs, we analytically obtain:

$$T_{\mathrm{WTA}} \overset{N\gg 1}{\sim} \log\left[\frac{b}{\Delta}\right] . \tag{3}$$

As for the case with weak inhibition, T_{WTA} depends on Δ. Notably, however, T_{WTA} is asymptotically independent of N, Figure 1e (simulations dashed dark gray, Equation (3) solid black), meeting the parallelism benchmark of a factor-N speed-up compared to the serial strategy.

To address whether the N-independent decision time holds beyond the quasi-2D case, we test the fully N-dimensional case of uniform inputs. Even then T_{WTA} is determined essentially by Δ (see S1.3, Figure S1d for details), and most of all asymptotically independent of N (Figure 1e, simulations in light gray).

In sum, WTA networks of leaky neurons with strong inhibition can fully and efficiently trade space for time, permitting a factor-N speed-up in computing **max**, **argmax** relative to serial strategies when the input is non-noisy.

Noisy max

We turn now to the more realistic scenario central to the rest of this work: solving a noisy version of the **max**, **argmax** problems. Suppose the inputs to the decision circuit are noisy, fluctuating over time about their mean values ($\tilde{b}_i(t) = b_i + \eta_i(t)$, where b_i is the fixed mean and $\eta_i(t)$ are zero-mean fluctuations, see Methods). This noise may be attributed to noise in the inputs or to stochastic activity in neurons of the decision circuit, or both. (The results on the existence of WTA states and convergence time are agnostic to the source of noise. For direct comparisons of accuracy with non-neural benchmarks, however, the noise should be interpreted as being in the inputs. A neural decision circuit with additional intrinsic noise would also follow the same results, but at a correspondingly higher total noise variance.) The goal is to identify the input with the largest true mean. Obtaining a correct answer involves collecting information for long enough to gain a good estimate of the mean values of each input, and then performing a **max** operation on the estimated means.

Any strategy with a finite decision time will have a non-zero error probability on noisy **max** and we expect a *speed-accuracy tradeoff* where efficient solutions involve setting an acceptable error probability then finding the fastest way to make a decision, or setting the observation time T and finding a way to make a maximally accurate decision within T. Clearly, the appropriate averaging time and decision error will depend on the ambiguity or separation of the inputs from each other and on the amplitude of noise. As we noted earlier — even in the deterministic setting where gap size was computationally irrelevant — the neural WTA dynamics exhibited an inherent dependence on the gap size, or more precisely on the signal-to-noise ratio Δ/b, suggesting that neural dynamics may be naturally suited to solving the noisy **max** problem.

We next characterize the time-complexity of noisy **max**, **argmax** in a serial framework and from it characterize the parallelism benchmark, then turn to neural WTA solutions to the same problem.

Serial strategy Consider the set of $N-1$ summed differences $\delta_i^T = \sum_{t=1}^{T}(\tilde{b}_1(t) - \tilde{b}_i(t))$ between the fluctuating highest-mean input, and each of the others. The probability that the wrong element is selected by **argmax** after averaging for time T is bounded by the sum of the probabilities that any of these individual δ_i's is greater than zero. The quantities δ_i concentrate around the true gaps ($\Delta_i = b_1 - b_i$), with the probability of error-inducing fluctuations about the mean decaying as $e^{-\Delta_i^2 T}$ across a wide set of possible input distributions. The waiting time $T \sim \log(N)$ depresses individual error probabilities so they scale as $1/N$, keeping the total error probability constant as N is varied (see Figure 2f, 3f for traces and S3.1 for a more detailed analysis). Thus, the time for a serial strategy to achieve a constant decision accuracy across N noisy inputs with top gap Δ is $T_S \sim N\log(N)/\Delta^2$.

We can instead consider how accuracy varies with N if the averaging time T is held fixed. The accuracy will decline as N increases, because there are more non-top inputs that could transiently appear to be larger than the top input. The error probability can be computed using results on the distributions of extreme order statistics (S3.2). For Gaussian noise and a fixed top gap, the error probability for fixed T is a sigmoidal function of log(N), with a power-law dependence on N for large N (S3.2).

WTA networks A network that converges to a unique WTA state with non-noisy inputs and non-noisy internal dynamics need not do the same when driven by noise. Noisy inputs or internal noise kick the state around and the system generally cannot remain at a single point. Nevertheless, there can still be a sense in which the noise-driven network state flows toward and remains in the neighborhood of a fixed point in the corresponding deterministic system (Figure S1a,b; S1.5, Figure S2a–d). We will refer to such behavior in the noise-driven WTA networks as successful WTA dynamics, with the neighborhood defined in terms of one neuron reaching a criterion distance from the deterministic WTA high-activity attractor (set by the dynamical system, not an external threshold) while the rest are strongly suppressed (Methods). We then examine the existence and decision time of WTA dynamics in neural networks with strong and weak inhibition.

Strong inhibition: breakdown of WTA dynamics The parallelism benchmark was previously attained (in the non-noisy case) with strong inhibition, thus we begin there. For a given N, a network with strong inhibition can evolve to a WTA state (according to our criterion), if the noise amplitude is sufficiently small relative to the top gap, Figure 2a. However, as N grows (while holding the gap and noise amplitude fixed; also see S1.8 and Figure S1i–q for a gap that shrinks as $\Delta \sim 1/N$), the network entirely fails to reach a WTA state, Figure 2b. Equivalently, if N and the gap are fixed, WTA breaks down as the noise amplitude is increased, Figure 2c. This failure is to be distinguished from an error: The network does not select the wrong winner, it simply fails to arrive at any winner.

We can understand the failure as follows. Unbiased (zero-mean) noise in the inputs, when thresholded, produces a biasing effect: Neurons receiving below-zero mean input will nevertheless exhibit non-zero mean activity because of input fluctuations (Figure 2a). Thus, even neurons with input smaller than their deterministic thresh-

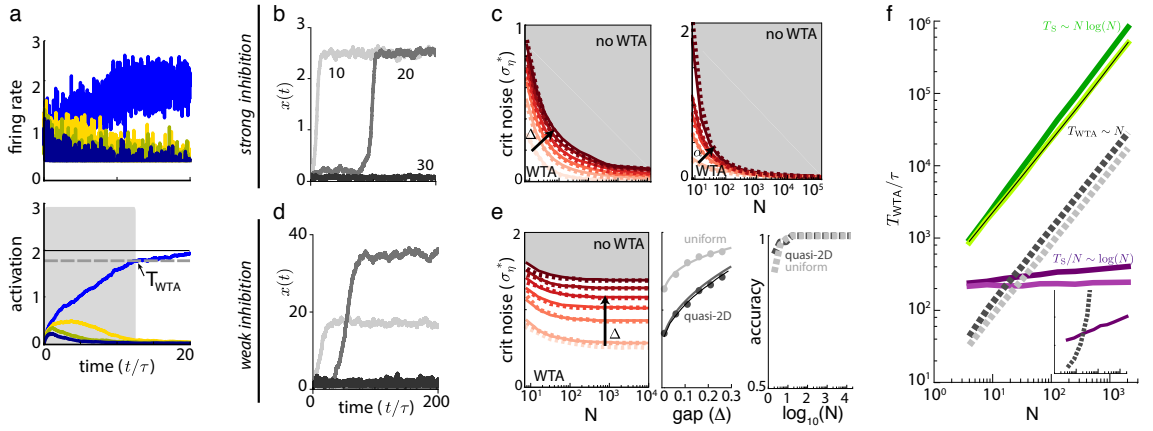

Figure 2: **WTA with noisy inputs: strong and weak inhibition** a) Firing rates (top) and activations (bottom) of neurons with noisy inputs. Dashed gray line: The convergence criterion for the top neuron to be declared a winner (defined in text); black line: activation of this neuron if it were the winner when the network was run without noise. $[b_i = \{1, 0.9, 0.8, 0.7\}$; $\sigma_\eta = 0.6$; $\alpha = 0.5$; $\beta = 0.6]$ (b-c) Results from network with strong inhibition. b) Activity dynamics of the most-active neuron for networks of size $N = 10, 20, 30$ (light to dark gray), respectively. $[\alpha = 0.6$; $\beta = 1$; $\Delta = 0.1$; $\sigma_\eta = 0.35]$ c) Critical noise amplitude versus N: WTA dynamics exists below a given curve and fails above it (dashed: numerical simulation; solid: analytical). Darker curves correspond to a widening top gap (left; $\Delta = \{0.01, 0.06, \ldots, 0.26\}$; $\alpha = 0.6$) or more self-excitation (right; $\alpha = \{0.1, 0.6, 0.9\}$; $\Delta = 0.1$). $[\beta = 1$ in all curves.] (d-g) Results from network with weak inhibition. d) Same as b). e) Left: Same as c). Middle: Variation of critical noise amplitude with the top gap Δ for quasi-2D (black) and uniform (gray) input drives. Dots: simulation; thick lines: best fit (dark gray: $f(\Delta) \sim \sqrt{\Delta}$, light gray: $f(\Delta) \sim \Delta\log(\Delta)$; thin black line: theoretical prediction. Right: Average accuracy as function of N. f) Decision time of the network (gray; colors as in Figure 1d-f), the serial strategy (dark green: quasi-2D; light green: uniform, black line: theory), and the parallelism benchmark (T_S/N; purple shades). Inset: the same curve on a semi-log scale, to make apparent the different scalings of the benchmark ($\log(N)$) and the faster growth of T_{WTA}. $[\tau_\eta = 0.005\tau$; $b = 1 - \Delta]$

olds continue to contribute an inhibition term with strength of $O(1)$ to the circuit. The total inhibition in the circuit remains of order N over time, and increases with noise amplitude, preventing any neuron from breaking away from the rest to become a winner for sufficiently large N. This problem holds for both the quasi-2D and uniform input cases.

