Spike-Based Winner-Take-All Computation: Fundamental Limits and Order-Optimal Circuits

Lili Su
CSAIL, MIT
lilisu@mit.edu

Chia-Jung Chang
Brain and Cognitive Sciences, MIT
chiajung@mit.edu

Nancy Lynch
CSAIL, MIT
lynch@csail.mit.edu

Abstract

Winner-Take-All (WTA) refers to the neural operation that selects a (typically small) group of neurons from a large neuron pool. It is conjectured to underlie many of the brain's fundamental computational abilities. However, not much is known about the robustness of a spike-based WTA network to the inherent randomness of the input spike trains. In this work, we consider a spike-based k–WTA model wherein n randomly generated input spike trains compete with each other based on their underlying statistics, and k winners are supposed to be selected. We slot the time evenly with each time slot of length $1\,ms$, and model the n input spike trains as n independent Bernoulli processes. The Bernoulli process is a good approximation of the popular Poisson process but is more biologically relevant as it takes the refractory periods into account. Due to the randomness in the input spike trains, no circuits can guarantee to successfully select the correct winners in finite time. We focus on analytically characterizing the minimal amount of time needed so that a target minimax decision accuracy (success probability) can be reached.

We first derive an information-theoretic lower bound on the decision time. We then design a simple WTA circuit whose decision time, for any fixed $\delta \in (0,1)$, is order-optimal in terms of its scaling in the network size n, the number of true winners k, and the task complexity $T_{\mathcal{R}}$.

1 Introduction

Humans and animals can form a stable perception and make robust judgments under ambiguous conditions. For example, we can easily recognize a dog in a picture regardless of its posture, hair color, and whether it stands in the shadow or is occluded by other objects. One fundamental feature of brain computation is its robustness to the randomness introduced at different stages, such as sensory representations, feature integration, decision formation, and motor planning. It has been shown that neurons encode information in a stochastic manner in the brain; even when the exact same sensory stimulus is presented or when the same kinematics are achieved, no deterministic patterns in the spike trains exist. Facing environmental ambiguity, humans and animals adaptively refine their behaviors by incorporating prior knowledge with their current sensory measurements. Nevertheless, it remains relatively unclear how neurons carry out robust computation facing ambiguity. Sparse coding is a common strategy in brain computation; to encode a task-relevant variable, often only a small group of neurons from a large neuron pool are activated. Understanding the underlying neuron selection mechanism is highly challenging.

Winner-Take-All (WTA) is a hypothesized mechanism to select proper neurons from a competitive network of neurons, and is conjectured to be a fundamental primitive of cognitive functions such as attention and object recognition. Among these studies, it is commonly assumed that neurons transmit information with a continuous variable such as the firing rate. This assumption, however, ignores how temporal coding may additionally contribute to cortical computations. For example, some neurons in the auditory cortex will respond to auditory events with bursts at a fixed latency. This phase-locking property is also observed in the hippocampus as well as the prefrontal cortex. Another feature that has been neglected in a rate-based model is the inherent noise in the inputs. Although some studies used additive Gaussian noise to account for input randomness, such WTA circuits are very sensitive to noise and could not successfully select even a single winner unless extra robustness strategy such as an additional nonlinearity is introduced into the dynamics. Last but not least, neurons have a refractory period, which prevents spikes from back propagating in axons, and such a feature is usually neglected in the rate-based models. In contrast, a spike-based model may capture these neglected features. Nevertheless, how WTA computation can be implemented and its algorithmic characterization remains relatively under-explored.

2 Computational Model: Spiking Neuron Networks

In this section, we provide a general description of the computation model used. There is much freedom in choosing the detailed specification of the model. Later, we provide a circuit construction (for solving the k–WTA competition) under this computation model.

