The Temporal Winner-Take-All Readout

Maoz Shamir*

Department of Physiology and the Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Abstract

How can the central nervous system make accurate decisions about external stimuli at short times on the basis of the noisy responses of nerve cell populations? It has been suggested that spike time latency is the source of fast decisions. Here, we propose a simple and fast readout mechanism, the temporal Winner-Take-All (tWTA), and undertake a study of its accuracy. The tWTA is studied in the framework of a statistical model for the dynamic response of a nerve cell population to an external stimulus. Each cell is characterized by a preferred stimulus, a unique value of the external stimulus for which it responds fastest. The tWTA estimate for the stimulus is the preferred stimulus of the cell that fired the first spike in the entire population. We then pose the questions: How accurate is the tWTA readout? What are the parameters that govern this accuracy? What are the effects of noise correlations and baseline firing? We find that tWTA sensitivity to the stimulus grows algebraically fast with the number of cells in the population, N, in contrast to the logarithmic slow scaling of the conventional rate-WTA sensitivity with N. Noise correlations in first-spike times of different cells can limit the accuracy of the tWTA readout, even in the limit of large N, similar to the effect that has been observed in population coding theory. We show that baseline firing also has a detrimental effect on tWTA accuracy. We suggest a generalization of the tWTA, the n-tWTA, which estimates the stimulus by the identity of the group of cells firing the first n spikes and show how this simple generalization can overcome the detrimental effect of baseline firing. Thus, the tWTA can provide fast and accurate responses discriminating between a small number of alternatives. High accuracy in estimation of a continuous stimulus can be obtained using the n-tWTA.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: shmaoz@bgu.ac.il

Introduction

In recent years, there has been growing interest in coding information about external stimuli by the fine temporal structure of the neural dynamic response [1–18]. Several studies have shown that response latency is modulated by external stimuli [1–4]. Many cells in the middle temporal (MT) cortex code for the direction of motion of visual stimuli, and can be characterized by a preferred direction of the stimulus, to which they respond maximally, see e.g., [19,20]. Osborne et al. [1] reported that the MT cells respond with the shortest delay when the stimulus is in their preferred direction and that the response delay increases as the stimulus direction diverges from the preferred direction of the cell. In the auditory system of the ferret, Nelken et al. [2] showed response-latency tuning in primary auditory cortex cells to the direction of a virtual sound source. In a recent work Gollisch and Meister [18] showed that relative first-spike times of retinal ganglion cells carry considerable information about the external stimulus, but they did not suggest a concrete readout mechanism.

Here we study the accuracy of a simple readout mechanism, the temporal-Winner-Take-All (tWTA), which extracts information from response latency. The tWTA estimates the stimulus by the identity of the cell that fired the *first* spike in a population of cells, in contrast to the conventional rate-Winner-Take-All (WTA), which estimates the stimulus by the identity of the cell that fired the *most* spikes. For example, the tWTA estimate for the direction of motion of a visual stimulus from the responses of a population of MT cells would be the preferred direction of the cell that fired the first spike in the entire population.

Considerable theoretical effort has been devoted to the study of the accuracy of population code readout mechanisms, such as the population-vector, optimal-linear and ideal observer readouts. Of particular interest in the investigation of these readouts was the dependence of the readout accuracy on the population size and the effects of noise correlations in the neuronal responses. In this work, we quantify tWTA accuracy. To this end, we address three specific questions. One, what are the essential features of the neuronal dynamic response to the stimulus to which the tWTA is sensitive? Two, how does the tWTA accuracy depend on the population size? Three, what are the effects of noise correlations and baseline firing on tWTA accuracy?

These questions are addressed in the framework of a statistical model for the dynamic response of MT cells to a moving visual stimulus. In the first part of the results section we investigate tWTA accuracy in a two-column competition model, and in the second part we study tWTA accuracy in the framework of a hypercolumn model. Both parts start by defining the statistical model of the neuronal dynamic response and then follow with an investigation of tWTA accuracy in the absence of noise correlations and baseline firing. In the final stage of each part, correlations and baseline firing are introduced and their effect on tWTA accuracy is investigated.

Results

tWTA Readout Accuracy in a Two Competing Columns Model

The model. We study tWTA accuracy in a model of two competing MT columns coding for the direction of motion of visual stimuli. Each column is comprised of N homogeneous cells. We denote the preferred direction of the cells in column 1 by ϕ_1

Author Summary

Considerable experimental as well as theoretical effort has been devoted to the investigation of the neural code. The traditional approach has been to study the information content of the total neural spike count during a long period of time. However, in many cases, the central nervous system is required to estimate the external stimulus at much shorter times. What readout mechanism could account for such fast decisions? We suggest a readout mechanism that estimates the external stimulus by the first spike in the population, the tWTA. We show that the tWTA can account for accurate discriminations between a small number of choices. We find that the accuracy of the tWTA is limited by the neuronal baseline firing. We further find that, due to baseline firing, the single first spike does not encode sufficient information for estimating a continuous variable, such as the direction of motion of a visual stimulus, with fine resolution. In such cases, fast and accurate decisions can be obtained by a generalization of the tWTA to a readout that estimates the stimulus by the first n spikes fire by the population, where n is larger than the mean number of baseline spikes in the population.

and the preferred direction of the cells in column 2 by ϕ_2. Without loss of generality, we take $\phi_1 = 0$, which is equivalent to measuring all angles with respect to ϕ_1. We denote the probability density of a single cell k ($k = 1, \ldots N$) in column i with preferred direction ϕ_i to fire its first spike at time $t_k^{(i)}$ given that stimulus θ was presented at time $t_o = 0$ by $f\left(t_k^{(i)}|\theta - \phi_i\right)$. Assuming that first-spike times are statistically independent, the probability density of the first spike in the entire column i at time t is given by the product of three terms: the probability density of a specific cell to fire its first at time t, $f(t|\theta - \phi_i)$, the probability that that the first spike times of the rest $(N-1)$ cells in the population occurred after time t, $\left(1 - \int_0^t dt' f(t'|\theta - \phi_i)\right)^{N-1}$, and the N different possibilities of choosing the cell that fired the first spike:

$$P\left(\min_k\{t_k^{(i)}\} = t\right) = N f(t|\theta - \phi_i) e^{[N-1]\Phi(t|\theta - \phi_i)} \quad (1)$$

$$\Phi(t|\theta - \phi_i) = \log\left(1 - \int_0^t dt' f(t'|\theta - \phi_i)\right) \quad (2)$$

The function Φ is the logarithm of the probability of a single cell firing its first spike after time t, and it has the following properties: $\Phi(0) = 0$, $\Phi(t \to \infty) = -\infty$ and $\frac{d\Phi(t)}{dt} < 0$. Equation (1) can also be obtained by taking the derivative of the probability that the first spike in the column occurred after time t: $P\left(\min_k\{t_k^{(i)}\} > t\right) = e^{N\Phi(t|\theta - \phi_i)}$, with respect to first spike time, t.

Throughout this section, we will quantify tWTA accuracy by using the two alternative forced choice (2AFC) paradigm. In a 2AFC discrimination task, the system is given a stimulus, either θ_1 or θ_2, randomly with equal probabilities. Presentation of the stimulus generates a population response in the two columns, $\mathbf{t}^{(1)}$ and $\mathbf{t}^{(2)}$, which are distributed as defined above. The task of the readout is to infer, on the basis of these spike times, whether the stimulus was θ_1 or θ_2. We will use the probability of correct discrimination, P_C, and the error rate, $P_\varepsilon = 1 - P_C$, as measures of the tWTA performance. We will use the term *sensitivity* to designate

the inverse of the stimulus difference, $\Delta\theta$, at which P_C crosses a certain threshold, P_{th}. This latter measure is related to the '*just noticeable difference*' used in psychophysics.

tWTA accuracy in the absence of correlations. Assuming that column 1 responds faster to a stimulus in direction θ_1 than column 2 and vice versa for stimulus θ_2, we define the tWTA readout in the 2AFC task as follows:

$$\text{decide first stimulus was} \begin{cases} \theta_1 & \min\{t_k^{(1)}\} < \min\{t_k^{(2)}\} \\ \theta_2 & \min\{t_k^{(2)}\} < \min\{t_k^{(1)}\} \end{cases} \quad (3)$$

For the sake of convenience, we take $\theta_1 = \phi_1 = 0$ and $\theta_2 = \phi_2 = \delta\theta$. This choice equalizes the probability of correct response given stimulus θ_1 and given stimulus θ_2. The probability of correct response, P_C, is given by the probability that population 1 fired the first spike, given stimulus θ. Thus, P_C can be written as the integration over all possible first-spike times, t, of the probability density that population 1 fired its first spike at time t multiplied by the probability that the first spike time of population 2 is large than t:

$$P_C = N \int_0^\infty dt f(t|0) e^{[N-1]\Phi(t|0) + N\Phi(t|\delta\theta)} \quad (4)$$

In the limit of large populations, $N \to \infty$, the integral in the right-hand-side of equation (4) will be dominated by the region in which the exponent obtains its maximum. Since Φ is a monotonically decreasing function of t, this region is the region of small t. For small t, we approximate f by:

$$f(t|\theta) \approx r(\theta)^{\alpha(\theta)} [t - \tau(\theta)]_+^{\alpha(\theta)-1} \quad (5)$$

where r is the scale parameter, α is the shape parameter, τ is the delay parameter and $[x]_+ = x$ for $x > 0$ and 0 otherwise.

Relation to the peri stimulus time histogram (PSTH) in an inhomogeneous Poisson process (IHPP). The IHPP is widely used to model the stochastic nature of the neural temporal response [21,22] and is fully defined by the PSTH. In the context of first spike-time distribution, the choice of an IHPP model does not limit the generality of the model, since every PSTH, $\rho(t)$, of an IHPP could be mapped to first spike time distribution, $f(t)$, and vice versa. For a given IHPP with PSTH, $\rho(t)$, the first spike time distribution is given by (see e.g., [21,22])

$$f(t) = \rho(t) e^{-\int_0^t \rho(t') dt'} \quad (6)$$

In the other direction, we want to obtain the PSTH, $\rho(t)$, that will yield a specific first spike time distribution, $f(t)$, in an IHPP model. The probability density that the first spike has occurred in time $t + dt$ in an IHPP model, can be written as the product of the probability density of spiking at that time, $\rho(t)$, multiplied by the probability that there were no prior spikes, $1 - \int_0^t f(t') dt'$; hence, $f(t + dt) = \left(1 - \int_0^{t+dt} f(t') dt'\right) \rho(t + dt)$. Thus we obtain the reciprocal relation

$$\rho(t) = \frac{f(t)}{1 - \int_0^t f(t') dt'} \quad (7)$$

which could be verified by substituting equation (7) into equation (6). For small t: $f(t) \approx \rho(t)$. Thus, the scale parameter corresponds to the scale of the PSTH, the shape parameter governs the initial

acceleration of the PSTH, and the delay parameter measures the temporal shift of the PSTH. Figure 1 illustrates how the different parameters that characterize the initial neural response: scale, shape and delay, affect the first spike probability density and the corresponding PSTH. Note that whereas $f(t)$ and $\rho(t)$ are very similar for small t, for large t, $f(t)$ decays to zero while $\rho(t)$ may continue to increase. Below we study the different effects of the tuning of these parameters on the accuracy of the tWTA.