The breakdown of the WTA-regime coincides analytically, for the quasi-2D case, with the onset of stability of a non-WTA branch of self-consistent solutions in the coupled dynamics of the noisy version of Equation (1) (see Methods, Equation (6) and S1.5).

The resulting predictions for critical N and noise amplitude at WTA breakdown closely match numerical simulation results (Figure S2a–c) and can be used to determine the feasibility of WTA computation in large networks in the presence of

noise. Qualitatively, the uniform input case exhibits similar breakdowns, but at higher noise levels, because fewer neurons contribute noise-fluctuations above the activation threshold (not shown).

The parameter regime for WTA states is shown in Figure 2c: WTA solutions exist below a given curve, and break down above it (dashed lines: numerical simulation; solid lines: analytical results). Increasing self-excitation or the gap expands the WTA regime (Figure 2c, left and right). Nevertheless, at any self-excitation strength and gap size, the WTA regime shrinks rapidly as N grows. In the limit $N \to \infty$ our numerical results suggest that WTA will fail at any finite noise level.

In summary, strong inhibition networks, which met the parallelism benchmark when the inputs were deterministic, are not capable of finding a winner in large networks with even slightly noisy inputs.

Weak inhibition: accurate but slow WTA after extreme fine-tuning The weak inhibition regime ($\beta \sim 1/N$) might permit WTA behavior in the noisy case, because the excess inhibition that prevented strongly inhibiting large networks from reaching a WTA state is greatly reduced. As before, we set the parameters to be: $1 - \beta < \alpha < 1$ and $\beta = \beta_0/N$, again requiring exceedingly fine tuning.

The weak inhibition network is capable of WTA-like dynamics for sufficiently small noise and N (Figure 2d), and as in the strong inhibition network, the breakdown of the WTA regime with quasi-2D inputs can be predicted analytically (S1.5, Figure S2b). The key difference is that the WTA regime persists for finite noise levels even for $N \to \infty$, Figure 2e (left; simulations: dashed, analytics: solid lines). The noise level up to which the network exhibits WTA dynamics can be substantially larger than the gap (Figure 2e, left, middle) and in fact the network exhibits WTA behavior even for $\Delta = 0$, (lightest line in Figure 2e, left), selecting a random neuron as the winner.

Weakly inhibiting networks continue to exhibit WTA dynamics for large N because the total amount of inhibition at each neuron remains roughly independent of N ($\beta \sim 1/N$ cancels $\sim N$-fold inhibitory contribution), while simultaneously, α increases towards 1 from fine-tuning as N grows. As a result, inhibition does not swamp self-excitation even in asymptotically large networks, and a neuron can break free to win the competition. As before, the WTA regime (Figure 2e, middle) is larger and T_{WTA} smaller (Figure 2f) in the uniform input case compared to the quasi-2D case. In general, the quasi-2D case serves as an upper bound for T_{WTA} and as a lower bound on the critical noise amplitude (σ_η^*) in noisy WTA; thus, to be conservative, we will only show results for quasi-2D in what follows.

WTA accuracy and speed with weak inhibition The choice of winner is random when $\Delta = 0$. Even for $\Delta > 0$ identifying the correct **max**, **argmax** element becomes harder with increasing N as more elements are competing, and noise fluctuations might permit any of these competitors to win, if the network does not spend enough time averaging its inputs. The question is where the network falls on the tradeoff between accuracy and speed, and whether the tradeoff is controllable and efficient.

First, note that wherever the weak inhibition WTA network falls in the tradeoff, it is not controllable because it has no free parameters: β scales as $1/N$ by definition, and α is then automatically determined through the fine-tuning condition. As N is varied, the network therefore exhibits an inherent (non-adjustable) scaling of accuracy and convergence time, which we characterize next.

Interestingly, the network exhibits perfect asymptotic accuracy in its WTA computations (the network size above which perfect accuracy is obtained depends on the exact choice of Δ and σ_η; for parameters in Figure 2e (right), accuracy ~ 1 for $N \gtrsim 10$), which suggests that the network must be averaging over longer times than necessary, favoring accuracy over speed at all N.

Conditioned on the existence of WTA dynamics, the decision time with fluctuating inputs (Figure 2f) exhibits the same linear scaling as when the inputs are constant (Figure 1d). Convergence is thus not qualitatively slowed by the noisy dynamics (see Figure S1h,k for similar results in networks with a shrinking gap).

To understand why averaging over a time $\sim N$ is sufficient to obtain essentially perfect accuracy, note that the serial strategy for fixed accuracy has time-complexity $N\log(N)$ and the parallelism benchmark is $\sim \log(N)$. The weak inhibition network therefore takes $N/\log(N)$ times longer than strictly required for a fixed, imperfect accuracy. This excess time produces a computation with near-perfect accuracy.

Mechanistically, the increase in the network's decision time results from the growth of the time-constant of the network's WTA mode: As N increases, the positive self-feedback term, $\alpha + \beta$, approaches 1 from above as $1 + 1/N(1-K)$ (where $\beta = 1/N, \alpha = 1 - K/N$ for some $K < 1$) and the time-constant of the WTA mode, $\tau/(\alpha + \beta - 1) \sim \tau N(1-K)$, therefore grows linearly with N, strongly filtering the noisy input fluctuations.

In sum, conventional WTA networks, described by Equation (1), can only select a winner from among a large number of noisy options when inhibition is weak and excitation is extremely fine-tuned. In that case, they exhibit a decision time of $T_{\mathrm{WTA}} \sim N$ for a fixed top gap Δ. This represents a modest speed-up of a factor $\log(N)$ relative to the serial strategy for noisy inputs, but does not come close to the factor of N speed-up desired for an efficient parallel strategy.

This result is pessimistic, and raises the question of whether networks of forgetful neurons can ever implement parallel computation that is efficient, fully trading serial time for space. In the next section, we find an affirmative answer to the question, under a modified model for winner-take-all neural computation.

The nWTA network: fast, robust WTA with noisy inputs and an inhibitory threshold

We motivate the construction of a new model for neural WTA from the successes and failings of the existing models. As we have seen, large networks with weak inhibition and fine-tuning can perform WTA computation on noisy inputs and are accurate, but too slow, because inhibition is not strong enough to enforce a rapid separation between winner and losers. Networks with strong inhibition achieve WTA with a full parallelism speed-up for constant inputs, but they entirely fail to perform (accurate or inaccurate) noisy WTA for large N, because most of the nearly-losing neurons

continue to be weakly noise-driven and contribute an amount of inhibition to the circuit that prevents any neuron from becoming a breakaway winner. (In Figure S2e we show that a simple upshift of the activation threshold from 0 to $\theta_{\text{act}} > 0$ for all neurons does not fix the failure of WTA dynamics for large N). In addition, the residual noise-driven inhibition also decreases the average asymptotic activity of the near-winner, thus the true value of **max** will be underestimated.

Ideally, a WTA network would express strong competitive inhibition early and weak inhibition later, so that the network can initially compare the different inputs while allowing the top neuron to later take off unimpeded. Thus, we consider a model where neurons contribute strong inhibition, but an individual neuron can only do so, if its activation level exceeds a threshold θ. Effectively, the linear sum over activations in the inhibitory term of Equation (1) is replaced by a set of individually thresholded terms (see Discussion for biological candidates for this inhibition):

$$\tau \frac{dx_i}{dt} + x_i = \left[b_i + \eta_i + \alpha x_i - \beta \sum_{j \neq i} [x_j - \theta]_+ \right]_+ . \tag{4}$$

The inhibitory threshold θ is set in the range $0 < \theta < b_1/(1-\alpha)$ (no fine tuning), and the strength of self-excitation is in the range $1 - \beta < \alpha < 1$ (untuned). In this new WTA network construction, the expected asymptotic state of the winning neuron, if the ith neuron wins, is $b_i/(1-\alpha)$: the proportionality to the true **max** is recovered.

In this *nonlinear-inhibition WTA* (nWTA) network, every neuron contributes an inhibition of strength ~ 1 when it is highly active (above threshold θ), ensuring robust competition. However, the threshold on inhibitory contributions causes neurons with decreasing activations to effectively drop out of the circuit when their activity level is sufficiently low. The diminishing inhibition in the circuit over time permits the leading neuron to break away and win, Figure 3a. The losing neurons continue to receive inhibitory drive from the remaining highly active neuron(s), which suppresses their activations. This network exhibits WTA states well into the noisy regime, and for asymptotically many neurons, Figure 3b, without fine-tuning. Interestingly, the maximal number of neurons that actively provide inhibition at a given time is small and depends only very weakly on N, which partly explains why the inhibitory threshold need not be tuned when N is varied (see S1.5 and Figure S2f,g for details).