Network Structure We assume that a SNN can be partitioned into three non-overlapping layers: *input layer* N_{in}, *hidden layer* N_h, and *output layer* N_{out}. The synapses E are essentially *directed* edges, i.e, $E := \{(\nu, \nu') : \ \nu, \nu' \in U\}$. For each $\nu \in U$, define $\mathsf{PRE}_\nu := \{\nu' : (\nu', \nu) \in E\}$ and $\mathsf{POST}_\nu := \{\nu' : (\nu, \nu') \in E\}$. Intuitively, PRE_ν is the collection of neurons that can directly influence neuron ν; similarly, POST_ν is the collection of neurons that can be directly influenced by neuron ν. [1] We assume that the input neurons cannot be influenced by other neurons in the network, i.e., $\mathsf{PRE}_\nu = \varnothing$ for all $\nu \in N_{in}$. Each edge (ν, ν') in E has a *weight*, denoted by $\mathsf{w}(\nu, \nu')$. The family of WTA circuits under consideration is rather generic. We only assume that $|N_{in}| = |N_{out}| = n$ the numbers of the input neurons and of the output neurons are equal. Denote $N_{in} = \{u_1, \cdots, u_n\}$, and $N_{out} = \{v_1, \cdots, v_n\}$. The hidden neuron subset N_h can be arbitrary. The output neurons and the hidden neurons may be connected to each other in an arbitrary manner.

Network State The communication among neurons is abstracted as spikes. We assume each neuron ν has two local variables: *spiking state* variable $S(\nu)$ and *memory state* variable $M(\nu)$. Nevertheless, for input neurons, we only consider their spiking states, assuming that their memory states are not influenced by the dynamics of the spiking neuron network under consideration. We slot the time evenly with each time slot of length $1\,ms$. Let $t = 1, 2, \cdots$ be the indices of the time slots. For $t \ge 1$, let $S_t(\nu) \in \{0, 1\}$ be the spiking state of neuron ν at time t indicating whether neuron ν spikes at time t or not. By convention, $S_0(\nu) := 0$. For a non-input neuron ν and for $t \ge 1$, let $M_t(\nu)$ be the memory state of neuron ν at time t summarizing the cumulative influence caused by the spikes of the neurons in PRE_i during the most recent m times, i.e., times $t-1, t-2, \cdots, t-m$. Concretely, let $V_t(\nu)$ be the charge of (non-input) neuron ν at time t (for $t \ge 1$) defined as $V_t(\nu) := \sum_{\nu' \in \mathsf{PRE}_\nu} w(\nu', \nu) S_t(\nu')$. Clearly, $V_0(\nu) = 0$. Let $\boldsymbol{V}_t^\nu$ be the sequence of length m such that $\boldsymbol{V}_t^\nu := [V_t(\nu), \cdots, V_{t-m+1}(\nu)]$, and let $\boldsymbol{S}_t(\nu)$ be the sequence of length m such that $\boldsymbol{S}_t^\nu := [S_t(\nu), \cdots, S_{t-m+1}(\nu)]$. For $t \ge 1$, define the memory variable $M_t(\nu)$ as a pair of vectors $\boldsymbol{S}_{t-1}^\nu$ and $\boldsymbol{V}_{t-1}^\nu$, i.e., $M_t(\nu) := \left(\boldsymbol{S}_{t-1}^\nu, \boldsymbol{V}_{t-1}^\nu\right)$. By convention, let $M_0(\nu) := (\mathbf{0}, \mathbf{0})$, where $\mathbf{0}$ is the length m zero vector.

At time $t+1$, the memory variable $M_{t+1}(\nu)$ is updated by shifting the two sequences forwards by one time unit – fetching in $S_t(\nu)$ and $V_t(\nu)$, respectively, and removing $S_{t-m}(\nu)$ and $V_{t-m}(\nu)$, respectively. The memory state $M_t(\nu)$ is known to neuron ν only, and it can influence the probability of generating a spike at time t through an activation function ϕ_ν, i.e., $S_t(\nu) = \phi_\nu\left(M_t(\nu)\right), \forall\ t \ge 0$. Notably, ϕ_ν might be random.

