Supplementary material - Beyond exploding and vanishing gradients: analysing RNN training using attractors and smoothness

Antônio H. Ribeiro
Federal University of Minas Gerais

Koen Tiels
Eindhoven University of Technology

Luis A. Aguirre
Federal University of Minas Gerais

Thomas B. Schön
Uppsala University

A Computing the derivatives

Let the Jacobian matrices of $\mathbf{f}(\mathbf{x}, \mathbf{u}; \boldsymbol{\theta})$ with respect to $\mathbf{x}$ and to $\boldsymbol{\theta}$ evaluated at the point $(\mathbf{x}_t, \mathbf{u}_t; \boldsymbol{\theta})$ be denoted, respectively, as A_t and B_t. Similarly, the Jacobian matrices of $\mathbf{g}(\mathbf{x}, \mathbf{u}_t; \boldsymbol{\theta})$ are denoted as C_t and F_t. A direct application of the chain rule to (1) gives a recursive formula for computing the derivatives of the predicted output in relation to the parameters in the interval $1 \leq t \leq N$:

$$\begin{aligned} D_{t+1} &= A_t D_t + B_t \text{ for } D_0 = \mathbf{0} \\ J_t &= C_t D_t + F_t, \end{aligned} \tag{S1}$$

where we denote the Jacobian matrices of $\hat{\mathbf{y}}_t$ and $\mathbf{x}_t$ with respect to $\boldsymbol{\theta}$ respectively as J_t and D_t.

For the cost function V defined as in Eq. (2), its gradient ∇V is given by:

$$\nabla V = \tfrac{1}{N} \sum_{t=1}^{N} J_t \; l'(\hat{\mathbf{y}}_t, \mathbf{y}_t), \tag{S2}$$

where $l'(\hat{\mathbf{y}}_t, \mathbf{y}_t)$ denotes the gradient of the loss function with respect to its first argument, evaluated at $(\hat{\mathbf{y}}_t, \mathbf{y}_t)$.

B Proofs

B.1 Entropy lower bound

Theorem 2. *Let $\mathbf{f}(\cdot, \mathbf{u}_t; \boldsymbol{\theta})$ in Eq. (1) be a one-to-one continuous differentiable map, and let $\mathbf{f}(\mathbf{x}, \mathbf{u}; \boldsymbol{\theta})$ be Lipschitz in $(\mathbf{x}, \boldsymbol{\theta})$ with constant L_f on a compact and convex set $\Omega = (\Omega_{\mathbf{x}}, \Omega_{\mathbf{u}}, \Omega_{\boldsymbol{\theta}})$. Then the entropy H_t in Eq. (3) with $x_t \in \mathbb{R}^{N_x}$ satisfies:*

$$H_t + N_x \log L_f \leq H_{t+1}. \tag{S3}$$

Proof. Under the assumption that $\mathbf{f}(\cdot, \mathbf{u}_t; \boldsymbol{\theta})$ is a 1-1 continuous differentiable map (cf. Theorems 3-13 and 3-14 by Spivak (1998)), applying the change of variable $\mathbf{x}_{t+1} = \mathbf{f}(\mathbf{x}_t, \mathbf{u}_t; \boldsymbol{\theta})$ we get:

$$\begin{aligned} H_t &= -\int_{\mathbf{f}(\Omega_x, \mathbf{u}_t; \boldsymbol{\theta})} p_{t+1}(\mathbf{x}_{t+1}) \log \left(p_{t+1}(\mathbf{x}_{t+1}) \left| \det \frac{\partial \mathbf{x}_{t+1}}{\partial \mathbf{x}_t} \right| \right) \mathrm{d}\mathbf{x}_{t+1} \\ &= H_{t+1} - \int_{\mathbf{f}(\Omega_x, \mathbf{u}_t; \boldsymbol{\theta})} p_{t+1}(\mathbf{x}_{t+1}) \log \left(\left| \det \frac{\partial \mathbf{x}_{t+1}}{\partial \mathbf{x}_t} \right| \right) d\mathbf{x}_{t+1}, \end{aligned}$$

where $\frac{\partial \mathbf{x}_{t+1}}{\partial \mathbf{x}_t}$ is the Jacobian matrix of $\mathbf{f}(\cdot, \mathbf{u}_t; \boldsymbol{\theta})$. Using Hadamard's inequality:

$$\log \left| \det \frac{\partial \mathbf{x}_{t+1}}{\partial \mathbf{x}_t} \right| \leq \sum_{i=1}^{N_x} \log \|\mathbf{v}_i\|_2, \tag{S4}$$

where $\mathbf{v}_i$ is the i-th column of the Jacobian matrix and $\log \|\mathbf{v}_i\|_2 \leq \log \left\| \frac{\partial \mathbf{x}_{t+1}}{\partial \mathbf{x}_t} \right\|_2 \leq \log L_f$. Hence, we have that

$$H_t + N_x \log L_f \leq H_{t+1}. \tag{S5}$$

□

B.2 Preliminary results

Lemma 1. *For $i = 1, \cdots, n$, let $\mathbf{f}_i$ be a Lipschitz function on Ω with constants L_i. Then,*

1. *$\sum_{i=1}^{n} \mathbf{f}_i$ is also a Lipschitz function on Ω with Lipschitz constant upper bounded by $\sum_{i=1}^{n} L_i$;*
2. *if, additionally, $\mathbf{f}_i$ are bounded by M_i on Ω, then $\prod_{i=1}^{n} \mathbf{f}_i$ is also a Lipschitz function on Ω with Lipschitz constant upper bounded by $\left(\sum_{i=1}^{n} M_1 \cdots M_{i-1} L_i M_{i+1} \cdots M_n\right)$.*

