Midterm 4: presentation on paper n.3

BO CHANG, MINMIN CHEN, ELDAD HABER, ED H. CHI, ICLR 2019

ANTISYMMETRIC-RNN: A DYNAMICAL SYSTEM VIEW ON RECURRENT NEURAL NETWORKS

Francesca Poli
f.poli12@studenti.unipi.it

M.Sc. Computer Science - Artificial intelligence
Intelligent Systems for Pattern Recognition, a.y. 2022/23

INTRODUCING THE INITIAL ISSUE

VANISHING AND EXPLODING GRADIENT IN RNN AND ITS VARIANTS

We are challenged with **modeling complex temporal dependencies in sequential data using RNNs**, especially the long-term dependencies. Here the gradient problems originates as the error signal back-propagates through time (BPTT) and suffers from exponential **growth** or **decay**. Gated variants (LSTM, GRU) are born to take care of this issue, but they need additional techniques to achieve good performances.

<u>Other solutions that didn't work out:</u>

- Identity and orthogonal initialization
- Orthogonal weight matrices throughout the entire learning process.

- **vanishing gradient** → lossy system
- **exploding gradient** → unstable system

dynamical systems viewpoint

A **dynamical system** is a system whose state is uniquely specified by a set of variables and whose behavior is described by predefined rules. Connections between RNNs and the ordinary differential equation theory help us design new recurrent architectures by **discretizing ODEs**.

The **antisymmetricRNN** is a special form of RNN that can capture longterm dependencies in the inputs. We can build one starting from **ordinary differential equations**: an ODE is an equality involving a function and its derivatives, here considered as a dynamical system with a single variable time ***t*** trying to solve the **BPTT** problem.

MODEL DESCRIPTION: STABILITY OF ORDINARY DIFFERENTIAL EQUATIONS

- **Initial value problem:** $t \geq 0,\ h(t) \in \mathbb{R}^{K},\ f : \mathbb{R}^n \rightarrow \mathbb{R}^n \Rightarrow h'(t) = f(h(t))$
  Applying the **Forward Euler Method** (approximation relying on discretization):

$$\frac{h_{(t)} - h_{(t-1)}}{\epsilon} = f(h_{(t-1)}) \Rightarrow h_t = h_{t-1} + \epsilon f(h_{t-1})$$

- An ODE solution is stable if the long-term behavior of the system does not depend significantly on the initial conditions, meaning if:

$$\max_{i=1,2,\ldots,n} Re(\lambda_i(\boldsymbol{J}(t))) \leq 0, \quad \forall t \geq 0$$

where *Re(.)* is the real part of a complex number, *J(t)* the Jacobian matrix of *f*, and λ_i the i-th eigenvalue of *J(t)*. A perturbation of size δ on h(0) must be: **0 ≤ δ ≤ ε** where ε is the **step size** of the Forward Euler method while moving along the tangential direction to the exact trajectory of h_{t-1} .

- When satisfying the **critical criterion:** $Re(\lambda_i(\boldsymbol{J}(t))) \approx 0, \quad \forall i = 1, 2, \ldots, n$
  the system preserves the long-term dependencies of the inputs while being stable.

antisymmetric formulation is a sufficient condition of stability, not a necessary one

FORWARD EULER METHOD

Inputs: initial condition y_0, end time t_{end}.

Initialize: $t_0 = 0, y_0$.

for $n = 0 : t_{end}/h$ **do**

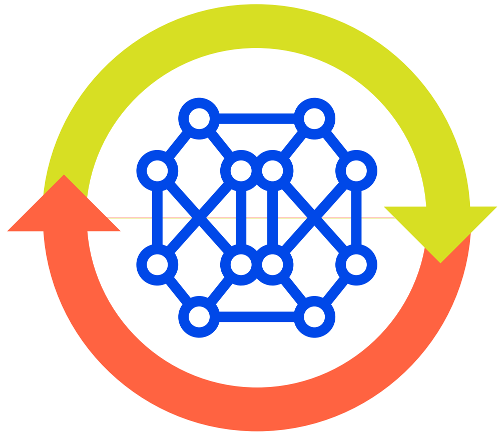
$t_n = nh$

$y_{n+1} \leftarrow y_n + hf(t_n, y_n)$

end for

return y vector.

