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

Lili Su

Computer Science & Artificial Intelligence Laboratory

Massachusetts Institute of Technology

lilisu@mit.edu

Chia-Jung Chang

Brain and Cognitive Sciences

Massachusetts Institute of Technology

chiajung@mit.edu

Nancy Lynch

Computer Science & Artificial Intelligence Laboratory

Massachusetts Institute of Technology

lynch@csail.mit.edu

August 6, 2019

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 firing rates, 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. We analytically characterize the minimum waiting 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 show that to guarantee a (minimax) decision error $\leq \delta$ (where $\delta \in (0,1)$), the waiting time of any WTA circuit is at least

$$((1-\delta)\log(k(n-k)+1)-1)T_{\mathcal{R}},$$

where $\mathcal{R} \subseteq (0,1)$ is a finite set of rates, and $T_{\mathcal{R}}$ is a difficulty parameter of a WTA task with respect to set $\mathcal{R}$ for independent input spike trains. Additionally, $T_{\mathcal{R}}$ is independent of δ, n, and k. We then design a simple WTA circuit whose waiting time is

$$O\left(\left(\log\left(\frac{1}{\delta}\right)+\log k(n-k)\right)T_{\mathcal{R}}\right),$$

provided that the local memory of each output neuron is sufficiently long. It turns out that for any fixed δ, this decision time is order-optimal (i.e., it

matches the above lower bound up to a multiplicative constant factor) in terms of its scaling in n, k, and $T_{\mathcal{R}}$.

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 (Kinoshita & Komatsu, 2001; Hubel & Wiesel, 1959), feature integration (Kourtzi, Tolias, Altmann, Augath, & Logothetis, 2003; Majaj, Carandini, & Movshon, 2007), decision formation (Platt & Glimcher, 1999; Shadlen & Newsome, 2001), and motor planning (Harris & Wolpert, 1998; N. Li, Chen, Guo, Gerfen, & Svoboda, 2015). It has been shown that neurons encode information in a stochastic manner in the brain (Baddeley et al., 1997; Kara, Reinagel, & Reid, 2000; Maimon & Assad, 2009; Ferrari, Deny, Marre, & Mora, 2018); 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 (Faisal, Selen, & Wolpert, 2008; Knill & Pouget, 2004; Stocker & Simoncelli, 2006; Ernst & Banks, 2002; Körding & Wolpert, 2004). 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 (Olshausen & Field, 2004; Perez-Orive et al., 2002; Hromádka, DeWeese, & Zador, 2008; Quiroga, Kreiman, Koch, & Fried, 2008; Karlsson & Frank, 2008; Redgrave, Prescott, & Gurney, 1999). 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 (Riesenhuber & Poggio, 1999; Itti, Koch, & Niebur, 1998; Yuille & Geiger, 1998; Maass, 2000; Hertz, Krogh, Palmer, & Horner, 1991; Shamir, 2006). 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 (Gerstner, Kempter, van Hemmen, & Wagner, 1996; Nelken, 2004). This phase-locking property is also observed in the hippocampus as well as the prefrontal cortex (Siapas, Lubenov, & Wilson, 2005; Hahn, Sakmann, & Mehta, 2006; Buzsáki & Chrobak, 1995). 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 (Kriener, Chaudhuri, & Fiete, 2017; S. Li, Li, & Wang, 2013; Lee, Itti, Koch, & Braun, 1999; Rougier & Vitay, 2006) 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 (Kriener et al., 2017). Last but not least, neurons have a refractory period, which prevents spikes from back propagating in axons (Berry II & Meister, 1998), 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 (Shamir, 2006, 2009).

In this paper, we study a spike-based k-WTA model wherein n randomly generated input spike trains are competing with each other with their underlying firing rates, and the true winners are the k input spike trains whose underlying firing rates are higher than others (Hertz et al., 1991). A desired WTA circuit should quickly respond to these random input spike trains and should successfully select the k true winners with high probability. We analytically characterize the minimum amount of waiting time needed so that a target minimax decision accuracy (defined in Section 3.2) can be reached. More precisely, we slot the time evenly with each time slot of length $1\,ms$, and assume that these n input spike trains are generated by n independent Bernoulli processes with different rates. We use Bernoulli processes to capture the randomness in the input spike trains rather than using the popular Poisson processes because a Bernoulli process can be viewed as the time-slotted version of a refractory-period-modified Poisson process. Notably, a Bernoulli process with $1\,ms$ time slot is just a simplified approximation to the real dynamics in the brain, given that, in the brain, the refractory period varies across neurons and the refractory period of some neuron could extend beyond $1\,ms$. In our model, we implicitly assume that the absolute refractory period is $1\,ms$, a value commonly reported in the literature (Teleńczuk, Kempter, Curio, & Destexhe, 2017; Nicholls, Martin, Wallace, & Fuchs, 2001). [1] A WTA circuit contains n output neurons, each of which is paired with an input spike train. What's more, the behaviors (spike patterns) of these output neurons encode which input spike trains are declared to be the winners. For special case where $k = 1$, different winner declaration strategies are considered in the literature (Shamir, 2006, 2009; Lynch, Musco, & Parter, 2016; Kriener et al., 2017), such as the identity of output neuron that spikes much more frequently than the other output neurons (Kriener et al., 2017), of the neuron that fires the first spike in a population of neurons (Shamir, 2009, 2006), and of the output neuron that fires alone for a sufficiently long time (Lynch et al., 2016). Clearly, the minimum amount of waiting time needed to achieve a given accuracy varies with the choice of winner declaration strategy. Nevertheless, in order to derive a lower bound that holds for all winner declaration strategies, at this point, we do not specify the winner declaration strategy used in our circuit construction – this specification is postponed to Section 5. In this paper, we investigate the following two closely related problems: (1) the fundamental limits of any WTA circuit in selecting k true winners from n independent Bernoulli input spike trains (in terms of waiting time to achieve a target accuracy), and (2) the existence of WTA circuits that can achieve the

[1] We plan to investigate the impact of the heterogeneity in the refractory period on waiting time in our future work.

above fundamental limits.

To answer the first question, we consider a general model (formally described in Section 2) without restricting the adopted network architectures, activation functions, winner declaration strategies, etc. so that the derived lower bound can provide guidance and insight for constructing a large family of WTA circuits. We derive a lower bound on the waiting/decision time in order to achieve a given decision accuracy. We show that no WTA circuit can have a waiting time strictly less than

$$((1-\delta)\log(k(n-k)+1)-1)T_{\mathcal{R}}, \tag{1}$$

where $\mathcal{R} \subseteq [c, C] \subseteq (0,1)$ is a finite set of rates, $T_{\mathcal{R}}$ is a difficulty parameter of a WTA task with respect to set $\mathcal{R}$ for independent input spike trains, n is the number of input spike trains, k is the number of winners, and $(1-\delta)$ is the given target decision accuracy. Here c, C are two absolute constants such that $0 < c < C < 1$, and $\delta \in (0,1)$. Moreover, $T_{\mathcal{R}}$ is independent of δ, n, and k. In many practical settings we care about the sparse coding region where $k \ll n$. Not surprisingly, the above lower bound grows with the network size n when other parameters are fixed. This is because the larger n, the noisier the WTA competition. Similarly, when n and k are fixed, the easier to distinguish two independent spike trains with different rates (i.e., the smaller $T_{\mathcal{R}}$), the shorter the necessary decision time is. Our lower-bound is obtained by an information-theoretic argument, and holds for all WTA circuits without restricting their winner declaration strategies, circuit architectures, and the adopted activation functions. Throughout this paper, we are interested in the decision time's scaling in n, k, and $T_{\mathcal{R}}$, while treating $\delta \in (0,1)$

as a small but fixed constant.

To answer the second question, we construct a simple circuit whose decision time is

$$O\left(\left(\log\left(\frac{1}{\delta}\right) + \log k(n-k)\right) T_{\mathcal{R}}\right),$$

provided that the local memory of each output neuron is sufficiently long [2]. In this circuit, there are n pairs of input and output neurons, and no hidden neurons. Each input neuron is connected to the corresponding output neuron, and the n output neurons mutually inhibit each other. Each output neuron has a local memory of length m (formally defined in Section 2.2), and adopts a simple threshold activation function (specified in Section 5.1.3). The first k output neurons that spike in the same time slot are declared to be the winners; the identities of such k output neurons are the circuit's estimate of the k true winners. The formal circuit construction can be found in Section 5. We show that for any fixed $\delta \in (0,1)$, provided that

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

with probability at least $1-\delta$ it holds that: By time m^*, there exist exactly k output neurons that spike in the same time slot, and the first set of such k output

[2] In this paper, the notations $O(\cdot)$ and $\Omega(\cdot)$ are used to describe the limiting behavior of a function when the argument tends towards a particular value or infinity. In our case the waiting time can be viewed as a "function" of several other parameters such as δ, $\mathcal{R}$, n, and k. Formally, for any sequences $\{a_N\}$ and $\{b_N\}$, we say $a_N = O(b_N)$ if there exists an absolute constant $c > 0$ such that $a_N \le c \times b_N$. Similarly, we say $a_N = \Omega(b_N)$ is there exists an absolute constant $c > 0$ such that $a_N \ge c \times b_N$.

neurons are indeed the true winners. It turns out that this decision time (m^*) is order-optimal in terms of its scaling in n, k, and $T_{\mathcal{R}}$; m^* matches the lower-bound in (1) up to a constant multiplicative factor. The formal argument showing order-optimality can be found in Remark 11. In a sense, the local memory of each output neuron plays a crucial role in "denoising" the randomness in the input spike trains. In practice, an output neuron's local memory might not satisfy the above condition in (2). Nevertheless, this does not exclude the application of our WTA circuit to the contexts where m is small. This is because the memory variable might be implemented via some neural code near an output neuron. The detailed implementation of the local memory only affects the circuit's architecture; it does not affect the order optimality of our WTA circuit. The typical dynamics of our circuit are: The number of output neurons that spike simultaneously (i.e., spike at the same time) increases monotonically until exactly k output neurons spike simultaneously. The simultaneous spikes of these k output neurons cause strong inhibition of other output neurons; in particular, no other output neuron can spike within a sufficiently long period $\Omega\left(\left(\log\left(\frac{1}{\delta}\right) + \log k(n-k)\right) T_{\mathcal{R}}\right)$.

In addition, our results also give a set of testable hypotheses on neural recordings and human/animal behaviors in decision-making; detailed discussion can be found in Section 6.

2 Computational Model: Spiking Neuron Networks

In this section, we provide a general description of our computation model; there is much freedom in choosing the detailed specification of the model. We consider such a general model so that our derived lower bound applies to WTA circuits with many alternative network architectures, activation functions, winner declaration strategies (i.e., the desired behaviors of the output neurons), etc. In Section 5 we provide a circuit construction (for solving the k–WTA competition) under this computation model but with specific choices for the adopted network architecture, activation function, and winner declaration strategy.

2.1 Network Structure

A *spiking neuron network* (SNN) $\mathcal{N} = (U, E)$ consists of a collection of neurons U that are connected through synapses E. We assume that a SNN can be conceptually partitioned into three non-overlapping layers: *input layer* N_{in}, *hidden layer* N_h, and *output layer* N_{out}; the neurons in each of these layers are referred to as *input neurons*, *hidden neurons*, and *output neurons*, respectively. 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. [3] 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 strength of the interaction between neuron ν and neuron ν' is captured as $|\mathsf{w}(\nu, \nu')|$. The sign of $\mathsf{w}(\nu, \nu')$ indicates whether neuron ν excites or inhibits neuron ν': In particular, if neuron ν excites neuron ν', then $\mathsf{w}(\nu, \nu') > 0$; if neuron ν inhibits neuron ν', then $\mathsf{w}(\nu, \nu') < 0$. The set E might contain self-loops with $\mathsf{w}(\nu, \nu)$ capturing the self-excitatory/self-inhibitory effects. Typically, in neuroscience a neuron is either excitatory or inhibitory, i.e., $\mathsf{sign}(\mathsf{w}(\nu, \nu_1)) = \mathsf{sign}(\mathsf{w}(\nu, \nu_2))$ for all $\nu_1, \nu_2 \in \mathsf{POST}_\nu$. Our order-optimal WTA circuit in Section 5 indeed assumes this common sign restriction. Nevertheless, our lower bound holds even for the general case where there exist $\nu_1, \nu_2 \in \mathsf{POST}_\nu$ such that $\mathsf{sign}(\mathsf{w}(\nu, \nu_1)) \neq \mathsf{sign}(\mathsf{w}(\nu, \nu_2))$.

Generic network structure for WTA circuits 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. For ease of exposition, denote

$$N_{in} = \{u_1, \cdots, u_n\}, \text{ and } N_{out} = \{v_1, \cdots, v_n\}.$$

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.

