Dale's principle preserves sequentiality in neural circuits

Alberto Bernacchia[1], József Fiser[2], Guillaume Hennequin[1*], Máté Lengyel[1,2*]

[1]Computational & Biological Learning Lab, Department of Engineering, University of Cambridge, UK
[2]Department of Cognitive Science, Central European University, Hungary
*equal contributions

Cortical circuits obey Dale's principle: each neuron either excites or inhibits all its postsynaptic targets. There is no known principled justification for why this must be so; in fact, Dale's principle is considered – if at all – a mere constraint in neural network models. Here we provide a novel rationale for Dale's principle: networks with separate excitatory (E) and inhibitory (I) populations preserve the temporal relationships between their inputs, thus preventing spurious temporal correlations that could mislead spike timing-dependent plasticity (STDP). To show this, we study a recurrent firing rate network model with arbitrary nonlinear response functions. We assume that, in line with known Hebbian mechanisms at both excitatory and inhibitory synapses, the magnitudes of recurrent synaptic weights are proportional to the covariance of pre- and postsynaptic rates, while their sign is determined by the E/I identity of the presynaptic cell. We show that this connectivity pattern is both necessary and sufficient to ensure that neural circuit output will be non-sequential, if the input has no specific temporal ordering of its elements. Conversely, if there is some specific temporal ordering of inputs to different neurons, then the neural circuit output will also have sequences that reproduce those of the input. Our theory predicts the relative degree of sequentiality of V1 responses to visual stimuli with different statistics, which we confirmed in cortical recordings: stimuli that are similar in lacking temporal ordering evoke responses that differ in their sequentiality, depending on whether V1 has been adapted to them. Our results suggest a novel and unexpected connection between the ubiquitous Dale's principle and STDP, namely that Dale's principle acts as a control mechanism to guarantee that STDP will act only on input-driven temporal sequences, rather than on internally generated ones.

We consider a set of n neurons with internal activities (e.g. subthreshold membrane potentials) v_i, $i = 1, \dots, n$, and the following standard dynamics:

$$\tau \frac{dv_i}{dt} = -v_i(t) + \sum_{j=1}^{n} J_{ij}\, f_j\left[v_j(t)\right] + \xi_i(t) \qquad (1)$$

where J_{ij} is the synaptic weight between presynaptic neuron j and postsynaptic neuron i, $f_j[v]$ is the transfer function of neuron j mapping its internal activity into a firing rate, ξ_i is the external input received by neuron i and $\tau = 20ms$ is a single neuron time constant. We assume that the external input is a stationary Gaussian process, with some mean and covariance, $\Sigma^{\text{in}}_{ij}(s) = \langle \delta\xi_i(t+s)\, \delta\xi_j(t) \rangle$, where angular brackets denote averaging over different trials, δ denotes deviations from the mean, and s is the time lag. We assume that the system relaxes to a unique stationary distribution over $\mathbf{v}(t)$ which is approximately multivariate normal.

We consider the case when each synaptic weight J_{ij} is the product of the integrated covariance between the inputs to the pre- and postsynaptic neurons, $\bar{\Sigma}^{\text{in}}_{ij} = \int \Sigma^{\text{in}}_{ij}(s)\, ds$, and a presynaptic factor Δ_j whose sign depends on whether the presynaptic cell is excitatory (>0) or inhibitory (<0):

$$J_{ij} = \bar{\Sigma}^{\text{in}}_{ij}\, \Delta_j \qquad (2)$$

We further assume that all input correlations are positive, which in turn guarantees that Eq. 2 is consistent with Dale's principle. An alternative form for the synaptic matrix is $J_{ij} = \bar{\Sigma}^{\text{out}}_{ij} \Delta_j$, where $\bar{\Sigma}^{\text{out}}_{ij} = \int \Sigma^{\text{out}}_{ij}(s)\, ds$ and $\Sigma^{\text{out}}_{ij}(s) = \langle \delta v_i(t+s)\, \delta v_j(t) \rangle$ is the covariance of the two neurons' responses. This form is consistent with Hebbian learning and empirical observations [Cossell et al, Nature (2015)]. While both forms of synaptic matrix are consistent with our results, we use Eq. 2, as the input is given and we can fix the synaptic strengths accordingly.

Our first main result is the following: the network will generate non-sequential activity, i.e. all output cross-covariances will be temporally symmetric ($\Sigma^{\text{out}}_{ij}(s) = \Sigma^{\text{out}}_{ij}(-s)$) if, and only if, the input too is time-reversible ($\Sigma^{\text{in}}_{ij}(s) = \Sigma^{\text{in}}_{ij}(-s)$). Secondly, if the input has some degree of sequentiality, expressed as the matrix of "expected pair-wise time lags" in the network, then the output will show the same sequential activation up to a linear spatial transformation: $\int s\, \Sigma^{\text{out}}(s)\, ds = \mathbf{A} \left[\int s\, \Sigma^{\text{in}}(s)\, ds\right] \mathbf{A}^{\mathsf{T}}$. Finally, the form of the synaptic matrix we consider in Eq. 2 is both necessary and sufficient for the conservation of sequentiality.