Speed and accuracy of nWTA dynamics The second nonlinearity in the nWTA dynamics makes it difficult to analytically evaluate the model's behavior. Nevertheless, we can obtain a good estimate of its behavior and of quantities like the decision time and critical noise amplitude from simulation.

When the nWTA network is presented with constant inputs, it again meets the parallelism benchmark, converging in a factor of N less time than the serial strategy, similar to conventional WTA networks with strong inhibition. Thus, the second nonlinearity does not degrade performance on the deterministic problem (not shown).

The network exhibits a broad tradeoff between speed and accuracy, Figure 3d (top curves: lower noise, bottom curves: higher noise; see also S1.6, Figure S3a–g), in contrast with the weakly inhibiting conventional WTA circuit. Starting at high accuracy and holding noise amplitude fixed, the accuracy of computation can be decreased, and

speed increased, by increasing α (for fixed β; darker gray circles along a curve correspond to increasing α). Alternatively, it is possible to move along the speed-accuracy tradeoff by varying the strength of β while holding α fixed (Figure S3g,h): speed increases and accuracy decreases with increasing inhibition. The overall integration time of the network is generally set by the combination of α and β, with high accuracy and low speed achieved as $\alpha + \beta$ approaches 1. When $\alpha + \beta$ are increased away from 1, speed increases and accuracy decreases. Conveniently therefore, a top-down neuromodulatory or synaptic drive can control where the network lies on the speed-accuracy curves, with many mechanistic knobs for control, including synaptic gain control of all (excitatory and inhibitory) synapses together (resulting in covariation of α, β), neural gain control of principal cells (also resulting in effective covariation of α, β), or a threshold control of inhibitory cells (modulation of inhibition).

The nWTA network exhibits an interesting non-monotonic dependence of accuracy on noise level: except for very small N, the accuracy minimum occurs not at the highest but intermediate noise-levels, Figure 3c (horizontal slices, middle panel; see also S3j). This improvement in performance at some higher noise-level is a form of stochastic resonance [45]. The left-half of the stochastic resonance effect, that accuracy declines as noise increases, is easily understood. The improvement in performance as noise continues to increase, however, runs counter to intuition and requires explanation. We find that large noise effectively extends the network's integration window, thus allowing it more time to average the noisy inputs and arrive at a correct decision (S1.7, Figure S3i–n; also see below, Multi-alternative forced-choice decision making). Consequently, the outcome of the computation more frequently reflects the largest mean input. By contrast, conventional WTA networks, which converge only when inhibition is weak, integrate for a sub-optimally long time and produce such accurate results across noise levels that stochastic resonance is either not visible or highly marginal (data not shown).

(There is a different non-monotonic effect in the speed-accuracy curve at the lowest accuracies: further increasing self-excitation produces decreasing accuracy, as discussed above, but also decreasing speed. This happens because the asymptotic firing rate of the winning neuron, which grows as $1/(1-\alpha)$, diverges as α approaches 1, and the network thus takes increasingly long to converge to this diverging steady-state value. In short, at low accuracy with α approaching 1, the network rapidly makes an "effective" low-accuracy decision but then actually converges to its steady-state value increasingly slowly.)

For a fixed amount of noise per input or neuron, the decision time to reach a fixed accuracy scales as $T_{\mathrm{WTA}} \sim \log(N)$ (Figure 3f; also see Figure S3e,h), compared to the serial time-complexity of $T_{\mathrm{S}} \sim N \log(N)$. The nWTA network therefore achieves a fully efficient tradeoff of space for time, matching the parallelism benchmark of T_{S}/N at fixed accuracy, for noisy inputs.

Not only does the *scaling* of decision time with N in the nWTA network match the functional form of the parallelism benchmark, the prefactor is nearly optimal too: it takes only a factor of 2-3 more time steps (in units of the biophysical time-constant of single neurons) to converge than the parallelism benchmark (Figure 3f, black vs purple curves).

On the other hand, if we evaluate accuracy at fixed T_{WTA} in Figure 3d (here at

$T_{\mathrm{WTA}} = 28\tau$), and compare performance with the parallelism benchmark at some fixed $T_S/N = kT_{\mathrm{WTA}}$ (here, $k = 0.34$), we see that accuracy at large N is almost identical to that of the parallelism benchmark, again given just a two to three-fold absolute increase in decision time (Figure 3e).

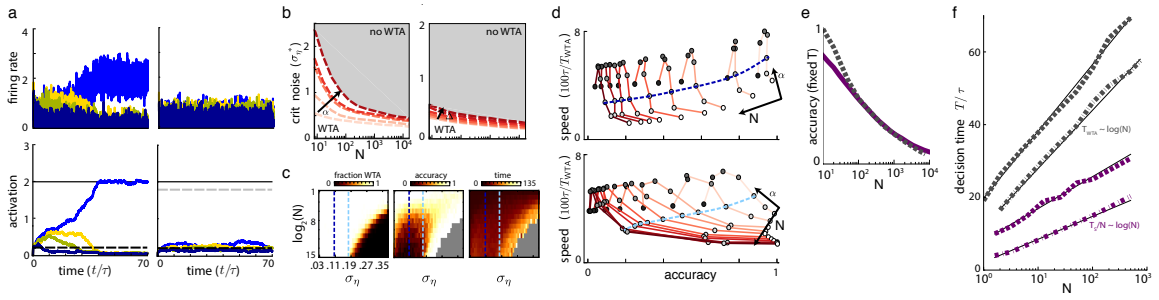

Figure 3: **WTA dynamics for networks with strong nonlinear inhibition is robust to finite noise** a) Firing rates (upper panels) and activations (lower panels) in noisy WTA networks with strong inhibition. Left column: network with additional nonlinear threshold on the ability of neurons to contribute inhibition to the network (threshold depicted as black dashed line) Right column: conventional WTA network without inhibitory threshold can fail to exhibit WTA dynamics (gray dashed line as in Figure 2a). [$\theta = 0.2$; $\alpha = 0.5$; $\beta = 0.6$] b) Critical noise amplitude σ_η^* as a function of N for varying $\alpha = \{0.5, 0.7, 0.9, 1.1, 1.5\}$, $\Delta = 0.1$ (left) and $\Delta = \{0, 0.0125, 0.05, 0.075, 0.1, 0.15\}$, $\alpha = 0.6$ (right). Below each curve, WTA behavior exists, while above it does not. c) Heatmaps showing fraction trials with a WTA solution (single winner; left), accuracy of the WTA solution (middle) and convergence time T_{WTA}/τ (right) as function of network size N and noise amplitude σ_η. Dashed lines denote $\sigma_\eta = 0.1$ (dark blue) and $\sigma_\eta = 0.17$ (light blue), which are the noise amplitudes used in the upper or lower panels of (d), respectively. d) Speed-accuracy curves for $\sigma_\eta = 0.1$ (upper) and $\sigma_\eta = 0.17$ (lower panel) for varying $N = \{10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 7500, 10000\}$ (light to dark red) and $\alpha = \{0.42, 0.45, 0.5, 0.6, 0.7, 0.8, 0.9\}$ (light to dark gray circles). Dashed lines indicate $\alpha = 0.5$ used in (c). Only trials that produced a WTA solution were included. e) N-scaling of accuracy at fixed decision time. Gray dashed line: WTA dynamics; purple: parallelism benchmark. f) N-scaling of decision time at fixed accuracy for WTA (dashed gray curves) and the parallelism benchmark (purple) for accuracy 0.6 (dash-dotted) and 0.8 (dashed). Solid lines are logarithmic fits [$\alpha = 0.5$; $\beta = 0.6$; $\Delta = 0.05$; $\tau_\eta = 0.05\tau$; $\theta = 0.2$; $\sigma_\eta = 0.2$; see S1.7, Figure S3a–h for similar results with different parameters and noise levels].

In summary, the nWTA network can perform the **max**, **argmax** operations on noisy inputs with comparable accuracy as the optimal serial strategy, but with a full factor-N parallelism speed-up, even though the constituent neurons are leaky. It does so with network-level integration and competition, but does not require fine-tuning of network parameters. Finally, the suggested form of additional nonlinearity is likely not unique: it might be possible to replace the threshold-nonlinearity in the contribution of individual neurons to the inhibitory drive with other forms of nonlinearity in either the excitatory or inhibitory units (see Discussion).

Multi-alternative forced-choice decision making

Thus far, we have focused on the efficiency of neural WTA in computing **max**, **argmax** when the number of alternatives or competitors is large, equating each competitor with an individual neuron or small pool of neurons during a highly distributed internal computation. The brain also makes explicit judgements between small numbers ($N \in \{1, \cdots, 10\}$) of externally presented alternative objects or actions in the world, as widely studied under the rubric of multi-alternative forced-choice (multi-AFC) tasks [23] in human and non-human psychophysics.