TRAINABILITY OF THE RNN PRODUCED BY DISCRETIZATION



STABILITY OF THE ODE

TRAINABILITY OF AN ANTISYMMETRIC-RNN

Sensitivity analysis (stability of a solution if *h(0)* changes) with **chain rules:** $\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \boldsymbol{h}(t)}{\partial \boldsymbol{h}(0)}\right) = \boldsymbol{J}(t)\frac{\partial \boldsymbol{h}(t)}{\partial \boldsymbol{h}(0)}$

For notational simplicity: $\boldsymbol{A}(t) = \partial \boldsymbol{h}(t)/\partial \boldsymbol{h}(0) \Rightarrow \frac{\mathrm{d}\boldsymbol{A}(t)}{\mathrm{d}t} = \boldsymbol{J}(t)\boldsymbol{A}(t), \quad \boldsymbol{A}(0) = \boldsymbol{I}$

A(T) = JACOBIAN OF A HIDDEN STATE W.R.T THE INITIAL HIDDEN STATE
Λ(J) = EIGENVALUES OF *J*
COLUMNS ***P*** = EIGENVECTORS OF *Λ(J)*

$$\boldsymbol{A}(t) = e^{\boldsymbol{J}\cdot t} = \boldsymbol{P}e^{\boldsymbol{\Lambda}(\boldsymbol{J})t}\boldsymbol{P}^{-1}$$

When *A(t)* meets the critical criterion, its **magnitude** is costant in time → **NO vanishing/ exploding gradient!**

BUILDING THE ANTISYMMETRIC-RNN

- Let's consider an **antisymmetric matrix M** such that: $\boldsymbol{M}^T = -\boldsymbol{M}$
  Property: eigenvalues *Λ(M)* are **imaginary (or 0)** → Satisfies the **critical criterion!** → **Stability**

- $$\boldsymbol{h}'(t) = \tanh\left(\underbrace{(\boldsymbol{W}_h - \boldsymbol{W}_h^T)}_{\text{ANTISYMMETRIC MATRIX}}\boldsymbol{h}(t) + \boldsymbol{V}_h\boldsymbol{x}(t) + \boldsymbol{b}_h\right)$$

  $h(t) \in \mathbb{R}^n$ $\quad W_h \in \mathbb{R}^{n\times n}$ $\quad b_h \in \mathbb{R}^n$
  $x(t) \in \mathbb{R}^m$ $\quad V_h \in \mathbb{R}^{n\times m}$

  $$\boldsymbol{J}(t) = \mathrm{diag}\left[\tanh'\left((\boldsymbol{W}_h - \boldsymbol{W}_h^T)\boldsymbol{h}(t) + \boldsymbol{V}_h\boldsymbol{x}(t) + \boldsymbol{b}\right)\right](\boldsymbol{W}_h - \boldsymbol{W}_h^T)$$

  JACOBIAN OF $\boldsymbol{h}'(t)$

  ENTRIES OF DIAG [] BOUNDED IN [0,1] → J(T) CHANGES SMOOTHLY OVER TIME
  ODE MOSTLY AFFECTED BY ANTISYMMETRIC MATRIX

- Discretizing with Forward Euler Method:

ANTISYMMETRIC RNN

$$\boldsymbol{h}_t = \boldsymbol{h}_{t-1} + \epsilon\tanh\left((\boldsymbol{W}_h - \boldsymbol{W}_h^T)\boldsymbol{h}_{t-1} + \boldsymbol{V}_h\boldsymbol{x}_t + \boldsymbol{b}_h\right)$$

PARAMETER EFFICIENT MODEL THANKS TO $\boldsymbol{W}_h$ THAT CAN BE PARAMETERIZED AS A STRICTLY UPPER TRIANGULAR MATRIX (ALL DIAGONAL ENTRIES = 0)

STABILITY OF THE FORWARD EULER METHOD: DIFFUSION

The forward Euler method, and consequently of the final AntisymmetricRNN formula, are stable if:

$$\max_{i=1,2,\ldots,n} |1+\epsilon\lambda_i(\boldsymbol{J}_t)| \leq 1$$

LEFT HAND IS ALWAYS >1 SINCE THE EIGENVALUES OF THE JACOBIAN ARE ALL IMAGINARY, MAKING THE FINAL FORMULA **UNSTABLE**.