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

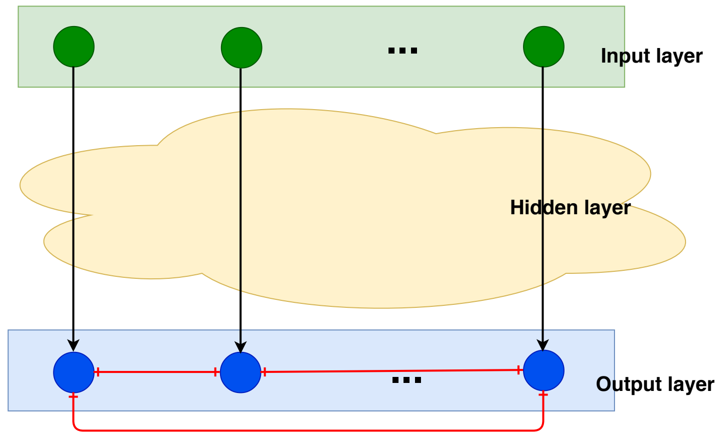

Figure 1: A SNN consists of three layers: the input layer, the output layer, and the hidden layer. The hidden neurons might connect to both the input neurons and the output neurons to assist the computation of the neuron network. Neurons are connected through synapses. WTA circuits is a family of SNNs in which the number of output neurons equals the number of the input neurons.

2.2 Network State

In a SNN, 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. Henceforth, by saying time t, we mean the time interval $[t-1, t)\,ms$. For $t \geq 1$, 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. For a non-input neuron ν and for $t \geq 1$, 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 $t \geq 1$) defined as

$$V_t(\nu) := \sum_{\nu' \in \mathsf{PRE}_\nu} w(\nu', \nu) S_t(\nu').$$

Let $\boldsymbol{V}_t^\nu$ be the sequence of length m such that

$$\boldsymbol{V}_t^\nu := \left[V_t(\nu), \cdots, V_{t-m+1}(\nu)\right],$$

and let $\boldsymbol{S}_t(\nu)$ be the sequence of length m such that

$$\boldsymbol{S}_t^\nu := \left[S_t(\nu), \cdots, S_{t-m+1}(\nu)\right].$$

By convention, when $1 \leq t \leq m$, let

$$\boldsymbol{V}_t^\nu := [V_t(\nu), \cdots, V_1(\nu), 0, \cdots, 0]$$

and

$$\boldsymbol{S}_t^\nu := \left[S_t(\nu), \cdots, S_1(\nu), 0, \cdots, 0\right].$$

For $t \geq 2$, 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 $M_1(\nu) := (\mathbf{0}, \mathbf{0})$, where $\mathbf{0}$ is the length m zero vector. Notably, as can be seen from our analysis, our lower bound holds provided that $M_1(\nu)$ does not contain any information about the circuit's dynamics for time $t \leq 0$, i.e., no information on the past $t \leq 0$ is used in determining the generation of a spike at time $t \geq 1$.

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.,

$$S_t(\nu) = \phi_\nu \left(M_t(\nu) \right), \forall\ t \geq 1. \tag{3}$$

Notably, ϕ_ν might be a random function.

In most neurons, the synaptic plasticity time window is about 80 -120 msec, but could also vary across brain regions, and vary across different time scales under different behavioral contexts. In a sense, the synaptic plasticity time window is closely related to m. As can be seen in Section 5, our order-optimal WTA circuit construction requires m to be sufficiently high. Nevertheless, this does not exclude the application of our WTA circuit to the contexts where m is small. This is because the memory variable can be implemented by a chain of hidden neurons near neuron ν. The detailed implementation of the local memory does not affect the order optimality of our WTA circuit.

3 Minimax Decision Accuracy/Success Probability

3.1 Random Input Spike Trains

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 (Verzi et al., 2018; Maass, 1997; Lynch et al., 2016) assume deterministic input spike trains.

Recall that time is slotted into intervals of length $1\,ms$. 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* of the WTA competition for a given external stimulus. For example, suppose an external stimulus induces 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. Notably, different external stimuli induce different rate assignment vectors $\boldsymbol{p}$'s. Henceforth, we use the terms "rate assignment" and "external stimulus" interchangeably.

Note that in the most general scenario the spikes of the input neurons might be correlated; see Section 6 for detailed comments. We would like to explore the more general input spikes in our future work.

3.2 Minimax Performance Metric

We adopt the minimax framework (Wu, 2017) 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 $\boldsymbol{p}$ (i.e., an external stimulus) is chosen by nature from $\mathcal{R}^n$ for which there exists a subset of $[n] := \{1, \cdots, n\}$, denoted by $\mathcal{W}(\boldsymbol{p})$, such that

$$|\mathcal{W}(\boldsymbol{p})| = k, \quad \text{and } p_i > p_j \ \ \forall\, i \in \mathcal{W}(\boldsymbol{p}), j \notin \mathcal{W}(\boldsymbol{p}) \tag{4}$$

– recall that $|\cdot|$ is the cardinality of a set. That is, $\mathcal{W}(\boldsymbol{p})$ is the set of true winners that should be selected when the external stimulus that induces $\boldsymbol{p}$ is presented. We refer to set $\mathcal{W}(\boldsymbol{p})$ as the true winners with respect to the rate assignment $\boldsymbol{p}$. For example, suppose $n = 5$, $k = 2$, and

$$\boldsymbol{p} \;=\; [p_1 = 0.2,\ p_2 = 0.1,\ p_3 = 0.2,\ p_4 = 0.8,\ p_5 = 0.85]\,.$$

Here the true winners are 4 and 5, i.e., $\mathcal{W}(\boldsymbol{p}) = \{4, 5\}$. In this paper, we consider the following collection of rate assignments, denoted by $\mathcal{AR} \subset \mathcal{R}^n$:

$$\mathcal{AR} := \{\boldsymbol{p} : \exists \mathcal{W}(\boldsymbol{p}) \subseteq [n]\, s.t.\ |\mathcal{W}(\boldsymbol{p})| = k, \text{and } p_i > p_j \ \forall\, i \in \mathcal{W}(\boldsymbol{p}), j \notin \mathcal{W}(\boldsymbol{p})\}\,. \tag{5}$$

Intuitively, $\mathcal{AR}$ corresponds to the collection of external stimuli considered. For ease 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 $\boldsymbol{p}$.[4] Later, we use the same notation to denote the n spike trains with random rate assignment, i.e., where $\boldsymbol{p}$ is randomly generated. Nevertheless, this abuse of notation significantly simplifies the exposition without sacrificing clarity.

Under minimax framework, we are interested in the minimax error probability

$$\min_{\widehat{\boldsymbol{win}}} \max_{\boldsymbol{p} \in \mathcal{AR}} \mathbb{P} \left\{ \widehat{\boldsymbol{win}} \left(\boldsymbol{S} \right) \neq \mathcal{W}(\boldsymbol{p}) \right\}. \tag{6}$$

For a given deterministic WTA circuit $\widehat{\boldsymbol{win}}$ (i.e., the activation functions used are deterministic), the probability in $\mathbb{P} \left\{ \widehat{\boldsymbol{win}} \left(\boldsymbol{S} \right) \neq \mathcal{W}(\boldsymbol{p}) \right\}$ is taken w.r.t. the randomness in the stochastic spikes of each input neuron; for a randomized WTA circuit $\widehat{\boldsymbol{win}}$ (i.e., the activation functions are stochastic), in addition to the aforementioned source of randomness, the probability in $\mathbb{P} \left\{ \widehat{\boldsymbol{win}} \left(\boldsymbol{S} \right) \neq \mathcal{W}(\boldsymbol{p}) \right\}$ is also taken w.r.t. the randomness in the activation functions. In (6), the performance metric of a WTA circuit is the worst-case error probability

$$\max_{\boldsymbol{p} \in \mathcal{AR}} \mathbb{P} \left\{ \widehat{\boldsymbol{win}} \left(\boldsymbol{S} \right) \neq \mathcal{W}(\boldsymbol{p}) \right\}.$$

[4]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.

4 Information-Theoretic Lower Bound on Decision Time

In this section, we 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, activation functions, and winner declaration strategies.

One observation is that the decision time is naturally lower bounded by the sample complexity, which is closely related to the Kullback-Leibler (KL) divergence[5] between two Bernoulli distributions. The KL divergence between Bernoulli random variables with parameters r and r', respectively, is defined as

$$d(r \parallel r') \;:=\; r\log\left(\frac{r}{r'}\right) + (1-r)\log\left(\frac{1-r}{1-r'}\right), \qquad (7)$$

where, by convention, $0\log\frac{0}{0} := 0$. Notably, $d(\cdot \parallel \cdot)$ is not symmetric in r and r'. In addition, if $r \in (0,1)$ and $r' \in \{0,1\}$, then $d(r \parallel r') = \infty$. Recall that set $\mathcal{R}$ is an arbitrary but finite set that are contained in the interval $[c, C]$, where $c, C \in (0,1)$ are two constants. It holds that $d(r \parallel r') < \infty$ for all $r, r' \in \mathcal{R}$. For the more general distributions over a common discrete alphabet $\mathcal{A}$, say distributions P and Q, the Kullback-Leibler (KL) divergence between P and Q is defined as follows.

Definition 1 (KL-divergence)**.** Let $\mathcal{A}$ be a discrete alphabet (finite or countably infinite), and P and Q be two distributions over $\mathcal{A}$. Then define

$$D(P \parallel Q) := \sum_{a\in\mathcal{A}} P(a)\log\left(\frac{P(a)}{Q(a)}\right),$$

[5]The Kullback-Leibler (KL) divergence gauges the **dissimilarity** between two distributions.

where $0 \cdot \log\left(\frac{0}{0}\right) = 0$ by convention.

Note that $D(P \parallel Q) \geq 0$ and $D(P \parallel Q) = 0$ if and only if $P = Q$ except for measure 0. Similar to $d(\cdot \parallel \cdot)$, $D(P \parallel Q)$ is not symmetric in P and Q. In this paper, we choose the base to be 2. [6] Recall that the set of admissible rate assignments $\mathcal{AR}$ is defined in (5).

Lemma 2. *Fix a finite set $\mathcal{R}$ of rates. Let $\boldsymbol{p} = [p_1, \cdots, p_n]$ and $\boldsymbol{q} = [q_1, \cdots, q_n]$ be two rate assignments in $\mathcal{AR}$. Let $P_{\boldsymbol{S}}$ and $Q_{\boldsymbol{S}}$ be the distributions of the n spike sequences of the input neurons under rate assignments $\boldsymbol{p}$ and $\boldsymbol{q}$, respectively. Then,*

$$D(P_{\boldsymbol{S}} \parallel Q_{\boldsymbol{S}}) = T \sum_{i=1}^{n} d(p_i \parallel q_i).$$

The two different rate assignments $\boldsymbol{p}$ and $\boldsymbol{q}$ correspond to two different external stimuli, and $D(P_{\boldsymbol{S}} \parallel Q_{\boldsymbol{S}})$ is the "distance" between the n input spike trains of length T induced by the first external stimulus and those induced by the second external stimulus. Lemma 2 is proved in Appendix B.

For a given $\mathcal{R}$, define task complexity $T_{\mathcal{R}}$ as

$$T_{\mathcal{R}} := \max_{r_1, r_2 \in \mathcal{R} \, s.t. \; r_1 \neq r_2} \frac{1}{d(r_2 \parallel r_1) + d(r_1 \parallel r_2)}. \tag{8}$$

It is closely related to the smallest KL divergence between two distinct rates in $\mathcal{R}$. The task complexity $T_{\mathcal{R}}$ kicks in due to the adoption of minimax decision framework (6). It turns out that if the input spike train length T is not sufficiently large (specified in Theorem 3), no matter how elegant the design of a WTA circuit is (no matter which activation function we choose, how many hidden neurons we

[6] Note that any base would work, see (Polyanskiy & Wu, 2014, Chapter 1.1).

use, and how we connect the hidden neurons and output neurons), its minimax decision accuracy is always lower than the target decision accuracy $(1-\delta)$.

Theorem 3. *For any $1 \le k \le n-1$ and any set $\mathcal{R}$ and any $\delta \in (0,1)$, if*

$$T \le \left((1-\delta)\log(k(n-k)+1)-1\right) T_{\mathcal{R}},$$

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 possible WTA circuits with different choices of activation functions and circuit architectures.

Theorem 3 says that if $T < \left((1-\delta)\log(k(n-k)+1)-1\right) T_{\mathcal{R}}$, the worst case probability error of any WTA circuit is greater than δ, i.e.,

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

Theorem 3 is proved in Appendix C.

Remark 4 (Tightness of the lower bound in Theorem 3)**.** The proof of Theorem 3 uses a technical supporting lemma – Lemma 16 (presented in Appendix C). Following our line of argument, by considering a richer family of critical rate assignments in Lemma 16, we might be able to obtain a tighter lower bound. Nevertheless, the constructed WTA circuit in Section 5 turn out to be order-optimal – its decision time matches the lower bound in Theorem 3 up to a multiplicative constant factor. This immediately implies that the lower bound obtained in Theorem 3 is tight up to a multiplicative constant factor.