Effect of scale parameter tuning. We first consider a simple model in which the scale is the only parameter that is tuned to the stimulus. In this case, we can write f near $t=0$ as the product of a function of the stimulus and a function of time:

$$f(t|\theta) = r(\theta)^\alpha [t]_+^{\alpha-1} \tag{8}$$

where α is independent of θ. Expanding Φ in small t, $\Phi(t|\theta) \approx -r(\theta)t^\alpha/\alpha$ and substituting in equation (4), we obtain

to a leading order in $1/N$

$$P_C = \frac{1}{1 + \left(\frac{r(\delta\theta)}{r(0)}\right)^\alpha} \tag{9}$$

Hence, in this case, the probability of correct response is at chance level, $P_C^{\text{chance}} = 1/2$, when the neural response has the same scale for the two alternatives, $r(0) = r(\delta\theta)$, and increases monotonically in the ratio $\frac{r(0)}{r(\delta\theta)}$. The accuracy of the tWTA is not improved by increasing N: The same accuracy will be obtained with $N=1$ and $N=1000$ cells, but, somewhat faster for the $N=1000$ case. Figure 2a shows the probability of correct discrimination as a function of N for different values of $\frac{r(\delta\theta)}{r(0)} = 0, 0.2 \ldots 1$ from top to bottom. The open circles are estimates of the tWTA accuracy obtained by averaging the tWTA accuracy over 10^6 realizations of
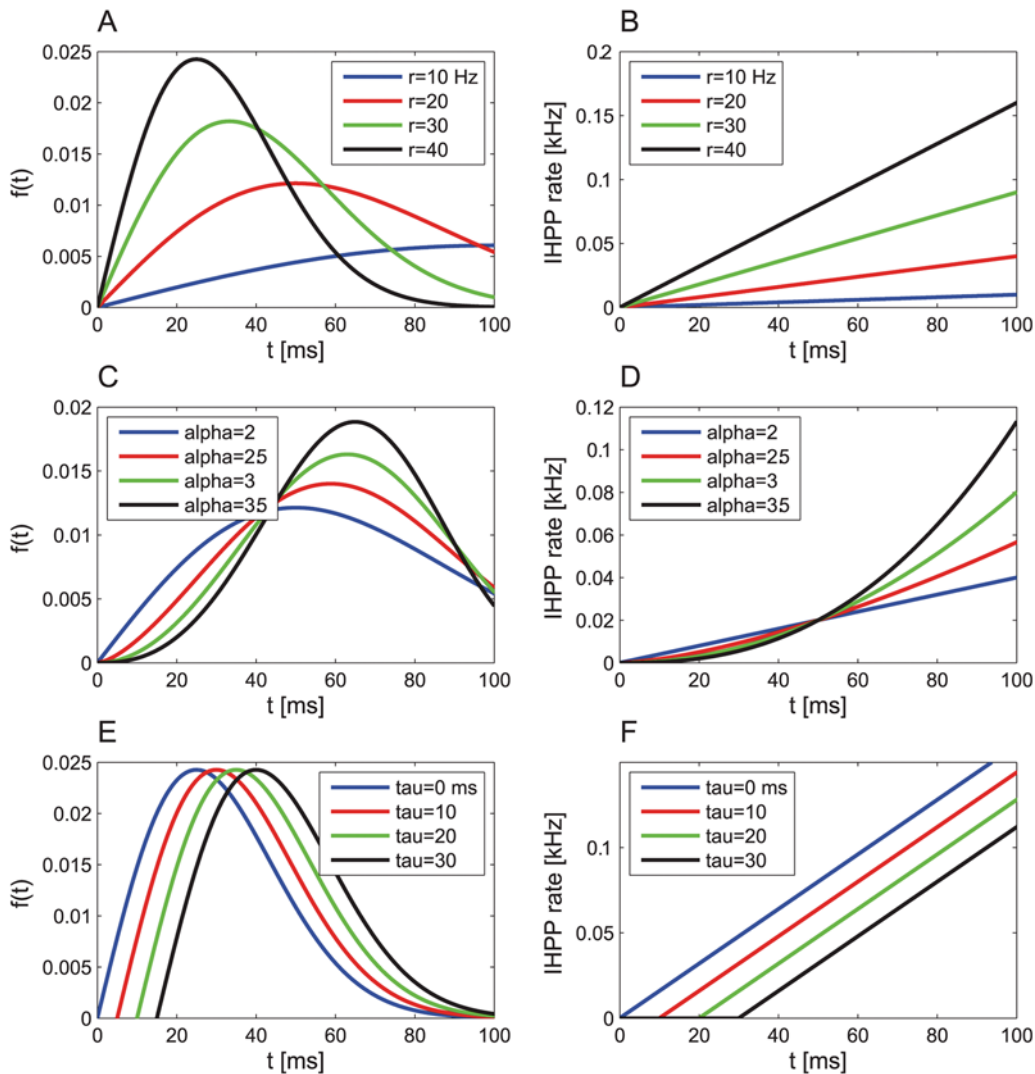

Figure 1. Three examples showing the effects of the scale parameter (a,b), the shape parameter (c,d), and the delay parameter (e,f) on the first spike time probability density, $f(t)$, (left column) and the PSTH rate, $\rho(t)$, of a corresponding inhomogeneous Poisson process (right column). The PSTHs were taken to be of the form of $\rho(t) = r^\alpha[t-\tau]_+^{\alpha-1}$ (compare with equation 5), and $f(t)$ is obtained via the relation of equation (6). The parameters used to generate the plots are as follows. For a and b: $\tau=0$ ms, $\alpha=2$, and r as appears on the figure. For (c,d): $\tau=0$ ms, $r=20$ Hz, and α as appears on the figure. For (e,f): $\alpha=$, $r=40$ Hz, and τ as appears on the figure.
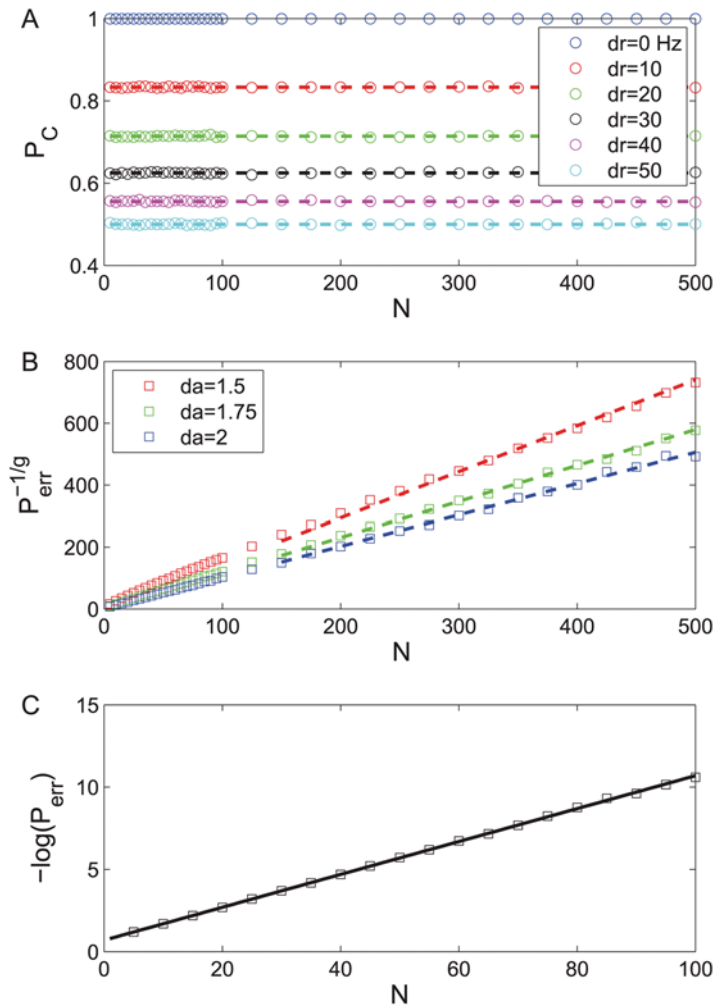doi:10.1371/journal.pcbi.1000286.g001

Figure 2. tWTA performance in a 2AFC discrimination task between stimulus 0° and $\delta\theta$ in a two-column model as function of the number of cells in the population. Open symbols show numerical estimation of the tWTA performance as obtained by averaging the probability of correct discrimination over 10^6 realizations of the stochastic neural responses. Probability distribution of first spike times followed an IHPP with the following PSTHs. (a) Scale parameter tuning: $\rho(t|\theta - \phi_i) = r(\theta - \phi_i)\Theta(t)$ with $r(0) = 50$ Hz and $dr = \frac{r(\delta\theta)}{r(0)} = 0, 0.2 \ldots 1$ from top to bottom. The dashed lines show the analytical prediction of equation (9). (b) Shape parameter tuning: $\rho(t|\theta - \phi_i) = r\Theta(t)t^{\alpha(\theta - \phi_i)}$ with $r(0) = 50$ Hz, $\alpha(0) = 1$ and $da = \frac{\alpha(\delta\theta)}{\alpha(0)} = 1.5, 1.75, 2$ from top to bottom. The tWTA performance is shown in terms of $P_\varepsilon^{-1/g}$ where $g = \frac{\alpha(\delta\theta)}{\alpha(0)} - 1$. The dashed lines show linear regression lines in keeping with the prediction of equation (12). (c) Delay parameter tuning: $\rho(t|\theta - \phi_i) = r\Theta(t - \tau(\theta - \phi_i))$ with $r(0) = 50$ Hz, $\tau(0) = 0$ and $\tau(\delta\theta) = 2$ ms. The tWTA performance is shown in terms of minus the log of the error rate. The solid line shows the analytical prediction of equation (15).
doi:10.1371/journal.pcbi.1000286.g002

the neural stochastic response. The dashed line shows the analytical prediction of equation 9 with $\alpha = 1$.