Lemma 2. *Let us define the properties:*

1. *$|l(\hat{\mathbf{y}}, \mathbf{y}) - l(\hat{\mathbf{z}}, \mathbf{y})| < (K_1 \|\mathbf{y}\| + K_2 \max(\|\hat{\mathbf{y}}\|, \|\hat{\mathbf{z}}\|)) \|\hat{\mathbf{y}} - \hat{\mathbf{z}}\|$,*
2. *$l'(\hat{\mathbf{y}}, \mathbf{y}) = \Psi(\hat{\mathbf{y}}) - K_3 \mathbf{y}$,*

where $l'(\hat{\mathbf{y}}, \mathbf{y})$ denotes the first derivative of the loss function with respect to its first argument, evaluated at $(\hat{\mathbf{y}}, \mathbf{y})$. There exist constants K_1, K_2, and K_3 and a function Ψ that is Lipschitz continuous with constant K_4 and for which Ψ such that these properties hold for both: a) $l(\hat{\mathbf{y}}, \mathbf{y}) = \|\hat{\mathbf{y}} - \mathbf{y}\|^2$; b) $l(\hat{\mathbf{y}}, \mathbf{y}) = -\mathbf{y}^T \log(\sigma(\hat{\mathbf{y}})) - (1 - \mathbf{y})^T \log(1 - \sigma(\hat{\mathbf{y}}))$. In (b), the sigmoid function, $\sigma(x) = \frac{1}{1+\exp(-x)}$, and the logarithm are evaluated element-wise. We assume in (b) that the elements in $\mathbf{y}$ are either 0 or 1.

Proof. For (a) and (b), property 2 follows from differentiating $l(\hat{\mathbf{y}}, \mathbf{y})$ with respect to its first argument. For (a), $\phi(\hat{\mathbf{y}}) = 2\hat{\mathbf{y}}$ and $K_3 = 2$; for (b), $\phi(\hat{\mathbf{y}}) = \sigma(\hat{\mathbf{y}})$ and $K_3 = 1$.

For loss function (a), property 1 holds due to the following inequalities

$$\begin{aligned} \left| \|\hat{\mathbf{y}} - \mathbf{y}\|^2 - \|\hat{\mathbf{z}} - \mathbf{y}\|^2 \right| &= \left| \|\hat{\mathbf{y}}\|^2 - \|\hat{\mathbf{z}}\|^2 - 2\mathbf{y}^T (\hat{\mathbf{y}} - \hat{\mathbf{z}}) \right| \leq \\ &\leq \left| (\|\hat{\mathbf{y}}\| - \|\hat{\mathbf{z}}\|) (\|\hat{\mathbf{y}}\| + \|\hat{\mathbf{z}}\|) - 2\mathbf{y}^T (\hat{\mathbf{y}} - \hat{\mathbf{z}}) \right| \leq \\ &\leq (2\|\mathbf{y}\| + 2\max(\|\hat{\mathbf{y}}\|, \|\hat{\mathbf{z}}\|)) \|\hat{\mathbf{y}} - \hat{\mathbf{z}}\|. \end{aligned}$$

For loss function (b), let $\hat{y}$ and $\hat{z}$ be two scalar values. Furthermore, consider, without loss of generality, that $\hat{y} \geq \hat{z}$. Then:

$$0 \leq \log(\sigma(\hat{y})) - \log(\sigma(\hat{z})) = (\hat{y} - \hat{z}) - \log \left(\frac{\exp(\hat{y}) + 1}{\exp(\hat{z}) + 1} \right) \leq (\hat{y} - \hat{z}) \tag{S6}$$

The first inequality follows from the fact that $\log(\sigma(\cdot))$ is a monotonically increasing function. The last inequality holds because $\log \left(\frac{\exp(\hat{y})+1}{\exp(\hat{z})+1} \right) \geq 0$. Analogously,

$$\begin{aligned} 0 &\leq \log(1 - \sigma(\hat{z})) - \log(1 - \sigma(\hat{y})) \\ &= (\hat{y} - \hat{z}) - \log \left(\frac{\exp(-\hat{z}) + 1}{\exp(-\hat{y}) + 1} \right) \\ &\leq (\hat{y} - \hat{z}). \end{aligned} \tag{S7}$$

For $l(\mathbf{y},\hat{\mathbf{y}})$ defined as in (b), it follows from Eq. (S6), Eq. (S7), and the fact that $\mathbf{y}$ contains only values in the set $\{0,1\}$, that:

$$|l(\hat{\mathbf{y}},\mathbf{y}) - l(\hat{\mathbf{z}},\mathbf{y})| \leq \|\hat{\mathbf{y}}-\hat{\mathbf{z}}\|_1 \leq \sqrt{N_y}\|\hat{\mathbf{y}}-\hat{\mathbf{z}}\|_2 \quad \text{(S8)}$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ denote l_1 and l_2 norm of a vector and N_y is the number of outputs. □

B.3 Proof of Theorem 1 (a)

Assume two different trajectories resulting from simulating the system in Eq (1) with parameters and initial conditions $(\mathbf{x}_0,\boldsymbol{\theta})$ and $(\mathbf{w}_0,\boldsymbol{\phi})$, respectively. We denote the corresponding trajectories by $\mathbf{x}$ and $\mathbf{w}$. Let us call:

$$\|\Delta\hat{\mathbf{y}}_t\| = \|\mathbf{g}(\mathbf{x}_t,\mathbf{u}_t;\boldsymbol{\theta}) - \mathbf{g}(\mathbf{w}_t,\mathbf{u}_t;\boldsymbol{\phi})\|. \quad \text{(S9)}$$

