Supplemental material

A. MinimalRNN Architecture

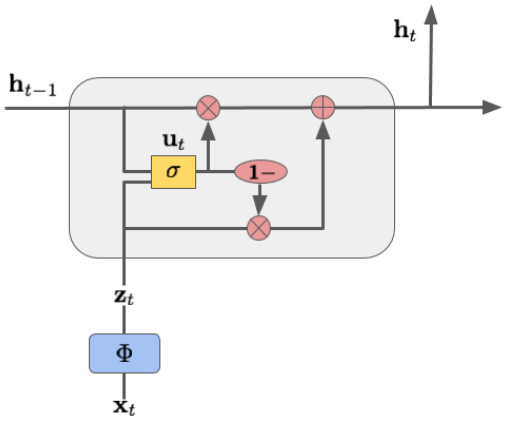

Figure Supp.1. Model architecture of minimalRNN.

B. Diagonal Recurrence Relation

Here we analyze the mean field dynamics of the minimalRNN. The minimalRNN features a hidden state $\boldsymbol{h}^t \in \mathbb{R}^N$ and inputs $\boldsymbol{x}^t$. The inputs are transformed via a fully-connected network $\boldsymbol{z}^t = \boldsymbol{\Phi}(\boldsymbol{x}^t) \in \mathbb{R}^M$ before being fed into the network. The RNN cell is then described by the equations,

$$v^t_{i;a} = \sum_j W_{ij} h^{t-1}_{j;a} + \sum_j V_{ij} z^t_{j;a} + b_i \tag{16}$$

$$u^t_{i;a} = \sigma(v^t_{i;a}) \tag{17}$$

$$h^t_{i;a} = u^t_{i;a} h^{t-1}_{i;a} + (1 - u^t_{i;a}) z^t_{i;a}. \tag{18}$$

Here i denotes the (pre)-activation and a denotes an input to the network.Thus, u^t_i acts as a gate on the t'th step. We take $W_{ij} \sim \mathcal{N}(0, \sigma_w^2/N)$, $V_{ij} \sim \mathcal{N}(0, \sigma_v^2/M)$ and $b_i \sim \mathcal{N}(\mu_b, \sigma_b^2)$.

By the CTL we can make a mean field assumption that $v^t_{i;a} \sim \mathcal{N}(\mu_b, q^t_{ab})$ where,

$$q^t_{ab} = \sigma_w^2 \mathbb{E}[h^{t-1}_{i;a} h^{t-1}_{i;b}] + \sigma_v^2 \mathbb{E}[z^t_{i;a} z^t_{i;b}] + \sigma_b^2 = \sigma_w^2 Q^{t-1}_{ab} + \sigma_v^2 R^t_{ab} + \sigma_b^2 \tag{19}$$

where we have defined $Q^t_{ab} = [h^t_{i;a} h^t_{i;b}]$ and $R^t_{ab} = \mathbb{E}[z^t_{i;a} z^t_{i;b}]$. We note that R^t_{ab} is fixed by the input, but it remains for us to work out Q^t_{ab}. We find that,

$$Q^t_{ab} = \mathbb{E}[h^t_{i;a} h^t_{i;b}] = \mathbb{E}[u^t_{i;a} h^{t-1}_{i;a} u^t_{i;b} h^{t-1}_{i;b}] + \mathbb{E}[u^t_{i;a} h^{t-1}_{i;a} (1 - u^t_{i;b}) z^t_{i;b}] \tag{20}$$

$$+ \mathbb{E}[(1 - u^t_{i;a}) z^t_{i;a} u^t_{i;b} h^{t-1}_{i;b}] + \mathbb{E}[(1 - u^t_{i;a}) z^t_{i;a} (1 - u^t_{i;b}) z^t_{i;b}] \tag{21}$$

$$\approx \mathbb{E}[u^t_{i;a} u^t_{i;b}] \mathbb{E}[h^{t-1}_{i;a} h^{t-1}_{i;b}] + \mathbb{E}[(1 - u^t_{i;a})(1 - u^t_{i;b})] \mathbb{E}[z^t_{i;a} z^t_{i;b}] \tag{22}$$

where we have assumed that the expectation factorizes so that $h^{t-1}_{i;a}$ and $u^t_{i;a}$ are approximately independent.