An influential result in multi-AFC decision making, used as far afield as commercial marketing and design to improve the presentation of choices [36], is Hick's law [27, 35]: the time to reach an accurate decision increases with the number of alternatives N, as $\log(N+1)$.

This observed increase in decision time with alternatives is reproduced by accumulate-to-bound (AB) models, in which the evidence for each option is integrated by either perfect or leaky accumulators, and an N-dependent threshold (specifically, one that increases as $\log(N)$ to maintain a fixed decision-making accuracy across N) is then applied to the integrated evidence [68, 46].

We investigate the behavior of WTA networks that arrive at a decision on multi-AFC tasks through self-terminating dynamics. Consider the case of $N-1$ noisy alternatives with true input means b and a final noisy alternative with input mean $b+\Delta$ (quasi-2D input), consistent with the setup of most multi-AFC psychophysics studies [10, 40]. As we will see, Hick's law is a natural byproduct of efficient parallel computation through WTA dynamics in a neural circuit.

For a small number of alternatives N, it is not meaningful to define a "scaling" of inhibition strength with N. Instead, we consider the speed and accuracy of WTA computation across strengths of self-excitation and inhibition in the interval $[0, 1]$ for both α and β (β could be chosen, in principle, to be arbitrarily large, but the outcomes of interest can be found with $\beta \leq 1$: increasing β further decreases integration time, thus increasing speed while decreasing accuracy to a degree not consistent with behavior; these trends asymptote by about $\beta \approx 3$, not shown), Figure 4a. For multi-AFC tasks, with their relatively small numbers of alternatives, the performance of networks with conventional WTA and nWTA networks is qualitatively similar. For simplicity, we describe the nWTA network results here; comparisons between the conventional and nWTA results are in S2, Figure S4a–d.

As before, the condition for WTA dynamics and the emergence of a winner is that $\alpha + \beta > 1$. Within this constraint, decision accuracy is maximized as $\alpha + \beta$ approaches 1 (diagonal band, Figure 4a panel 1), consistent with the increase in the network integration time ($\tau_{\text{integ}} = \tau/(1-(\alpha+\beta))$; see S1.1) for the evidence-accumulating differential modes: The longer the network can integrate information, the more accurate its eventual decision. (The limit $\alpha = 1$ and $\beta = 0$ corresponds to a non-leaky, non-competitive integration process and, thus, is equivalent to a non-leaky AB model, if supplied with an explicit bound. The non-leaky AB model, in turn, is equivalent to the serial strategy from previous sections, but operated in parallel; thus its performance defines the parallelism benchmark.)

Decision speed, on the other hand — computed by averaging all trials that pro-

duced a winner (correct or wrong) — is maximized when inhibition is large but self-excitation is intermediate in size, Figure 4a (panel 2). The network exhibits speed-accuracy tradeoffs, controllable by top-down influences that simply modulate the strengths of inhibition, excitation, or both, Figure 4a (panels 1,2).

Reward rate, the product of accuracy and an offset decision speed (offset decision speed is the inverse of the sum of decision time with a constant offset T_0, reflecting a baseline of internally or externally imposed response latency that does not depend on the decision-making computation), is a key quantity for psychophysics experiments since subjects are taught the task using reward contingencies. Because speed varies more steeply with parameters than does accuracy at zero latency ($T_0 = 0\,\text{ms}$), the maximum in the speed landscape determines the maximum reward rate (Figure 4a, panel 3), which is therefore achieved for strong inhibition and intermediate self-excitation. With no latency, responding fast and less accurately yields a higher reward rate than waiting longer to be more accurate.

With the addition of a non-zero latency, the reward rate becomes less sensitive to time or speed and relatively more sensitive to accuracy; hence, the reward rate maximum occurs at higher accuracy. For even a modest $T_0 = 300\,\text{ms}$ latency (chosen from a range of latency estimates from psychophysics results [61, 7]), near-perfect accuracy (closer to the $\alpha + \beta = 1$ diagonal, Figure 4b panel 4) is required to achieve the maximum in reward rate. Thus, at modest-to-high response latencies the maximal reward rate is achieved by waiting longer to be more accurate. Qualitatively similar results hold for different numbers of alternatives N (see S2, Figure S4e,f for how α, β values for maximal reward rate depend on N and T_0).

The WTA condition $\alpha + \beta > 1$ corresponds to unstable network dynamics. As a result, the network is generically impulsive, more strongly weighting early inputs relative to late ones [29, 71, 32], as seen in the decision-triggered average input curves of Figure 4b (top row; blue curves). However, the network can achieve uniform integration over longer decision time windows when tuned, with $\alpha + \beta$ set close to 1, Figure 4b (top row, gray curves). To obtain a uniform weighting of evidence over the relatively short 1-2 second integration time-windows tested in existing experiments [8, 31, 59], the tuning of $\alpha + \beta$ to 1 need not be finer than $\sim 2\%$, however, even if the biophysical time-constant of single neurons or synapses is as short as $20 - 50$ ms. (See the energy landscape of the network dynamics in Figure S1a,b, which shows that the landscape is quite flat early on, consistent with the network evenly integrating its inputs rather than being pulled strongly by the WTA attractor states, for $\alpha + \beta = 1.05$). The circuit tuning required to arrive at this or a more finely specified parameter setting is likely achieved through plasticity during task training. AB models, by contrast, do not naturally display impulsive dynamics, instead requiring the addition of another dynamical process, such as an "urgency signal" or collapsing decision thresholds over time during the trial, to reproduce impulsive behavior [12, 10].

Neural responses The outputs of neurons during the integration period vary enough from trial to trial for fixed parameter settings and across parameter settings to look variously more step-like or ramp-like (compare curves within and across Figure 4b-d). Different choices of α, β modulate the neural response curves, shifting them from more ramp-like to more step-like even as the network integration time

$(\tau/(1-(\alpha+\beta))$ and mean inputs are held fixed, Figure 4c. Thus, the sharp distinctions drawn by statistical models that delineate and interpret step- versus ramp-like response curves as supporting binary or graded evidence accumulation [34] are not meaningful in dynamical neural models of noisy choice behavior, at least on the level of individual neural responses during the decision period [70].

When parameters and numbers of options are held fixed, correct WTA trials terminate faster than wrong ones (Figure 3g), as observed in other attractor dynamics-based decision models [72, 21] and consistent with the psychophysics literature [56, 40] (and a drawback of AB models because these do not produce the faster-more-accurate result [64, 53, 44] without additional modifications [54]).

In experiments, the decision threshold or pre-decision activity level of the winning choice neurons increases under pressure to respond rapidly [26]. This result has been noted as counterintuitive from the perspective of AB models that increase speed by lowering the bound [26]. In the WTA networks, the asymptotic activity of the winning neuron is proportional to $1/(1-\alpha)$, while speed increases with $\alpha+\beta$. Starting from parameters consistent with a high reward rate (high β at zero latency or intermediate α, β at non-zero latency, Figure 4a, panels 3-4), a speed-up can be achieved by increasing α or β or both (panel 2). Thus, except in the special case where only inhibition is allowed to increase, the asymptotic activity level of the winning choice neurons is predicted to increase under speed pressure, as seen in experiments [26].

Performance comparison: WTA versus benchmark models We next compare neural WTA performance against a phenomenological model with perfect integration: non-leaky AB (Figure 4e, gray and purple curves, respectively) [68, 6, 46]. There is no simple decision model (i.e., there is no simple approximation to the full Bayesian expression) known to be optimal for multi-AFC tasks with more than two alternatives [13, 46, 66]. However, non-leaky AB is a commonly used benchmark.

The reward rate on multi-AFC tasks achieved by neural WTA (Figure 4e; throughout we assume a response latency of $T_0 = 300$ ms) is competitive with AB at small N (see also [46] and Figure S4b), even outperforming it for high accuracy (compare gray and purple in Figure 4e). With more alternatives, neural WTA becomes competitive with the benchmark across a wider range of accuracy levels. Note that AB does not make an optimal decision, as can be seen from the slightly better performance of neural WTA at high accuracies (Figure 4e; SI 2.1; also see [46] for a similar observation about integration to threshold in the presence of inhibition).

If the decision time is held fixed as N varies, our analytical results for large N suggest that accuracy should asymptotically decay as a power law in N. Indeed, the response accuracy declines with N (Figure 4f), but the asymptotic scaling does not apply for small numbers of alternatives. For very small N ($N = 2-3$), WTA accuracy at fixed time matches that of the AB strategy (Figure 4f), if the matched noise in both cases is purely in the inputs. However, WTA accuracy decreases faster than AB accuracy as a function of number of alternatives (Figure 4f).