Since $\mathbf{f}$ and $\mathbf{g}$ are Lipschitz in $(\mathbf{x},\boldsymbol{\theta})$ we have:

$$\|\mathbf{f}(\mathbf{x},\mathbf{u}_t,\boldsymbol{\theta}) - \mathbf{f}(\mathbf{w},\mathbf{u}_t,\boldsymbol{\phi})\|^2 \leq L_f^2\left(\|\mathbf{x}-\mathbf{w}\|^2 + \|\boldsymbol{\theta}-\boldsymbol{\phi}\|^2\right),$$

$$\|\mathbf{g}(\mathbf{x},\mathbf{u}_t,\boldsymbol{\theta}) - \mathbf{g}(\mathbf{w},\mathbf{u}_t,\boldsymbol{\phi})\|^2 \leq L_g^2\left(\|\mathbf{x}-\mathbf{w}\|^2 + \|\boldsymbol{\theta}-\boldsymbol{\phi}\|^2\right),$$

for all $(\mathbf{x},\mathbf{u}_t,\boldsymbol{\theta})$ and $(\mathbf{w},\mathbf{u}_t,\boldsymbol{\phi})$ in $(\Omega_{\mathbf{x}},\Omega_{\mathbf{u}},\Omega_{\boldsymbol{\theta}})$. Applying these relations recursively we get that:

$$\|\Delta\hat{\mathbf{y}}_t\|^2 \leq L_g^2 L_f^{2t}\|\boldsymbol{x}_0-\boldsymbol{w}_0\|^2 + L_g^2\left(\sum_{\ell=0}^{t} L_f^{2\ell}\right)\|\boldsymbol{\theta}-\boldsymbol{\phi}\|^2.$$

Since L_f is positive, the constant multiplying the second term in the above equation is always larger than the constant multiplying the first term. Hence, taking the square root on both sides of the above inequality and after simple manipulations, we obtain:

$$\|\Delta\hat{\mathbf{y}}_t\| \leq L_g S(t)\|[\boldsymbol{\theta},\boldsymbol{x}_0]^T - [\boldsymbol{\phi},\boldsymbol{w}_0]^T\|, \quad \text{(S10)}$$

where

$$S(t) = \sqrt{\sum_{\ell=0}^{t} L_f^{2\ell}} = \begin{cases} \sqrt{t+1} & \text{if } L_f = 1, \\ \sqrt{\frac{L_f^{2t+2}-1}{L_f^2-1}} & \text{if } L_f \neq 1. \end{cases} \quad \text{(S11)}$$

Since Ω is compact and $\hat{\mathbf{y}}_t$ is a (Lipschitz) continuous function of the parameters and initial conditions, then $\hat{\mathbf{y}}_t$ is bounded in Ω, i.e. $\|\hat{\mathbf{y}}_t\| \leq M(t)$. Furthermore, it follows from Eq. (S10) and from the existence of an invariant set[4] in Ω that $M(t) = \mathcal{O}(S(t))$.

The following inequality follows from Eq. (2), and from property 1 in Lemma 2:

$$|V(\boldsymbol{\theta},\boldsymbol{x}_0) - V(\boldsymbol{\phi},\boldsymbol{w}_0)| \leq \tfrac{1}{N}\sum_{t=1}^{N}(K_1L_y + K_2M(t))\|\Delta\hat{\mathbf{y}}_t\|, \quad \text{(S12)}$$

where $L_y = \max_{1\leq t\leq N}\|\mathbf{y}_t\|$. Now, assembling Eq. (S12) and Eq. (S10) we obtain

$$|V(\boldsymbol{\theta},\boldsymbol{x}_0) - V(\boldsymbol{\phi},\boldsymbol{w}_0)| \leq L_{V_1}\left\|[\boldsymbol{x}_0,\boldsymbol{\theta}]^T - [\boldsymbol{w}_0,\boldsymbol{\phi}]^T\right\|,$$

for $L_V = \left(\frac{L_g}{N}\sum_{t=1}^{N}(K_1L_y + K_2M(t))S(t)\right)$. The asymptotic analysis of this expression with respect to N yields Eq. (6).

[4] There are multiple ways to guarantee the invariant set premise will hold, but a very simple way is to just choose $\mathbf{f}$ such that $\mathbf{f}(\mathbf{0},\mathbf{u}_t;\mathbf{0}) = \mathbf{0}$. In this case, $\{\mathbf{0}\}$ is an invariant set and if $\Omega_{\boldsymbol{\theta}}$ contains this point the premise is satisfied. For this specific case, one can just choose $[\boldsymbol{\phi},\boldsymbol{w}_0] = \mathbf{0}$ and it follows from Eq. (S10) that $\|\hat{\mathbf{y}}_t\| \leq L_g S(t)\|[\boldsymbol{\theta},\boldsymbol{x}_0]\| = \mathcal{O}(S(t))$. The more general case, for any invariant set, follows from a similar deduction.