We choose to normalize the data so that $R^t_{aa} = R^t_{bb} = R$ independent of time. An immediate consequence of this normalization is that $Q^t_{aa} = Q^t_{bb} = Q^t$ and $q^t_{aa} = q^t_{bb} = q^t$. We then write $R^t_{ab} = R\Sigma^t$, $Q^t_{ab} = Q^t C^t$ and $q^t_{ab} = q^t c^t$ where Σ^t, C^t, and c^t are cosine similarities between the inputs, the hidden states, and the $v^t_{a,b}$ respectively. With this normalization, we can work out the mean-field recurrence relation characterizing the covariance matrix for the minimalRNN.

We begin by considering the diagonal recurrence relations. We find that the dynamics are described by the equation,

$$Q^t = Q^{t-1} \int \mathcal{D}z\sigma^2(\sqrt{q^t}z + \mu_b) + R \int \mathcal{D}z \left[1 - \sigma(\sqrt{q^t}z + \mu_b)\right]^2 \tag{23}$$

$$q^t = \sigma_w^2 Q^{t-1} + \sigma_v^2 R + \sigma_b^2 \tag{24}$$

As expected, the first and second integrands determine how much of the update of the random network is controlled by the norm of the hidden state and how much is determined by the norm of the input. Since $\sigma(z) = 1 - \sigma(-z)$ it follows that when $\mu_b = 0$ the first and second term will be equal and so,

$$Q^t = (Q^{t-1} + R) \int \mathcal{D}z\sigma^2(\sqrt{q^t}z). \tag{25}$$

In general, μ_b will therefore control the degree to which the hidden state of the random minimalRNN is updated based on the previous hidden state or based on the inputs with $\mu_b = 0$ implying parity between the two. This is reflected in eq. (25).

C. Existence of a Q^* Fixed Point

In the event that the norm of the inputs is time-independent, $R^t = R$ for all t, then the minimalRNN will have a fixed point provided there exists a Q^* that satisfies a transcendental equation, namely that

$$\mathcal{F}(Q^*) \equiv \frac{\int \mathcal{D}_{q^*} z \left[1 - \sigma\left(z\right)\right]^2}{\int \mathcal{D}_{q^*} z \left[1 - \sigma^2\left(z\right)\right]} - \frac{Q^*}{R} = 0\,. \tag{26}$$

It is easy to see that such a solution always exists. When $Q^* \to \infty$ the first term of $\mathcal{F}(Q^*)$ approaches 1 while the magnitude of the second increases without bound and so $\mathcal{F}(Q^*) < 0$. Conversely, when $Q^* \to 0$ the first term is positive while $Q^*/R \to 0$ and so $\mathcal{F}(Q^*) > 0$. The existence of a Q^* satisfying the transcendental equation then follows directly from the intermediate value theorem.

D. Q^* Dynamics

We can now investigate the dynamics of the norm of the hidden state in the vicinity of Q^*. To do this suppose that $Q^t = Q^* + \epsilon^t$ with $\epsilon \ll 1$. Our goal is then to expand eq.(23) about Q^*. First, we note that,

$$\sigma(\sqrt{q^t}z + \mu_b) = \sigma(\sqrt{q^* + \sigma_w^2\epsilon^t}z + \mu_b) \tag{27}$$

$$\approx \sigma\left(\sqrt{q^*}z + \mu_b + \frac{1}{2\sqrt{q^*}}\sigma_w^2\epsilon^t z\right) \tag{28}$$

$$\approx \sigma\left(\sqrt{q^*}z + \mu_b\right) + \frac{1}{2\sqrt{q^*}}\sigma_w^2\epsilon^t z\sigma'\left(\sqrt{q^*}z + \mu_b\right) + \mathcal{O}((\epsilon^t)^2). \tag{29}$$

Letting $\zeta(z) = \sqrt{q^*}z + \mu_b$ this implies that,

$$Q^t = Q^{t-1} \int \mathcal{D}z\sigma^2(\sqrt{q^* + \sigma_w^2\epsilon^{t-1}}z + \mu_b) + R \int \mathcal{D}z \left[1 - \sigma(\sqrt{q^* + \sigma_w^2\epsilon^{t-1}}z + \mu_b)\right]^2 \tag{30}$$

$$Q^* + \epsilon^t = Q^* \int \mathcal{D}z\sigma^2\left(\zeta(z)\right) + R \int \mathcal{D}z \left[1 - \sigma(\zeta(z))\right]^2 + \epsilon^{t-1}\Bigg[\int \mathcal{D}z\sigma^2\left(\zeta(z)\right) \tag{31}$$

