Goal-Directed Decision Making with Spiking Neurons – Convergence Proof –

Johannes Friedrich, Máté Lengyel

Computational and Biological Learning Laboratory
Department of Engineering, University of Cambridge, Cambridge, UK

January 29, 2016

In the main text we have shown that at the fixed point(s) the neural activity is, up to constants, the solution of the Bellman optimality equation. Here we proof convergence to such a fixed point for $\gamma < 1$. To reduce notational clutter, without loss of generality we set the constants k and λ_r equal to 1 and θ to 0, which just amounts to using appropriate units. In the dynamics $\tau_\mathrm{m}\dot{\mathbf{u}} = -\mathbf{u}+\mathbf{W}\boldsymbol{\lambda}-\eta\,\boldsymbol{\lambda}+\mathbf{w}^\mathrm{r}$ (Eq. 5, main text) $\boldsymbol{\lambda}$ depends implicitly on $\mathbf{u}$. An equivalent explicit expression is

$$\tau_\mathrm{m}\dot{\mathbf{u}} = (\mathbf{W}^\eta(\mathbf{u}) - \mathbb{1})\mathbf{u} + \mathbf{w}^\mathrm{r} = \mathbf{X}(\mathbf{u})\mathbf{u} + \mathbf{w}^\mathrm{r}. \tag{1}$$

Here we defined a new set of effective weights $\mathbf{W}^\eta(\mathbf{u})$, obtained by taking the matrix $\mathbf{W} - \eta\mathbb{1}$ and setting column i to zero if $u_i < 0$, and defined $\mathbf{X}(\mathbf{u}) := \mathbf{W}^\eta(\mathbf{u}) - \mathbb{1}$. Within each orthant of $\mathbf{u}$-space the matrix $\mathbf{W}^\eta(\mathbf{u})$ is constant and the dynamics described by a linear ODE with solution

$$\mathbf{u}(t) = e^{\mathbf{X}(t-t')/\tau_\mathrm{m}}\mathbf{u}(t') + \int_{t'}^{t} e^{\mathbf{X}(t-\tau)/\tau_\mathrm{m}}\mathbf{w}^\mathrm{r}\mathrm{d}\tau, \tag{2}$$

where t' is the time of initiation or when $\mathbf{u}$ entered the orthant.

We proceed to show that no eigenvalue μ of $\mathbf{X}(\mathbf{u})$ has positive real part for any $\mathbf{u}$. We arrange the matrices blockwise with actions varying within and states between blocks. Using abusive but concise notation, we write for the transition probability $P(s'|s,a) = p_{ij}$ with indices $i = As + a$ and $j = s'$, where $a \in \{1, 2, ..., A\}$ and $s \in \{1, 2, ..., S\}$. This also defines the matrix $\mathbf{P} = (p_{ij})$. Let further $\mathbf{E} = (e_{ij})$ be the $SA \times S$ matrix with $e_{ij} = 1$ if $i \in \{(j-1)A + 1, ..., jA\}$, i.e. $s = s'$, else 0. If $u_i \geq 0$ for all i then the matrix $\mathbf{X}$, which we denote in this case by $\mathbf{X}^+$, can be expressed based on Eq. (7-9, main text) as $\mathbf{X}^+ = (1+\eta)(\gamma\mathbf{P} - \mathbf{E})\mathbf{E}^T$. For the sake of illustration, e.g. with $S = A = 2$ and $\eta = 0$ one has:

$$\overbrace{\begin{pmatrix} \gamma p_{11}-1 & \gamma p_{11}-1 & \gamma p_{12} & \gamma p_{12} \\ \gamma p_{21}-1 & \gamma p_{21}-1 & \gamma p_{22} & \gamma p_{22} \\ \gamma p_{31} & \gamma p_{31} & \gamma p_{32}-1 & \gamma p_{32}-1 \\ \gamma p_{41} & \gamma p_{41} & \gamma p_{42}-1 & \gamma p_{42}-1 \end{pmatrix}}^{\mathbf{X}^+} = \underbrace{\left(\gamma \overbrace{\begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \\ p_{31} & p_{32} \\ p_{41} & p_{42} \end{pmatrix}}^{\mathbf{P}} - \overbrace{\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}}^{\mathbf{E}}\right)}_{\mathbf{C}} \overbrace{\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}}^{\mathbf{E}^T},$$