B.4 Proof of Theorem 1 (b)

It follows from Eq. (S2), and from property 2 in Lemma 2, that:

$$\|\nabla V(\boldsymbol{\theta},\boldsymbol{x}_0) - \nabla V(\boldsymbol{\phi},\boldsymbol{w}_0)\| \leq \tfrac{1}{N}\sum_{t=1}^{N} K_3L_y\|\Delta J_t\| + \|\Delta(J_t\Psi(\hat{\mathbf{y}}_t))\|, \quad \text{(S13)}$$

where we have used the notation ΔJ_t to denote the difference between J_t evaluated at $(\boldsymbol{\theta},\boldsymbol{x}_0)$ and $(\boldsymbol{\phi},\boldsymbol{w}_0)$. Analogously, $\Delta(J_t\Psi(\hat{\mathbf{y}}_t))$ denotes the difference between $J_t\Psi(\hat{\mathbf{y}}_t)$ evaluated at the two distinct points, where Ψ is the Lipschitz continuous function with constant K_4 defined in Lemma 2.

From Eq. (S1) it follows that:

$$J_t = C_t\sum_{\ell=1}^{t}\left(\prod_{j=1}^{t-\ell} A_{t-j+1}\right)B_\ell + F_t; \quad \text{(S14)}$$

Since the Jacobian of $\mathbf{f}$ is Lipschitz with Lipschitz constant L_f', it follows that:

$$\|\Delta A_j\|^2 \leq (L_f')^2\left(\|\mathbf{x}_j - \mathbf{w}_j\|^2 + \|\boldsymbol{\theta}-\boldsymbol{\phi}\|^2\right). \quad \text{(S15)}$$

Using a procedure analogous to the one used to obtain Eq. (S10), it follows that:

$$\|\Delta A_j\| \leq L_f' S(j) \quad \|[\boldsymbol{\theta},\boldsymbol{x}_0]^T - [\boldsymbol{\phi},\boldsymbol{w}_0]^T\|, \quad \text{(S16)}$$

where $S(j)$ is defined as in Eq. (S11). An identical formula holds for B_j and a similar formula, replacing L_f' with L_g', holds for C_j and F_j.

Since $\mathbf{f}$ and $\mathbf{g}$ are Lipschitz with Lipschitz constants L_f and L_g it follows that $\|A_j\| \leq L_f$, $\|B_j\| \leq L_f$, $\|C_j\| \leq L_g$ and $\|F_j\| \leq L_g$. Hence, it follows from Eq. (S10), Eq. (S14), Eq. (S16) and the repetitive application of Lemma 1 that $\|\Delta J_t\|$ and $\|\Delta(J_t\Psi(\hat{\mathbf{y}}_t))\|$ are upper bounded by $\|[\boldsymbol{\theta},\boldsymbol{x}_0]^T - [\boldsymbol{\phi},\boldsymbol{w}_0]^T\|$ multiplied by the following constants:

$$L_{J,t} = \sum_{\ell=1}^{t} P(t,\ell) + L_g'S(t),$$

$$L_{J\hat{y},t} = \sum_{\ell=1}^{t} Q(t,\ell) + T(t)S(t),$$

where $T(t) = K_4(L_g'M(t) + L_g^2)$ and:

$$P(t,\ell) = L_f^{t-\ell}\left(L_gL_f'\sum_{j=\ell}^{t} S(j) + L_fL_g'S(t)\right),$$

$$Q(t,\ell) = L_f^{t-\ell}\left(K_4M(t)L_gL_f'\sum_{j=\ell}^{t} S(j) + L_fT(t)S(t)\right).$$

Hence,

$$\|\nabla V(\boldsymbol{\theta},\boldsymbol{x}_0) - \nabla V(\boldsymbol{\phi},\boldsymbol{w}_0)\| \leq L_V'\|[\boldsymbol{\theta},\boldsymbol{x}_0]^T - [\boldsymbol{\phi},\boldsymbol{w}_0]^T\|,$$

where:

$$L_V' = \tfrac{1}{N}\sum_{t=1}^{N}\left(K_3L_yL_{J,t} + L_{J\hat{y},t}\right).$$

Combining the above, the asymptotic analysis of L_V' results in Eq. (7).

C Chaotic LSTM example

An LSTM with zero input and without bias terms is considered:

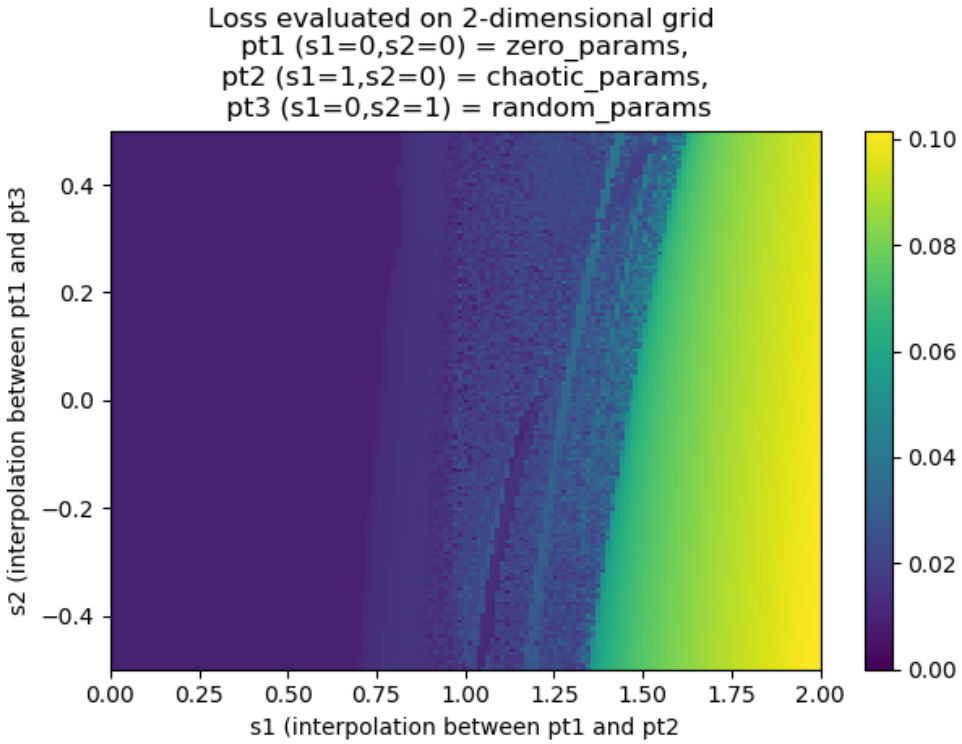
$$c_t = \underbrace{\sigma\left(W_{\mathrm{hf}} h_{t-1}\right)}_{\text{forget gate}} * c_{t-1} + \underbrace{\sigma\left(W_{\mathrm{hi}} h_{t-1}\right)}_{\text{input gate}} * \underbrace{\tanh\left(W_{\mathrm{hg}} h_{t-1}\right)}_{\text{cell gate}} \tag{S17}$$