To deal with this other level of instability, we use a stabilization technique: *diffusion*

$$\boldsymbol{h}_t = \boldsymbol{h}_{t-1} + \epsilon \tanh\left((\boldsymbol{W}_h - \boldsymbol{W}_h^T - \gamma\boldsymbol{I})\boldsymbol{h}_{t-1} + \boldsymbol{V}_h\boldsymbol{x}_t + \boldsymbol{b}_h\right)$$

We inserted a number γ>0 to be subtracted from the diagonal elements of the transition matrix, so that the $\lambda_i(\boldsymbol{J}_t)$ have some negative parts and the stability of the forward Euler method holds.

THE **HADAMARD PRODUCT** IS A BINARY OPERATION THAT TAKES IN TWO MATRICES OF THE SAME DIMENSIONS AND RETURNS A MATRIX OF THE MULTIPLIED CORRESPONDING ELEMENTS.

GATING MECHANISM

$$\boldsymbol{z}_t = \sigma\left((\boldsymbol{W}_h - \boldsymbol{W}_h^T - \gamma\boldsymbol{I})\boldsymbol{h}_{t-1} + \boldsymbol{V}_z\boldsymbol{x}_t + \boldsymbol{b}_z\right)$$

INPUT GATE TO CONTROL THE FLOW OF INFORMATION INTO THE HIDDEN STATES

σ: SIGMOID FUNCTION $\circ$: HADAMARD PRODUCT

Gating in RNNs offers flexible control of two salient features of the collective dynamics: **timescales** and **dimensionality.** We introduce $\boldsymbol{z}_t$, an input gate that controls information flow. We use **Hadamard product** so that a diagonal matrix multiplied by an antisymmetric matrix make the Jacobian matrix of this gated model similar to the one of $\boldsymbol{h}'(t)$. The real parts of the eigenvalues of the Jacobian matrix are still close to zero, and the critical criterion remains satisfied.

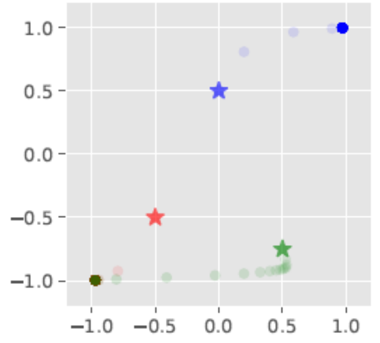
$$\boldsymbol{h}_t = \boldsymbol{h}_{t-1} + \epsilon\boldsymbol{z}_t \circ \tanh\left((\boldsymbol{W}_h - \boldsymbol{W}_h^T - \gamma\boldsymbol{I})\boldsymbol{h}_{t-1} + \boldsymbol{V}_h\boldsymbol{x}_t + \boldsymbol{b}_h\right)$$

SIMULATIONS

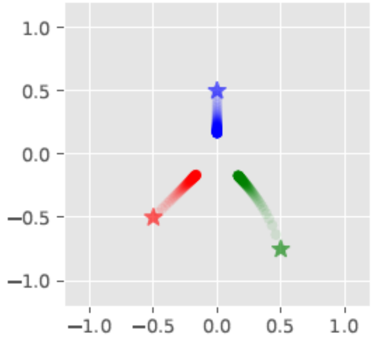
2-DIM VANILLA RNN $\boldsymbol{h}_t = \tanh(\boldsymbol{W}\boldsymbol{h}_{t-1})$

Initial states: (0,0.5); (-0.5,-0.5) and (0.5,-0.75)