Effect of shape parameter tuning. In the case where only the shape parameter, α, is tuned to the stimulus, we write:

$$f(t|\theta) = r^{\alpha(\theta)}[t]_+^{\alpha(\theta) - 1} \qquad (10)$$

where r is independent of θ. We assume that population 1, with preferred direction $\phi_1 = 0$, fires faster than population 2, with preferred direction $\phi_2 = \delta\theta$, given stimulus $\theta = 0$, in the sense that for short times the probability of firing of cell in population 2 is larger than that in population 1; hence, $\alpha(0) < \alpha(\delta\theta)$. To compute P_C in the limit of large populations, equation (4), it is convenient to make a change of variables to $u(t) = -[\Phi(t|0) + \Phi(t|\delta\theta)]$, yielding:

$$P_C = N \int_0^\infty du \frac{1}{1 + \frac{\dot\Phi(t(u)|\delta\theta)}{\dot\Phi(t(u)|0)}} e^{-Nu} \qquad (11)$$

where $\dot\Phi$ is the time derivative of Φ, $\dot\Phi(t|\theta) \equiv \frac{d}{dt}\Phi(t|\theta) = -f_0(t)e^{-\Phi(t|\theta)}$. To leading order in small t for $\alpha(0) < \alpha(\delta\theta)$, $t(u) \approx \frac{(\alpha(0)u)^{1/\alpha(0)}}{r}$. Applying Watson's Lemma [23] we obtain the asymptotic approximation for the error rate:

$$P_\varepsilon \propto N^{-\left[\frac{\alpha(\delta\theta)}{\alpha(0)} - 1\right]} \qquad (12)$$

Hence, in this case, the probability of error decays algebraically with N to zero. This scaling of the readout accuracy with population size is similar to the scaling of the conventional rate-WTA accuracy with population size [24]. For small $\delta\theta$, $\alpha(\delta\theta) \approx \alpha(0) + \delta\theta\alpha'(0)$, we obtain:

$$\Delta\theta^{-1} \approx \frac{\alpha(0)\log(N)}{\alpha'(0)|\log(1-P_{th})|} \quad (13)$$

Thus, although in this case tWTA sensitivity improves by utilizing larger populations, this logarithmic improvement is extremely slow. Figure 2b shows the discrimination error rate to the power of $-1/\gamma = -\left(\frac{\alpha(\delta\theta)}{\alpha(0)} - 1\right)^{-1}$ as a function of N for different values of $\frac{\alpha(\delta\theta)}{\alpha(0)} = 1.5, 1.75, 2$ from top to bottom. The open squares are estimates of the tWTA accuracy obtained by averaging tWTA accuracy over 10^6 realizations of the neural stochastic response. The dashed lines show linear regression fits to the curves, in keeping with the asymptotic relation of equation (12).

Effect of delay parameter tuning. In the case where the delay parameter, τ, is the only the parameter that is tuned to the stimulus, we write:

$$f(t|\theta) = f_o(t-\tau(\theta))\Theta(t-\tau(\theta)) \quad (14)$$

where $\Theta(x)$ is the Heavyside function: $\Theta(x) = 1$ for $x > 0$ and 0 otherwise. In this case. we find (see Methods) that the probability of error decays exponentially fast with the population size, N:

$$P_\varepsilon \equiv 1 - P_C = A(\alpha)N^{-B(\alpha)}e^{-N|\Phi(\tau|0)|} \quad (15)$$

$$B(\alpha) = \begin{cases} 0 & 0 < \alpha \le 1 \\ \alpha-1 & 1 < \alpha \end{cases} \quad (16)$$

where $A(\alpha)$ is defined in Methods. Hence, in this case, the tWTA error rate decays to zero exponentially with N, in contrast to the slow algebraic scaling of the conventional rate-WTA accuracy with the population size [24]. For small $\delta\theta$, we can expand the delay parameter, τ, in $\delta\theta$ and approximate $\tau(\delta\theta) \approx \tau'\delta\theta$; for small $\delta\theta$, we thus find that tWTA sensitivity grows algebraically with N:

$$\Delta\theta^{-1} = \frac{\log(1-P_{th})}{\dot{\Phi}(\tau|0)}N \quad (17)$$

in contrast to the logarithmic scaling of the conventional rate-WTA sensitivity with population size [24]. Figure 2c shows minus the logarithm of the discrimination error rate in the case of delay parameter tuning to the stimulus. The open squares are estimates of the tWTA accuracy obtained by averaging tWTA accuracy over 10^6 realizations of the neural stochastic response. The solid line shows the analytical prediction of equation (15).

The different effects exerted by scale, shape and delay parameters on the scaling of the tWTA accuracy with the population size highlights the sensitivity of the tWTA to fine details of the first-spike-time distribution. Nevertheless, in the general case, all parameters will be tuned to the stimulus. The dominant contribution to the tWTA accuracy will result from the tuning of the delay parameter. Hence, the tWTA error rate will decay exponentially fast to zero with N, and the sensitivity will scale algebraically with N. We will therefore focus hereafter on models in which the delay parameter is tuned to the stimulus and ignore the tuning of other parameters to the stimulus.

Two important factors may have a considerable effect on the tWTA accuracy are addressed below. The first is noise correlations in the fluctuations of first spike times of different cells. It has been shown that noise correlations have a considerable effect on population code readout accuracy [25–29]. The second factors is nonzero baseline firing rate.

Effect of correlations on the tWTA accuracy. How should the covariance between first spike times of different cells be modeled? One possible mechanism that can cause correlated firing is having a shared input. Two cells that receive a common input that fluctuates above its mean will integrate it over time and reach spiking threshold sooner than their average first spike time. If the common input fluctuates below its average value, spike time of both cells will be delayed. It is reasonable to assume that cells that are *functionally* close, i.e., have similar preferred directions, will have more common input. Hence, their first spike times are expected to be more positively correlated. motivated by this intuition, we model correlations by adding a uniform random shift, $t^{(i)c}$, to the spike times of the cells in column $i = 1, 2$, which represents the effect of fluctuations in shared inputs to cells in every column. Thus, we write the first spike time $t_k^{(i)}$ of neuron k in population i as the sum of a correlated term and an independent term:

$$t_k^{(i)} = t^{(i)c} + t_k^{(i)u} \quad (18)$$

where $\left\{t_k^{(i)u}\right\}$ are statistically independent, given the stimulus, with probability distribution $f_i^u(t|\theta)$. We assume that, given stimulus θ, $f_2^u(t|\theta)$ is delayed relative to $f_1^u(t|\theta)$ by τ, i.e., $f_2^u(t|\theta) = 0$ for $t < \tau$ whereas $f_1^u(t|\theta) > 0$ for $t > 0$. The correlated components, $t^{(1)c}$ and $t^{(1)c}$, are independent, with probability distribution $f^c(t)$. In the limit of large N, the probability of correct discrimination is given by (see Methods):

$$\lim_{N\to\infty} P_C = \int_0^\infty dt f^c(t) \int_{t+\tau}^\infty dt' f^c(t') \quad (19)$$

Hence, for large populations, the uncorrelated fluctuations can be ignored, and the probability of correct discrimination saturates to a size-independent limit. Figure 3 shows the performance of the tWTA, in terms of percent correct discrimination, as a function of the number of cells in each column, N, for increasing values of $\tau_c = 0, 1, 2, 3 \; ms$ from top to bottom. In the simulations, we used a model in which only the delay parameter is tuned to the stimulus. Specifically we took: $f_i^u(t|\theta) = re^{-r(t-\tau(\theta-\phi_i))}\Theta(t-\tau(\theta-\phi_i))$ with $\tau(0) = 0$ and $\tau(\delta\theta) \equiv \tau = 2$ ms. For the correlated, part we used $f^c(t) = e^{-t/\tau_c}/\tau_c$. In this case we obtain (see Methods):

$$P_C = 1 - \frac{1}{2}\left(\frac{1}{1-\left(\frac{N}{N_{eff}}\right)^{-2}}e^{-N_{eff}r\tau} + \frac{1}{1-\left(\frac{N}{N_{eff}}\right)^2}e^{-Nr\tau}\right) \quad (20)$$

$$N_{eff} = \frac{1}{r\tau_c} \quad (21)$$

In the absence of correlations, $\tau_c \to 0 \Rightarrow N_{eff} \to \infty$, equation (20) converges to equation (15) with $\alpha = 0$ and $\Phi(\tau|0) = -r\tau$. The error rate, $P_{err} = 1 - P_C$, decays to zero exponentially with the number of cells, N. In the presence of correlations, $\tau_c > 0$, for small populations, $N \ll N_{eff}$, the tWTA error rate decays exponentially with N, as in the uncorrelated case, equation (15). When $N \sim N_{eff}$,
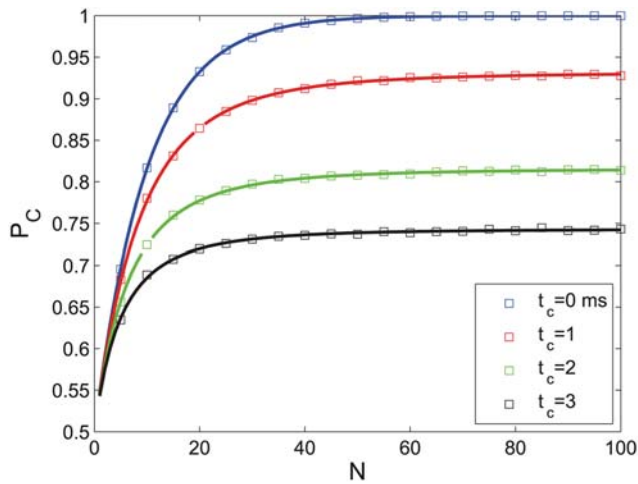
Figure 3. Effect of correlations on the tWTA readout accuracy.
The probability of tWTA correct response, P_C, in the presence of noise
correlations is shown as a function of the population size, N. Open
squares show numerical estimation of the probability of correct
response by averaging over 10^5 trials of simulating the network
stochastic response. The model was defined as in section 'effect of
correlations on the tWTA accuracy'. We write the first spike time $t_k^{(i)}$ of
neuron k in population i as the sum of a correlated term and an
independent term: $t_k^{(i)} = t^{(i)c} + t_k^{(i)u}$ (see equation 18), where $\{t_k^{(i)u}\}$ are
statistically independent, given the stimulus, with probability distribution
$f_i^u(t|\theta)$. Specifically, here we took: $f_i^u(t|\theta) = re^{-r(t-\tau(\theta-\phi_i))}\Theta(t-\tau(\theta-\phi_i))$
with $\tau(0)=0$ and $\tau(\delta\theta)\equiv\tau=2$ ms. The probability density of the
correlated part, $t^{(i)c}$, is given by $f^c(t)=e^{-t/t_c}/t_c$. The parameters that
were used for the simulations are: $r=50$ Hz, $\tau=2$ ms and $\tau_c=0,1,2,3$ ms
from top to bottom. The solid lines show the analytical result of
equation (20).
doi:10.1371/journal.pcbi.1000286.g003

tWTA performance reaches the saturation regime, and tWTA
accuracy converges to a finite limit for $N\to\infty$:

$$\lim_{N\to\infty} P_C = 1 - \frac{1}{2}e^{-N_{eff}r\tau} = 1 - \frac{1}{2}e^{-\frac{\tau}{\tau_c}}, \quad (\tau_c>0) \quad (22)$$

Hence, in the presence of correlations for large N, the tWTA error
rate is an increasing function of τ_c, which saturates to chance level
(chance level: $P_C^{chance} = 1 - P_\varepsilon^{chance} = \frac{1}{2}$) in the limit of $\tau_c\to\infty$.