$$h_t = \underbrace{\sigma\left(W_{\mathrm{ho}} h_{t-1}\right)}_{\text{output gate}} * \tanh(c_t). \tag{S18}$$

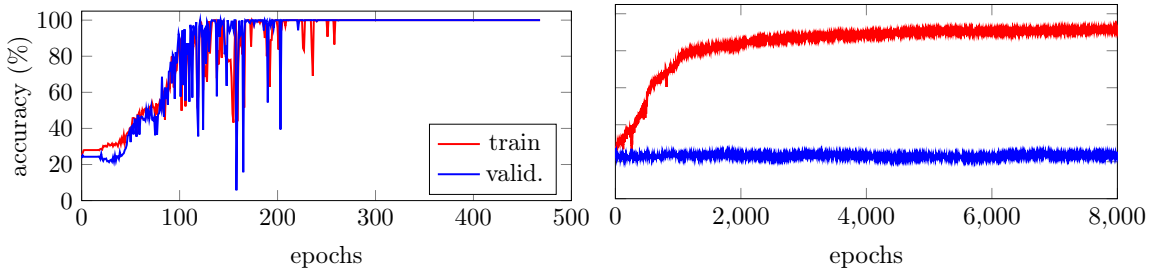
The sigmoids $\sigma(\cdot)$ and hyperbolic tangents $\tanh(\cdot)$ operate element-wise. The symbol $*$ indicates an element-wise product. The hidden and cell state have initial conditions $h_0 = c_0 = [0.5 \quad 0.5]^{\mathsf{T}}$. The hidden state h_t is also the output of the model. The weight matrices are put equal to $W_{\mathrm{hi}} = \begin{bmatrix} -1 & 4 \\ -3 & -2 \end{bmatrix}$, $W_{\mathrm{hf}} = \begin{bmatrix} -2 & 6 \\ 0 & -6 \end{bmatrix}$, $W_{\mathrm{hg}} = \begin{bmatrix} -1 & -6 \\ 6 & -9 \end{bmatrix}$, and $W_{\mathrm{ho}} = \begin{bmatrix} 4 & 1 \\ -9 & 7 \end{bmatrix}$ to generate the data. The values in the weight matrices are stacked on top of each other in a parameter vector θ_{true}.

Figure 1 shows the mean-square loss evaluated on data generated by the same LSTM model with the same initial conditions, but with different parameter values. A two-dimensional grid of parameter values $\theta(s_1, s_2)$ is used with values interpolated (and extrapolated) between θ_{true}, zero parameter values, and a randomly chosen parameter vector θ_{random}, i.e. in this case, $\theta(s_1, s_2) = s_1\theta_{\mathrm{true}} + s_2\theta_{\mathrm{random}}$. Again, the region in the parameter space around θ_{true} is intricate. There is an entire region where the cost function is intricate and has many neighboring local minima.