On the other hand, if the desired accuracy is fixed at a high value (again, moving through speed-accuracy space by sweeping (α, β) for each N to achieve this accuracy, similar to the threshold adjustment to maintain fixed accuracy with varying N in the AB models; note that for small N $\alpha+\beta$ moves toward 1 to maintain fixed accuracy

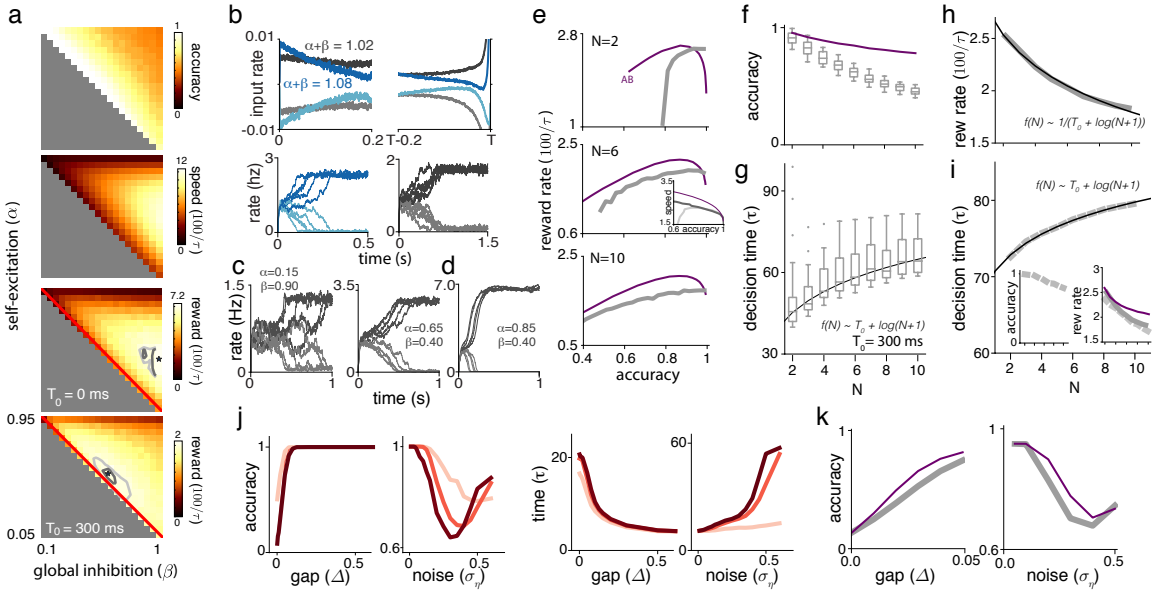

Figure 4: **Self-terminating WTA dynamics as a minimal-parameter, neural model of multi-AFC decision-making** a) Accuracy, speed and reward rate as a function of β and α [$N = 6$]. Gray: tuples excluded by constraints $\alpha < 1$, $\beta > 1 - \alpha$. Reward rates shown at non-decision latencies of $T_0 = 0\,\text{ms}$ (top) and $T_0 = 300\,\text{ms}$ (bottom). Stars: optimal (α, β); gray lines: iso-reward contours at 98% (dark) and 95% (light) of maximum. b) Top: average input for winning (dark gray, blue) and losing (light gray, blue) neurons in a WTA network performing a 2-AFC task with zero gap for two parameter settings (gray versus blue). Bottom: example firing rates from trials used for the averages in the top panel. c) Example firing rate trajectories for WTA networks with different (α, β) but with $\alpha + \beta$ (i.e., integration time-constant) held fixed. Left: lower self-excitation and higher inhibition; right: vice versa. d) Example firing rate trajectories for network with same inhibition as c) (right panel) but stronger self-excitation. Trials are faster, but final activation of winner is higher. e) Reward rate vs. accuracy for $N = \{2, 6, 10\}$. Thick gray: WTA (using best α, β for given accuracy); thin purple: AB model (note that AB model takes time T_S/N and is thus the parallelism benchmark). Inset for $N = 6$ panel shows speed-accuracy curve (dark gray: α fixed, β varied; light gray: β fixed, α varied, $T_0 = 300\,\text{ms}$). f) Accuracy at fixed decision time, $T_{\text{WTA}} = 90\,\text{ms}$. Box plots: distribution of accuracies across different network parameter settings that reach a decision at this time. Thin purple: AB model accuracy at same decision time. g) Decision time at fixed accuracy, $A = 0.99$. Box plots: distribution of times across different network parameter settings that reach this accuracy. Solid lines: fit of distribution medians to $\log(N + 1)$. h) Reward rate at (α, β) settings individually optimized for each N. Thin solid line: fit to $(T_0 + \log(N+1))^{-1}$. i) Convergence time at best shared (α, β) across N. Thin solid line: fit to $(T_0 + \log(N + 1))$. Inset: accuracy (left) and reward rate (right, dashed) for shared parameters across N remain high and reward rate is comparable to when parameters are optimized for each N (thick solid gray) and to the AB strategy (thin purple). For $T_0 = 300\,\text{ms}$: $\bar{\alpha} = 0.41, \bar{\beta} = 0.7$. j) Accuracy (plots 1-2) and decision time (plots 3-4) in WTA network with fixed parameters as gap size Δ (fixed $\sigma_\eta = 0.2$) or noise amplitude σ_η (fixed $\Delta = 0.05$) are varied; light to dark curves: $N = \{2, 6, 10\}$. k) Accuracy for WTA network ($N = 6$) compared to AB model for matched time as Δ and σ_η are varied. In j,k $\alpha = 0.6$, $\beta = 0.6$.

as N increases, see S1.7, Figure S3h), the resulting decision time increases weakly with the number of alternatives and the increase is well-fit by $\log(N+1)$, Figure 4g. This result shows that a self-terminating dynamical decision process in a competitive network of leaky neurons with noisy input produces behavior consistent with Hick's law.

The Hick's law-like scaling holds across a range of fixed accuracy values (S2, Figure S4g–j). Note that the agreement with Hick's law does not depend on the choice of decision latency T_0, because different values of T_0 merely shift the decision time curve up or down without affecting its functional form. The reward rate achievable by neural WTA across N is comparable to the benchmarks, Figures 4h,i and S4.

Interestingly, in the neural WTA network, it is possible to maintain near-constant, high accuracy while holding all parameters fixed as N is varied, Figure 4i (inset 1, see also Figure S4g). When we examine the decision time of the network as a function of number of options, it once again produces a $\log(N+1)$, or Hick's law-like scaling, Figure 4i (also Figure S4h,j), but this time without any parameter readjustments as a function of N. In other words, unlike in the AB models and the neural WTA result above where Hick's law is recovered when decision or circuit parameters are adjusted by hand as a function of N, the neural WTA model maintains a high response accuracy and produces Hick's law behavior even when the network parameters are held fixed as the number of options is varied.

Finally, if the top gap is held fixed while the remaining inputs are drawn uniformly, the decision time is predicted to become very weakly dependent on N, since the inputs beyond the first two are smaller, and thus largely irrelevant, in the competition (not shown). To our knowledge, this experiment has not been done. Similarly, if the top gap is made very large (with all the other inputs equal), decision time is predicted to become independent of N (not shown).

Performance: Network re-tuning for high reward-rate across alternatives? As we have seen, the neural WTA network reproduces Hick's law whether or not parameters are re-optimized when the network solves tasks with different numbers of alternatives. In Hick's experiments, subjects were trained on blocks of trials with a fixed number of alternatives, then retrained on a new block with a different number of alternatives, presumably allowing for the possibility of network parameter re-tuning. Yet it is unclear if the brain does re-tune its parameters, or even whether such retuning is necessary in principle, to achieve high reward rates across varying numbers of alternatives.

To answer the latter question, we set (α, β) to a single $(\alpha, \beta)^{\text{opt}}$ value that maximizes the summed reward across all N-alternative tasks, with N ranging from 2 to 10; the setting corresponds to intermediate self-excitation and stronger inhibition (Figure 4 caption and Figure S4e). We recomputed the reward rate across N with this fixed parameter setting, and found that performance closely matched that achieved when the network parameters were separately optimized at each N, Figure 4i (second inset). In other words, a WTA network with a fixed strength of self-excitation and global inhibition can achieve good performance across decision tasks with varying numbers of alternatives, without parameter re-tuning.

The network achieves this high reward rate across N without parameter retuning

because its dynamics are (partially) self-adjusting to the difficulty of the task, automatically slowing down as the noise increases or the gap shrinks, Figure 4j, while remaining competitive with non-leaky integration for the resulting decision time, Figure 4k. It has recently been shown in experiment that decision circuits can be trained to adapt their integration time to the time-varying statistics of the input data [51]. The present result shows how neural circuits, if similar to our WTA model, may be able to automatically and instantly (from trial to trial), without plasticity, partially adjust to the statistics of the input stimulus.

Discussion

Our work has focused on the question of efficient parallel computation of **max, argmax** in neural network models with leak and in noisy settings that include fluctuating inputs and potentially noisy neural dynamics. As noted in the Introduction, **max, argmax** are elemental operations in inference, optimization, decision making, action selection, consensus, error-correction, and foraging computations. We showed that conventional WTA networks are either too slow or fail altogether in concluding the computation in the presence of noise. We introduced the nWTA network, consisting of neurons competing to win, with each contributing strong inhibition (synaptic strengths do not scale down as the number of competitors, N, is increased) to the circuit, but only if their individual activations are higher than a threshold. With this second non-linearity, networks converge to a WTA state even in the presence of noise, and the accuracy and speed of the computation matches the benchmark for efficient parallel computation — N times faster than the optimal serial strategy. Specifically, we showed that neural nWTA networks can accurately determine and report the maximum of a set of N inputs with an asymptotically constant decision time in the noiseless case and with time that grows as $O(\log(N))$ in the presence of noise. This type of efficiency is a necessary condition for the hypothesis that the brain performs massively parallel computations.