where we newly defined the matrix $\mathbf{C}$. The columns in each $A \times A$-block are identical, hence $S(A-1)$ eigenvalues are zero. In general, not all u_i are greater or equal zero, but some entries are negative. Let Z denote their index set. Therefore the columns of $\mathbf{W}^\eta(\mathbf{u})$ with indices z in Z are zero and the respective columns in $\mathbf{X}(\mathbf{u})$ have the entry -1 on the diagonal, $a_{zz} = -1$, and zeros elsewhere. By inspection $\mathbf{X}(\mathbf{u})$ has eigenvalue $\mu = -1$ with multiplicity $|Z|$ and eigenvectors with components $v_i = \delta_{iz}$. Characteristic of these negative eigenvalues is the exponential decay towards the fixed point, as seen in Fig. 2F, main text.

It will turn out useful to consider matrices $\mathbf{B}(\mathbf{u})$ where these diagonal elements are also set to zero. In order to see, how this affects the eigenvalues, imagine deriving the characteristic polynomial using the Laplace expansion of the determinant on the considered columns. For the original matrix $\mathbf{X}(\mathbf{u})$ the characteristic polynomial is $\det(\mathbf{X}(\mathbf{u}) - \mu\mathbb{1}) = (-1-\mu)M_{zz}$ with some minor M_{zz}. Setting the diagonal entry in the considered column to zero we obtain $(0-\lambda)M_{zz}$ with the same minor M_{zz}, hence one eigenvalue changes from -1 to 0 whereas the others remain the same. Sequentially setting

all $|Z|$ diagonal entries a_{zz} to zero yields $\mathbf{B}(\mathbf{u})$. In summary, the already found eigenvalue -1 with multiplicity $|Z|$ changes to an eigenvalue 0 with multiplicity $|Z|$, but all other eigenvalues, the ones we are still looking for, are left invariant.

The idea behind introducing $\mathbf{B}(\mathbf{u})$ and underlying the following is that $\mathbf{B}(\mathbf{u})$ has low rank, can be expressed as a product of the $SA \times S$ matrix $\mathbf{C}$ and a $S \times SA$ matrix $\mathbf{D}(\mathbf{u})$, and the non-zero eigenvalues of $\mathbf{CD}(\mathbf{u})$ are the eigenvalues of $\mathbf{D}(\mathbf{u})\mathbf{C}$. Let $\mathbf{D}(\mathbf{u})$ be obtained by setting columns of $\mathbf{E}^T$ with indices in Z to zero. With this definitions we have $\mathbf{B}(\mathbf{u}) = (1+\eta)\mathbf{CD}(\mathbf{u})$. We are interested in the eigenvalues of $\mathbf{Y} := \mathbf{D}(\mathbf{u})(\gamma\mathbf{P} - \mathbf{E})$, where we dropped the constant prefactor $1+\eta$, which is positive by assumption and thus does not affect the sign of the eigenvalues. Assuming e.g. $Z = \{2\}$ our example reads:

$$\overbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}}^{\mathbf{D}(\mathbf{u})} \left(\gamma \overbrace{\begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \\ p_{31} & p_{32} \\ p_{41} & p_{42} \end{pmatrix}}^{\mathbf{P}} - \overbrace{\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}}^{\mathbf{E}} \right) = \overbrace{\begin{pmatrix} \gamma p_{11} - 1 & \gamma p_{12} \\ \gamma(p_{31}+p_{41}) & \gamma(p_{32}+p_{42}) - 2 \end{pmatrix}}^{\mathbf{Y}}$$