D Additional Experiments



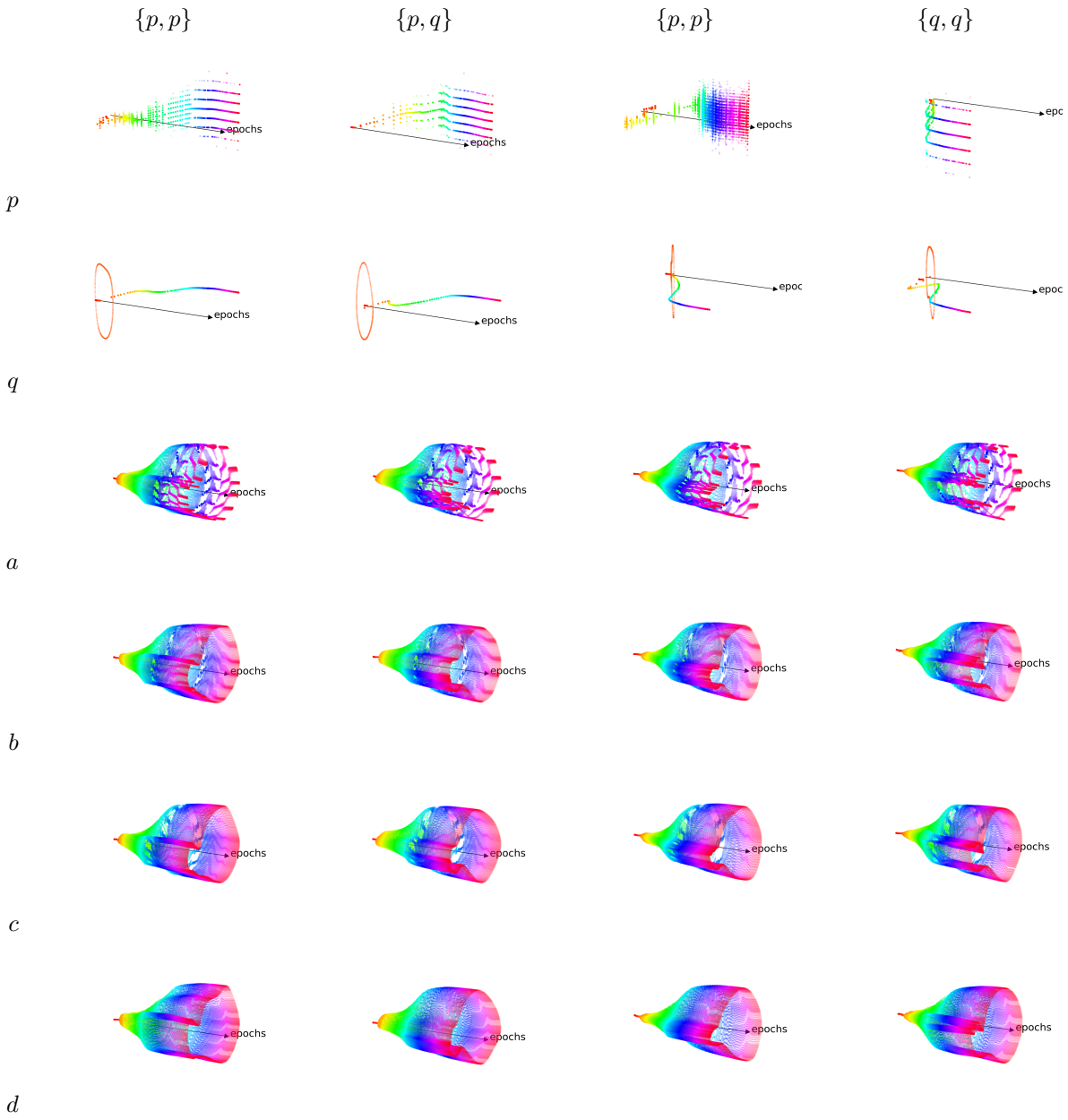
Supplementary Figure 1: **Chaotic LSTM.** Mean-square cost function (2) for LSTM models with parameter vectors $\theta(s_1, s_2) = s_1\theta_{\text{true}} + s_2\theta_{\text{random}}$.