5 Order-Optimal WTA Circuits

In Section 2, we provided a general description of the computation model we are interested in. In this section, we construct a specific WTA circuit whose decision time is order-optimal among the WTA circuits under the general computation model. 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. In the constructed circuit, we declare the first k output neurons that spike simultaneously as winners.

5.1 Circuit Design

In our designed circuit, there are four parameters $\mathcal{R}$, m, b, and δ, where $\mathcal{R} \subseteq [c, C]$ [7] is a finite set of rates from which the p_i's of the input spike trains are chosen, m is the memory range and b is the bias at the non-input neurons, and $(1-\delta)$ is the target decision accuracy (i.e., success probability). Here, we assume that every non-input neuron has the same bias, i.e., $b_\nu = b$ for all non-input neurons ν. The four parameters $\mathcal{R}$, m, b, and δ can be viewed as some prior knowledge of the WTA circuit; they might be learned through some unknown network development procedure which is outside the scope of this work. In Sections 5.1.1, 5.1.3, and 5.1.4, we present the network structure and the activation functions adopted, and

[7] Recall that $c, C \in (0,1)$ are two absolute constants, i.e., they do not change with other parameters of the WTA circuit such as n, k, and δ.

the requirement on m. For completeness, we specify the local memory update (in particular the vector $\boldsymbol{V}$) separately in Section 5.1.2. The dynamics of our WTA circuit is summarized in Section 5.1.5.

5.1.1 Network structure:

We propose a WTA circuit with the following network structure:

- All output neurons are connected to each other by a complete graph. That is, $(v_i, v_j) \in E$ for all $v_i, v_j \in N_{out}$ such that $v_i \neq v_j$;
- Each edge from an input neuron to an output neuron has weight 1, i.e., $\mathsf{w}(u_i, v_i) = 1$ for all $u_i \in N_{in}, v_i \in N_{out}$.
- All edges among the output neurons have weights $-\frac{1}{k}$. That is, $\mathsf{w}(v_i, v_j) = -\frac{1}{k}$ for all $v_i, v_j \in N_{out}$ such that $v_i \neq v_j$.
- There are no hidden neurons, i.e., $N_h = \emptyset$;

5.1.2 Update local charge vector:

With the above choice of network structure, the charge $V_{t-1}(v_i)$ at the output neuron v_i at time $t-1$ is

$$V_{t-1}(v_i) \;=\; S_{t-1}(u_i) - \frac{1}{k} \sum_{j:1\le j\le n, \&\ j\neq i} S_{t-1}(v_j). \tag{9}$$

When $k = 1$, the above update becomes

$$V_{t-1}(v_i) \;=\; S_{t-1}(u_i) - \sum_{j:1\le j\le n, \&\ j\neq i} S_{t-1}(v_j).$$

which can be viewed as a spike model counterpart of the potential update under the traditional continuous rate model (Kriener et al., 2017; Mao & Massaquoi, 2007) with lateral inhibition.

It is easy to see the following claims hold. For brevity, their proofs are omitted.

Claim 5. For $t \geq 1$ and for $i = 1, \cdots, n$, $V_{t-1}(v_i) > 0$ if and only if $S_{t-1}(u_i) = 1$ and $\sum_{j:1\leq j\leq n,\&\ j\neq i} S_{t-1}(v_j) \leq k-1$, i.e., at time $t-1$, input neuron u_i spikes, and fewer than $k-1$ other output neurons spike.

Claim 6. For $t \geq 1$ and for $i = 1, \cdots, n$, $V_{t-1}(v_i) \leq -1$ only if $\sum_{j:1\leq j\leq n,\&\ j\neq i} S_{t-1}(v_j) \geq k$, i.e., at time $t-1$, more than k other output neurons spike.

Note that $\sum_{j:1\leq j\leq n,\&\ j\neq i} S_{t-1}(v_j) \geq k$ is not a sufficient condition to have $V_{t-1}(v_i) \leq -1$. To see this, suppose $\sum_{j:1\leq j\leq n,\&\ j\neq i} S_{t-1}(v_j) = k$ and $S_{t-1}(u_i) = 1$. In this case it holds that $V_{t-1}(v_i) = 0$.

Claim 7. For $t \geq 1$ and for $i = 1, \cdots, n$, if $V_{t-1}(v_i) = 0$, one of the following holds:

(1) $S_{t-1}(u_i) = 1$ and $\sum_{j:1\leq j\leq n,\&\ j\neq i} S_{t-1}(v_j) = k$, i.e., at time $t-1$, input neuron u_i spikes, and exactly k other output neurons spike;

(2) $S_{t-1}(u_i) = 0$ and $\sum_{j:1\leq j\leq n,\&\ j\neq i} S_{t-1}(v_j) = 0$, i.e., at time $t-1$, input neuron u_i does not spike, and no other output neurons spike.

5.1.3 Activation functions:

There are many different choices of activation functions; see (*Activation function*, n.d.) for a detailed list. In our construction, we use a simple threshold activation function, i.e.,

$$S_t(v_i) = \begin{cases} 1, & \text{if } (b-1)\mathbf{1}_{\{S_{t-1}(v_i)=1\}} + \left[\sum_{r=1}^m \mathbf{1}_{\{V_{t-r}(v_i)>0\}} - m\sum_{r=1}^m \mathbf{1}_{\{V_{t-r}(v_i)\leq -1\}}\right]_+ \geq b; \\ 0, & \text{otherwise,} \end{cases}$$

where $[\cdot]_+ = \max[\cdot, 0]$, and $b \geq 1$ is the bias at the output neuron v_i for $i = 1, \cdots, n$. It is easy to see that this activation function falls under the general form given by (3).

Remark 8. If the output neuron v_i does not spike at time $t-1$, i.e., $S_{t-1}(v_i) = 0$, then in order for v_i to spike at time t, the following needs to hold:

$$\left[\sum_{r=1}^m \mathbf{1}_{\{V_{t-r}(v_i)>0\}} - m\sum_{r=1}^m \mathbf{1}_{\{V_{t-r}(v_i)\leq -1\}}\right]_+ \geq b.$$

In contrast, if the output neuron v_i does spike at time $t-1$, i.e., $S_{t-1}(v_i) = 1$, then

$$\left[\sum_{r=1}^m \mathbf{1}_{\{V_{t-r}(v_i)>0\}} - m\sum_{r=1}^m \mathbf{1}_{\{V_{t-r}(v_i)\leq -1\}}\right]_+ \geq 1$$

is enough for v_i to spike at time t. That is, under our activation rule, $S_{t-1}(v_i) = 1$ makes the activation of v_i much easier in the next round. However, if there exists $r \in \{1, 2, \cdots, m\}$ such that

$$\mathbf{1}_{\{V_{t-r}(v_i)\leq -1\}} = 1,$$

then

$$\begin{aligned} \sum_{r=1}^m \mathbf{1}_{\{V_{t-r}(v_i)>0\}} - m\sum_{r=1}^m \mathbf{1}_{\{V_{t-r}(v_i)\leq -1\}} &\leq \sum_{r=1}^m \mathbf{1}_{\{V_{t-r}(v_i)>0\}} - m \\ &\leq m - m = 0. \end{aligned}$$

Thus,

$$
\begin{aligned}
&(b-1)\mathbf{1}_{\{S_{t-1}(v_i)=1\}} + \left[\sum_{r=1}^{m} \mathbf{1}_{\{V_{t-r}(v_i)>0\}} - m\sum_{r=1}^{m} \mathbf{1}_{\{V_{t-r}(v_i)\le -1\}}\right]_+ \\
&= (b-1)\mathbf{1}_{\{S_{t-1}(v_i)=1\}} + 0 \\
&\le b-1 < b,
\end{aligned}
$$

i.e., the output neuron v_i does not spike at time t. In other words, provided that there exists $r \in \{1, 2, \cdots, m\}$ such that $\mathbf{1}_{\{V_{t-r}(v_i)\le -1\}} = 1$, the activation of v_i is inhibited at time t.

5.1.4 Local memorization capability:

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

$$
m \ge \frac{8C^2(1-c)}{c^2(1-C)} \left(\log\left(\frac{3}{\delta}\right) + \log k(n-k)\right) T_{\mathcal{R}} \; := \; m^* \tag{10}
$$

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 and upper bounds, respectively, of any $\mathcal{R}$.

Intuitively, when other parameters are fixed, the higher the desired accuracy (i.e., the smaller δ) , the larger the required minimum memory m^*, i.e., the more memory is needed for selecting the winners in our WTA circuit. Similarly, the easier to distinguish two independent spike trains with different rates (i.e., the lower $T_{\mathcal{R}}$), the smaller m^*. Interesting, with other parameters fixed, m^* depends on k as follows: m^* is increasing in k when $k \in \left\{1, \cdots, \lfloor \frac{n}{2} \rfloor\right\}$, and m^* is decreasing in k when $k \in \left\{\lceil \frac{n}{2} \rceil, \cdots, n-1\right\}$. In many practical settings we care about the region where $k \ll n$. Besides, with the choice of bias $b = cm^*$, the larger m^* also implies longer time is needed for our WTA circuit to declare k winners; details can be found (1) in Theorem 9.

On the other hand, in most neurons the synaptic plasticity time window is about 80-120 ms, and it is unclear whether (10) can be immediately satisfied or not. Fortunately, even if (10) is not immediately satisfied by a neuron due to its local bio-plausibility, it is possible that its local memory might be realized via some population codes such as a chain of hidden neurons.

5.1.5 Algorithm 1

The dynamics of our WTA circuit is summarized in Algorithm 1, which is fully determined by what has been described in Sections 5.1.1, 5.1.2, 5.1.3, and 5.1.4. For Algorithm 1, we declare the first k output neurons that spike simultaneously as winners.

5.2 Circuit Performance

Recall that $\mathcal{W}(\boldsymbol{p})$ and m^* are defined in (4) and (10), respectively.

Theorem 9. *Fix $\delta \in (0, 1]$, and $1 \le k \le n-1$. Choose $m \ge m^*$ 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,

Algorithm 1: k–WTA

Input: $\mathcal{R}$, m, b, and δ.

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

    $S_t(v_i) \leftarrow 0$.

these k output neurons spike consecutively for at least b times, and no other output neurons can spike within b times.

The proof of Theorem 9 can be found in Appendix D. The first bullet in Theorem 9 implies that our WTA circuit can provide an output (a selection of k output neurons) by time m^*; the second bullet in Theorem 9 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 9 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 10 (Controlling stability)**.** As can be seen from the proof of Theorem 9, in the activation function of Algorithm 1

$$(b-1)\mathbf{1}_{\{S_{t-1}(v_i)=1\}} + \left[\sum_{r=1}^{m} \mathbf{1}_{\{V_{t-r}(v_i)>0\}} - m\sum_{r=1}^{m} \mathbf{1}_{\{V_{t-r}(v_i)\le -1\}}\right]_+ \ge b$$

the first term $(b-1)\mathbf{1}_{\{S_{t-r}(v_i)=1\}}$ is crucial in achieving (3) in Theorem 9. In fact, we can increase the stability period by introducing a stability parameter s such that $1 < s \le m$ and modifying the activation rule. Details can be found in Algorithm 2. It is easy to see that the activation function falls under the general form in (3). In the new activation function in Algorithm 2, for output neuron v_i, once it spikes, it continues to spike for at least s times. Following our line of analysis in the proof of Theorem 9, it can be seen that the declared k winners, from the first time they spike simultaneously, continue to spike consecutively for at least s times.

Algorithm 2: k–WTA

Input: $\mathcal{R}$, m, b, δ, and s where $1 < s \leq m$.

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 $\left[\sum_{r=1}^m \mathbf{1}_{\{V_{t-r}(v_i)>0\}} - m\sum_{r=1}^m \mathbf{1}_{\{V_{t-r}(v_i)\leq -1\}}\right]_+ \geq b$ **then**

        $S_t(v_i) \leftarrow 1$.

    else

        if $S_{t-1}(v_i) = 1$ *and* $\exists\, r \in \{2, \cdots, s\}$ *such that* $S_{t-r}(v_i) = 0$ **then**

            $S_t(v_i) \leftarrow 1$.

        else

            $S_t(v_i) \leftarrow 0$.

Remark 11 (Order-optimality)**.** The decision time performance stated in (1) of

Theorem 9 matches the information-theoretical lower bound in Theorem 3 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 is given next.

Suppose that δ is sufficiently small and does not depend on n, k, $T_{\mathcal{R}}$, c, and C Here, for ease of exposition, we illustrate the order-optimality with a specific choice of δ. In fact, the order-optimality holds generally for constant $\delta \in (0,1)$ provided that it does not depend on n, k, $T_{\mathcal{R}}$, c, and C.