Minimax Decision Accuracy/Success Probability We study the k–WTA model, wherein n randomly generated input spike trains are competing with each other, and, as a result of this competition, k out of them are selected to be the winners. In contrast, most existing works assume deterministic input spike trains. We assume that the n input spike trains are generated from n independent Bernoulli processes with unknown parameters $p_1, \cdots, p_n$, respectively. We refer to $\boldsymbol{p} = [p_1, \cdots, p_n]$ as a *rate assignment*. For example, suppose there are 2 input spike trains with rates 0.6 and 0.8, respectively, i.e., $n = 2$ and $\boldsymbol{p} = [0.6, 0.8]$. In each time, with probability 0.6 the first input spike train has a spike independently from whether the second input spike train has a spike or not; similarly for the second input spike train.

We adopt the minimax framework (in which the circuit designer and nature play games against each other) to evaluate the performance (decision accuracy versus decision time) of a WTA circuit.

Let $\mathcal{R} \subseteq [c, C]$ be an arbitrary but finite set of rates where c and C are two absolute constants such that $0 < c < C < 1$. A rate assignment $\boldsymbol{p}$ is chosen by nature from $\mathcal{R}^n$ for which there exists a subset of

[1] In the languages of computational neuroscience, the incoming neighbors and outgoing neighbors are often referred to as pre-synaptic units and post-synaptic units.

$[n] := \{1, \cdots, n\}$, denoted by $\mathcal{W}(\boldsymbol{p})$, such that

$$|\mathcal{W}(\boldsymbol{p})| = k, \quad \text{and } p_i > p_j \;\; \forall\, i \in \mathcal{W}(\boldsymbol{p}), j \notin \mathcal{W}(\boldsymbol{p}) \tag{1}$$

– recall that $|\cdot|$ is the cardinality of a set. We refer to set $\mathcal{W}(\boldsymbol{p})$ as the true winners with respect to the rate assignment $\boldsymbol{p}$. In this paper, we consider the following collection of rate assignments, denoted by $\mathcal{AR} \subset \mathcal{R}^n$:

$$\mathcal{AR} := \{\boldsymbol{p} : \exists \mathcal{W}(\boldsymbol{p}) \subseteq [n] \, s.t. \; |\mathcal{W}(\boldsymbol{p})| = k, \text{and } p_i > p_j \;\forall\, i \in \mathcal{W}(\boldsymbol{p}), j \notin \mathcal{W}(\boldsymbol{p})\}. \tag{2}$$

For each of reference, we refer to an element in $\mathcal{AR}$ as an admissible rate assignment. Recall that the input of a WTA circuit is a collection of n independent spike trains. For a given rate assignment $\boldsymbol{p}$, let $\{S_t(u_i)\}_{t=1}^T$ denote the spike train of length T at input neuron u_i. The circuit designer wants to design a WTA circuit that outputs a good guess/estimate $\widehat{\boldsymbol{win}}$ of $\mathcal{W}(\boldsymbol{p})$ for any choice of rate assignment $\boldsymbol{p}$ in $\mathcal{AR}$. Note that conditioning on

$$\boldsymbol{S} := \left[\{S_t(u_1)\}_{t=1}^T, \cdots, \{S_t(u_n)\}_{t=1}^T\right],$$

the estimate $\widehat{\boldsymbol{win}}$ is independent of $\boldsymbol{p}$. Here $\boldsymbol{S}$ is used with a little abuse of notation as this notation hides its connection with T and the rate parameter $\boldsymbol{p}$.[2]

3 Main Results

Information-theoretic lower bound We first provide a lower bound on the decision time for a given decision accuracy. The lower bounds derived in this section hold universally for all possible network structures (including the hidden layer), synapse weights, and the activation functions.