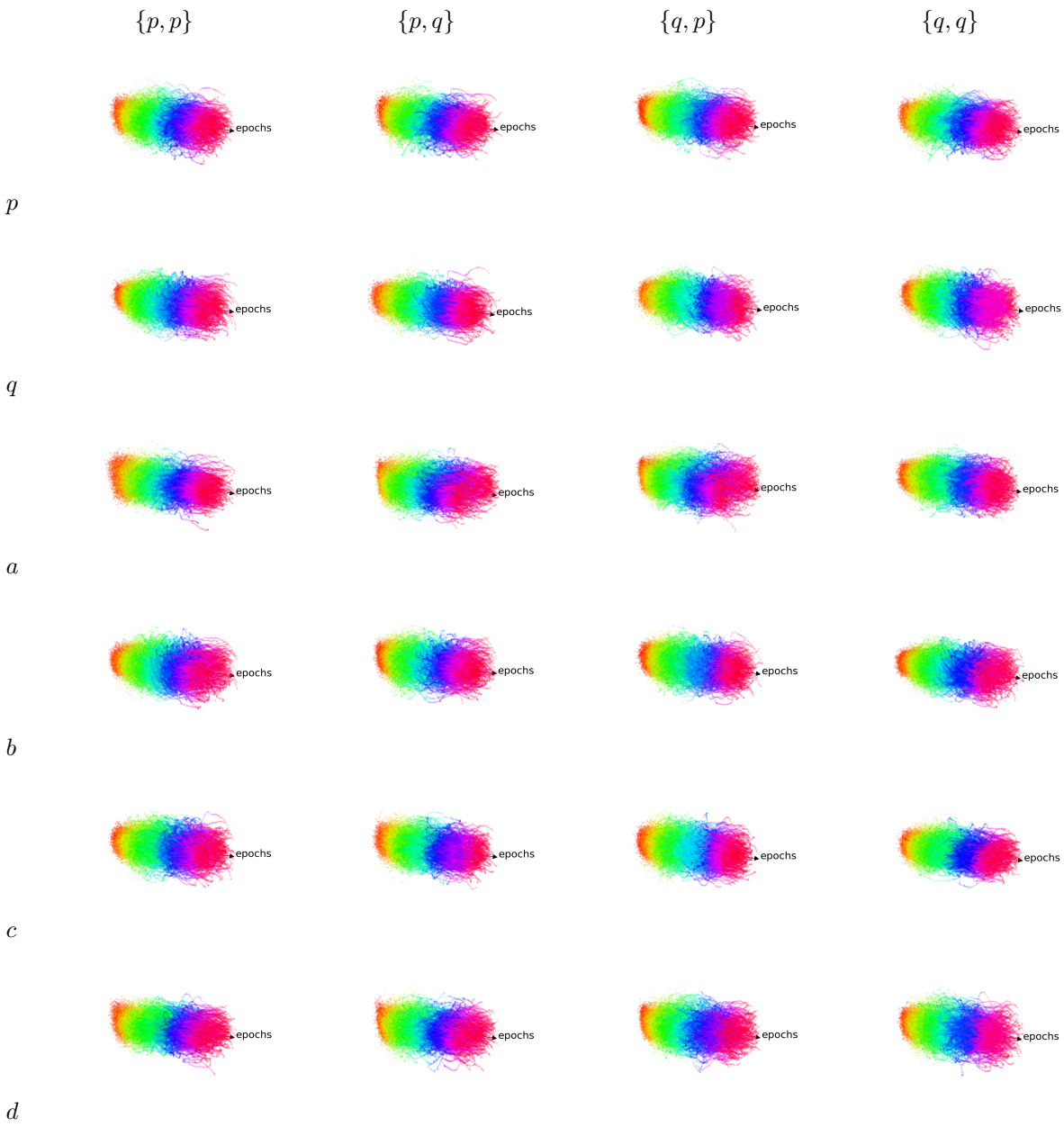
(a) good performance solution

(b) bad performance solution

Supplementary Figure 2: **The effect of the initial conditions.** Accuracy on training and validation data on the **symbol classification task** for the **LSTM** model in identical scenarios but with different initial random parameter initialization (from different random seeds). In (a), the optimization procedure abruptly finds a solution that has good accuracy on both training and validation; while, in (b), the convergence is slow and steady towards a solution that has good accuracy on the training set (above 90%) but is no better than random guessing on the validation set.

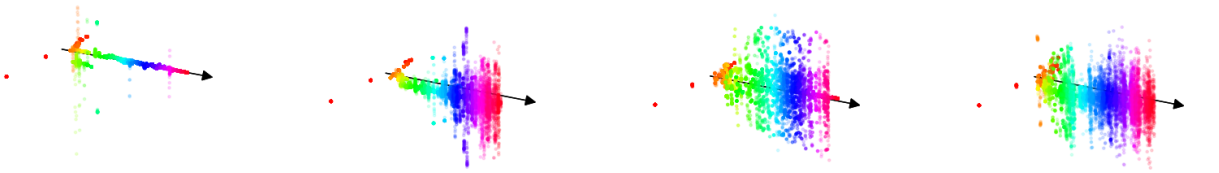


Supplementary Figure 3: **The LSTM learns to classify sequences.** Bifurcation diagram for the *LSTM* model in *symbol classification task* for sequences of length 100. It shows the steady-state of the output y_t and its first difference $y_t - y_{t-1}$. The arrows point towards the evolution of the number of epochs, that vary from 0 to 400.



Supplementary Figure 4: **The oRNN learns to classify sequences.** Bifurcation diagram for the *oRNN* model in *symbol classification task* for sequences of length 100. It shows the steady-state of the output y_t and its first difference $y_t - y_{t-1}$. The arrows point towards the evolution of the number of epochs, that vary from 0 to 400.

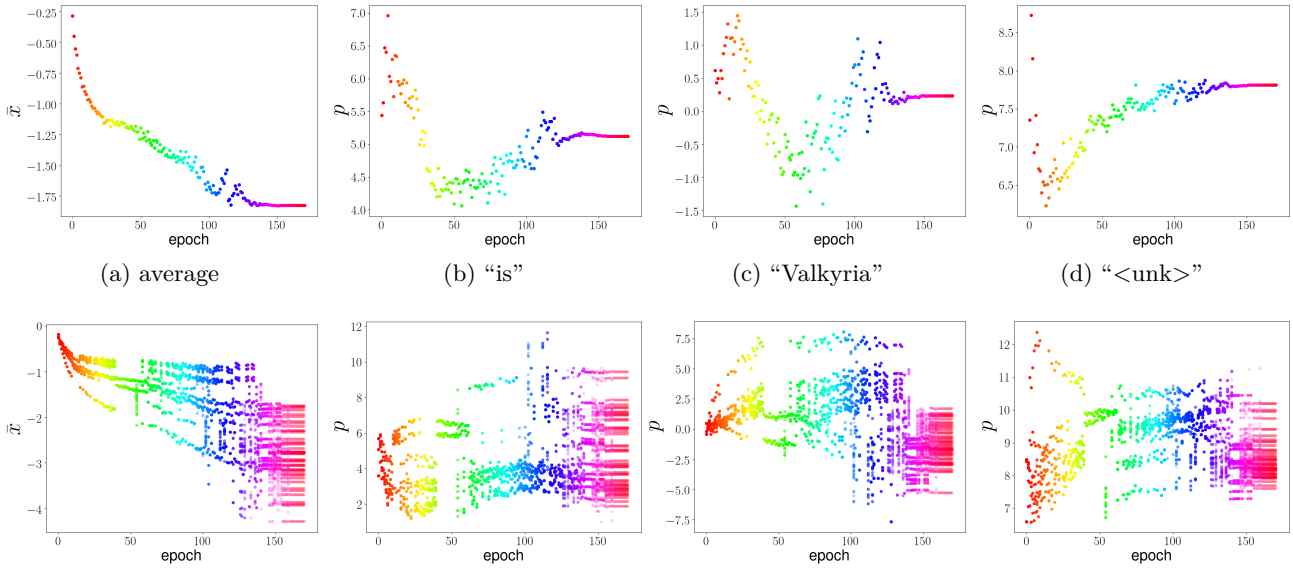
(a) average (b) average (c) average, with feedback (d) average, with feedback