Suppose the target decision accuracy is $1-\delta = 0.9$, i.e., $\delta = 0.1$. Then provided that $n \geq 31$, for any $1 \leq k \leq n-1$,

$$m^* = \frac{8C^2(1-c)}{c^2(1-C)}\left(\log\frac{3}{0.1} + \log k(n-k)\right)T_{\mathcal{R}} \leq \frac{16C^2(1-c)}{c^2(1-C)}\log k(n-k)T_{\mathcal{R}}.$$

On the other hand, recall from Theorem 3 that to have $\delta = 0.1$, the decision time is no less than

$$((1-\delta)\log(k(n-k)+1)-1)\,T_{\mathcal{R}} \geq \frac{1}{2}\log(k(n-k)+1)T_{\mathcal{R}} \geq \frac{1}{2}\log k(n-k)T_{\mathcal{R}}$$

where the first inequality holds provided that $n \geq 8$. Thus, when $n \geq 31$, in order to achieve the decision accuracy $1-\delta = 0.9$, the decision time of our WTA circuit is on the same order of the information-theoretic lower bound in Theorem 3.

Suppose δ decays to zero at a moderate speed The decision time of our WTA circuit is order-optimal even for diminishing decision error δ provided that $\delta = \Omega(\frac{3}{(k(n-k))^{c_0}})$ where $c_0 > 0$ – it does not decay to zero "too fast" in $k(n-k)$.

To see this, let $\delta = \frac{3}{(k(n-k))^{c_0}}$ for some constant $c_0 > 0$. We have

$$\frac{8C^2(1-c)}{c^2(1-C)}\left(\log\left(\frac{3}{\frac{3}{(k(n-k))^{c_0}}}\right) + \log k(n-k)\right) T_{\mathcal{R}}$$
$$= \frac{8C^2(1-c)(c_0+1)}{c^2(1-C)} \log k(n-k) T_{\mathcal{R}}. \tag{11}$$

Resetting circuit when the input spike trains become quiescent In Algorithm 1, if the input spike trains become quiescent, then the corresponding circuits also become quiescent despite some delay in this response.

Lemma 12. *If all input neurons are quiescent at time t_0, and remain to be quiescent for all $t \geq t_0$, then $V_t(v_i) = 0$ and $S_t(v_i) = 0$ for any $t > t_0 + m$.*

Lemma 12 is proved in Appendix E.

6 Discussion

In this paper, we investigated how k-WTA computation is robustly achieved in the presence of inherent noise in the input spike trains. In a spike-based k-WTA model, n randomly generated input spike trains are competing with each other, and the neurons with the top k highest underlying firing rates are the true winners. Given the stochastic nature of the spike trains, it is not trivial to properly select winners among a group of neurons. We derived an information-theoretic lower bound on the decision time for a given decision accuracy. Notably, this lower bound holds universally for any WTA circuit that falls within our model framework, regardless of their circuit architectures or their adopted activation functions. Furthermore, we constructed a circuit whose decision time matches this lower bound up to a constant multiplicative factor, suggesting that our derived lower bound is order-optimal. Here the order-optimality is stated in terms of its scaling in n, k, and $T_{\mathcal{R}}$. In addition, our results also give a set of testable hypotheses on neural recordings and humans'/animals' behaviors in decision-making.

6.1 Comparison to previous WTA models

Randomness is introduced at different stages of brain computation and the stochastic nature of the spike trains are well observed (Baddeley et al., 1997; Kara et al., 2000; Maimon & Assad, 2009; Shamir, 2009, 2006; Hertz et al., 1991; Ferrari et al., 2018). In our work, we focused on how to robustly achieve k-WTA computation in face of the intrinsic randomness in the spike trains. A common WTA model assumes that neurons transmit information by a continuous variable such as firing rate (Dayan & Abbott, 2001; Hertz et al., 1991), which often ignores the intrinsic randomness in spiking trains. Although some studies used additive Gaussian noise (Kriener et al., 2017; S. Li et al., 2013; Lee et al., 1999; Rougier & Vitay, 2006) in their rate-based WTA circuits to account for input randomness, these circuits are usually very sensitive to noise and could not successfully select even a single winner unless additional non-linearity is added (Kriener et al., 2017). In fact, a neuron with a second non-linearity is similar to an output neuron in our constructed WTA circuit in that they both integrate their local inputs. Unfortunately, only simulation results were provided in (Kriener et al., 2017); a theoretical justification of why such second non-linearity makes their WTA circuit robust to input noise is lacking. Random response of rate-based WTA is also considered in

(Shamir, 2006) with a focus on characterizing the scaling of WTA accuracy with the population size for a two-interval, two-alternative forced choice (2I2AFC) discrimination task. Though we focused on spike-based model, we hope our results can provide some insight for the rate-based model as well. On top of that, a rate-based model would require a high communication bandwidth, yet communication bandwidth is limited in the brain. Our spiking neural network model captures this feature by having a low communication cost, since it broadcasts 1 bit only. However, we did not try to model every biologically relevant feature. In several studies using spiking network models, individual units are often modeled with details like ion channels and specific synaptic connectivity. Though more biologically relevant than our spiking neuron network model, those details significantly complicate the analysis. In fact, it could be challenging and intricate to move beyond computer simulation to characterize the model dynamics (such as the spiking nature of each unit, the time it takes to stabilize, etc.) analytically.

Spike-based WTA is also considered in the insightful work (Shamir, 2009) under a statistical model for a two-alternative forced choice (2AFC) discrimination task. In particular, Shamir (Shamir, 2009) undertook an elegant study on the accuracy of his WTA mechanism focusing on the effects of population size, noise correlations, and baseline firing. Compared to (Shamir, 2009), our model is more restrictive in the sense that we do not consider the effects of population size, noise correlations, and baseline firing, yet is more general in the sense that we consider $n \geq 2$ alternatives. Additionally, we take a slightly different but closely related angle; instead of focusing on characterizing the accuracy w. r. t. a particular WTA circuit, we provide a general lower bound that provides insights on the fundamental limits of a WTA circuit on the waiting time in deciding among independent Bernoulli input spike trains. Nevertheless, all of the features studied in (Shamir, 2009) (i.e., population size, noise correlations, and baseline firing) are interesting, and we definitely would like to try to extend our results to incorporate these features in our future work.

6.2 Potential applications for physiological experiments

Our work might further provide hypotheses on inferring the changes of the network sizes, of the similarities between input spike trains, and of the synaptic memory capacities base on the changes of the performance accuracy. For example, in behavioral experiments using electrolytic lesions or pharmacological inhibition (Clark, Manes, Antoun, Sahakian, & Robbins, 2003; Hanks, Ditterich, & Shadlen, 2006; Yttri, Liu, & Snyder, 2013; Katz, Yates, Pillow, & Huk, 2016), the changes in performance are often highly variable and nonlinear. Such variability and nonlinearity might arise from the experimental difficulties in precisely manipulating network size and in disentangling sensory perception and motor planning from a core decision-making (winner-selecting) process. With an analytical characterization, one might be able to estimate changes in the network size given its performance changes. Several pioneering works studied the impact of the network size on accuracy (Seung & Sompolinsky, 1993; Shamir, 2009, 2006). While these works characterized this trade-off based on investigating specific WTA circuits, our work provides a complementary viewpoint by characterizing a lower bound on a large family of WTA circuits.

Besides the effect of network size, the distribution of feature representations (i.e., different set $\mathcal{R}$s of different individual animals) could be used to account for between-subject variability in decision making. Consider a random-dot coherent motion task where animals need to decide which of two directions the majority of dots are moving (Shadlen & Newsome, 2001). In this task, performance accuracy and reaction time vary across animals. If we perform neural recordings in their visual cortex (i.e., to record their $\mathcal{R}$s), we might be able to decode their reaction time or accuracy, given population representations of dot motion in these cortical neurons (Shadlen & Newsome, 1996; Jazayeri & Movshon, 2006). For example, an animal whose stimulus-evoked responses are more heterogeneous in the visual cortex might be able to react faster given the same accuracy, governed by our derived lower-bound.

Last but not least, our work also offer predictions on how local memory capacity could affect performance in decision-making. For example, when there is more ambiguity in input representations, to achieve the same accuracy, a larger minimum time window for memory storage in synapses (Knoblauch, Palm, & Sommer, 2010) is required. From previous experimental work (Bittner, Milstein, Grienberger, Romani, & Magee, 2017), we know that synaptic plasticity has time scale ranging from milliseconds to seconds across different brain regions, and such plasticity could efficiently store entire behavioral sequences within synaptic weights. Combining with our analytical characterization, when performance accuracy changes over time, assuming other parameters such as input rates, decision time and network size are fixed, one might be able to predict how synaptic plasticity changes.

6.3 Limitations and extensions

When δ is a constant, our lower bound is order-optimal in terms of its scaling in n, k, and $T_{\mathcal{R}}$. Nevertheless, the scaling of the derived lower bound in terms of δ is not tight. It would be interesting to know the optimal scaling in δ when other parameters (n, k, and $T_{\mathcal{R}}$) are fixed. We leave it as one future direction.

To simplify complexity, our model poses a few assumptions that ignored some features in the brain (Shamir, 2009). One of these assumptions is that each input neuron is independent. However, various degrees of average noise correlations between cortical neurons have been reported. For example, average noise correlations in primary visual cortex could be close to 0.1 (Schölvinck, Saleem, Benucci, Harris, & Carandini, 2015), 0.18 (Smith & Kohn, 2008), or even much larger as 0.35 (Gutnisky & Dragoi, 2008). Similarly, noise correlations have been observed in other sensory brain regions (Cohen & Kohn, 2011). In our work, we ignore correlations between these neurons, but it would be interesting as a future direction to extend in our spiking network model. Unfortunately, the impact of the noise-correlation on the lower bound is unclear at first glance. One of the challenges in answering such question is, in general, the details of correlations might matter – especially when there is more than one true winner, and it is unclear whether general statements such as "correlations always hurt" or "correlations always help" can be concluded in the end. Specifically, on the one hand, the insightful work (Shamir, 2009) showed that, similar to the effect of noise-correlation that has been observed in population coding theory, noise correlations in their proposed temporal Winner-Take-All (tWTA) limits and harms the accuracy of the tWTA readout. In fact, in population coding theory, it is commonly reported that noise correlation harms decoding accuracy (Eyherabide & Samengo, 2013). On the other hand, correlations in the variabilities of neuronal firing rates do not, in general, limit the increase in coding accuracy; in some cases, but not all, correlations improve the accuracy of a population code (Abbott & Dayan, 1999; Averbeck, Latham, & Pouget, 2006). Additionally, for the problem of k-WTA where $k \geq 2$, it could be possible that the noise correlation is neither purely positively corrected nor purely negatively corrected. In particular, it could be possible that one true winner is *positively correlated* with other true winners and is *negatively correlated* with non-winners, and another true winner is *negatively correlated* with other true winners and is *positively correlated* with non-winners. Thus, extra care is needed when one is trying to make claims on the impact of noise-correlation on a WTA circuit.

Second, our model uses a threshold activation function by assuming the synaptic transmission is basically noise-free and that the only noise source comes from the input in this paper. However, synaptic transmission is highly unreliable in biological networks (Allen & Stevens, 1994; Faisal et al., 2008; Borst, 2010), and a deterministic activation function would fail to capture this feature compared to a stochastic activation function. Nevertheless, our lower bound in Theorem 3 holds even if the activation functions are random. This is because the probability in $\mathbb{P}\left\{\widehat{\boldsymbol{win}}\left(\boldsymbol{S}\right) \neq \mathcal{W}(\boldsymbol{p})\right\}$ incorporates the possible randomness in the activation functions, and our lower bound characterization is independent of the activation functions used.

Another assumption in our circuit is that the output neurons can inhibit each other. In common scenarios, an output neuron is usually excitatory, and does not inhibit other neurons directly without recruiting inhibitory cells. We incorporate stability in these output neurons by assuming they can inhibit each other in our circuit implementation. For a model where an output neuron is limited to be excitatory only, we can add a chain of inhibitory neurons to achieve stability WTA computation.

Additionally, for our lower bound to hold we need that the initial memory of each neuron, i.e., $M_1(\nu)$, contain no information about the system's state in the past $t \leq 0$. That is, except for the input spike trains, no side information (especially the one on previous network dynamics) is available at a WTA circuit, and nothing happens before the start of WTA competition to affect the WTA dynamics. We impose this assumption on $M_1(\nu)$ in order to derive an information-theoretic lower bound on the observation time. On the other hand, spontaneous firings before the presence of an external stimulus might affect the initial states of neurons' local memory. For those scenarios, our results are applicable provided that the spontaneous firings are very sparse or even negligible. Nevertheless, it would be interesting to relax this assumption, and study how the spontaneous firings of the neurons in the past (i.e., $t \leq 0$) could affect $M_1(\nu)$ in general.