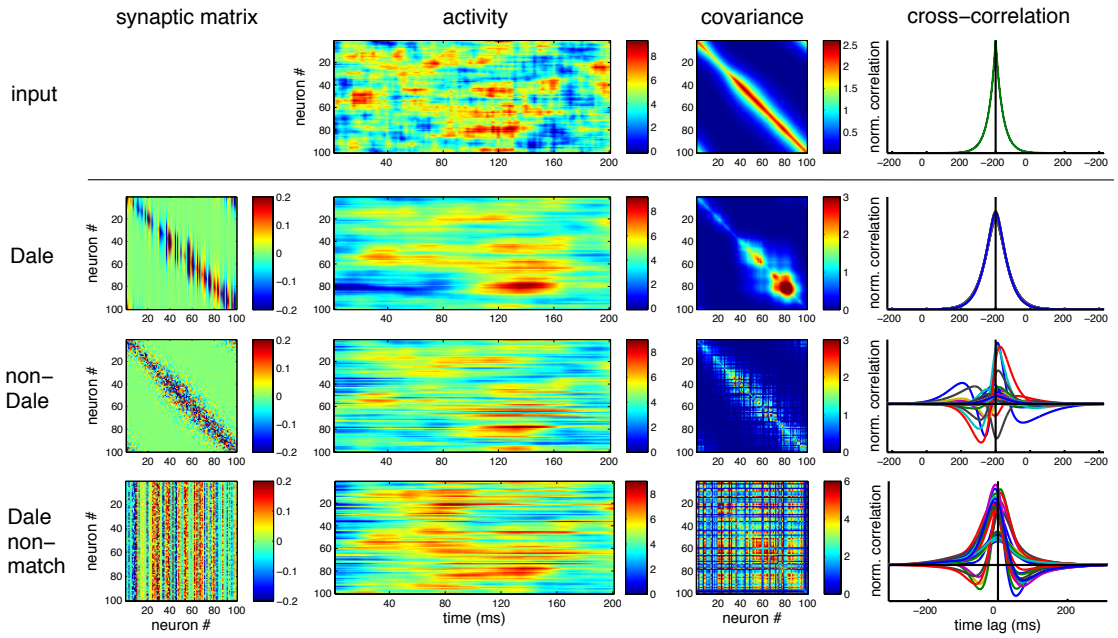

Figure 1. From top to bottom: Input current; network activity in the case of a Dale synaptic matrix; a non-Dale synaptic matrix; and Dale synaptic matrix with non-matching covariance. From left to right: the synaptic matrix; time series of network activity for all neurons (200ms sample); spatial covariance (at $s = 0$ time lag) between all neuron pairs; and normalized cross-correlations as a function of time lag. Cross-correlations of the input are time-symmetric, they remain time-symmetric for a Dale synaptic matrix (despite substantial changes in spatial covariances), but they become asymmetric for a non-Dale matrix or a Dale matrix with non-matching covariance.

Fig. 1 shows the dynamics of the model for an example non-sequential input (top). Cross-correlations of network activity are time-symmetric for a Dale synaptic matrix (2nd row), satisfying Eq. 2, they are not time-symmetric for a non-Dale synaptic matrix (3rd row), and they are also not time-symmetric for a Dale matrix that doesn't match the input covariance (4th row). Thus, unless the synaptic matrix follows Eq. 2, the network creates spurious sequentiality. STDP acting on these cross-covariances would lead to plasticity that does not reflect any input sequentiality.

Our theory makes non-trivial predictions concerning the temporal reversibility of V1 activity under different stimulation conditions. If the recurrent connectivity satisfies Eq. 2 for input covariances Σ^{in} corresponding to natural movies (to which it has been adapted), then responses to such movies should have the same degree of reversibility as the input. As natural movies can be modelled as being largely non-sequential at the level of features V1 extracts [Clopath et al, 2010] we expect weakly sequential responses. Noisy stimuli, to which V1 is not adapted, should instead elicit sequential activity. We tested these predictions in multiunit recordings from V1 of awake ferrets in different age groups spanning from eye opening (P30) to full maturity (P130+). Fig. 2 shows that temporal asymmetries are small but they are indeed larger for noise vs movie stimuli ($p=3\cdot10^{-4}$). Furthermore, temporal asymmetries grow with postnatal age (youngest age group vs rest, p=0.004). Our results suggest a novel and unexpected connection between the ubiquitous Dale's principle and STDP: preservation of sequentiality from input to output guarantees that STDP will only learn genuine input correlations.

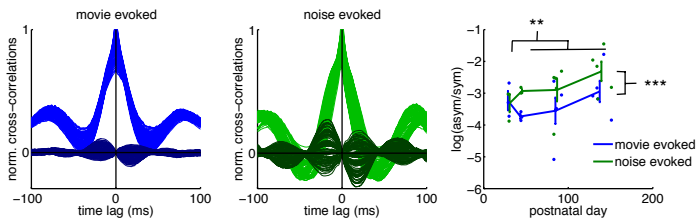

Figure 2. Left: Cross-correlations for movie- (light blue) and noise-evoked activities (light green) in multi-unit recordings from awake ferret V1, postnatal day 129; and their odd part extracted (dark blue and dark green). Right: Asymmetry (log area under odd part / area under even part) as a function of postnatal day in the two cases. Asymmetry is small, larger for noise-evoked activity and the difference from movie-evoked increases with age.