(e) “is” (f) “is” (g) “is”, with feedback (h) “is”, with feedback

(i) “Valkyria” (j) “Valkyria” (k) “Valkyria”, with feedback (l) “Valkyria”, with feedback

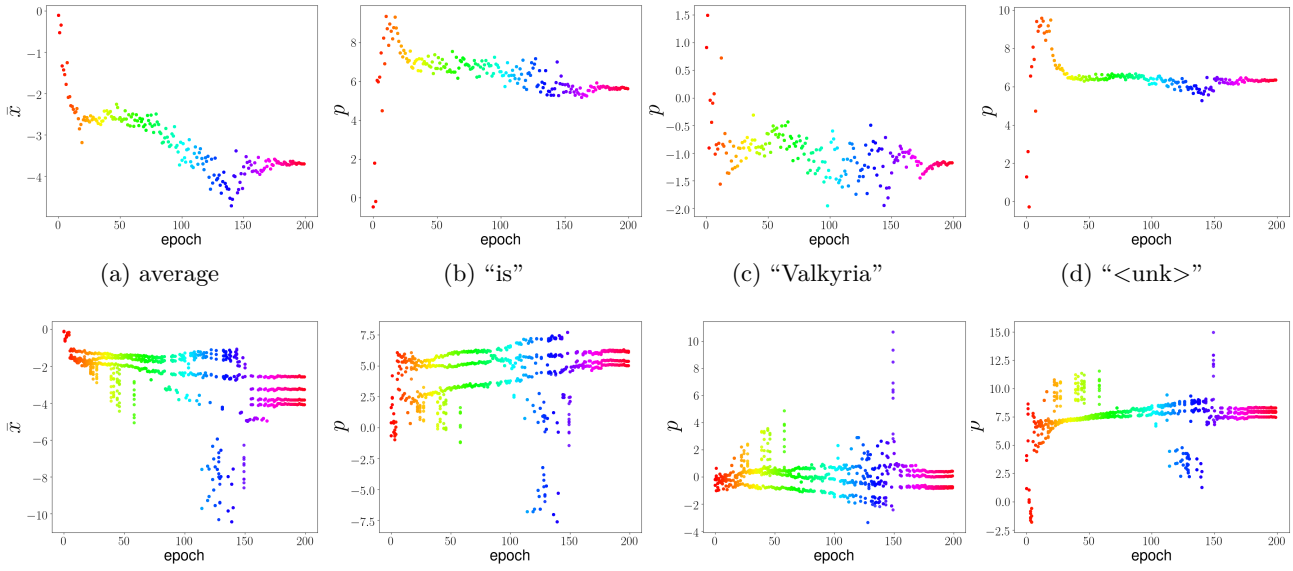
Supplementary Figure 5: **The LSTM mechanism for learning a language model.** Bifurcation diagram for the *LSTM world-level language model*. For each epoch, the plot shows values visited by the projections of the internal state $p(\mathbf{x}_t)$ and its first difference $p(\mathbf{x}_t) - p(\mathbf{x}_{t-1})$ after a burnout period of 1500 samples. This burnout period is used to remove the transient response and yields a visualization of the system attractors, per epoch. The arrow point towards the evolution of the number of epochs, that varies from 0 to 150. In (a) and (b), we have two different realizations of the bifurcation diagram obtained from constant inputs. In (c) and (d), the diagram is generated using as input the word predicted with the highest probability at the previous time instant, and using as first input to the sequence the same input as in (a) and (b), respectively. The subplots (a) to (d) use the average of internal states as projections, i.e. $p(\mathbf{x}_t) = \bar{x}_t$. The second row, (e) to (h), and third row, (i) to (l), show the exact same experiments but for the projections in the direction of the tokens “is” and “Valkyria”, respectively.



(a) average (b) “is” (c) “Valkyria” (d) “<unk>”

(e) average, with feedback (f) “is”, with feedback (g) “Valkyria”, with feedback (h) “<unk>”, with feedback

Supplementary Figure 6: **The sLSTM mechanism for learning a language model.** Bifurcation diagram for the *stable LSTM world-level language model*. For each epoch, the plot shows values visited by the projections of the internal state $p(\mathbf{x}_t)$ after a burnout period of 1500 samples. This burnout period is used to remove the transient response and yields a visualization of the system attractors, per epoch. In the displays (a) to (d), the diagram is generated for the same constant input. In (e) to (h), the diagram is generated using as input the word predicted with the highest probability at the previous time instant, and using as first input to the sequence the same input as in the first row. The projections are: the average of internal states as projections, i.e. $p(\mathbf{x}_t) = \bar{x}_t$; and, projections into the direction of the tokens “is”, “Valkyria” and “<unk>”.



(a) average (b) “is” (c) “Valkyria” (d) “<unk>”

(e) average, with feedback (f) “is”, with feedback (g) “Valkyria”, with feedback (h) “<unk>”, with feedback

Supplementary Figure 7: **The oRNN mechanism for learning a language model.** Bifurcation diagram for the *orthogonal RNN world-level language model*. For each epoch, the plot shows values visited by the projections of the internal state $p(\mathbf{x}_t)$ after a burnout period of 1500 samples. This burnout period is used to remove the transient response and yields a visualization of the system attractors, per epoch. In the displays (a) to (d), the diagram is generated for the same constant input. In (e) to (h), the diagram is generated using as input the word predicted with the highest probability at the previous time instant, and using as first input to the sequence the same input as in the first row. The projections are: the average of internal states as projections, i.e. $p(\mathbf{x}_t) = \bar{x}_t$; and, projections into the direction of the tokens “is”, “Valkyria” and “<unk>”.