The decision time is naturally lower bounded by the sample complexity, which is closely related to the Kullback-Leibler (KL) divergence between two Bernoulli distributions. The KL divergence between Bernoulli random variables with parameters r and r', respectively, is defined as $d(r \parallel r') := r \log\left(\frac{r}{r'}\right) + (1-r)\log\left(\frac{1-r}{1-r'}\right)$, where, by convention, $\log \frac{0}{0} := 0$. For the given $\mathcal{R}$, define task complexity $T_{\mathcal{R}}$ as

$$T_{\mathcal{R}} := \max_{r_1, r_2 \in \mathcal{R} \, s.t. \; r_1 \neq r_2} \frac{1}{d(r_2 \parallel r_1) + d(r_1 \parallel r_2)}. \tag{3}$$

It is closely related to the smallest KL divergence between two distinct statistics in $\mathcal{R}$. The task complexity $T_{\mathcal{R}}$ kicks in due to the adoption of minimax decision framework.

It turns out that if the input spike train length T is not sufficiently large (specified in Theorem 1), no matter how elegant the design of a WTA circuit is (no matter which activation function we choose, how many hidden neurons we use, and how we connect the hidden neurons and output neurons), its actual decision accuracy is always lower than the target decision accuracy $1-\delta$.

Theorem 1. *For any $1 \le k \le n-1b$, set $\mathcal{R}$ and $\delta \in (0,1)$, if $T \le ((1-\delta)\log(k(n-k)+1) - 1)\, T_{\mathcal{R}}$, then*

$$\min_{\widehat{\boldsymbol{win}}} \max_{\boldsymbol{p} \in \mathcal{AR}} \mathbb{P}\left\{\widehat{\boldsymbol{win}}\,(\boldsymbol{S}) \neq \mathcal{W}(\boldsymbol{p})\right\} \;\ge\; \delta,$$

where the min is taken over all WTA circuits with different choices of activation functions and architectures.

Theorem 1 says that if $T < ((1-\delta)\log(k(n-k)+1) - 1)\, T_{\mathcal{R}}$, the worst case probability error of any WTA circuit is greater than δ, i.e., $\max_{\boldsymbol{p} \in \mathcal{AR}} \mathbb{P}\left\{\widehat{\boldsymbol{win}}\,(\boldsymbol{S}) \neq \mathcal{W}(\boldsymbol{p})\right\} > \delta$. Following our line of argument, by considering a richer family of critical rate assignments, we might be able to obtain a tighter lower bound. Nevertheless, the constructed WTA circuit presented later turn out to be order-optimal – its decision time matches the lower bound in Theorem 1 up to a multiplicative constant factor. This immediately implies that the lower bound obtained in Theorem 1 is tight up to a multiplicative constant factor.

[2] A more rigorous notation should be $\boldsymbol{S}(T, \boldsymbol{p}) := \left[\{S_t(u_1)\}_{t=1}^T, \cdots, \{S_t(u_n)\}_{t=1}^T\right]$. We use $\boldsymbol{S}$ for $\boldsymbol{S}(T, \boldsymbol{p})$ for ease of exposition.

Order-Optimal WTA Circuits In Section 2, we provided a general description of the computation model we are interested in. Next, we construct a specific WTA circuit under this general computation model. This WTA circuit turns out to be order-optimal in terms of decision time – the decision time of our WTA circuit matches the lower bound in Theorem 1 up to a multiplicative constant factor. To do that, we need to specify (1) the network structure, including the number of hidden neurons, the collection of synapses (directed communication links) between neurons, and the weights of these synapses; (2) the memorization capability of each neuron, i.e., the magnitude of m; and (3) ϕ_ν – the activation function used by neuron ν. The dynamics of our WTA circuit is summarized in Algorithm 1.