Last but not least, in our k-WTA circuit, the number of output neurons that spike simultaneously increases monotonically until there are exactly k output neurons that spike simultaneously. We acknowledge that this might not be biologically plausible in most cases in the brain – especially considering the possibility of spontaneous firings. From large-scale neural recordings, we know that the number of neurons that spike simultaneously is usually variable, so this could be a future direction to construct a circuit that better matches experimental observations.

Acknowledgement

We would like to thank Christopher Quinn at Purdue University and Zhi-Hong Mao at University of Pittsburgh for the helpful discussions and references.

References

Abbott, L. F., & Dayan, P. (1999). The effect of correlated variability on the accuracy of a population code. *Neural computation*, *11*(1), 91–101. 37

Activation function. (n.d.). https://en.wikipedia.org/wiki/Activation_function. (Accessed: 2018-08-08) 23

Allen, C., & Stevens, C. F. (1994). An evaluation of causes for unreliability of synaptic transmission. *Proceedings of the National Academy of Sciences*, *91*(22), 10380–10383. 37

Averbeck, B. B., Latham, P. E., & Pouget, A. (2006). Neural correlations, population coding and computation. *Nature reviews neuroscience*, *7*(5), 358. 37

Baddeley, R., Abbott, L. F., Booth, M. C., Sengpiel, F., Freeman, T., Wakeman, E. A., & Rolls, E. T. (1997). Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proceedings of the Royal Society of London B: Biological Sciences*, *264*(1389), 1775–1783. 3, 32

Berry II, M. J., & Meister, M. (1998). Refractoriness and neural precision. In *Advances in neural information processing systems* (pp. 110–116). 5

Bittner, K. C., Milstein, A. D., Grienberger, C., Romani, S., & Magee, J. C. (2017). Behavioral time scale synaptic plasticity underlies ca1 place fields. *Science*, *357*(6355), 1033–1036. 35

Borst, J. G. G. (2010). The low synaptic release probability in vivo. *Trends in neurosciences*, *33*(6), 259–266. 37

Buzsáki, G., & Chrobak, J. J. (1995). Temporal structure in spatially organized neuronal ensembles: a role for interneuronal networks. *Current opinion in neurobiology*, *5*(4), 504–510. 4

Clark, L., Manes, F., Antoun, N., Sahakian, B. J., & Robbins, T. W. (2003). The contributions of lesion laterality and lesion volume to decision-making impairment following frontal lobe damage. *Neuropsychologia*, *41*(11), 1474–1483. 34

Cohen, M. R., & Kohn, A. (2011). Measuring and interpreting neuronal correlations. *Nature neuroscience*, *14*(7), 811. 36

Dayan, P., & Abbott, L. F. (2001). Theoretical neuroscience: computational and mathematical modeling of neural systems.

32

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429. 3

Eyherabide, H. G., & Samengo, I. (2013). When and why noise correlations are important in neural decoding. *Journal of Neuroscience*, *33*(45), 17921–17936. 37

Faisal, A. A., Selen, L. P., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature reviews neuroscience*, *9*(4), 292. 3, 37

Ferrari, U., Deny, S., Marre, O., & Mora, T. (2018). A simple model for low variability in neural spike trains. *Neural Computation*, *30*(11), 3009-3036. Retrieved from https://doi.org/10.1162/neco_a_01125 doi: 10.1162/neco_a_01125 3, 32

Gerstner, W., Kempter, R., van Hemmen, J. L., & Wagner, H. (1996). A neuronal learning rule for sub-millisecond temporal coding. *Nature*, *383*(6595), 76. 4

Gutnisky, D. A., & Dragoi, V. (2008). Adaptive coding of visual information in neural populations. *Nature*, *452*(7184), 220. 36

Hahn, T. T., Sakmann, B., & Mehta, M. R. (2006). Phase-locking of hippocampal interneurons' membrane potential to neocortical up-down states. *Nature neuroscience*, *9*(11), 1359. 4

Hanks, T. D., Ditterich, J., & Shadlen, M. N. (2006). Microstimulation of macaque area lip affects decision-making in a motion discrimination task. *Nature neuroscience*, *9*(5), 682. 34

Harris, C. M., & Wolpert, D. M. (1998). Signal-dependent noise determines motor planning. *Nature*, *394*(6695), 780. 3

Hertz, J., Krogh, A., Palmer, R. G., & Horner, H. (1991). Introduction to the theory of neural computation. *Physics Today*, *44*, 70. 4, 5, 32

Hromádka, T., DeWeese, M. R., & Zador, A. M. (2008). Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS biology*, *6*(1), e16. 4

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, *148*(3), 574–591. 3

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, *20*(11), 1254–1259. 4

Jacobs, I., & Berlekamp, E. (1967). A lower bound to the distribution of computation for sequential decoding. *IEEE Transactions on Information Theory*, *13*(2), 167–174. 53

Jazayeri, M., & Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nature neuroscience*, *9*(5), 690. 35

Kara, P., Reinagel, P., & Reid, R. C. (2000). Low response variability in simultaneously recorded retinal, thalamic, and cortical neurons. *Neuron*, *27*(3), 635–646. 3, 32

Karlsson, M. P., & Frank, L. M. (2008). Network dynamics underlying the formation of sparse, informative representations in the hippocampus. *Journal of Neuroscience*, *28*(52), 14271–14281. 4

Katz, L. N., Yates, J. L., Pillow, J. W., & Huk, A. C. (2016). Dissociated functional significance of decision-related activity in the primate dorsal stream. *Nature*, *535*(7611), 285. 34

Kinoshita, M., & Komatsu, H. (2001). Neural representation of the luminance and brightness of a uniform surface in the macaque primary visual cortex.

Journal of neurophysiology, *86*(5), 2559–2570. 3

Knill, D. C., & Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, *27*(12), 712–719. 3

Knoblauch, A., Palm, G., & Sommer, F. T. (2010). Memory capacities for synaptic and structural plasticity. *Neural Computation*, *22*(2), 289–341. 35

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244. 3

Kourtzi, Z., Tolias, A. S., Altmann, C. F., Augath, M., & Logothetis, N. K. (2003). Integration of local features into global shapes: monkey and human fmri studies. *Neuron*, *37*(2), 333–346. 3

Kriener, B., Chaudhuri, R., & Fiete, I. (2017). How fast is neural winner-take-all when deciding between many options? *bioRxiv*, 231753. 4, 5, 6, 23, 32

Lee, D. K., Itti, L., Koch, C., & Braun, J. (1999). Attention activates winner-take-all competition among visual filters. *Nature neuroscience*, *2*(4), 375. 4, 32

Li, N., Chen, T.-W., Guo, Z. V., Gerfen, C. R., & Svoboda, K. (2015). A motor cortex circuit for motor planning and movement. *Nature*, *519*(7541), 51. 3

Li, S., Li, Y., & Wang, Z. (2013). A class of finite-time dual neural networks for solving quadratic programming problems and its k-winners-take-all application. *Neural Networks*, *39*, 27–39. 4, 32

Lynch, N., Musco, C., & Parter, M. (2016). Computational tradeoffs in biological neural networks: Self-stabilizing winner-take-all networks. *arXiv preprint* *arXiv:1610.02084*. 6, 15

Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. *Neural networks*, *10*(9), 1659–1671. 15

Maass, W. (2000). On the computational power of winner-take-all. *Neural computation*, *12*(11), 2519–2535. 4

Maimon, G., & Assad, J. A. (2009). Beyond poisson: increased spike-time regularity across primate parietal cortex. *Neuron*, *62*(3), 426–440. 3, 32

Majaj, N. J., Carandini, M., & Movshon, J. A. (2007). Motion integration by neurons in macaque mt is local, not global. *Journal of Neuroscience*, *27*(2), 366–370. 3

Mao, Z.-H., & Massaquoi, S. G. (2007). Dynamics of winner-take-all competition in recurrent neural networks with lateral inhibition. *IEEE transactions on neural networks*, *18*(1), 55–69. 23

Nelken, I. (2004). Processing of complex stimuli and natural scenes in the auditory cortex. *Current opinion in neurobiology*, *14*(4), 474–480. 4

Nicholls, J. G., Martin, A. R., Wallace, B. G., & Fuchs, P. A. (2001). *From neuron to brain* (Vol. 271). Sinauer Associates Sunderland, MA. 6

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current opinion in neurobiology*, *14*(4), 481–487. 4

Perez-Orive, J., Mazor, O., Turner, G. C., Cassenaer, S., Wilson, R. I., & Laurent, G. (2002). Oscillations and sparsening of odor representations in the mushroom body. *Science*, *297*(5580), 359–365. 4

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*(6741), 233. 3

Polyanskiy, Y., & Wu, Y. (2014). Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, *6*, 2012–2016. 19, 47, 48, 51

Quiroga, R. Q., Kreiman, G., Koch, C., & Fried, I. (2008). Sparse but not ?grandmother-cell?coding in the medial temporal lobe. *Trends in cognitive sciences*, *12*(3), 87–91. 4

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

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, *2*(11), 1019. 4

Rougier, N. P., & Vitay, J. (2006). Emergence of attention within a neural population. *Neural Networks*, *19*(5), 573–581. 4, 32

Schölvinck, M. L., Saleem, A. B., Benucci, A., Harris, K. D., & Carandini, M. (2015). Cortical state determines global variability and correlations in visual cortex. *Journal of Neuroscience*, *35*(1), 170–178. 36

Seung, H. S., & Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences*, *90*(22), 10749–10753. Retrieved from https://www.pnas.org/content/90/22/10749 doi: 10.1073/pnas.90.22.10749 34

Shadlen, M. N., & Newsome, W. T. (1996). Motion perception: seeing and deciding. *Proceedings of the national academy of sciences*, *93*(2), 628–633. 35

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area lip) of the rhesus monkey. *Journal of neurophysiology*, *86*(4), 1916–1936. 3, 35

Shamir, M. (2006). The scaling of winner-takes-all accuracy with population size. *Neural Computation*, *18*(11), 2719-2729. Retrieved from https://doi.org/10.1162/neco.2006.18.11.2719 doi: 10.1162/neco.2006.18.11.2719 4, 5, 6, 32, 33, 34

Shamir, M. (2009, 02). The temporal winner-take-all readout. *PLOS Computational Biology*, *5*(2), 1-13. Retrieved from https://doi.org/10.1371/journal.pcbi.1000286 doi: 10.1371/journal.pcbi.1000286 5, 6, 32, 33, 34, 36

Siapas, A. G., Lubenov, E. V., & Wilson, M. A. (2005). Prefrontal phase locking to hippocampal theta oscillations. *Neuron*, *46*(1), 141–151. 4

Smith, M. A., & Kohn, A. (2008). Spatial and temporal scales of neuronal correlation in primary visual cortex. *Journal of Neuroscience*, *28*(48), 12591–12603. 36

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature neuroscience*, *9*(4), 578. 3

Teleńczuk, B., Kempter, R., Curio, G., & Destexhe, A. (2017). Refractoriness accounts for variable spike burst responses in somatosensory cortex. *eNeuro*, *4*(4). 6

Verzi, S. J., Rothganger, F., Parekh, O. D., Quach, T.-T., Miner, N. E., Vineyard, C. M., . . . Aimone, J. B. (2018). Computing with spikes: The advantage of fine-grained timing. *Neural computation*, *30*(10), 2660–2690. 15

Wu, Y. (2017). Lecture notes on information-theoretic methods for high-dimensional statistics. *Lecture Notes for ECE598YW (UIUC)*. 16

Yttri, E. A., Liu, Y., & Snyder, L. H. (2013). Lesions of cortical area lip affect reach onset only when the reach is accompanied by a saccade, revealing an active eye–hand coordination circuit. *Proceedings of the National Academy of Sciences*, *110*(6), 2371–2376. 34

Yuille, A., & Geiger, D. (1998). *The handbook of brain theory and neural networks.* Chapter Winner-Take-All Networks. 4

Appendices

A Preliminaries

In this section, we present some preliminaries on information measures and Fano's inequality. Interested readers are referred to (Polyanskiy & Wu, 2014) for comprehensive background.

A.1 Information Measures

Let X and Y be two random variables. The mutual information between X and Y, denoted by $I(X;Y)$, measures the dependence between X and Y, or, the information about X (resp. T) provided by Y (resp. X).

Definition 13 (Mutual information)**.** Let X and Y be two random variables.

$$I(X;Y) := D(P_{XY} \parallel P_X P_Y), D(P \parallel Q) := \sum_{a \in \mathcal{A}} P(a) \log \frac{P(a)}{Q(a)},$$

where P_{XY} denotes the joint distribution of X and Y, and $P_X P_Y$ denotes the product of the marginal distributions of X and Y.