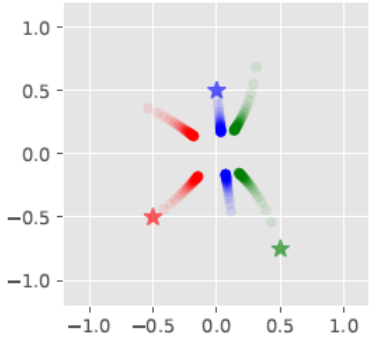
Total time steps *T* = 50



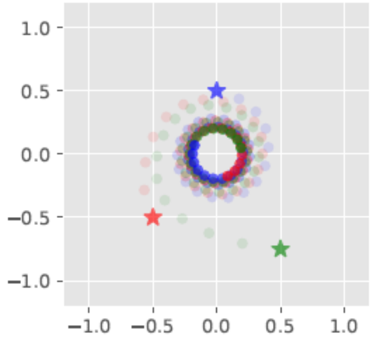
RANDOM WEIGHT MATRIX (UNSTRUCTURED): UNPREDICTABLE BEHAVIOUR



IDENTITY WEIGHT MATRIX: *TANH* BEING CONTRACTIVE MAPPING, HAS UNIQUE FIXED POINT IN ORIGIN



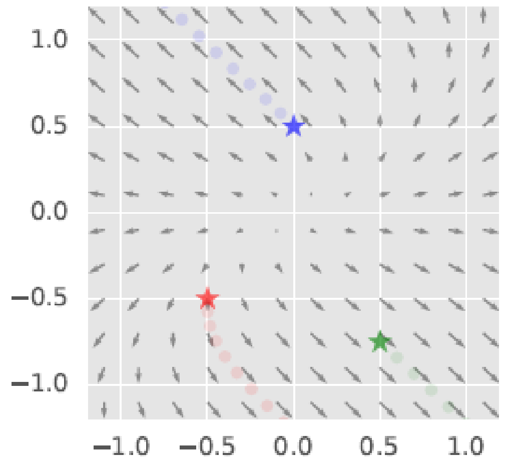
ORTHOGONAL WEIGHT MATRIX (**REFLECTION**)



ORTHOGONAL WEIGHT MATRIX (**ROTATION**)

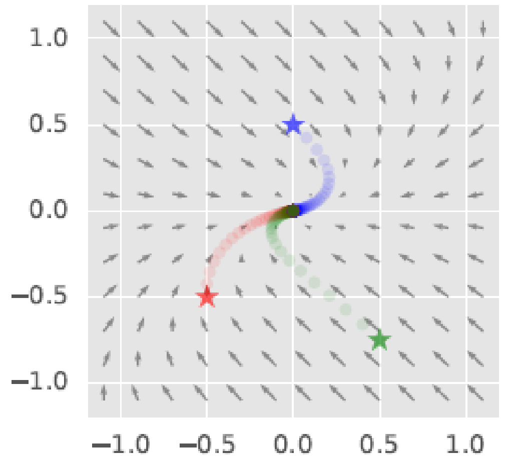
DETERMINANT OF THE MATRIX IS 1 OR -1. CONVERGES TO ORIGIN BECAUSE OF *TANH'*

RNN WITH FEEDBACK $\boldsymbol{h}_t = \boldsymbol{h}_{t-1} + \epsilon \tanh(\boldsymbol{W}\boldsymbol{h}_{t-1})$