Effect of baseline firing on tWTA accuracy. In the above
analysis we assumed zero baseline firing for all cells. However,
nonzero baseline firing may have a significant effect on the tWTA
accuracy. To incorporate baseline firing into our model, it is most
convenient to use the framework of the IHPP, which is defined by
the PSTH. The PSTHs of the two populations are modeled by:

$$\rho_i(t|\theta) = r_o + (r-r_o)\Theta(t-T-\tau_i(\theta-\phi_i)) \quad (23)$$

where, r_o is the baseline firing rate ($r_o<r$) and T is the duration in
which both columns fire at baseline prior to responding selectively
to the stimulus. The function $\tau(\theta)$ is the tuning of the delay
parameter. As above we take $\tau(0)=0$ and $\tau(\delta\theta)\equiv\tau$. In this case,
we find:

$$P_C = \frac{1}{2} + a(e^{-b_1 N} - e^{-b_2 N}) \quad (24)$$

$$a = \frac{1}{1+\frac{r_o}{r}} - \frac{1}{2} \quad (25)$$

$$b_1 = 2Tr_o \quad (26)$$

$$b_2 = 2Tr_o + \tau(r_o+r) \quad (27)$$

Figure 4 shows the probability of correct discrimination as a
function of N for different values of $T=0,1,5,10$ ms from top to
bottom. For any positive T, the probability of correct discrimi-
nation, P_C, decays to chance level, $P_C^{chance}=0.5$, exponentially fast
with N for large N. This decay results from the fact that the
probability of not spiking in the time interval before time T decays
to zero exponentially with N. For $T=0$, the probability of correct
response will saturate exponentially to $P_C^{max} = \frac{1}{1+\frac{r_o}{r}}$ (compare
with equation (9)) which can be high for low baseline firing rate,
$r_o/r\ll1$. For small N, there exists a region,
$N < N_{max} = \frac{\log\left(1+\frac{\tau}{T}\left[1+\frac{r}{r_o}\right]\right)}{\tau(r+r_o)}$, in which P_C increases with N.

The temporal n Winners-Take-All (n-tWTA). To
overcome the detrimental effect of baseline firing we generalize
the tWTA to a family of readouts, n-tWTA, that are determined
by the subgroup of cells that fired the first n spikes. In a 2AFC
competition between two homogeneous columns, the n-tWTA
estimates the stimulus by the preferred direction of the column
that fired the first n spikes. In the model of delayed step function
response PSTH, equation (23), spikes that are fired in the absolute
delay period, from time 0 to time T, are independent of the
stimulus and hence carry no information. The informative time of
spiking is that from time T to time $T+\tau$, where firing rates of the
cells depend on the stimulus. For a given population size, N, the
mean number of spikes fired during the absolute delay time is
Nr_oT. During the informative period, an average of $Nr\tau$ spikes is
being fired by the informative group. Taking $Nr\tau > n \gg Nr_oT$
diminishes the detrimental effect of baseline firing and conserves
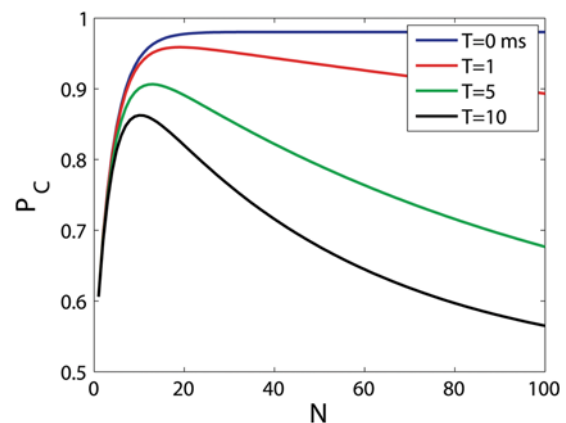the essential information embedded in the temporal order of the

Figure 4. Effect of baseline firing on tWTA readout accuracy.
The probability of tWTA correct response, P_C, in the case of nonzero
baseline firing is shown as a function of the population size, N,
equation (24), for $T=0,1,5,10$ ms from top to bottom. Parameters used
for this graph are: $\tau=5$ ms, $r=50$ Hz and $r_o=1$ Hz.
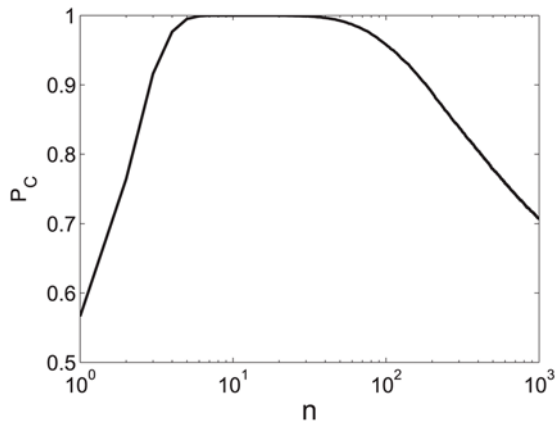doi:10.1371/journal.pcbi.1000286.g004

Figure 5. Performance of the *n*-tWTA readout in a 2AFC discrimination task in a two-column model. The probability of correct discrimination of the *n*-tWTA readout is shown as function of *n*. The probability of correct discrimination was estimated by averaging over 10^5 realizations of the neural stochastic response in an IHPP model for spike time distribution as defined in equation (23) with: $N=100$, $r=50$ Hz, $r_o=1$ Hz, $T=10$ ms and $\tau=5$ ms.
doi:10.1371/journal.pcbi.1000286.g005

neural responses. Figure 5 shows the percent correct discrimination of the *n*-tWTA, as a function of *n*. In this case, the average number of baseline spikes fired during the absolute delay time is $Nr_oT=1$, and P_C does indeed increase as *n* is increased from $n=1$ and to almost perfect discrimination at about $n=5$. During the informative period of spiking, an average of $Nr\tau=25$ spikes are fired by the 'correct' group. As expected, the probability of correct discrimination deteriorates for $n>Nr\tau=25$. In this example, the performance of the *n*-tWTA will decay to chance level in the limit of large *n*, since we did not incorporate any scale differences in the firings of the two populations. Thus, a reasonable choice of *n* can eliminate the effect of baseline firing and greatly improve the performance of the tWTA.

Note that the optimal region for *n*, depends on the population size. For any fixed *n*, increasing the population size increases the number of baseline spikes fired during the absolute delay period, Nr_oT. Hence, for $N>n/r_oT$ the *n*-tWTA performance will decay to chance level. An alternative *n*-tWTA generalization is to estimate the stimulus by the preferred direction of the first *single* cell that fired *n* spikes, see [2]. Results for this later generalization are qualitatively similar to those of the former in this model.

tWTA Estimation Accuracy in a Hypercolumn Model

The model. We study the tWTA estimation accuracy in a hypercolumn model of N cells coding for an angular variable, θ, such as the coding for the direction of motion of a visual stimulus by MT cells. Each cell $k\in\{1,\dots N\}$ is characterized by its preferred direction $\phi_k=\frac{2\pi k}{N}$ to which it responds fastest. Spike time distributions of different cells are modeled by independent IHPPs with PSTH $\rho_k(t|\theta)=\rho(t|\theta-\phi_k)$, for cell k, $k=1,\dots N$.

The tWTA estimate of the stimulus is given by the preferred direction, ϕ_k, of the cell k that fired the first spike

$$\hat{\theta}=\arg\min_{\phi_k}\{t(\phi_k)\} \qquad (28)$$

where $t(\phi_k)$ denotes the time of the first spike of cell k, following presentation of the stimulus. Throughout this section, we quantify tWTA sensitivity by the inverse of the root-mean-square estimation error, $\Delta\theta^{-1}\equiv 1\Big/\sqrt{\langle(\theta-\hat{\theta})^2\rangle}$, where $\langle X\rangle$ denotes

averaging of X over the distribution of spike times for a given external stimulus θ.

tWTA accuracy in the absence of correlations. The probability of the tWTA estimator to be $\hat{\theta}\in\{\phi_k\}_{k=1}^N$, is given by the probability that the first spike in the population was fired by the cell with preferred direction $\hat{\theta}$:

$$P(\hat{\theta}|\theta)=\int_0^\infty dt\rho(t|\theta-\hat{\theta})e^{-\sum_{k=1}^N\int_0^t dt'\rho(t'|\theta-\phi_k)} \qquad (29)$$

Empirical examples of first spike time tuning to an angular external stimulus is shown for example in [1,2]. Since tuning of the delay parameter makes the dominant contribution to the tWTA accuracy (see above), we now analyze the case of a delayed step function PSTH model with stimulus-independent scale and shape parameters. Specifically, we take the instantaneous firing rate of cell k with preferred direction ϕ_k, given that stimulus θ was presented at time $t_o=0$, to be:

$$\rho_k(t)=r\Theta(t-\tau(\phi_k-\theta)) \qquad (30)$$

This simple choice of PSTH does not limit the generality of our results but rather clarifies the analysis such that our conclusions are not obscured by non-relevant parameters. Figure 6 shows typical population response to stimulus $\theta=0$. The dots on row ϕ show the spike times of a single cell with preferred direction ϕ. The dashed line shows the delay tuning function, $\tau(\phi_k-\theta)=5[1-\cos(\phi_k-\theta)]$ ms, which yields the minimum possible spike time for each preferred direction.