When applied to psychophysical decision-making tasks [47, 27, 58, 10] involving much smaller numbers of alternatives ($N \leq 10$), the model provides a neural circuit-level explanation for Hick's law [27], the observation that the time taken for perceptual decision-making scales as the logarithm of the number of options. Thus, we interpret Hick's law as a signature of efficient parallel computation in neural circuits.

Our work additionally reproduces a number of (sometimes counterintuitive) psychophysical and neural observations, including faster performance on correct than error trials, a higher pre-decision neural activity level when subjects are pressured to make faster decisions, and a natural tendency to weight early information over late (however, the extent of this tendency to impulsivity is tunable).

In this way, our work provides a single umbrella under which systems neuroscience questions about parallel computation distributed across large numbers of individual neurons and psychophysics questions about explicit decision making can be answered.

Relationship to past work Parallel and network implementations of **max, argmax** have been variously studied in computer science [3, 16] and in both artificial [65] and biologically plausible neural networks [24, 30, 71, 57].

In neuroscience, various models of WTA dynamics and aspects of their computational properties have been fruitfully considered in previous work, including: Conditions for the emergence of a unique winner [73] or groups of winners [25], how the computation depends on gap size [20], stability of deterministic dynamics [43, 74], dependence on different initial conditions [65, 73, 43], the relative strengths of inhibition and excitation [37, 20, 57], and the scaling of computation time in deterministic, not fully-neural models [65]. Some models even considered nonlinear inhibitory contributions, in the quite different form of shunting inhibition [74], with encouraging results in the noise-free setting for at least medium-sized N, but high sensitivity to fluctuations, which prohibits accurate WTA for small and large N. The leaky, competing accumulator model [68, 46] would be mathematically equivalent to the conventional WTA model, if $\tau_\eta = \tau$ [48] and if the asymptotic state of the dynamics were used as the decision criterion; instead, the works use the crossing of a pre-determined threshold not tied explicitly to the asymptotic states as the decision criterion, while not considering the feasibility of WTA for large N or the possible breakdown of WTA states in the presence of noise. Collectively, these works provide insights into many individual aspects of WTA dynamics, usually in the noise-free case, and almost invariably, for small N.

Our work builds on this body of work, extending it along several directions while unifying previous results in the context of a single, neurally plausible network model with self-terminating dynamics in which the network's asymptotic state is its own readout. We study WTA performance for both very large numbers of competitors (in the thousands), and for small numbers (in the range 1-10). We examine network performance with a focus on the time-complexity of the operations (speed) together with accuracy. Our treatment centrally considers the role of noise in the dynamics, as noise is an inescapable property of neural dynamics, sensory processing, and real-world inputs. Here we find that conventional WTA models fail for large numbers of competitors, and propose a new model, with a second neural nonlinearity, that succeeds and matches the efficiency of a parallelism benchmark. The strength of inhibition in individual synapses remains fixed, and does not decrease with N, allowing the same circuit to be used for different N without changing the scaling of synaptic strength. We ahow that the network automatically (partially) adjusts for gap size and noise level, increasing its decision time as the ratio of gap to noise, or the signal-to-noise ratio, shrinks. Moreover, we show what we believe is the first demonstration of Hick's law within a neural network decision making model with self-terminating dynamics.

Biological mechanisms for thresholding inhibitory contributions. WTA networks converge toward a state with a single winner for large numbers of competitors only if the conventional models are modified with a second neural or synaptic nonlinearity that prevents weakly active principal cells from contributing inhibitory feedback to the circuit (Equation (4)). How could this nonlinearity be implemented, given that inhibitory neurons are not believed to possess nonlinear dendritic mechanisms to differentially gate different inputs?

In circuits with separate excitatory and inhibitory neurons [14], there are multiple candidate mechanisms for nonlinear inhibition. These can be divided by whether inhibitory interneurons are selectively tuned to particular principal cell inputs, or

whether they are non-selective and pool inputs from many principal cells. In the former case, to maintain global inhibition in the circuit, the inhibitory neurons would have to send outputs broadly across the network.

If inhibitory neurons are selective, then a simple threshold nonlinearity in its input-output transfer function, like the type-II firing rate responses in inhibitory neurons [28], is sufficient. A similar effect could be achieved by fast-spiking inhibitory interneurons that act as coincidence detectors rather than integrators [2]: these cells would respond at most weakly to low firing-rate inputs, while firing reliably when the inputs have a high rate, transmitting an inhibitory output based on effectively thresholding the activity of the input principal cell. Finally, if interneurons target pyramidal cells dendrites, then dendritic nonlinearities [39] could gate the selective inhibitory input, effectively only transmitting the inhibition when it exceeds a threshold.

If inhibitory neurons are non-selective, then the nonlinearity must be present in the excitatory-to-inhibitory synapse so that the drive from the low-activity principal cells is specifically ignored. If the excitatory to inhibitory synapses have low release probability and are strongly facilitating, only high firing rate inputs would make it through [75].

Since WTA is likely performed in many neural circuits across the brain, different circuits may use different mechanisms for thresholding inhibitory contributions, especially if they have different evolutionary histories.

Structure and properties of neural networks for noisy WTA The parallelism speed-up for **max**, **argmax** relative to the optimal serial strategy in the neural WTA circuits is achieved by trading time for space: Specifically, the network size (in neurons) and number of memory states grows linearly with N – each neuron or neuron group integrates its input over time, but the computation is performed N times faster. By contrast, the optimal serial strategy requires holding only a single item in memory, since each input item is integrated, compared to the single item in memory, and then discarded (if it is smaller) or used to update the item in memory (if it is larger).

Another form of spatial complexity is in the structure of synaptic connections. The neural WTA model can be viewed as N principal cells or groups, all inhibiting each other (e.g. through a local interneuron private to each cell), which requires $\sim N^2$ synapses. This connectivity is both dense and global. Alternatively, one can think of all cells as driving a single global inhibitory neuron, which requires only $\sim N$ synapses, a much sparser connectivity (only $O(1)$ synapse per neuron), that is nevertheless still global. We wonder whether it is possible to replace the global sum of excitatory activities by a set of local inhibitory neurons pooling local excitatory inputs. However, local inhibition typically produces pattern formation [15], and consensus formation with local connections can be unstable [4], thus it is an interesting open question whether WTA across a population of neurons can be implemented with purely local connectivity.

The neural circuit model considered here produces impulsive or uniform integration over time. It cannot, without modification, arrive at a self-terminating WTA state while performing leaky integration: Leaky integration requires $\alpha + \beta < 1$, which violates the condition for the existence of WTA states. Similarly, in tasks where evidence is provided in discrete pulses, the WTA network can integrate across pulses

as the interval between pulses is varied [31], if it is tuned for near-perfect integration without lateral competition ($\alpha \approx 1, \beta \approx 0$; data not shown); however, it does not then display WTA dynamics (or convergence to a WTA state is slow). Different nonlinearities could permit the co-existence of leaky or near-perfect long-time integration with self-terminating decision dynamics, for instance: starting each trial in the leaky or near-perfect integration setting, but using a generalized urgency signal to increase the strength of feedback over time, so that the network becomes more competitive and performs WTA later in the trial; or modeling the system as 2-stage network with leaky integration in the first stage and decision dynamics (e.g. through WTA) in the second. In addition, a network could be simultaneously leaky and display WTA dynamics, if the threshold-linear neural activations are replaced by a more nonlinear function [72].

It will be interesting to more explicitly relate the neural architectures required in decision-making models with impulsive, near-ideal, or leaky integration with the distributed architecture of decision circuits in the brain [23].

Extensions and generalizations Distributed decision making is a feature of many collective societies or colonies, including quorum sensing in bacteria [49], foraging and house-hunting in ants and bees [5, 18], social and political consensus formation [38], and various economic choice behaviors. While our model is based on neural dynamics, the ingredients (self-amplification; recurrent nonlinear inhibition) are simple and should have analogues in other distributed decision-making systems. Thus, although we are not aware of theoretical studies in these systems that investigate how decision time scales with N, our results suggest a scaling of $O(\log(N))$, where N is the number of options, and the existence of a thresholded or otherwise nonlinear inhibition if N is large.

In the present work, we have assumed high bandwidth communication: neurons exchanging analog signals in continuous time. Real world systems are bandwidth limited: Neurons communicate with spikes emitted at a finite rate; scout insects achieve consensus on decisions about competing food sites or competing nesting sites through brief interactions with subsets of others in the hive [17, 18]. Nevertheless, the principal cells in our model do not communicate their individual activation levels to all other cells; other principal cells receive information only about global activity in the network in the form of a single inhibitory signal, a form of limited communication. In this sense, our results should generalize to the lower-bandwidth case. Studying the impact of low-bandwidth communication on WTA and parallel decision making in more detail, together with constraints on global connectivity as discussed above, is an interesting direction for future work.