$$+ \frac{Q^*}{\sqrt{q^*}}\sigma_w^2 \int \mathcal{D}zz\sigma(\zeta(z))\sigma'(\zeta(z)) - R\sqrt{q^*}\sigma_w^2 \int \mathcal{D}zz(1 - \sigma(\zeta(z)))\sigma'(\zeta(z))\Bigg] \tag{32}$$

$$\epsilon^t = \epsilon^{t-1}\Bigg[\int \mathcal{D}z\sigma^2\left(\zeta(z)\right) + \frac{Q^*}{\sqrt{q^*}}\sigma_w^2 \int \mathcal{D}zz\sigma(\zeta(z))\sigma'(\zeta(z)) - R\sqrt{q^*}\sigma_w^2 \int \mathcal{D}zz(1 - \sigma(\zeta(z)))\sigma'(\zeta(z))\Bigg] \tag{33}$$

$$= \epsilon^{t-1} \int \mathcal{D}z \left[\sigma^2(\zeta(z)) + \frac{\sigma_w^2}{\sqrt{q^*}}\left\{(Q^* + R)\sigma(\zeta(z)) - R\right\} z\sigma'(\zeta(z))\right] \tag{34}$$

$$= \epsilon^{t-1} \int \mathcal{D}z \left[\sigma^2(\zeta(z)) + \frac{\sigma_w^2}{\sqrt{q^*}}\left\{(Q^* + R)\sigma(\zeta(z)) - R\right\} z\sigma(\zeta(z))(1 - \sigma(\zeta(z)))\right] \tag{35}$$

$$= \epsilon^{t-1} \int \mathcal{D}z \left[\sigma^2(\zeta(z)) + \frac{\sigma_w^2}{\sqrt{q^*}}\left\{(Q^* + R)\sigma(\zeta(z)) - R\right\} z\sigma(\zeta(z))(1 - \sigma(\zeta(z)))\right] \tag{36}$$

It follows that $q^t \to q^*$ as,

$$|q^t - q^*| \sim e^{-t/\xi_Q} \tag{37}$$

with

$$\xi_Q^{-1} = -\log\left(\int \mathcal{D}z \left[\sigma^2(\zeta(z)) + \frac{\sigma_w^2}{\sqrt{q^*}}\left\{(Q^* + R)\sigma(\zeta(z)) - R\right\}\sigma'(\zeta(z))\right]\right) \tag{38}$$

as expected.

E. Off-Diagonal Recurrence Relation

We now turn our attention to the off-diagonal term. From eq. (9) it follows that,

$$Q^t C^t = Q^{t-1}C^{t-1}\int \mathcal{D}z_1 \mathcal{D}z_2 \sigma(u_1^t)\sigma(u_2^t) + R\Sigma^t \int \mathcal{D}z_1 \mathcal{D}z_2 (1 - \sigma(u_1^t))(1 - \sigma(u_2^t)) \tag{39}$$

$$q^t c^t = \sigma_w^2 Q^{t-1} C^{t-1} + \sigma_v^2 R\Sigma^t + \sigma_b^2 \tag{40}$$

where

$$u_1^t = \sqrt{q^t} z_1 + \mu_b \quad \text{and} \quad u_2^t = \sqrt{q^t}\left(c^t z_1 + \sqrt{1 - (c^t)^2} z_2\right) + \mu_b. \tag{41}$$

By expanding eq (39) as $c^t = c^* + \epsilon^t$ we find $\epsilon^{t+1} = \chi_{c^*}\epsilon^t$ where,

$$\chi_{c^*} = \int \mathcal{D}z_1 \mathcal{D}z_2 \sigma(u_1)\sigma(u_2) + q^*(c^* + J_-)\int \mathcal{D}z_1 \mathcal{D}z_2 \sigma'(u_1)\sigma'(u_2). \tag{42}$$

We note that when $c^* = 1$ it follows that $\chi_{c^*} = \chi_1$.

F. Additional Hyperparameter Ranges

We tune the learning hyper-parameters in the following ranges for all the models:

- learning rate: $\{0.1, 0.2, 0.3, 0.5, 1, 2\}$
- max-epoch: $\{4, 7, 11\}$
- decay: $\{0.5, 0.65, 0.8\}$
- dropout: $\{0.0, 0.2, 0.3, 0.5\}$