The delay tuning function, $\tau(\phi)$, is assumed to be a periodic function of ϕ. We further assume that the delay function, $\tau(\phi)$, is a continuous, even function of its argument with a single minimum at $\phi=0$. For cells with preferred directions close to the stimulus,
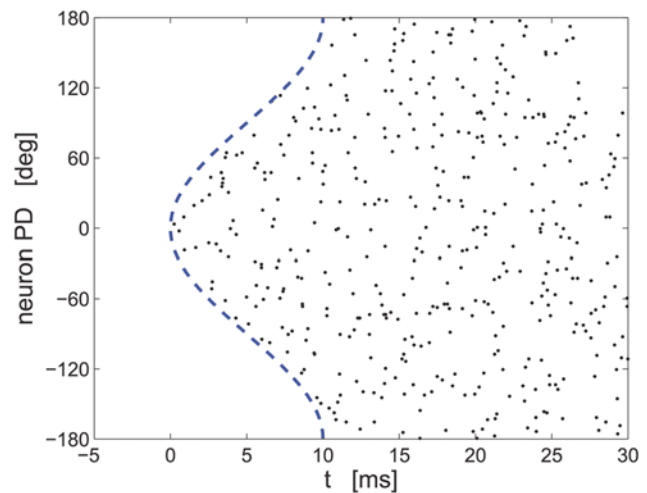

Figure 6. Simulation of a hypercolumn population raster plot. Spiking responses of 360 cells coding for an external stimulus $\theta=0°$ during a single trial are shown. Each line shows the spike-times of a single cell. The cells are arranged according to their preferred directions. Spike times of cell with preferred direction ϕ was modeled by an IHPP with PSTH $\rho(t)=r\Theta(t-\tau(\phi-\theta))$, where $r=50$ Hz is the rate and the latency function is $\tau(\varphi)=5[1-\cos(\varphi)]$ ms. The dashed line shows $\tau(\phi-\theta)$.
doi:10.1371/journal.pcbi.1000286.g006

we can approximate the delay function by:

$$\tau(\phi_k - \theta) \approx \beta|\phi_k - \theta|^\alpha \qquad (31)$$

where $\alpha \in [1,2]$ characterizes the delay tuning function near its unique minimum, for a smooth delay function $\alpha = 2$, and β is a constant in units of time. Since the tWTA is affected only by the fastest cells, we can use the approximation of equation (31) to describe the delay function of the entire hypercolumn, bearing in mind that the likelihood of cells with preferred directions that are far removed from the stimulus to affect the tWTA decays exponentially fast with N.

Using the continuum limit approximation for the exponent in the right hand side of equation (29), we evaluate the conditional probability density of the estimation error of $\delta\hat{\theta} \equiv \theta - \hat{\theta}$ and obtain:

$$P\left(\delta\hat{\theta}\right) \sim \int_{|\delta\theta|^\alpha}^{\pi^\alpha} dt \exp\left\{-\frac{\alpha}{\alpha+1}\beta N t^{\frac{\alpha+1}{\alpha}}\right\} + e^{-\frac{\alpha}{\alpha+1}\beta N} \qquad (32)$$

In the limit of large N, $P\left(\delta\hat{\theta}\right)$ is of $\mathcal{O}(1)$ in N for $\delta\hat{\theta} = \mathcal{O}\left(N^{-\frac{1}{1+\alpha}}\right)$ and decays exponentially with N for $\delta\hat{\theta} \gg N^{-\frac{1}{1+\alpha}}$. Hence, we obtain the following scaling law for the tWTA accuracy:

$$\Delta\theta^{-1} \propto N^{1/(1+\alpha)} \qquad (33)$$

As in the two-column competition in the 2AFC paradigm, the sensitivity of the tWTA readout in a hypercolumn model scales *algebraically* fast with N, in the absence of noise correlations and in the limit of low baseline firing. This fast scaling is in contrast to the slow logarithmic scaling of the conventional rate-WTA readout accuracy wih the population size [24]. Figure 7 shows tWTA sensitivity, in terms of the inverse root mean square estimation error, as a function of the population size in a hypercolumn model for $\alpha = 1, 1.2 \ldots 2$ from top to bottom. The open squares show numerical estimation of the sensitivity as obtained by averaging tWTA error over 10^4 realizations of simulating the network stochastic response. The solid lines show fits using the analytical result of equation (33) with $\alpha = 1, 1.2 \ldots 2$ from top to bottom.

Effect of correlations on the tWTA estimation accuracy. To model first spike time correlations in a hypercolumn, we write the spike times of each cell as the sum of correlated and uncorrelated parts

$$t_k = t_k^u + x\cos(\phi_k) + y\sin(\phi_k) + z \qquad (34)$$

where the uncorrelated parts, t_k^u, are taken to be distributed according to an IHPP with a PSTH $\rho_k(t) = r\Theta(t - \tau(\phi_k - \theta))$. For the sake of simplicity, we take $\tau(\phi) = \tau(1 - \cos(\phi))$. The terms x, y and z are the correlated components of the spike times. The z term represent the effect of shared input to the entire hypercolumn, whereas, x and y represent the effect of shared input that is stronger for columns that are functionally closer, i.e., have smaller preferred directions difference. We assume that the correlated noise has zero means $\langle x \rangle = \langle y \rangle = \langle z \rangle = 0$ and variance $\langle x^2 \rangle = \langle y^2 \rangle = \tau_c^2$ and $\langle z^2 \rangle = \overline{\tau}^2$. Figure 8 shows typical realizations of the population response during a single trial of presenting stimulus $\theta = 0$ in the presence of noise correlations. In Figure 8b $\tau_c = 0$ and $\overline{\tau} = 10$ ms. The uniform correlations generates collective fluctuations that shift the entire population response right (as in the specific realization in the figure) and left of the dashed line that shows $\tau(\phi_k - \theta)$. Nevertheless, this fluctuation
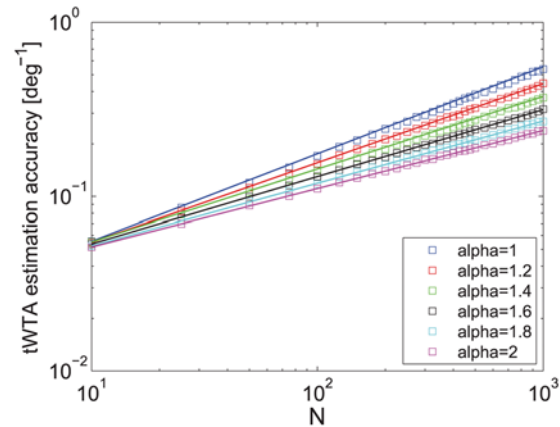

Figure 7. Estimation accuracy of the tWTA readout in a hypercolumn model. The accuracy of the tWTA readout, in terms of one over the squared estimation error of estimating $\theta = 0$, is plotted as a function of the population size, in an IHPP hypercolumn population model, equation(30). The latency tuning was modeled by $\tau(\Delta\theta) = |\Delta\theta|^\alpha$ ms (where $\Delta\theta$ is measured in radian) with $\alpha = 1, 1.2, \ldots, 2$ from top to bottom. Accuracy was measured by averaging the squared estimation error over 10,000 trials of simulating the neuronal stochastic response (squares). The solid lines show the analytical fit using equation (33).
doi:10.1371/journal.pcbi.1000286.g007

exists in a collective mode of the neural responses that does not alter the order of firing and hence does not affect tWTA accuracy. In Figure 8a $\tau_c = 10$ ms and $\overline{\tau} = 0$. In this case, the collective fluctuations shift the response of the entire population up and down (as in the specific realization in the figure). These fluctuations limit the accuracy in which the tWTA can estimate the stimulus. In the limit of large N, the error is dominated by the correlated response. Neglecting the uncorrelated part of the fluctuations, we obtain (see Methods):

$$\Delta\theta = \frac{\tau_c}{\tau} \qquad (35)$$

where $\Delta\theta$ is measured in radians. Note that equation (35) takes the form of a signal-to-noise ratio, where the signal is the modulation amplitude of the delay function, τ, and the noise is the component of collective noise correlations that affect the tWTA estimation, τ_c. The tWTA sensitivity, equation (35), is independent of the collective fluctuations in the uniform direction, $\overline{\tau}$.

Figure 9 shows the asymptotic accuracy of the tWTA as a function of the noise-to-signal ratio $\frac{\tau_c}{\tau}$. The solid line shows the analytical result of equation (35) in the limit of large N. The open squares show numerical estimation of asymptotic accuracy using a population of size $N = 10^5$ cells. The finite size of the network limits the ability of the numerical estimate to follow the analytic curve at high accuracy (low noise levels). To compensate somewhat for this effect, an extremely high firing rate was used in the simulations.

Effect of baseline firing on the tWTA accuracy. The effect of nonzero baseline firing on tWTA estimation accuracy is studied in the framework of a hypercolumn IHPP model with a delayed step function PSTH. Specifically, we took the following PSTH for the response of cell k with preferred direction ϕ_k:

$$\rho_k(t) = r_o + (r - r_o)\Theta(t - T - \tau(\phi_k - \theta)) \qquad (36)$$

where T is the absolute delay, $\tau(\theta) \geq 0$ is the tuning of the delay
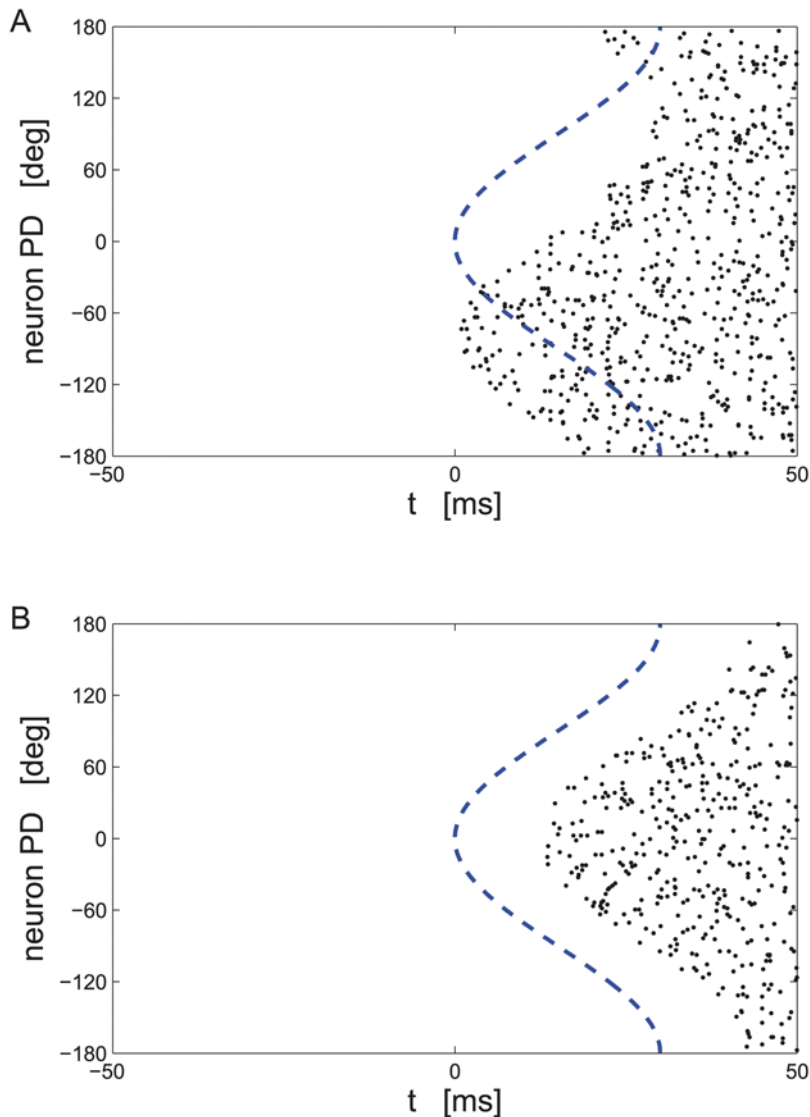
Figure 8. Simulation of a hypercolumn population raster in the presence of correlations. Spiking responses of 360 cells coding for an external stimulus $\theta = 0°$ during a single trial are shown. Every line shows the spike-times of a single cell. The cells are arranged according to their preferred directions. Spike times are distributed as defined in the section 'effect of correlations on tWTA accuracy', see equation (34), with $r = 50$ Hz and $\tau(\varphi) = 15(1 - \cos(\varphi))$ ms. For the correlated part: (a) $\tau_c = 10$ ms, $\bar{\tau} = 0$ ms; (b) $\tau_c = 0$ ms, $\bar{\tau} = 10$ ms. The dashed line shows $\tau(\phi - \theta)$.
doi:10.1371/journal.pcbi.1000286.g008

parameter with $\tau(0) = 0$. For $T = 0$ in the limit of large N, we can approximate the probability of the tWTA estimator to be $\hat{\theta} \in \{\phi_k\}_{k=1}^N$, equation (29), by:

$$P_o\left(\hat{\theta}|\theta\right) = \frac{1}{\mathcal{Z}}\left(1 + \frac{r - r_o}{r_o}e^{-Nr_o\tau(\hat{\theta} - \theta)}\right) \qquad (37)$$

where $\mathcal{Z}$ is a normalizing factor of the probability distribution. Figure 10a and 10b show histograms of tWTA estimations of stimulus $\theta = 0$ for $N = 360$ and $N = 3600$, respectively, in this model with $T = 0$. The solid line shows the analytical approximation, equation (37). The distribution is characterized by a narrow peak around zero error, with a width that decreases to zero as N grows to infinity and a uniform probability for large errors. The ratio of the peak distribution of the zero error (at $\hat{\theta} = \theta$) to the distribution of a specific large error is given by r/r_o. However, since the width of the peak decreases as N increases

(compare Figure 10a and 10b), the average squared estimation error increases for large N, even in for $T = 0$, in contrast to the effect of baseline firing in the 2AFC, where at $T = 0$ the probability of correct response is an increasing function of N. A hallmark of the tWTA readout is the high kurtosis of the estimation error.

In the case of $T > 0$, using equation (37), one obtains

$$P_T\left(\hat{\theta}|\theta\right) = \left(1 - e^{-Nr_oT}\right) + e^{-Nr_oT}P_o\left(\hat{\theta}|\theta\right) \qquad (38)$$

Hence, in this case the peak to base ratio of the distribution is decreased and decays exponentially to zero with the product Nr_oT. This effect is shown by the histogram of tWTA estimation errors in Figure 10c where we took $N = 3600$ and $T = 1$ ms (compare with Figure 10b where $N = 3600$ and $T = 0$ ms). The solid line shows the analytical approximation of equation (38).
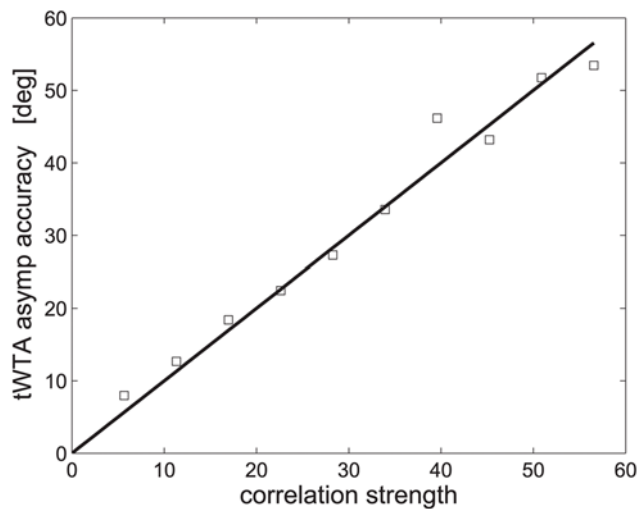
Figure 9. Effect of correlations on the asymptotic tWTA estimation accuracy in a hypercolumn model. tWTA accuracy, in terms of the root mean square estimation error, $\Delta\theta$, is shown as a function of the correlations' strength, $\frac{\tau_c}{\tau}\left(\frac{180}{\pi}\right)$, in a hypercolumn model, as defined in section 'effect of correlations on the tWTA estimation accuracy', see equation (34). The solid line shows the analytical asymptotic value, equation (35). Open squares show the numerical estimation of the asymptotic value as calculated by averaging the tWTA estimation error over 100 trials in a hypercolumn model of $N = 100{,}000$ cells. The latency function that was used was: $\tau(\varphi) = 2[1 - \cos(\varphi)]$ ms. To minimize the effect of finite N, an extremely high firing rate of $r = 5K$ Hz was used in the IHPP simulations.
doi:10.1371/journal.pcbi.1000286.g009

Discussion

At the time of the first spike, the tWTA is the ideal observer and, in the case of angle estimation, it is also the population vector readout. If a decision must be made at very short times, then the tWTA is the best readout. It is therefore important that we know and understand the capabilities and limitation of this readout. Scaling of the tWTA accuracy with the population size, N, can show a wide range of behaviors: from constant in N (equation 9), through logarithmic (equation 13) to algebraic (equation 17). These different scaling regimes depend on fine details of the tuning of the probability distribution of the first-spike-times or alternatively on the initial rise of the PSTH response to the stimulus. In the generic case in which scale, shape and delay parameters are all tuned to the stimulus, the tWTA accuracy will increase algebraically with N, in contrast to the expected logarithmic slow scaling of the conventional rate-WTA readout [24]. Nevertheless, the tWTA is expected to show high sensitivity to the inherent neuronal diversity at the level of single cell response properties (see e.g., [30]). This sensitivity of the tWTA predicts considerable subject-to-subject variability in psychophysical performance as well as large fluctuations in the psychophysical accuracy for the same subject under different stimuli conditions, such as discriminating θ_1 and $\theta_1 + \delta\theta$ versus discriminating θ_2 and $\theta_2 + \delta\theta$.

Noise correlations in the fluctuations of first-spike times of different cells have a drastically detrimental effect on the tWTA accuracy, limiting the effective number of degrees of freedom in the network and resulting in finite error levels, even in the limit of large N, see e.g., equations (21), (22) and (35) and Figures 3 and 9. This effect is similar to that has been reported in population coding literature [25–29,31], and depends on the correlations structure. Here we investigated the effect of correlations that had
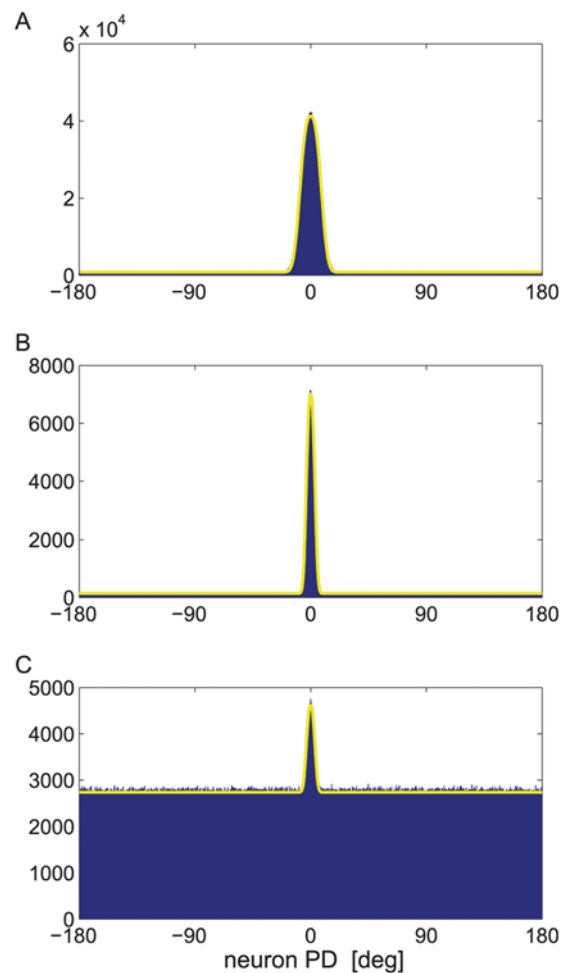

Figure 10. Effect of baseline firing on the tWTA estimation in a hypercolumn model. Histograms of tWTA estimation of stimulus $\theta = 0$ were obtained in a model of delayed step function response to the stimulus, equation (36), with $\tau(\theta) = \tau(1 - \cos(\theta))$, and parameters: $\tau = 50$ ms, $r = 50$ Hz and $r_o = 1$ Hz. Population size was $N = 360$ in (a) and $N = 3600$ in (b,c). Histograms were estimated using 10^6 repetitions in (a,b) and using 10^7 repetitions in (c). The solid lines are analytical approximations of equation (37) in (a,b) and equation (38) in (c).
doi:10.1371/journal.pcbi.1000286.g010

simple spatial structure and no temporal structure. A drastically detrimental effect on the tWTA accuracy is caused by neuronal response covariance which generates collective fluctuation that resembles the 'signal', i.e., similar to the tuning of the delay parameter (see Figure 8). For a detailed discussion on the effects of correlations structure see [27]. The temporal structure of response covariance may also have a considerable effect. For example, if the correlations depend on the absolute time, in a manner that they are negligible for small t and build up later in time, then they will not necessarily cause saturation of the tWTA accuracy. However, better empirical understanding of first spike time correlations is required to yield sufficient constraint for theoretical study of this issue. It is important to emphasize that by correlations we mean first spike time covariance of *simultaneously* recorded cells, in contrast to other types of correlations [5].