Methods

Network model and dynamics

Deterministic WTA network We first consider N coupled neurons with activations x_i, $i \in \{1, \ldots, N\}$, and dynamics given by [73]:

$$\tau \frac{dx_i}{dt} + x_i = \left[b_i + \alpha x_i - \beta \sum_{j \neq i} x_j \right]_+ =: r_i(t). \tag{5}$$

The neural nonlinearity is set to be the threshold-linear function: $[x]_+ = x$ when $x > 0$ and zero otherwise; τ is the neural time constant, α is the strength of self-excitation and β is the strength of global inhibition. The RHS of Equation (5) may be viewed as the instantaneous firing rate $r_i(t)$ of neuron i.

Each neuron receives a constant external drive b_i. Neurons are ordered so that $b_1 > b_2 \geq \ldots \geq b_N$, and it is assumed that the largest input drives just one neuron, see Figure 1 a. The goal of WTA is to amplify activity in the neuron with the largest input and to drive the activations and rates of the remaining neurons to zero, Figure 1 b,c. The number of choices b_i equals the number of neurons, and spatial complexity is thus $\sim N$.

The full coupling matrix $W = ((\alpha + \beta)\mathbf{I} - \beta \mathbf{1}\mathbf{1}^T)$ has one eigenvalue $\lambda_{\text{W,hom}} = \alpha - (N-1)\beta$ with uniform eigenvector $\mathbf{1} = (1, \ldots, 1)^\top$, and an $(N-1)$-fold degenerate eigenvalue $\lambda_{\text{W,diff}} = (\alpha + \beta)$ whose eigenvectors are difference modes with entries that sum to zero. When $\alpha + \beta > 1$ the difference modes grow through a linear instability, and the eventual (non-trivial) steady states involve only one active neuron. If $\alpha < 1$ and $\beta > 1 - \alpha$, the network will always select a unique winner for each $\mathbf{b}$ and initial condition [73]. For a discussion of more general constraints on α, β, see S1.1.

After meeting the conditions for stability and uniqueness ($\alpha < 1, \beta > (1 - \alpha)$), there is freedom in the choice of how the strength of global inhibition β scales with N: We may choose $\beta \sim O(1)$, which we call the *strong inhibition* condition, or $\beta \sim \beta_0/N$, the *weak inhibition* condition. Existing works on WTA dynamics variously use strong [73] or weak [37, 65] inhibition, usually without explicit justification or explanation for making one choice over the other. We consider both conditions. When in the weak inhibition regime, we set $\alpha = 1 - \frac{\beta_0}{kN}$ (with $k > 1$; specifically, we choose $\beta_0 = 1, k = 2$ throughout the paper) for stability.

For simplicity, throughout the paper we consider the case where all neurons start at the same resting activity level $\mathbf{x}(0) = (x_0, \ldots, x_0)^\top$. In this case, the winner is the neuron with the largest input. (For heterogeneous initial conditions the situation is more complex, since the wrong neuron can be pushed to be the winner just by starting at large enough activity to suppress all other neurons, see also the discussion in [73].) We further assume $x_0 = 0$ (if $x_0 > 0$ there is an initial transient that scales logarithmically with N but is unrelated to the actual WTA-computation, see S1.2).

WTA network with noisy inputs A second scenario we consider is the more biologically common case of noise-corrupted input, and we ask how this impacts the

performance of WTA. The model dynamics is identical to Equation (5), with the addition of a time-varying, zero-mean fluctuation term $\eta_i(t)$ in each input:

$$\tau \frac{dx_i}{dt} + x_i = \left[b_i + \eta_i + \alpha x_i - \beta \sum_{j \neq i} x_j \right]_+ \tag{6}$$

We model $\eta_i(t)$ by statistically identical Ornstein-Uhlenbeck processes, such that

$$\tau_\eta \frac{d\eta_i(t)}{dt} + \eta_i(t) = \sigma_\eta \sqrt{2\tau_\eta}\, \xi_i(t) \tag{7}$$

with Gaussian white noise $\xi_i(t)$, such that, $\langle \xi_i(t) \rangle = 0, \langle \xi_i(t) \xi_j(t') \rangle = \delta_{ij}\delta(t-t')$. It follows that $\langle \eta_i(t) \rangle = 0$ and $\langle \eta_i(t) \eta_j(t') \rangle = \sigma_\eta^2 e^{-\frac{|t-t'|}{\tau_\eta}} \delta_{ij}$. We set $\tau_\eta \ll \tau$ so that the fluctuations $\eta(t)$ are effectively uncorrelated over the timescale τ of single neurons. For numerical simulations of Ornstein-Uhlenbeck noise we make use of exact integration on a time grid with increment Δt [22], i.e.,

$$\eta(t + \Delta t) = \eta(t)\, e^{-\Delta t/\tau_\eta} + \sigma_\eta \sqrt{1 - e^{-2\Delta t/\tau_\eta}}\, \xi(t) \tag{8}$$

If the noise amplitude is below a critical value σ_η^* (see S1.5), the winner moves to a noisy high-activity state close to the deterministic attractor $x^\infty = b_w/(1-\alpha)$, where b_w is the input drive of the winner (not necessarily the largest b_i), while losers are strongly suppressed. For convenience, we define T_{WTA} as the time some neuron reaches an activation level greater than or equal to $c\frac{b_2}{(1-\alpha)}$ with $c \lesssim 1$ (we use $c = 0.88$). We emphasize that this convergence criterion is set by the dynamics and hence inherently different from an external arbitrary threshold.

References

[1] D.G. Amaral and M.P. Witter. The three-dimensional organization of the hippocampal formation: A review of anatomical data. *Neuroscience*, 31(3):571 – 591, 1989.

[2] Maria Cecilia Angulo, Jean Rossier, and Etienne Audinat. Postsynaptic glutamate receptors and integrative properties of fast-spiking interneurons in the rat neocortex. *Journal of Neurophysiology*, 82(3):1295–1302, 1999.

[3] Shay Assaf and Eli Upfal. Fault tolerant sorting networks. *SIAM Journal on Discrete Mathematics*, 4(4):472–480, 1991.

[4] Bassam Bamieh, Mihailo R Jovanovic, Partha Mitra, and Stacy Patterson. Coherence in large-scale networks: Dimension-dependent limitations of local feedback. *IEEE Transactions on Automatic Control*, 57(9):2235–2249, 2012.

[5] Ralph Beckers, Jean-Louis Deneubourg, Simon Goss, and Jacques M Pasteels. Collective decision making through food recruitment. *Insectes sociaux*, 37(3):258–267, 1990.

[6] Rafal Bogacz, Eric Brown, Jeff Moehlis, Philip Holmes, and Jonathan D Cohen. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113(4):700–765, 2006.

[7] Rafal Bogacz, Peter T Hu, Philip J Holmes, and Jonathan D Cohen. Do humans produce the speed–accuracy trade-off that maximizes reward rate? *The Quarterly Journal of Experimental Psychology*, 63(5):863–891, 2010.

[8] Bingni W. Brunton, Matthew M. Botvinick, and Carlos D. Brody. Rats and humans can optimally accumulate evidence for decision-making. *Science*, 340(6128):95–98, 2013.

[9] Lars Chittka, Peter Skorupski, and Nigel E Raine. Speed–accuracy tradeoffs in animal decision making. *Trends in Ecology & Evolution*, 24(7):400–407, 2009.

[10] Anne K Churchland, Roozbeh Kiani, and Michael N Shadlen. Decision-making with multiple alternatives. *Nature Neuroscience*, 11(6):693–702, 2008.

[11] Robert L. Coultrip, Richard H. Granger, and Gary Lynch. A cortical model of winner-take-all competition via lateral inhibition. *Neural Networks*, 5(1):47–54, 1992.

[12] Jochen Ditterich. Stochastic models of decisions about motion direction: behavior and physiology. *Neural Networks*, 19(8):981–1012, 2006.

[13] VP Dragalin, Alexander G Tartakovsky, and Venugopal V Veeravalli. Multihypothesis sequential probability ratio tests. I. asymptotic optimality. *IEEE Transactions on Information Theory*, 45(7):2448–2461, 1999.

[14] John C Eccles, P Fatt, and K Koketsu. Cholinergic and inhibitory synapses in a pathway from motor-axon collaterals to motoneurones. *The Journal of Physiology*, 126(3):524–562, 1954.

[15] Bard Ermentrout. Neural networks as spatio-temporal pattern-forming systems. *Reports on Progress in Physics*, 61(4):353, 1998.

[16] Uriel Feige, Prabhakar Raghavan, David Peleg, and Eli Upfal. Computing with noisy information. *SIAM Journal on Computing*, 23(5):1001–1018, October 1994.

[17] Nigel R Franks, François-Xavier Dechaume-Moncharmont, Emma Hanmore, and Jocelyn K Reynolds. Speed versus accuracy in decision-making ants: expediting politics and policy implementation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1518):845–852, 2009.

[18] Nigel R Franks, Stephen C Pratt, Eamonn B Mallon, Nicholas F Britton, and David JT Sumpter. Information flow, opinion polling and collective intelligence in house–hunting social insects. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 357(1427):1567–1583, 2002.