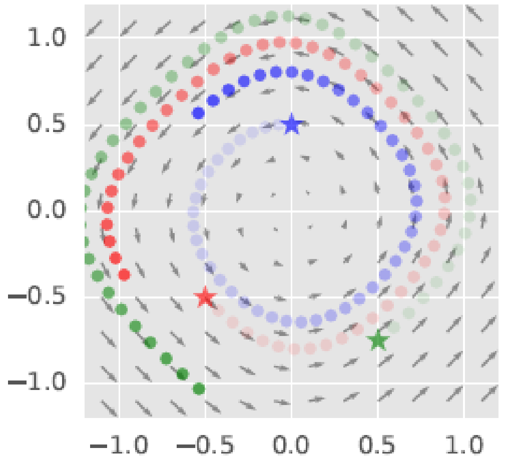
POSITIVE EIGENVALUES

$$\lambda_1(W_+) = \lambda_2(W_+) = 2$$



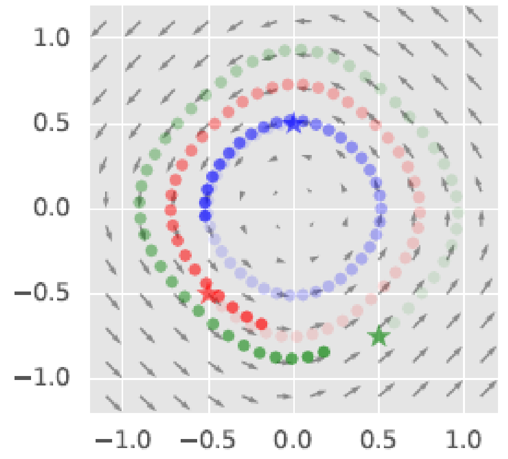
NEGATIVE EIGENVALUES

$$\lambda_1(W_-) = \lambda_2(W_-) = -2$$



IMAGINARY EIGENVALUES **(ANTISYMMETRIC RNN)**

$$\lambda_1(\boldsymbol{W}_0) = 2i$$
$$\lambda_2(\boldsymbol{W}_0) = -2i$$



IMAGINARY EIGENVALUES + DIFFUSION

$$\gamma = 0.15$$
$$\lambda_1(\boldsymbol{W}_{\text{diff}}) = -0.15 + 2i$$
$$\lambda_2(\boldsymbol{W}_{\text{diff}}) = -0.15 - 2i$$

$$\boldsymbol{W}_+ = \begin{pmatrix} 2 & -2 \\ 0 & 2 \end{pmatrix}, \boldsymbol{W}_- = \begin{pmatrix} -2 & 2 \\ 0 & -2 \end{pmatrix}, \boldsymbol{W}_0 = \begin{pmatrix} 0 & -2 \\ 2 & 0 \end{pmatrix}, \boldsymbol{W}_{\text{diff}} = \begin{pmatrix} -0.15 & -2 \\ 2 & -0.15 \end{pmatrix}$$

HIDDEN STATES ARE MOVING AWAY FROM AND TOWARDS THE ORIGIN RESPECTIVELY

VECTOR FIELD IS CIRCULAR, TRAJECTORIES ARE OUTWARD SPIRALS! MOVING ALONG THE TANGETIAL DIRECTION LEADS TO NUMERICAL INSTABILITY

CIRCULAR VECTOR FIELD IS TILTING TOWARD THE ORIGIN AND TRAJECTORY MAINTAINS A CONSTANT DISTANCE FROM ORIGIN

Hidden states of an AntisymmetricRNN have **predictable dynamics** without the complication of maintaining an orthogonal/unitary matrix thanks to the **eigenvalues** of the weight matrix ***W***!

SETUP
CROSS-ENTROPY LOSS
SGD WITH MOMENTUM AND ADAGRAD

EXPERIMENTAL WORK

Evaluation of the models on image classification tasks with long-range dependencies. The last hidden state of the models is fed to a a fully-connected layer and a softmax function.

PIXEL-BY-PIXEL MNIST

Benchmark task: predict the digit of the MNIST image (grayscale 28x28 pixels) after seeing all the permuted (shuffled) pixels. Orthogonal weights, even smoothed, are largely outperformed by AntisymmetricRNNs, corroborating that such constraints restrict the capacity of the learned model.

method	MNIST	pMNIST	# units	# params
LSTM (Arjovsky et al., 2016)[1]	97.3%	92.6%	128	68k
FC uRNN (Wisdom et al., 2016)	92.8%	92.1%	116	16k
FC uRNN (Wisdom et al., 2016)	96.9%	94.1%	512	270k
Soft orthogonal (Vorontsov et al., 2017)	94.1%	91.4%	128	18k
KRU (Jose et al., 2017)	96.4%	94.5%	512	11k
AntisymmetricRNN	98.0%	**95.8%**	128	10k
AntisymmetricRNN w/ gating	**98.8%**	93.1%	128	10k