In a 2AFC, task nonzero baseline firing has a twofold detrimental effect on the tWTA accuracy. The first is in the case in which the onset of the tWTA readout precedes the stimulus response of the fastest cell in the entire population, $T > 0$. In this case, the tWTA accuracy will decrease as N is increased beyond

some optimal value $N_{\max}$. This effect can be minimized by obtaining a more accurate estimate for the minimal response time of the cells in the population, i.e., effectively decreasing T [5]. The second effect is a saturating effect, which limits the maximal accuracy that can be obtained by the tWTA, $P_C^{\max} = \dfrac{1}{1+\frac{r_o}{r}}$, even for $T=0$. Note that, although $P_C^{\max}$ is less than 1, psychophysical accuracy is also finite. The value of $P_C^{\max}$ can be rather high in cases in which the baseline firing is small relative to the stimulus response. These effects can be decreased for any given N by using a generalized n-tWTA readout that makes a decision according to the population that fired the first n spikes, see Figure 5. Nevertheless, for any given fixed value of n, increasing the population size, N, will decrease the n-tWTA performance to chance level, for $T>0$. Hence, for fast decisions there are advantages to reading out the responses of small neuronal populations rather than larger populations.

Baseline firing has similar detrimental effects on the tWTA readout in estimation tasks (see Figure 10). A hallmark of the tWTA readout that can serve as a prediction is its high kurtosis. There are various ways to generalize the tWTA to more than one spike in order to overcome the detrimental effect of baseline firing. One option is that readout is determined by the preferred direction of the single cell that fired the first n spikes. An alternative generalization is to define the readout by a 'vote' of cells that fired the first n spikes in the population. In the later case, different weights may be assigned to the votes. The utility of the different possible generalizations is expected to depend largely on the structure of the correlations in the neuronal initial dynamic response to the stimulus.

In a series of highly influential papers, Thorpe and colleagues (see e.g., [12,14]), have highlighted the possible role of spike latency as primary source of information in the CNS and have shown, for example, how an image falling on the retina could be reconstructed from a spike latency (see also work of [15]). In the context of this work, their readout could be thought of as a specific choice for the n-tWTA generalization. Here, we presented a systematic investigation of the tWTA accuracy that allows for comparison with psychophysical accuracy; hence, enables testing of the hypothesis that tWTA is actually used by the CNS. In addition, our analysis provides a framework that allows for the understanding and the investigation of the effects of correlations and baseline firing on the tWTA accuracy.

Neural network implementations of the tWTA. Considerable theoretical effort has been devoted to the investigation of neural network models that can implement the conventional rate-WTA [32–43]. These studies have focused on inputs that are constant in time and differ by their scale. However, it has been acknowledged that the temporal structure of the inputs may have a considerable effect on the WTA readout [43]. This effect shows the sensitivity of existing rate-WTA neural network models to the order of firing and demonstrates the capability of neural networks to implement a tWTA computation. Indeed one can imagine the responses of the (assumed excitatory) hypercolumn network that code for the external stimulus by their spike time latency, being input to a n-tWTA readout layer of laterally all to all connected inhibitory neurons. Once, input to a inhibitory cell crosses firing threshold of n excitatory post synaptic potential, it will fire and silence the rest of the network. Investigation of various neural network implementations, their limitations and deviations from the mathematically ideal tWTA and the computational consequences of these deviations if exist is beyond the scope of the current work and will be addressed elsewhere.

The neural code. To what extent does the CNS use the tWTA as a readout mechanism? Readout mechanisms used by the CNS are necessarily dynamic processes that may involve inhibition and hence generate WTA-like competition between inputs from different columns. If fast decisions between a small number of alternatives are required, then the tWTA can provide the correct result with high probability. In such a case, we predict that the readout will be determined by competition between relatively *small* groups of cells rather than by the entire cell population that responds to the stimulus so as to decrease the effect of baseline firing. Such decisions include, for example, estimation of the direction of motion of a visual stimulus at a low resolution of $45°$. However, for discrimination between many alternatives the tWTA is limited by the baseline firing. Why is this task more sensitive to baseline firing? Consider an example in which estimation of the direction of motion of a visual stimulus is required at a precision of $3.6°$. For this angular resolution, a population of at least $N=100$ cells is needed. Let us assume that at the stimulus onset the 'correct' cell fires at a rate of $r=100$ Hz while the rest of the population fires at a baseline rate of $r_o=1$ Hz. During the first $\tau=10$ ms of stimulus presentation, the 'correct' cell will fire an average of $\tau r=1$ spike, while the rest of the cells will fire an average of $(N-1)\tau r_o=0.99$ spikes; thus, the tWTA is expected to err in more than $3.6°$ in about 50% of the cases. Hence, fine estimation tasks cannot rely on the *single* first spike, and our theory predicts that in these cases the first n spikes must be considered where n should be larger than the average number of baseline spikes. How should the n-tWTA combine the information from the first n spikes? The answer to this question depends on the temporal structure of correlations, fine details of the PSTH, and on our assumptions on the computational capabilities of this readout and is beyond the scope of the current paper. The current work provides the essential framework for addressing this question. To further study the hypothesis that the CNS actually uses the tWTA better empirical understanding of the tuning of first spike time distribution to the stimulus, baseline firing, and the spatial and temporal structure of noise correlations is required.

Methods

Calculation of tWTA Accuracy in 2AFC in the Case of Delay Parameter Tuning to the Stimulus

Substituting equation (14) into equation (4), we obtain the probability of correct discrimination as sum of two terms:

$$P_C = I_1 + I_2 \tag{39}$$

$$I_1 = -N \int_0^{\tau(\delta\theta)} dt\, \dot{\Phi}(t|0) e^{N\Phi(t|0)} \tag{40}$$

$$I_2 = -N \int_{\tau(\delta\theta)}^{\infty} dt\, \dot{\Phi}(t|0) e^{N(\Phi(t|0)+\Phi(t|\delta\theta))} \tag{41}$$

The integral I_1, equation (40), can be evaluated exactly, yielding the contribution of

$$I_1 = 1 - e^{N\Phi(\tau|0)} \tag{42}$$

where we have used τ as shorthand for $\tau(\delta\theta)$. The contribution of I_2 to P_C is positive; hence, the tWTA error rate, in this case, will decay to zero *exponentially* with the population size N.

For the calculation of I_2, equation (41), we change variables to $u(t) = -[\Phi(t|0) - \Phi(\tau|0) + \Phi(t|\delta\theta)]$, yielding:

$$I_2 = Ne^{-N|\Phi(\tau|0)|} \int_0^\infty du \frac{1}{1 + \frac{\dot{\Phi}(t(u)|\delta\theta)}{\dot{\Phi}(t(u)|0)}} e^{-Nu} du \quad (43)$$

where for a small positive $\bar{t} \equiv t - \tau$, $\dot{\Phi}(\bar{t}|\delta\theta) \approx -r^\alpha \bar{t}^{\alpha-1}$. Assuming $\dot{\Phi}(\tau|0) \neq 0$, the leading term in $t(u)$, for small u, is given by:

$$t(u) \approx \begin{cases} \dfrac{(\alpha u)^{1/\alpha}}{r} & 0 < \alpha < 1 \\[2mm] \dfrac{u}{|\dot{\Phi}(\tau|0)| + r} & \alpha = 1 \\[2mm] \dfrac{u}{|\dot{\Phi}(\tau|0)|} & 1 < \alpha \end{cases} \quad (44)$$

Using Watson's Lemma to evaluate I_2 to leading order in N, we obtain

$$P_C \approx \begin{cases} 1 - e^{-N|\Phi(\tau|0)|}\left(1 - A(\alpha)N^{-B(\alpha)}\right) & 0 < \alpha < 1 \\[2mm] 1 - e^{-N|\Phi(\tau|0)|}A(\alpha)N^{-B(\alpha)} & 1 \leq \alpha \end{cases} \quad (45)$$

where

$$A(\alpha) = \begin{cases} \dfrac{|\dot{\Phi}(\tau|0)|\Gamma\left(\frac{1}{\alpha}\right)\alpha^{\frac{1}{\alpha}-1}}{r} & 0 < \alpha < 1 \\[2mm] \dfrac{r}{r + |\dot{\Phi}(\tau|0)|} & \alpha = 1 \\[2mm] \dfrac{r^\alpha \Gamma(\alpha)}{|\dot{\Phi}(\tau|0)|} & 1 < \alpha \end{cases} \quad (46)$$

and

$$B(\alpha) = \begin{cases} \frac{1}{\alpha} - 1 & 0 < \alpha < 1 \\[2mm] 0 & \alpha = 1 \\[2mm] \alpha - 1 & 1 < \alpha \end{cases} \quad (47)$$

Calculation of the tWTA Accuracy in 2AFC with Correlations

For a given stimulus θ, the probability density, $P_1(t)$, that the first spike of population 1 occurred at time t is

$$P_1(t) = -N \int_0^t dt' f^c(t') \dot{\Phi}_1^u(t-t') e^{N\Phi_1^u(t-t')} \quad (48)$$

where $\Phi_i^u(t) = -\log\left[1 - \int_0^t dt f_i^u(t|\theta)\right]$. The probability density, $P_2(t)$, that population 2 did not fire before time t is given by:

$$P_2(t) = -N \int_t^\infty dt'' f^c(t'') \dot{\Phi}_2^u(t-t'') e^{N\Phi_2^u(t-t'')} \quad (49)$$

Assuming the tWTA decides the stimulus is θ if the first spike comes from population 1, the probability of correct response is:

$$P_C = \int_0^\infty dt P_1(t) P_2(t)$$
$$= \int_0^\infty dt \int_0^t dt' \int_t^\infty dt'' N\dot{\Phi}_1^u(t-t') e^{N\Phi_1^u(t-t')} N\dot{\Phi}_2^u(t-t'') e^{N\Phi_2^u(t-t'')} f^c(t') f^c(t'') \quad (50)$$