In the following, we use the notation $X \to Y$ to denote that Y is a (possibly random) function of X. Thus, $W \to X \to Y \to \widehat{W}$ means that X is a (possibly random) function of W; Y is a (possibly random) function of X; and $\widehat{W}$ is a (possibly random) function of Y. Fano's inequality:

Theorem 14. *(Polyanskiy & Wu, 2014, Corollary 5.1) Let $T : \Theta \to [M]$, and let $\theta \to X \to Y \to \widehat{T}(\theta)$ be an arbitrary Markov chain. Suppose both θ and $T(\theta)$ are uniformly distributed over a set of size M. Then*

$$P_e := \mathbb{P}\left\{T(\theta) \neq \widehat{T}(\theta)\right\} \geq 1 - \frac{I(X;Y) + 1}{\log M}.$$

Theorem 15 (Chernoff Bound)**.** *Let $X_1, \cdots, X_n$ be i.i.d. with $X_i \in \{0,1\}$ and $\mathbb{P}\{X_1 = 1\} = p$. Set $X = \sum_{i=1}^{n} X_i$. Then*

- *for any $t \in [0, 1-p]$, we have $\mathbb{P}\{X \geq (p+t)\, n\} \leq \exp\left(-nd(p+t \parallel p)\right)$.*
- *for any $t \in [0, p]$, we have $\mathbb{P}\{X \leq (p-t)\, n\} \leq \exp\left(-nd(p-t \parallel p)\right)$.*

B Proof of Lemma 2

Proof of Lemma 2. Lemma 2 follows easily from the independence between input spike trains and the assumption that the spikes in each input spike train are $i.i.d.$. For completeness, we present the proof as follows.

Recall that

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

Denote $\boldsymbol{s} = [s_1, \cdots, s_n]$ as one realization of $\boldsymbol{S}$, wherein each component s_i is a binary sequence of length T, i.e.,

$$s_i = \left[b_1^i, \cdots, b_T^i\right] \in \{0,1\}^T.$$

For each $i = 1, \cdots, n$, let $P_{\boldsymbol{S}}(\{S_t(u_i)\}_{t=1}^T)$ and $Q_{\boldsymbol{S}}(\{S_t(u_i)\}_{t=1}^T)$ be the marginal distributions of the i–th length T input spike train $\{S_t(u_i)\}_{t=1}^T$ under joint distributions $P_{\boldsymbol{S}}$ and $Q_{\boldsymbol{S}}$ respectively. Similarly, $P_{\boldsymbol{S}}(S_t(u_i))$ and $Q_{\boldsymbol{S}}(S_t(u_i))$ are the corresponding two marginal distributions of $S_t(u_i)$ – the spiking state of input neuron u_i at time t. Thus, we have

$$\begin{aligned}
& D\left(P_{\boldsymbol{S}}(\{S_t(u_i)\}_{t=1}^T) \parallel Q_{\boldsymbol{S}}(\{S_t(u_i)\}_{t=1}^T)\right) \\
& \overset{(a)}{=} \sum_{\left[b_1^i,\cdots,b_T^i\right]} P_{\boldsymbol{S}}(\{S_t(u_i)\}_{t=1}^T = \left[b_1^i, \cdots, b_T^i\right]) \log \frac{P_{\boldsymbol{S}}(\{S_t(u_i)\}_{t=1}^T = [b_1^i, \cdots, b_T^i])}{Q_{\boldsymbol{S}}(\{S_t(u_i)\}_{t=1}^T = [b_1^i, \cdots, b_T^i])} \\
& \overset{(b)}{=} \sum_{\left[b_1^i,\cdots,b_T^i\right]} \left(\prod_{t'=0}^{T-1} P_{\boldsymbol{S}}(S_{t'}(u_i) = b_{t'}^i)\right) \log \frac{\prod_{t=1}^T P_{\boldsymbol{S}}(S_t(u_i) = b_t^i)}{\prod_{t=1}^T Q_{\boldsymbol{S}}(S_t(u_i) = b_t^i)} \\
& = \sum_{t=1}^T \sum_{\left[b_1^i,\cdots,b_T^i\right]} \left(\prod_{t'=0}^{T-1} P_{\boldsymbol{S}}(S_{t'}(u_i) = b_{t'}^i)\right) \log \frac{P_{\boldsymbol{S}}(S_t(u_i) = b_t^i)}{Q_{\boldsymbol{S}}(S_t(u_i) = b_t^i)} \\
& = \sum_{t=1}^T \sum_{\left[b_1^i,\cdots,b_T^i\right]} \left(\prod_{t'=0\&t'\neq t}^{T-1} P_{\boldsymbol{S}}(S_{t'}(u_i) = b_{t'}^i)\right) P_{\boldsymbol{S}}(S_t(u_i) = b_t) \log \frac{P_{\boldsymbol{S}}(S_t(u_i) = b_t^i)}{Q_{\boldsymbol{S}}(S_t(u_i) = b_t^i)} \\
& \overset{(c)}{=} \sum_{t=1}^T \sum_{b_t^i} P_{\boldsymbol{S}}(S_t(u_i) = b_t^i) \log \frac{P_{\boldsymbol{S}}(S_t(u_i) = b_t^i)}{Q_{\boldsymbol{S}}(S_t(u_i) = b_t^i)} \\
& = \sum_{t=1}^T \left(p_i \log \frac{p_i}{q_i} + (1-p_i) \log \frac{1-p_i}{1-q_i}\right) \\
& = \sum_{t=1}^T d(p_i \parallel q_i) = T \cdot d(p_i \parallel q_i).
\end{aligned}$$

where $\sum_{\left[b_1^i,\cdots,b_T^i\right]}$ is the summation over all binary sequences of length T. In the last displayed equation, equality (a) follows from the definition of KL divergence; equality (b) is true because of independence of spikes; equality (c) follows from the fact that for any fixed b_t^i,

$$\sum_{\left[b_1^i,\cdots,b_T^i\right]\backslash\{t\}} \left(\prod_{t'=1\&t'\neq t}^{T} P_{\boldsymbol{S}}(S_{t'}(u_i)=b_{t'}^i)\right)=1,$$

where we use $\sum_{\left[b_1^i,\cdots,b_T^i\right]\backslash\{t\}}$ to denote the summation over all binary sequences of length T with the t–th entry fixed.

Similarly, we get

$$\begin{aligned}
D(P_{\boldsymbol{S}}\parallel Q_{\boldsymbol{S}}) &= \sum_{\boldsymbol{s}=[s_1,\cdots,s_n]} P_{\boldsymbol{S}}(\boldsymbol{S}=\boldsymbol{s})\log\frac{P_{\boldsymbol{S}}(\boldsymbol{S}=\boldsymbol{s})}{Q_{\boldsymbol{S}}(\boldsymbol{S}=\boldsymbol{s})}\\
&= \sum_{i=1}^{n} D\left(P_{\boldsymbol{S}}(\{S_t(u_i)\}_{t=1}^{T})\parallel Q_{\boldsymbol{S}}(\{S_t(u_i)\}_{t=1}^{T})\right)\\
&= \sum_{i=1}^{n} Td(p_i\parallel q_i) = T\sum_{i=1}^{n} d(p_i\parallel q_i),
\end{aligned}$$

proving the lemma. □

C Proof of Theorem 3

The following lemma is used in the proof of our information-theoretic lower bound. This is a technical supporting lemma, and the choice of the specific rate assignments is due to some technical convenience in proving Theorem 3. See Appendix A for definition of $I(\cdot\,;\,\cdot)$.

Lemma 16. *For any finite set $\mathcal{R}$, let $r_1, r_2 \in \mathcal{R}$ such that $r_1 \neq r_2$. Let $\boldsymbol{p}^0 =$*

$[p_1^0, \cdots, p_n^0]$ *be*

$$p_\ell^0 = \begin{cases} r_1, & \textit{if } \ell = 1, \cdots, k; \\ r_2, & \textit{otherwise.} \end{cases} \tag{12}$$

For $i = 1, \cdots, k$ *and* $j = k+1, \cdots, n$, *define rate assignment* $\boldsymbol{p}^{ij}$ *as*

$$p_\ell^{ij} = \begin{cases} p_\ell^0, & \textit{if } \ell \neq i, \neq j; \\ p_j^0, & \textit{if } \ell = i; \\ p_i^0, & \textit{if } \ell = j. \end{cases}$$

Let $X_{\boldsymbol{p}}$ *be a random rate assignment. If* $X_{\boldsymbol{p}}$ *is uniformly distributed over*

$$\{\boldsymbol{p}^0\} \cup \left\{\boldsymbol{p}^{ij} : \ i = 1, \cdots, k, \& \ j = k+1, \cdots, n\right\},$$

then the mutual information $I(X_{\boldsymbol{p}}; \boldsymbol{S})$ *satisfies the following:*

$$I(X_{\boldsymbol{p}}; \boldsymbol{S}) \leq T\left(d(r_2 \parallel r_1) + d(r_1 \parallel r_2)\right).$$

Proof. Since mutual information can be viewed as distance to product distributions, by (Polyanskiy & Wu, 2014, Theorem 3.4), we have

$$I(X_{\boldsymbol{p}}; \boldsymbol{S}) \ = \ \min_{Q_{X_{\boldsymbol{p}}} Q_{\boldsymbol{S}}} D\left(P_{X_{\boldsymbol{p}}, \boldsymbol{S}} \parallel Q_{X_{\boldsymbol{p}}} Q_{\boldsymbol{S}}\right).$$

where $P_{X_{\boldsymbol{p}}, \boldsymbol{S}}$ is the joint distribution of $X_{\boldsymbol{p}}$ and $\boldsymbol{S}$, and $Q_{X_{\boldsymbol{p}}}$ and $Q_{\boldsymbol{S}}$ are any distributions of $X_{\boldsymbol{p}}$ and $\boldsymbol{S}$, respectively.

For any fixed $Q_{\boldsymbol{S}}$, it holds that

$$\begin{aligned} \min_{Q_{X_{\boldsymbol{p}}}} D\left(P_{X_{\boldsymbol{p}}, \boldsymbol{S}} \parallel Q_{X_{\boldsymbol{p}}} Q_{\boldsymbol{S}}\right) &= \min_{Q_{X_{\boldsymbol{p}}}} D\left(P_{\boldsymbol{S}|X_{\boldsymbol{p}}} P_{X_{\boldsymbol{p}}} \parallel Q_{X_{\boldsymbol{p}}} Q_{\boldsymbol{S}}\right) \\ &\leq D\left(P_{\boldsymbol{S}|X_{\boldsymbol{p}}} P_{X_{\boldsymbol{p}}} \parallel P_{X_{\boldsymbol{p}}} Q_{\boldsymbol{S}}\right), \end{aligned}$$

where the equality follows from conditioning, and the inequality is true because the best choice over all $Q_{X_{\boldsymbol{p}}}$ cannot be worse than any specific choice of $Q_{X_{\boldsymbol{p}}}$. Here $\boldsymbol{S} \mid X_{\boldsymbol{p}}$ denotes the n input spike trains conditioning on the choice of rate assignment.

For any fixed $Q_{\boldsymbol{S}}$, we have

$$\begin{aligned}
&D\left(P_{\boldsymbol{S}|X_{\boldsymbol{p}}} P_{X_{\boldsymbol{p}}} \parallel P_{X_{\boldsymbol{p}}} Q_{\boldsymbol{S}}\right) \\
&= P_{X_{\boldsymbol{p}}}(X_{\boldsymbol{p}} = \boldsymbol{p}^0) \sum_{\boldsymbol{s}} P_{\boldsymbol{S}|X_{\boldsymbol{p}}=\boldsymbol{p}^0}(\boldsymbol{S} = \boldsymbol{s}) \left[\log \frac{P_{\boldsymbol{S}|X_{\boldsymbol{p}}=\boldsymbol{p}^0}(\boldsymbol{S} = \boldsymbol{s}) P_{X_{\boldsymbol{p}}}(X_{\boldsymbol{p}} = \boldsymbol{p}^0)}{Q_{\boldsymbol{S}}(\boldsymbol{S} = \boldsymbol{s}) P_{X_{\boldsymbol{p}}}(X_{\boldsymbol{p}} = \boldsymbol{p}^0)}\right] \\
&+ \sum_{i=1}^{k} \sum_{j=k+1}^{n} P_{X_{\boldsymbol{p}}}(X_{\boldsymbol{p}} = \boldsymbol{p}^{ij}) \sum_{\boldsymbol{s}} P_{\boldsymbol{S}|X_{\boldsymbol{p}}=\boldsymbol{p}^{ij}}(\boldsymbol{S} = \boldsymbol{s}) \left[\log \frac{P_{\boldsymbol{S}|X_{\boldsymbol{p}}=\boldsymbol{p}^{ij}}(\boldsymbol{S} = \boldsymbol{s}) P_{X_{\boldsymbol{p}}}(X_{\boldsymbol{p}} = \boldsymbol{p}^{ij})}{Q_{\boldsymbol{S}}(\boldsymbol{S} = \boldsymbol{s}) P_{X_{\boldsymbol{p}}}(X_{\boldsymbol{p}} = \boldsymbol{p}^{ij})}\right] \\
&= \frac{1}{k(n-k)+1} \sum_{\boldsymbol{s}} P_{\boldsymbol{S}|X_{\boldsymbol{p}}=\boldsymbol{p}^0}(\boldsymbol{S} = \boldsymbol{s}) \left[\log \frac{P_{\boldsymbol{S}|X_{\boldsymbol{p}}=\boldsymbol{p}^0}(\boldsymbol{S} = \boldsymbol{s})}{Q_{\boldsymbol{S}}(\boldsymbol{S} = \boldsymbol{s})}\right] \\
&+ \frac{1}{k(n-k)+1} \sum_{i=1}^{k} \sum_{j=k+1}^{n} \sum_{\boldsymbol{s}} P_{\boldsymbol{S}|X_{\boldsymbol{p}}=\boldsymbol{p}^{ij}}(\boldsymbol{S} = \boldsymbol{s}) \left[\log \frac{P_{\boldsymbol{S}|X_{\boldsymbol{p}}=\boldsymbol{p}^{ij}}(\boldsymbol{S} = \boldsymbol{s})}{Q_{\boldsymbol{S}}(\boldsymbol{S} = \boldsymbol{s})}\right] \\
&= \frac{1}{k(n-k)+1} D\left(P_{\boldsymbol{S}|X_{\boldsymbol{p}}=\boldsymbol{p}^0} \parallel Q_{\boldsymbol{S}}\right) + \frac{1}{k(n-k)+1} \sum_{i=1}^{k} \sum_{j=k+1}^{n} D\left(P_{\boldsymbol{S}|X_{\boldsymbol{p}}=\boldsymbol{p}^{ij}} \parallel Q_{\boldsymbol{S}}\right),
\end{aligned}$$