In our proposed circuit, we require that m satisfies the following:

$$m \geq \frac{8C^2(1-c)}{c^2(1-C)}\left(\log\left(\frac{3}{\delta}\right) + \log k(n-k)\right) T_{\mathcal{R}} \ := \ m^* \tag{4}$$

for target decision accuracy $1-\delta \in (0,1)$. In addition, we set $b = cm^*$. Recall that $c, C \in (0,1)$ are two absolute constants that are lower bound and upper bound of any $\mathcal{R}$, respectively. For Algorithm 1, we declare the first k output neurons that spike simultaneously to be winners.

```
Algorithm 1: $k$–WTA
Input: $\mathcal{R}$, $m$, $b$, and $\delta$.

for $t \geq 1$ do
    At output neuron $v_i$ for $i = 1, \cdots, n$: $V_{t-1}(v_i) \leftarrow S_{t-1}(u_i) - \frac{1}{k}\sum_{j:1\leq j\leq n, \& j\neq i} S_{t-1}(v_j)$;
      $\boldsymbol{V}_{t-1}(v_i) \leftarrow [V_{t-1}(v_i), V_{t-2}(v_i), \cdots, V_{t-m}(v_i)]$;
      $\boldsymbol{S}_{t-1}(v_i) \leftarrow [S_{t-1}(v_i), S_{t-2}(v_i), \cdots, S_{t-m}(v_i)]$;
      $M_t(v_i) \leftarrow (\boldsymbol{V}_{t-1}(v_i), \boldsymbol{S}_{t-1}(v_i))$;
    if $(b-1)\mathbf{1}_{\{S_{t-1}(v_i)=1\}} + \left[\sum_{r=1}^{m}\mathbf{1}_{\{V_{t-r}(v_i)>0\}} - m\sum_{r=1}^{m}\mathbf{1}_{\{V_{t-r}(v_i)\leq -1\}}\right]_+ \geq b$ then
        $S_t(v_i) \leftarrow 1$.
    else
        $S_t(v_i) \leftarrow 0$.
```

Recall that $\mathcal{W}(\boldsymbol{p})$ and m^* are defined in (1) and (4), respectively.

Theorem 2. *Fix $\delta \in (0,1]$, and $1 \leq k \leq n-1$. Choose $m \geq m^*$ and $b = \max\{cm^*, 2\}$. Then for any admissible rate assignment $\boldsymbol{p}$, with probability at least $1-\delta$, the following hold:*

(1) There exist k output neurons that spike simultaneously by time m^*.

(2) The first set of such k output neurons are the true winners $\mathcal{W}(\boldsymbol{p})$.

(3) From the first time in which these k output neurons spike simultaneously, these k output neurons spike consecutively for at least b times, and no other output neurons can spike within b times.

The first bullet in Theorem 2 implies that our WTA circuit can provide an output (a selection of k output neurons) by time m^*; the second bullet in Theorem 2 says that the circuit's output indeed corresponds to the k true winners; and the third bullet says that the k simultaneous spikes of the selected winners are stable – the k selected winners continue to spike consecutively for at least b times. The proof of Theorem 2 essentially says that with high probability, under Algorithm 1, the number of output neurons that spike simultaneously is monotonically increasing until it reaches k. Upon the simultaneous spike of k output neurons, by our threshold activation rule, we know that the other output neurons are likely to be inhibited. In particular, if these k output neurons are the first k output neurons that spike simultaneously, then the activation of the other output neurons are likely to be inhibited for at least b times.

Remark 3 (Order-optimality)**.** The decision time performance stated in (1) of Theorem 2 matches the information-theoretical lower bound in Theorem 1 up to a multiplicative constant factor both (a) when δ is sufficiently small and does not depend on n, k, $T_{\mathcal{R}}$, c, and C, and (b) when δ decays to zero at a speed at most $\frac{1}{(k(n-k))^{c_0}}$ where $c_0 > 0$ is some fixed constant. The detailed order-optimality argument can be found in the full version of this paper.