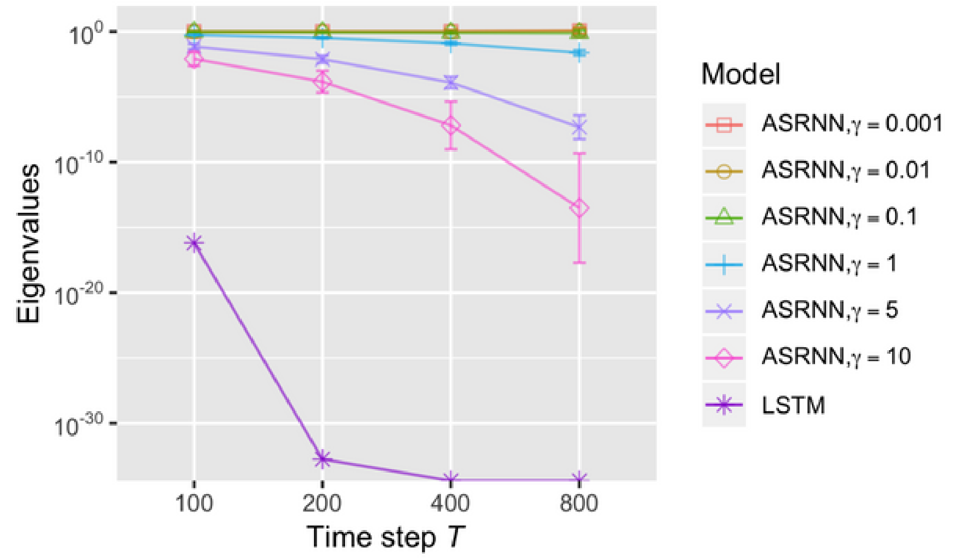
PIXEL-BY-PIXEL CIFAR-10

CIFAR-10: 32x32 colour images in 10 classes; feeding the three channels of a pixel into the model at each time step. Performances all similar, but A-RNNs use half parameters. Task mostly dominated by short term dependencies!
Ablation model: antisymmetric weight matrices substituted with unstructured weight matrices.

method	pixel-by-pixel	noise padded	# units	# params
LSTM	59.7%	11.6%	128	69k
Ablation model	54.6%	46.2%	196	42k
AntisymmetricRNN	58.7%	48.3%	256	36k
AntisymmetricRNN w/ gating	**62.2%**	**54.7%**	256	37k

NOISE PADDED CIFAR-10

Noise: after the first 32 time steps, we input independent standard Gaussian noise for the remaining time steps, meaning that all remaining 968 time steps are merely random and uninformative. Additional experiments varying the length of noise padding are shown in the figure.
Both LSTM and A-RNNs run into vanishing gradients, even with A-RNNs' eigenvalues centered around 1 and γ constats to balance.



MEAN AND STANDARD DEVIATION OF EIGENVALUES OF THE END-TO-END JACOBIAN MATRIX IN ANTISYMMETRICRNNS WITH DIFFERENT DIFFUSION CONSTANTS AND LSTMS

CONCLUSIONS

OVERALL VIEW

Proposition of *AntisymmetricRNN*, an innovative model that opens up to the exploitation of the computational and theoretical success from dynamical systems, as to improve trainability of RNNs.

NOVELTIES

- Conjunction of dynamical systems' view and machine learning techniques in the use of discretized ordinary differential equations for neural network construction
- Exploitation of imaginary eigenvalues of the antisymmetrix matrix to satisfy the critical criterion and improve stability

STRONG POINTS

- Optimization of Jacobian matrices of the hidden states of the model through numerical approximation using the forward Euler method
- Further level of stabilization for the imaginary eigenvalues using diffusion for a slightly negative part
- Better or similar results with less parametrization w.r.t. other RNNs

WEAKNESSES

- The forward Euler method itself does not guarantee stability and diffusion require a search for an optimal constant value
- For more complex tasks, there is no actual improvement in the accuracy of the model and the vanishing gradient problem is not eradicated