where $\sum_{\boldsymbol{s}}$ is summation over all possible n binary sequences of length T. Here $P_{\boldsymbol{S}|X_{\boldsymbol{p}}=\boldsymbol{p}^0}$ is the distribution of $\boldsymbol{S}$ with the rate assignment $\boldsymbol{p}^0$, and $P_{\boldsymbol{S}|X_{\boldsymbol{p}}=\boldsymbol{p}^{ij}}$ is the distribution of $\boldsymbol{S}$ with the rate assignment $\boldsymbol{p}^{ij}$. Choosing $Q_{\boldsymbol{S}}$ to be the distribution of $\boldsymbol{S}$ with rate assignment $\boldsymbol{p}^0$ defined in (12), then for any $i = 1, \cdots, k$ and $j = k+1, \cdots, n$, we have

$$D\left(P_{\boldsymbol{S}|X_{\boldsymbol{p}^{ij}}} \parallel Q_{\boldsymbol{S}}\right) = T(d(r_2 \parallel r_1) + d(r_1 \parallel r_2)).$$

Therefore,

$$\begin{aligned}
I\left(X_{\boldsymbol{p}} \parallel \boldsymbol{S}\right) &\leq \frac{1}{k(n-k)+1} \sum_{i=1}^{k} \sum_{j=k+1}^{n} T(d(r_2 \parallel r_1) + d(r_1 \parallel r_2)) \\
&\leq T(d(r_2 \parallel r_1) + d(r_1 \parallel r_2)).
\end{aligned}$$

$\square$

Proof of Theorem 3. We prove this via a genie-aided argument (Jacobs & Berlekamp, 1967) by assuming that there is a genie that can access the firing sequences of all the n input neurons. By assuming the existence of a genie, we are essentially considering the centralized setting. Clearly, if the error probability is high even in the centralized setting, then no SNNs (which are distributed algorithms) can achieve lower error probability.

Suppose that $T \leq ((1-\delta)\log(k(n-k)+1)-1)\, T_{\mathcal{R}}$. By (8) there exists r_1, r_2 such that $r_1 \neq r_2$ and

$$T \leq ((1-\delta)\log(k(n-k)+1)-1)\,\frac{1}{d(r_2 \parallel r_1) + d(r_1 \parallel r_2)}.$$

Without loss of generality, assume that $r_1 > r_2$.

Consider the $k(n-k)+1$ possible rate assignments defined in Lemma 16. Let $\mathcal{P}$ be the set of such rate assignments. By Yao's minimax principle, we know the minimax probability of error is always lower bounded by Bayes probability of error with any prior distribution:

$$\max_{\boldsymbol{p} \in \mathcal{AR}_k} \mathbb{P}\left\{\widehat{\boldsymbol{win}}\,(\boldsymbol{S}) \neq \mathcal{W}(\boldsymbol{p})\right\} \geq \mathbb{E}_{X_{\boldsymbol{p}} \sim Unif(\mathcal{P})}\left[\mathbb{P}\left\{\widehat{\boldsymbol{win}}\,(\boldsymbol{S}) \neq \mathcal{W}(X_{\boldsymbol{p}})\right\}\right],$$

where $X_{\boldsymbol{p}} \sim Unif(\mathcal{P})$ is uniformly distributed over set $\mathcal{P}$. In addition, by Fano's inequality (see Theorem 14), we have

$$\mathbb{E}_{X_{\boldsymbol{p}} \sim Unif(\mathcal{P})}\left[\mathbb{P}\left\{\widehat{\boldsymbol{win}}\,(\boldsymbol{S}) \neq \mathcal{W}(X_{\boldsymbol{p}})\right\}\right] \geq 1 - \frac{I(X_{\boldsymbol{p}}; \boldsymbol{S}) + 1}{\log(k(n-k)+1)}. \tag{13}$$

Applying Lemma 16, we get

$$\begin{aligned}
\max_{\boldsymbol{p}\in\mathcal{AR}_k} \mathbb{P}\left\{\widehat{\boldsymbol{win}}\,(\boldsymbol{S}) \neq \mathcal{W}(\boldsymbol{p})\right\} &\geq 1 - \frac{I(X_{\boldsymbol{p}};\boldsymbol{S})+1}{\log(k(n-k)+1)} \\
&\geq 1 - \frac{T\left(d(r_2 \parallel r_1) + d(r_1 \parallel r_2)\right)+1}{\log(k(n-k)+1)} \\
&\geq \delta.
\end{aligned}$$

The last inequality holds as $T \leq ((1-\delta)\log(k(n-k)+1)-1)\,T_{\mathcal{R}}$. □

D Proof of Theorem 9

The proof of Theorem 9 uses the following technical fact and lemma.

Fact 17. For any given $p \in (0,1)$ and $b > 0$, let $f_{p,b} : \mathbb{R} \to \mathbb{R}$, defined as: for all $t > 0$,

$$f_{p,b}(t) := \exp\left(-td\left(\frac{b}{t} \parallel p\right)\right).$$

Function $f_{p,b}(\cdot)$ is increasing when $t \in (0, \frac{b}{p})$ and decreasing when $t \geq \frac{b}{p}$.

This fact follows immediately from a simple algebra.

Lemma 18. *Assume $u, v \in [c, C] \subseteq (0,1)$. Then for any $\alpha \in (0,1)$,*

$$d\left((1-\alpha)u + \alpha v \parallel u\right) \geq \frac{\alpha^2 c(1-C)}{2C(1-c)}\left(d\left(u \parallel v\right) + d\left(v \parallel u\right)\right).$$

Proof. Note that for any fixed $q \in [c, C]$, $d\left(x \parallel q\right)$ is a function of x, where $x \in [c, C]$. In addition, by simple algebra, we have

$$d'\left(x \parallel q\right) = \log\frac{(1-q)x}{q(1-x)}, \text{ and } d''\left(x \parallel q\right) = \frac{1}{x(1-x)}. \tag{14}$$

By Taylor expansion, we have

$$
\begin{aligned}
d\left((1-\alpha)u+\alpha v \parallel u\right) = d\left(u \parallel u\right) + \left((1-\alpha)u+\alpha v-u\right) d'\left(u \parallel u\right) \\
+ \frac{\left((1-\alpha)u+\alpha v-u\right)^2}{2} d''\left(\xi \parallel u\right),
\end{aligned}
$$

where $\xi \in [\min\{u,(1-\alpha)u+\alpha v\},\ \ \max\{u,(1-\alpha)u+\alpha v\}]$. By (14),

$$
d\left((1-\alpha)u+\alpha v \parallel u\right) = 0+0+\frac{1}{\xi(1-\xi)}\frac{\alpha^2(u-v)^2}{2} \geq \frac{\alpha^2(u-v)^2}{2C(1-c)}.
$$

On the other hand, since $d\left(u \parallel v\right)+d\left(u \parallel v\right)$ is symmetric in u and v, without loss of generality, assume that $u \geq v$. We have

$$
\begin{aligned}
d\left(u \parallel v\right)+d\left(u \parallel v\right) &= (u-v)\log\frac{u(1-v)}{v(1-u)} \\
&= (u-v)\log\left(1+\frac{u-v}{v(1-u)}\right) \\
&\leq (u-v)\frac{u-v}{v(1-u)} = \frac{(u-v)^2}{v(1-u)} \leq \frac{(u-v)^2}{c(1-C)} \\
&\leq \frac{2C(1-c)}{c(1-C)\alpha^2} d\left((1-\alpha)u+\alpha v \parallel u\right),
\end{aligned}
$$

proving the lemma. □

Now we are ready to prove Theorem 9.

Proof of Theorem 9. Without loss of generality, assume that

$$
p_1 \geq \cdots \geq p_k > p_{k+1} \geq \cdots \geq p_n.
$$

For a given rate assignment $\boldsymbol{p} \in \mathcal{AR}$, define $\tau_1, \tau_2, \cdots, \tau_n$ as

$$
\tau_i := \inf_t \left\{ t : \sum_{r=1}^{\min\{t,m^*\}} S_r(u_i) \geq b \right\}, \quad \forall\ i = 1, \cdots, n.
$$

Notably, in the above definition $\left\{t: \sum_{r=1}^{\min\{t,m^*\}} S_r(u_i) \ge b\right\}$ could be empty. In that case, we define $\tau_i := \infty$ by convention. To show Theorem 9, it is enough to show that with probability $1-\delta$,

$$\tau_i < \tau_j \quad \forall\, i = 1, \cdots, k, \text{ and } j = k+1, \cdots, n; \tag{15}$$

$$\text{and} \quad \tau_i \le m^* \quad \forall\, i = 1, \cdots, k.. \tag{16}$$

Before diving into proving (15) and (16) hold with high probability, let's check the sufficiency of (15) and (16). Let $t_0 := \max_{1\le i\le k} \tau_i$. Let $\mathcal{E}$ be the event on which (15) and (16) hold. Clearly, conditioning on event $\mathcal{E}$, we have

$$\left(\max_{1\le i\le k} \tau_i \mid \mathcal{E}\right) = \; t_0 \mid \mathcal{E} \le m^* - 1 \le m - 1,$$

where the last inequality follows from the assumption in Theorem 9, and

$$\left(\max_{1\le i\le k} \tau_i \;\mid\; \mathcal{E}\right) = t_0 \mid \mathcal{E} \; < \; \tau_j \mid \mathcal{E} \;\; \forall j = k+1, \cdots, n.$$

Notably, for any $t \le t_0 \le m-1$ and for $i = 1, \cdots, n$,

$$\left[\sum_{r=1}^{t} \mathbf{1}_{\{V_r(v_i)>0\}} - m \sum_{r=1}^{t} \mathbf{1}_{\{V_r(v_i)\le -1\}}\right]_+ \le \sum_{r=1}^{t} \mathbf{1}_{\{V_r(v_i)>0\}} \le \sum_{r=1}^{t} S_r(u_i)$$

– recalling that $V_r(v_i)$ is defined in (9). Thus, conditioning on $\mathcal{E}$, at most $k-1$ output neurons ever spike by time t_0. So we have (1) $\mathbf{1}_{\{V_t(v_i)\le -1\}} = 0$, and (2) $\mathbf{1}_{\{V_t(v_i)>0\}} = S_t(u_i)$, for all $i = 1, \cdots, n$ and for all $t \le t_0$. In addition, we have for all $t \le t_0$,

$$\begin{aligned}
&(b-1)\mathbf{1}_{\{S_t(v_i)=1\}} + \left[\sum_{r=1}^{t} \mathbf{1}_{\{V_r(v_i)>0\}} - m \sum_{r=1}^{t} \mathbf{1}_{\{V_r(v_i)\le -1\}}\right]_+ \\
&= (b-1)\mathbf{1}_{\{S_t(v_i)=1\}} + \sum_{r=1}^{t} \mathbf{1}_{\{V_r(v_i)>0\}} \\
&= (b-1)\mathbf{1}_{\{S_t(v_i)=1\}} + \sum_{r=1}^{t} S_r(u_i).
\end{aligned}$$

By the activation rules in Algorithm 1, we know, conditioning on $\mathcal{E}$, at time $t_0 + 1 \le m^*$, output neurons $v_1, \cdots, v_k$ spike simultaneously, and output neurons $v_{k+1}, \cdots, v_n$ do not spike, proving (1) in Theorem 9. By the choice of t_0, we know that, on $\mathcal{E}$, $t_0 + 1$ is the first time that k output neurons spike simultaneously, and no other k output neurons ever spike simultaneously, proving (2) in Theorem 9.