The effect of left-multiplication with $\mathbf{D}(\mathbf{u})$ is the summation of the A rows corresponding to one state and all possible actions but skipping rows with index in Z. Note that, summarising the approximate values (i.e. the summed up activity over all possible actions for each state $\tilde{V}(s) + V_0 = \sum_{i|s_i=s} \lambda_i$) in one vector $\mathbf{v} = (\tilde{V}(1) + V_0, ..., \tilde{V}(S) + V_0)^T$, the resulting matrix $\mathbf{Y}$ has following interpretation: it is the matrix in the linear ODE $\tau_{\mathrm{m}}\dot{\mathbf{v}} = \mathbf{Y}\mathbf{v} + \mathbf{w}^{\mathrm{v}}$ describing the evolution of the approximate values. (The details of $\mathbf{w}^{\mathrm{v}}$ are not relevant.) $\mathbf{Y}$ has non-positive diagonal entries and non-negative entries elsewhere. Its row sums are in $(\gamma - 1)\{0, 1, ..., A\}$ because all row sums of $\gamma\mathbf{P} - \mathbf{E}$ are $\gamma - 1$ by construction as $\sum_j p_{ij} = \sum_{s'} P(s'|s, a) = 1$. Hence the matrix $\mathbf{Y}$ can be written as $A(\mathbf{M} - \mathbb{1})$ where $\mathbf{M}$ is a non-negative matrix with $\gamma \leq \sum_j m_{ij} \leq 1$ for all i. According to the Perron-Frobenius theorem [1] or the Gershgorin circle theorem [2], all eigenvalues μ of $\mathbf{M}$ have absolute values $|\mu| \leq \max_i \sum_j m_{ij}$, hence $|\mu| \leq 1$. The eigenvectors of $\mathbf{M}$ are also eigenvectors of the identity matrix with eigenvalue 1. Thus the eigenvalues of $\mathbf{M} - \mathbb{1}$ are $\mu - 1$ and none has positive real part.

Thus far, we have proven that none of the eigenvalues of $\mathbf{X}(\mathbf{u})$ has positive real part. However, some are identical zero and we still have to rule out linear divergence driven by $\mathbf{w}^{\mathrm{r}}$, cf. Eq. (1). Indeed, such a linear increase is observed in Fig. 2F of the main text for the activity of the blue coloured neuron, but only in some finite interval until the orthant of $\mathbf{u}$ changes.

To rule out divergence to $+\infty$ we consider the summed up activity over all possible actions for each state, $\tilde{V}(s)$. As pointed out, its dynamics is captured by $\mathbf{Y}$, which has no positive eigenvalues. By inspection we see, that $\mathbf{Y}$ has an eigenvalue zero only if none of the neurons within one block of $\mathbf{X}$ is active, $\exists\hat{s}\forall i|s_i = \hat{s} : u_i \leq 0$, hence $\tilde{V}(\hat{s}) = -V_0 < \infty$. The corresponding row and column of $\mathbf{Y}$ with index $\hat{s}$ contains only zeros. It does not interact with the other approximate values $\tilde{V}(s \neq \hat{s})$ and can be removed to obtain a smaller matrix for the dynamics of $\tilde{V}(s \neq \hat{s})$. The resulting matrix can be written as $A(\mathbf{N} - \mathbb{1})$ where $\mathbf{N}$ is a non-negative matrix with $\sum_j n_{ij} < 1$ for all i if $\gamma < 1$. Thus the eigenvalues of $\mathbf{N} - \mathbb{1}$ have strictly negative real part and $\tilde{V}(s \neq \hat{s})$ converges. Because activities are positive and the summed up activity converges, the individual summands cannot diverge either. They also don't oscillate while keeping the sum at its asymptotic value showing limit cycle or limit torus behaviour, because the matrix $\mathbf{X}$ doesn't have pure imaginary eigenvalues.

Next we rule out divergence to $-\infty$. A neuron i with negative membrane potential u_i does not spike, $\lambda_i = 0$. Its total input $I_i^{\mathrm{tot}} = \mathbf{w}_i\boldsymbol{\lambda} + \mathbf{w}_i^{\mathrm{r}}$ (cf. Eq. 5, main text), due to other neurons' spiking activity and the external input, can be negative. However, it is bounded, because we have already established that activities cannot diverge, and it further does not depend on u_i. Due to the presence of the leak term in the neural dynamics $\tau_{\mathrm{m}}\dot{u}_i = -u_i + I_i^{\mathrm{tot}}$ the membrane potential converges to $I_i^{\mathrm{tot}} > -\infty$.

Taken together, we have successfully proven convergence of $\mathbf{u}$.

References

[1] Oskar Perron. Zur Theorie der Matrices. *Math. Ann.*, 64:248–263, 1907.

[2] Semyon A. Gerschgorin. Über die Abgrenzung der Eigenwerte einer Matrix. *Izv. Akad. Nauk. USSR, Ser. Mat.*, 7(3):749–754, 1931.