In the limit of large N, $N \to \infty$, we obtain $N\dot{\Phi}_1^u(t-t') e^{N\Phi_1^u(t-t')} \to \delta(t-t')$, and due to the delay $N\dot{\Phi}_2^u(t-t'') e^{N\Phi_2^u(t-t'')} \to \Theta(\tau - [t-t''])$, we thus obtain

$$\lim_{N \to \infty} P_C = \int_0^\infty dt f^c(t)\left[1 - \int_\tau^\infty f^c(t'')\right] \quad (51)$$

In the specific example of Figure 3, the following model was used. The uncorrelated part of the first spike times was generated by an IHPP with a delayed step function PSTH, yielding the first-spike-time probability density: $f_i^u(t|\theta) = re^{-r(t-\tau(\theta-\phi_i))}\Theta(t-\tau(\theta-\phi_i))$ with $\tau(0) = 0$ and $\tau(\delta\theta) \equiv \tau$. For the correlated part, we used an IHPP model with PSTH, yielding: $f^c(t) = e^{-t/\tau_c}/\tau_c$. From equation (48), the probability density, $P_1(t)$, that the first spike in column 1 occurred at time t, is given by:

$$P_1(t) = Nr \int_0^t dt^{(1)c} \frac{1}{\tau_c} \exp\left\{ -\frac{t^{(1)c}}{\tau_c} - Nr\left(t - t^{(1)c}\right) \right\} \quad (52)$$

The probability density, $P_2(t)$, that no cell in column 2 had fired until time $t > \tau$, equation (49), is equal to the probability that the first spike of the cells in column 2 occurred at any time $t' > t$:

$$P_2(t)$$
$$= Nr \int_t^\infty dt' \int_0^{t'-\tau} dt^{(2)c} \frac{1}{\tau_c} \exp\left\{ -\frac{t^{(2)c}}{\tau_c} - Nr\left(t' - t^{(2)c} - \tau\right) \right\} \quad (53)$$

Note that for $t < \tau$, $P_2(t) = 1$. The probability of correct classification is given by:

$$P_C(t) = \int_0^\tau dt P_1(t) + \int_\tau^\infty dt P_1(t) P_2(t) \quad (54)$$

Substituting equations (52) and (53) into equation (54) and integrating one obtains the result of equation (20).

tWTA Accuracy in a Correlated Hypercolumn Model

We now turn to calculate tWTA asymptotic accuracy in the presence of correlations, see section 'Effect of correlations on the tWTA estimation accuracy'. In the limit of large N, the estimation error will be dominated by the correlated noise. We can, therefore, neglect the fluctuations of the uncorrelated part, t_{k1}^u (see equation (34)), replacing its distribution with a delta function at the first time the PSTH of cell k is larger than zero:

$$P\left(t_{k1}^u\right) = \delta\left(t_{k1}^u - \tau(\phi_k)\right) \quad (55)$$

In this case, for a specific realization of x, y and z we can write the first spike time of cell k as

$$t_{k1} = \tau[1 - \cos(\theta - \phi_k)] + x\cos(\phi_k) + y\sin(\phi_k) + z \quad (56)$$

Without loss of generality, we will assume $\theta = 0$. The tWTA estimate for the stimulus, $\hat{\theta}$ is obtained by taking the derivative of equation (56) with respect to ϕ_k and equating to zero at $\phi_k = \hat{\theta}$

$$(\tau - x)\sin\left(\hat{\theta}\right) + y\cos\left(\hat{\theta}\right) = 0 \quad (57)$$

For small errors, we can approximate:

$$\hat{\theta} \approx \frac{y}{\tau} \qquad (58)$$

and obtain:

$$\langle \delta \hat{\theta}^2 \rangle = \frac{\langle \delta y^2 \rangle}{\tau^2} = \left(\frac{\tau_c}{\tau} \right)^2 \qquad (59)$$

which is equivalent to the result of equation (35).

References

1. Osborne LC, Bialek W, Lisberger SG (2004) Time course of information about motion direction in visual area MT of macaque monkeys. J Neurosci 24: 3210–3222.
2. Nelken I, Chechik G, Mrsic-Flogel TD, King AJ, Schnupp JW (2005) Encoding stimulus information by spike numbers and mean response time in primary auditory cortex. J Comput Neurosci 19: 199–221.
3. Brugge JF, Reale RA, Hind JE (1996) The structure of spatial receptive fields of neurons in primary auditory cortex of the cat. J Neurosci 16: 4420–4437.
4. Brugge JF, Reale RA, Jenison RL, Schnupp J (2001) Auditory cortical spatial receptive fields. Audiol Neurootol 6: 173–177.
5. Chase SM, Young ED (2007) First-spike latency information in single neurons increases when referenced to population onset. Proc Natl Acad Sci U S A 104: 5175–5180.
6. Foffani G, Chapin JK, Moxon KA (2008) Computational role of large receptive fields in the primary somatosensory cortex. J Neurophysiol 100: 268–280.
7. Gütig R, Sompolinsky H (2006) The tempotron: a neuron that learns spike timing-based decisions. Nat Neurosci 9: 420–428.
8. Johansson RS, Birznieks I (2004) First spikes in ensembles of human tactile afferents code complex spatial fingertip events. Nat Neurosci 7: 170–177.
9. Furukawa S, Middlebrooks JC (2002) Cortical representation of auditory space: information-bearing features of spike patterns. J Neurophysiol 87: 1749–1762.
10. Jenison RL, Reale RA (2003) Likelihood approaches to sensory coding in auditory cortex. Network 14: 83–102.
11. Van Rullen R, Gautrais J, Delorme A, Thorpe S (1998) Face processing using one spike per neurone. Biosystems 48: 229–239.
12. Van Rullen R, Thorpe SJ (2001) Rate coding versus temporal order coding: What the retinal ganglion cells tell the visual cortex. Neural Comput 13: 1255–1283.
13. Thorpe S, Delorme A, Van Rullen R (2001) Spike-based strategies for rapid processing. Neural Netw 14: 715–725.
14. VanRullen R, Guyonneau R, Thorpe SJ (2005) Spike times make sense. Trends Neurosci 28: 1–4.
15. Wiener MC, Richmond BJ (2003) Decoding spike trains instant by instant using order statistics and the mixture-of-Poissons model. J Neurosci 23: 2394–2406.
16. Panzeri S, Petersen RS, Schultz SR, Lebedev M, Diamond ME (2001) The role of spike timing in the coding of stimulus location in rat somatosensory cortex. Neuron 29: 769–777.
17. Shamir M, Sen K, Colburn HS (2007) Temporal coding of time varying stimuli. Neural Comput 19: 3239–3261.
18. Gollisch T, Meister M (2008) Rapid neural coding in the retina with relative spike latencies. Science 319: 1108–1111.
19. Dubner R, Zeki SM (1971) Response properties and receptive fields of cells in an anatomically defined region of the superior temporal sulcus in the monkey. Brain Res 35: 528–532.
20. Britten KH, Shadlen MN, Newsome WT, Movshon JA (1992) The analysis of visual motion: a comparison of neuronal and psychophysical performance. J Neurosci 12: 4745–4765.
21. Tuckwell CH (1988) Introduction to Theoretical Neurobiology. New York: Cambridge University Press, Volume 2.
22. van Kampen NG (1997) Stochastic Processes in Physics and Chemistry. Amsterdam, The Netherlands: Elsevier Science B.V.
23. Orszag SA, Bender CM (1991) Advanced Mathematical Methods for Scientists and Engineers. New York: Springer-Verlag.
24. Shamir M (2006) The scaling of winner-takes-all accuracy with population size. Neural Comput 18: 2719–2729.
25. Zohary E, Shadlen MN, Newsome WT (1994) Correlated neuronal discharge rate and its implications for psychophysical performance. Nature 370: 140–143.
26. Abbott LF, Dayan P (1999) The effect of correlated variability on the accuracy of a population code. Neural Comput 11: 91–101.
27. Sompolinsky H, Yoon H, Kang K, Shamir M (2001) Population coding in neuronal systems with correlated noise. Phys Rev E Stat Nonlin Soft Matter Phys 64: 051904.
28. Mazurek ME, Shadlen MN (2002) Limits to the temporal fidelity of cortical spike rate signals. Nat Neurosci 5: 463–471.
29. Montani F, Kohn A, Smith MA, Schultz SR (2007) The role of correlations in direction and contrast coding in the primary visual cortex. J Neurosci 27: 2338–2348.
30. Ringach DL, Shapley RM, Hawken MJ (2002) Orientation selectivity in macaque V1: diversity and laminar dependence. J Neurosci 22: 5639–5651.
31. Averbeck BB, Latham PE, Pouget A (2006) Neural correlations, population coding and computation. Nat Rev Neurosci 7: 358–366.
32. Hertz J, Krogh A, Palmer RG (1991) Introduction to the Theory of Neural Computation. Redwood City, CA: Addison-Wesley Publishing Company.
33. Grossberg S (1973) Contour enhancement, short-term memory, and constancies in reverberating neural networks. Stud Appl Math L11: 213.
34. Amari S, Arbib MA (1977) In: Systems Neuroscience. Metzler J, ed. Boston: Academic Press. 119 p.
35. Ermentrout B (1992) Complex dynamics in WTA neural nets with slow inhibition. Neural Netw 5: 415–431.
36. Feng J, Hadeler KP (1996) Qualitative behaviour of some simple networks. J Phys A 29: 5019–5033.
37. Hahnloser RHR, Sarpeshkar R, Mahowald MA, Douglas RJ, Seung HS (2000) Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature 405: 947–951.
38. Ermentrout B, Wang JW, Flores J, Gelperin A (2001) Model for olfactory discrimination and learning in Limax procerebrum incorporating oscillatory dynamics and wave propagation. J Neurophysiol 85: 1444–1452.
39. Jin DZ, Seung HS (2002) Fast computation with spikes in a recurrent neural network. Phys Rev E 65: 051922.
40. Xie X, Hahnloser RH, Seung HS (2002) Selectively grouping neurons in recurrent networks of lateral inhibition. Neural Comput 14: 2627–2646.
41. Richards W, Seung HS, Pickard G (2006) Neural voting machines. Neural Netw 19: 1161–1167.
42. Fukai T, Tanaka S (1997) A simple neural network exhibiting selective activation of neuronal ensembles: from WTA to winners-share-all. Neural Comput 9: 77–97.
43. Lumer ED (2000) Effects of spike timing on WTA competition in model cortical circuits. Neural Comput 12: 181–194.