By a simple induction argument, it can be shown that conditioning on $\mathcal{E}$, in each of the time slot t such that $t_0 + 1 \le t \le m + 1$, output neurons $v_1, \cdots, v_k$ spike, and no other output neurons (i.e., output neurons $v_{k+1}, \cdots, v_n$ do not spike). Let's consider the case when $t = (m + 1) + 1$. As among output neurons, only $v_1, \cdots, v_k$ spike, and no other output neurons spike for any $t' \le m + 1$, it follows that

$$m \sum_{r=1}^{m} \mathbf{1}_{\{V_{t-r}(v_i) \le -1\}} = 0, \quad \forall\ v_1, \cdots, v_k.$$

Thus, for these k output neurons,

$$\begin{aligned}
&(b-1)\mathbf{1}_{\{S_{t-1}(v_i)=1\}} + \left[\sum_{r=1}^{m} \mathbf{1}_{\{V_{t-r}(v_i)>0\}} - m \sum_{r=1}^{m} \mathbf{1}_{\{V_{t-r}(v_i) \le -1\}} \right]_+ \\
&= (b-1) + \sum_{r=1}^{m} \mathbf{1}_{\{V_{t-r}(v_i)>0\}} \\
&= (b-1) + \sum_{r=1}^{m} \mathbf{1}_{\{V_{t-1-r}(v_i)>0\}} + \mathbf{1}_{\{V_{t-1}(v_i)>0\}} - \mathbf{1}_{\{V_{t-1-m}(v_i)>0\}} \\
&\ge b - 2 + \sum_{r=1}^{m} \mathbf{1}_{\{V_{t-1-r}(v_i)>0\}} \\
&= b - 2 + \sum_{r=1}^{m} S_r(u_i) \ge 2b - 2 \ge b,
\end{aligned}$$

where the last inequality holds provided that $b \ge 2$. For output neurons $v_{k+1}, \cdots, v_n$, we have

$$\begin{aligned}
&(b-1)\mathbf{1}_{\{S_{t-1}(v_i)=1\}} + \left[\sum_{r=1}^{m}\mathbf{1}_{\{V_{t-r}(v_i)>0\}} - m\sum_{r=1}^{m}\mathbf{1}_{\{V_{t-r}(v_i)\le -1\}}\right]_+ \\
&\le \sum_{r=1}^{m}\mathbf{1}_{\{V_{t-r}(v_i)>0\}} \\
&\overset{(a)}{=} \sum_{r=1}^{m}\mathbf{1}_{\{V_{t-1-r}(v_i)>0\}} + \mathbf{1}_{\{V_{t-1}(v_i)>0\}} - \mathbf{1}_{\{V_{t-1-m}(v_i)>0\}} \\
&= \sum_{r=1}^{m}\mathbf{1}_{\{V_{t-1-r}(v_i)>0\}} - \mathbf{1}_{\{V_{t-1-m}(v_i)>0\}} \\
&\le \sum_{r=1}^{m}\mathbf{1}_{\{V_{t-1-r}(v_i)>0\}} = \sum_{r=1}^{m}\mathbf{1}_{\{V_{r}(v_i)>0\}} < b.
\end{aligned}$$

Equality (a) follows because at time $t-1$, output neurons $v_1, \cdots, v_k$ spike, resulting in $\mathbf{1}_{\{V_{t-1-r}(v_i)>0\}} = 0$ for $i \neq 1, \cdots, k$. Thus, we know conditioning on event $\mathcal{E}$, at time $(m+1)+1$, the output neurons $v_1, \cdots, v_k$ spike, and no other output neuron spike. It can be shown by a simple induction that at each time t such that $t_0 + 1 \le t \le m + b$, the output neurons $v_1, \cdots, v_k$ spike, and no other output neurons spike. This proves (3) in Theorem 9.

Next we prove (15) and (16). By definition of τ_j, we know that $\tau_j \le m^*$ for all $j = 1, \cdots, n$. Thus, we only need to show that with probability $1-\delta$,

$$\tau_i < \tau_j \quad \forall\ i = 1, \cdots, k, \text{ and } j = k+1, \cdots, n,$$

which is the focus of the remainder of our proof.

Note that

$$\begin{aligned}
&\mathbb{P}\left\{\tau_i < \tau_j, \quad \forall i \in \{1, \cdots, k\}, \forall j \in \{k+1, \cdots, n\}\right\} \\
&= \mathbb{P}\left\{\tau_i < \tau_j, \& \ \tau_i < m^*, \quad \forall i \in \{1, \cdots, k\}, \forall j \in \{k+1, \cdots, n\}\right\} \\
&\geq 1 - \sum_{i=1}^{k} \sum_{j=k+1}^{n} \mathbb{P}\left\{\tau_i \geq \tau_j, \text{ or } \tau_i = m^*\right\}. \qquad (17)
\end{aligned}$$

For each term in the summation of (17), we have

$$\mathbb{P}\left\{\tau_i \geq \tau_j, \text{ or } \tau_i = m^*\right\} = \mathbb{P}\left\{\tau_i = m^*\right\} + \mathbb{P}\left\{\tau_i \geq \tau_j, \& \ \tau_i < m^*\right\}, \qquad (18)$$

which follows from the fact that $\mathbb{P}\{A \cup B\} = \mathbb{P}\{A\} + \mathbb{P}\{B - A\}$ for any sets A and B. Note that $m^* p_i \geq b$. By Chernoff bound (see Theorem 15), the first term in (18) is bounded as

$$\mathbb{P}\left\{\tau_i = m^*\right\} = \mathbb{P}\left\{\sum_{r=0}^{m^*} S_r(u_i) \leq b\right\} \leq \exp\left(-m^* \cdot d\left(\frac{b}{m^*} \parallel p_i\right)\right). \qquad (19)$$

For the second term in (18), we have

$$\begin{aligned}
\mathbb{P}\left\{\tau_i \geq \tau_j \ \text{ and } \tau_i < m^*\right\} &= \mathbb{P}\left\{\sum_{r=0}^{\tau_i} S_r(u_j) \geq b, \quad \text{and } \tau_i < m^*\right\} \\
&\leq \exp\left(-t^* \cdot d\left(\frac{b}{t^*} \parallel p_{k+1}\right)\right) + \exp\left(-t^* \cdot d\left(\frac{b}{t^*} \parallel p_k\right)\right),
\end{aligned}$$

where $t^* \in \left(\frac{b}{p_{k+1}}, \frac{b}{p_k}\right)$. Thus, (18) is upper bounded as

$$\begin{aligned}
\mathbb{P}\left\{\tau_i \geq \tau_j, \text{ or } \tau_i = m^*\right\} &\leq \exp\left(-m^* \cdot d\left(\frac{b}{m^*} \parallel p_{k+1}\right)\right) \\
&\quad + \exp\left(-t^* \cdot d\left(\frac{b}{t^*} \parallel p_{k+1}\right)\right) + \exp\left(-t^* \cdot d\left(\frac{b}{t^*} \parallel p_k\right)\right) \\
&\leq \exp\left(-t^* \cdot d\left(\frac{b}{t^*} \parallel p_{k+1}\right)\right) + 2\exp\left(-t^* \cdot d\left(\frac{b}{t^*} \parallel p_k\right)\right).
\end{aligned}$$

Eq (17) is bounded as

$$
\begin{aligned}
&\mathbb{P}\left\{\tau_i < \tau_j, \quad \forall i \in \{1, \cdots, k\}, \forall j \in \{k+1, \cdots, n\}\right\} \\
&\geq 1 - \sum_{i=1}^{k} \sum_{j=k+1}^{n} \mathbb{P}\left\{\tau_i \geq \tau_j, \text{ or } \tau_i = m^*\right\} \\
&\geq 1 - \sum_{i=1}^{k} \sum_{j=k+1}^{n} \left(\exp\left(-t^* \cdot d\left(\frac{b}{t^*} \parallel p_{k+1}\right)\right) + 2\exp\left(-t^* \cdot d\left(\frac{b}{t^*} \parallel p_k\right)\right)\right) \\
&= 1 - k(n-k)\left(\exp\left(-t^* \cdot d\left(\frac{b}{t^*} \parallel p_{k+1}\right)\right) + 2\exp\left(-t^* \cdot d\left(\frac{b}{t^*} \parallel p_k\right)\right)\right).
\end{aligned}
$$

Let $t^* = \frac{b}{(p_k+p_{k+1})/2}$, it holds that

$$
\begin{aligned}
\exp\left(-t^* \cdot d\left(\frac{b}{t^*} \parallel p_{k+1}\right)\right) &= \exp\left(-\frac{b}{(p_k+p_{k+1})/2} \cdot d\left(\frac{p_k+p_{k+1}}{2} \parallel p_{k+1}\right)\right), \\
2\exp\left(-t^* \cdot d\left(\frac{b}{t^*} \parallel p_k\right)\right) &= 2\exp\left(-\frac{b}{(p_k+p_{k+1})/2} \cdot d\left(\frac{p_k+p_{k+1}}{2} \parallel p_k\right)\right).
\end{aligned}
$$

By Lemma 18, we know

$$
d\left(\frac{p_k+p_{k+1}}{2} \parallel p_{k+1}\right) \geq \frac{c(1-C)}{8C(1-c)}\left(d(p_{k+1} \parallel p_k) + d(p_k \parallel p_{k+1})\right),
$$

and,

$$
d\left(\frac{p_k+p_{k+1}}{2} \parallel p_k\right) \geq \frac{c(1-C)}{8C(1-c)}\left(d(p_{k+1} \parallel p_k) + d(p_k \parallel p_{k+1})\right).
$$

Thus, we get

$$
\begin{aligned}
&\mathbb{P}\left\{\tau_i < \tau_j, \quad \forall i \in \{1, \cdots, k\}, \forall j \in \{k+1, \cdots, n\}\right\} \\
&\geq 1 - 3k(n-k)\exp\left(-\frac{2b}{p_k+p_{k+1}} \frac{c(1-C)}{8C(1-c)}\left(d(p_k \parallel p_{k+1}) + d(p_{k+1} \parallel p_k)\right)\right)
\end{aligned}
$$

Since $b = \frac{8C^2(1-c)}{c(1-C)}\left(\log\frac{3}{\delta} + \log k(n-k)\right)T_{\mathcal{R}}$, we have

$$
3k(n-k)\exp\left(-\frac{2b}{p_k+p_{k+1}} \frac{c(1-C)}{8C(1-c)}\left(d(p_k \parallel p_{k+1}) + d(p_{k+1} \parallel p_k)\right)\right) \leq \delta.
$$

Thus, $\mathbb{P}\left\{\tau_i < \tau_j, \quad \forall i \in \{1, \cdots, k\}, \forall j \in \{k+1, \cdots, n\}\right\} \leq 1-\delta$.

In addition,

$$t^* = \frac{2b}{p_k + p_{k+1}} \leq \frac{1}{c} b = m^* \leq m,$$

completing the proof of Theorem 9. $\square$

E Proof of Lemma 12

By the activation rules in Algorithm 1, we know that

$$S_{t_0+m} = \begin{cases} 1, \text{ if } \quad (b-1)\mathbf{1}_{\left\{S_{t_0+m-1}(v_i)=1\right\}} + \sum_{r=1}^{m} \left(\mathbf{1}_{\left\{V_{t_0+m-r}>0\right\}} - m\mathbf{1}_{\left\{V_{t_0+m-r}\leq -1\right\}}\right) > b; \\ \\ 0, \text{ otherwise.} \end{cases}$$

As all input neurons are quiescent at time t_0 and remain to be quiescent for all $t \geq t_0$, it follows that

$$\begin{aligned} &(b-1)\mathbf{1}_{\left\{S_{t_0+m-1}(v_i)=1\right\}} + \sum_{r=1}^{m} \left(\mathbf{1}_{\left\{V_{t_0+m-r}>0\right\}} - m\mathbf{1}_{\left\{V_{t_0+m-r}\leq -1\right\}}\right) \\ &= (b-1)\mathbf{1}_{\left\{S_{t_0+m-1}(v_i)=1\right\}} - m\sum_{r=1}^{m} \mathbf{1}_{\left\{V_{t_0+m-r}\leq -1\right\}} \\ &\leq b-1 < b. \end{aligned}$$

Thus, $S_{t_0+m}(v_i) = 0$ for all $i = 1, \cdots, n$. So we have $V_{t_0+m+1}(v_i) = 0$ for all $i = 1, \cdots, n$, which again implies that $S_{t_0+m+1}(v_i) = 0$ for all $i = 1, \cdots, n$. Therefore, we conclude that $S_t(v_i) = 0$ and $V_t(v_i) = 0$ for all $t > t_0 + m$.