[19] Tamas F Freund and Gyorgi Buzsaki. Interneurons of the hippocampus. *Hippocampus*, 6(4):347–470, 1996.

[20] T Fukai and S Tanaka. A simple neural network exhibiting selective activation of neuronal ensembles: from winner-take-all to winners-share-all. *Neural Computation*, 9:77–97, 1997.

[21] Moran Furman and Xiao-Jing Wang. Similarity effect and optimal control of multiple-choice decision making. *Neuron*, 60(6):1153–1168, 2008.

[22] Daniel T. Gillespie. Exact numerical simulation of the ornstein-uhlenbeck process and its integral. *Physical Review E*, 54:2084–2091, Aug 1996.

[23] Joshua I. Gold and Michael N. Shadlen. The neural basis of decision making. *Annual Review of Neuroscience*, 30(1):535–574, 2007. PMID: 17600525.

[24] Stephen Grossberg. Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, 52(3):213–257, 1973.

[25] Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947, 2000.

[26] Richard P Heitz and Jeffrey D Schall. Neural mechanisms of speed-accuracy tradeoff. *Neuron*, 76(3):616–628, 2012.

[27] W E Hick. On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, 4(1):11–26, 1952.

[28] Alan L Hodgkin. The local electric changes associated with repetitive action in a non-medullated axon. *The Journal of Physiology*, 107(2):165–181, 1948.

[29] Alexander C Huk and Michael N Shadlen. Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making. *Journal of Neuroscience*, 25(45):10420–10436, 2005.

[30] Dezhe Z Jin and H Sebastian Seung. Fast computation with spikes in a recurrent neural network. *Physical Review E*, 65(5):051922, 2002.

[31] Roozbeh Kiani, Anne K Churchland, and Michael N Shadlen. Integration of direction cues is invariant to the temporal gap between them. *Journal of Neuroscience*, 33(42):16483–16489, 2013.

[32] Roozbeh Kiani, Timothy D Hanks, and Michael N Shadlen. Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. *Journal of Neuroscience*, 28(12):3017–3029, 2008.

[33] Donald RJ Laming. *Information Theory of Choice-Reaction Times*. Wiley, New York, 1968.

[34] Kenneth W. Latimer, Jacob L. Yates, Miriam L. R. Meister, Alexander C. Huk, and Jonathan W. Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science*, 349(6244):184–187, 2015.

[35] Arne M Laursen. Task dependence of slowing after pyramidal lesions in monkeys. *Journal of Comparative and Physiological Psychology*, 91(4):897–906, 1977.

[36] William Lidwell, Kritina Holden, and Jill Butler. *Universal principles of design, revised and updated: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design.* Rockport Pub, 2010.

[37] R.P. Lippmann, B. Gold, M.L. Malpass, and Lincoln Laboratory. *A Comparison of Hamming and Hopfield Neural Nets for Pattern Classification.* Technical report (Lincoln Laboratory). Massachusetts Institute of Technology, Lincoln Laboratory, 1987.

[38] Christian List and Robert E Goodin. Epistemic democracy: generalizing the condorcet jury theorem. *Journal of Political Philosophy*, 9(3):277–306, 2001.

[39] Michael London and Michael Häusser. Dendritic computation. *Annual Review of Neuroscience*, 28:503–532, 2005.

[40] R Duncan Luce. *Response times: Their role in inferring elementary mental organization.* Number 8. Oxford University Press on Demand, 1986.

[41] Wolfgang Maass. On the computational power of winner-take-all. *Neural Computation*, 12(11):2519–2535, 2000.

[42] E Majani, Ruth Erlanson, and Yaser S Abu-Mostafa. On the K-winners-take-all network. In *Advances in Neural Information Processing Systems*, pages 634–642, 1989.

[43] Z. H. Mao and S. G. Massaquoi. Dynamics of winner-take-all competition in recurrent neural networks with lateral inhibition. *IEEE Transactions on Neural Networks*, 18(1):55–69, Jan 2007.

[44] Mark E Mazurek, Jamie D Roitman, Jochen Ditterich, and Michael N Shadlen. A role for neural integrators in perceptual decision making. *Cerebral Cortex*, 13(11):1257–1269, 2003.

[45] Mark D McDonnell and Lawrence M Ward. The benefits of noise in neural systems: bridging theory and experiment. *Nature Reviews Neuroscience*, 12(7):415, 2011.

[46] Tyler McMillen and Philip Holmes. The dynamics of choice among multiple alternatives. *Journal of Mathematical Psychology*, 50(1):30–57, 2006.

[47] Julius Merkel. Die zeitlichen verhältnisse der willensthäthigkeit. *Philisophische Studien*, 2:73–127, 1885.

[48] Ken Miller and Francesco Fumarola. Mathematical equivalence of two common forms of firing-rate models of neural networks. *Neural Computation*, 24(1):25–31, 2012.

[49] Melissa B Miller and Bonnie L Bassler. Quorum sensing in bacteria. *Annual Reviews in Microbiology*, 55(1):165–199, 2001.

[50] Jonathan W Mink. The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, 50(4):381 – 425, 1996.

[51] Alex Piet, Ahmed El Hady, and Carlos D. Brody. Rats optimally accumulate and discount evidence in a dynamic environment. *bioRxiv*, 2017.

[52] Cindy Poo and Jeffry S. Isaacson. Odor representations in olfactory cortex: sparse coding, global inhibition, and oscillations. *Neuron*, 62(6):850 – 861, 2009.

[53] Roger Ratcliff and Gail McKoon. The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, 20(4):873–922, 2008.

[54] Roger Ratcliff and Jeffrey N Rouder. Modeling response times for two-choice decisions. *Psychological Science*, 9(5):347–356, 1998.

[55] P. Redgrave, T.J. Prescott, and K. Gurney. The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, 89(4):1009 – 1023, 1999.

[56] Jamie D Roitman and Michael N Shadlen. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience*, 22(21):9475–9489, 2002.

[57] Ueli Rutishauser, Rodney J. Douglas, and Jean-Jacques Slotine. Collective stability of networks of winner-take-all circuits. *Neural Computation*, 23(3):735–773, March 2011.

[58] C Daniel Salzman and William T Newsome. Neural mechanisms for forming a perceptual decision. *Science*, 264(5156):231–238, 1994.

[59] Benjamin B Scott, Christine M Constantinople, Jeffrey C Erlich, David W Tank, and Carlos D Brody. Sources of noise during accumulation of evidence in unrestrained and voluntarily head-restrained rats. *eLife*, 4:e11308, dec 2015.

[60] G. M. Shepherd. Synaptic organization of the mammalian olfactory bulb. *Physiological Reviews*, 52(4):864–917, 1972.

[61] Patrick Simen, David Contreras, Cara Buck, Peter Hu, Philip Holmes, and Jonathan D Cohen. Reward rate optimization in two-alternative decision making: empirical tests of theoretical predictions. *Journal of Experimental Psychology: Human Perception and Performance*, 35(6):1865, 2009.

[62] Sameet Sreenivasan and Ila Fiete. Grid cells generate an analog error-correcting code for singularly precise neural computation. *Nature Neuroscience*, 14(10):1330–1337, 2011.

[63] Charles F. Stevens. What the flys nose tells the flys brain. *Proceedings of the National Academy of Sciences*, 112(30):9460–9465, 2015.

[64] Mervyn Stone. Models for choice-reaction time. *Psychometrika*, 25(3):251–260, 1960.

[65] Bruce W. Suter and Matthew Kabrisky. On a magnitude preserving iterative maxnet algorithm. *Neural Computation*, 4(2):224–233, 1992.

[66] S Tajima, J Drugowitsch, and A Pouget. Optimal policy leads to irrational choice behavior under multiple alternatives. *Cosyne Abstracts 2017, Salt Lake City, UT, USA.*, 2017.

[67] Warren H Teichner and Marjorie J Krebs. Laws of visual choice reaction time. *Psychological Review*, 81(1):75, 1974.

[68] Marius Usher and James L McClelland. The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, 108(3):550, 2001.

[69] Marius Usher, Zeev Olami, and James L McClelland. Hick's law in a stochastic race model with speed–accuracy tradeoff. *Journal of Mathematical Psychology*, 46(6):704–715, 2002.

[70] Xiao-Jing Wang. Neural dynamics and circuit mechanisms of decision-making. *Current Opinion in Neurobiology*, 22(6):1039–1046, 2012.

[71] Kong-Fatt Wong, Alexander Huk, Michael Shadlen, and Xiao-Jing Wang. Neural circuit dynamics underlying accumulation of time-varying evidence during perceptual decision making. *Frontiers in Computational Neuroscience*, 1:6, 2007.

[72] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006.

[73] X Xie, RHR Hahnloser, and SH Seung. Selectively grouping neurons in recurrent networks of lateral inhibition. *Neural Computation*, 14:2627–2646, 2002.

[74] Alan L. Yuille and Norberto M. Grzywacz. A winner-take-all mechanism based on presynaptic inhibition feedback. *Neural Computation*, 1(3):334–347, September 1989.

[75] Robert S Zucker and Wade G Regehr. Short-term synaptic plasticity. *Annual Review of Physiology*, 64(1):355–405, 2002.