Structure preservation in (some) deep learning architectures

Brynjulf Owren

Department of Mathematical Sciences, NTNU, Trondheim, Norway

LMS-Bath Symposium – 2020

Joint work with: Martin Benning, Elena Celledoni, Matthias Ehrhardt, Christian Etmann, Carola-Bibiane Schönlieb and Ferdia Sherry

- Benning, Martin; Celledoni, Elena; Ehrhardt, Matthias J.; Owren, Brynjulf; Schönlieb, Carola-Bibiane, *Deep Learning as Optimal Control Problems: Models and Numerical Methods* J. Comput. Dyn. 6 (2019), no. 2, 171–198.
- Elena Celledoni, Matthias J. Ehrhardt, Christian Etmann, Robert I McLachlan, Brynjulf Owren, Carola-Bibiane Schönlieb, Ferdia Sherry, *Structure preserving deep learning*, arXiv:2006.03364 (June 2020)

Neural network layers: $\phi^k : \mathcal{X}^k \times \Theta^k \to \mathcal{X}^{k+1}$,

Θ^k: Parameter space of layer k

$\mathcal{X}^k$ The kth feature space

The full neural network

$$\begin{aligned} \Psi : \mathcal{X} \times \Theta &\to \mathcal{Y} \\ (x, \theta) &\mapsto z^K \end{aligned}$$

can then be defined via the iteration

$$\begin{aligned} z^0 &= x \\ z^{k+1} &= \phi^k(z^k, \theta^k), \quad k = 0, \dots, K-1, \end{aligned}$$

Extra final layer may be needed: $\eta : \mathcal{X}^K \times \Theta^K \to \mathcal{Y}$.
In this talk, $\mathcal{X}^k = \mathcal{X}$ for all k.

Training data: $(x_n, y_n)_{n=1}^N \subset \mathcal{X} \times \mathcal{Y}$
Training the network amounts to minimising the loss function

$$\min_{\theta \in \Theta} \left\{ E(\theta) = \frac{1}{N} \sum_{n=1}^{N} L_n(\Psi(x_n, \theta)) + R(\theta) \right\},$$

where

- $L_n(y) : \mathcal{Y} \to \mathbb{R}_\infty$ is the loss for a specific data point
- $R : \Theta \to \mathbb{R}_\infty$ acts as a regulariser which penalises and constrains unwanted solutions.

We can define the loss over a batch of N data points in terms of the final layer as

$$E(z; \theta) = \frac{1}{N} \sum_{n=1}^{N} L_n(\eta(z_n), \theta) + R(\theta)$$

ResNet model (He et al. (2016))

$\Psi : \mathcal{X} \times \Theta \to \mathcal{X},\ \Psi(x,\theta) = z^K$ given by the iteration

$$\begin{aligned} z^0 &= x \\ z^{k+1} &= z^k + \sigma(A^k z^k + b^k), \quad k = 0, \ldots, K-1, \\ y &= \eta(w^T z^K + \mu) \end{aligned}$$

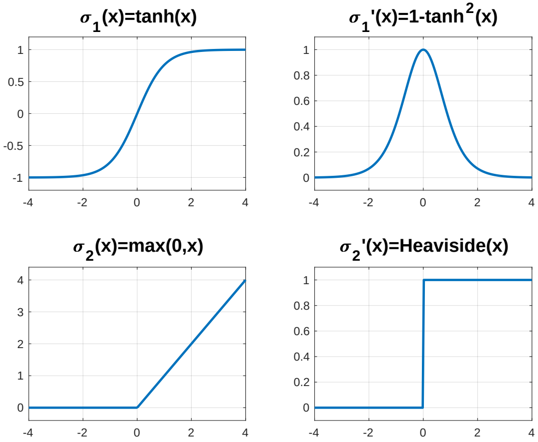
- σ is a nonlinear activation function, a scalar function acting element-wise on vectors.
- $\theta^k = (A^k, b^k),\ k \leq K-1$. $\theta^K = (w, \mu)$.

The ResNet layers can be seen as a time stepper for the ODE

$$\dot{z} = \sigma(A(t)z + b(t)),\ t \in [0, T]$$

It is the explicit Euler method with stepsize $h = 1$.

$$\sigma_1(x) = \tanh x$$
$$\sigma_2(x) = \max(0, x), \quad \text{(RELU)}$$



$$\min_{(\theta,z)\in\Theta\times\mathcal{X}^N}\left\{E(\theta,z)=\frac{1}{N}\sum_{n=1}^{N}L_n(z_n(T))+R(\theta)\right\}$$

$$\text{such that}\quad \dot{z}_n=f(z_n,\theta(t)),\quad z_n(0)=x_n,\quad n=1,\dots,N.$$

The first order optimality conditions can be phrased as a Hamiltonian Boundary Value Problem (Benning et al. (2020)). Define

$$H(z, p; \theta) = \langle p, f(z, p; \theta) \rangle$$

Solve

$$\dot{z} = \frac{\partial H}{\partial p}, \quad \dot{p} = -\frac{\partial H}{\partial z}, \quad 0 = \frac{\partial H}{\partial \theta}.$$

with boundary conditions

$$z(0) = x, \quad p(T) = \left.\frac{\partial L}{\partial z}\right|_{t=T}$$

For ResNet, $f(z, p; \theta) = \sigma(A(t)z + b(t))$, and we shall discuss other alternative vector fields f.

Standard procedure:

Initial guess $\theta^{(0)}$
while not converged
    Sweep forward $\dot{z} = f(z; \theta^{(i)})$ to get $z^1, \ldots, z^K$, $z^k = \phi(z^{k-1})$
    Backprop on $\dot{p} = -Df(z)^T p$ to obtain $\nabla_\theta E$
    Update by some descent method e.g. $\theta^{(i+1)} = \theta^{(i)} - \tau \nabla_\theta E(\theta^{(i)})$

- Chen et al (2018) suggest to use a black-box solver. Obtain $z(T)$ and then do $(z(t), p(t))$ backwards in time simultaneously to save memory usage.
- Problematic for various reasons. No explicit solver satisfying first order optimality conditions + stability issues.
- Gholami et al (2019) amend problem by a checkpointing method so only forward sweeps through feature spaces. Again: first order optimality is not so clear

Two options

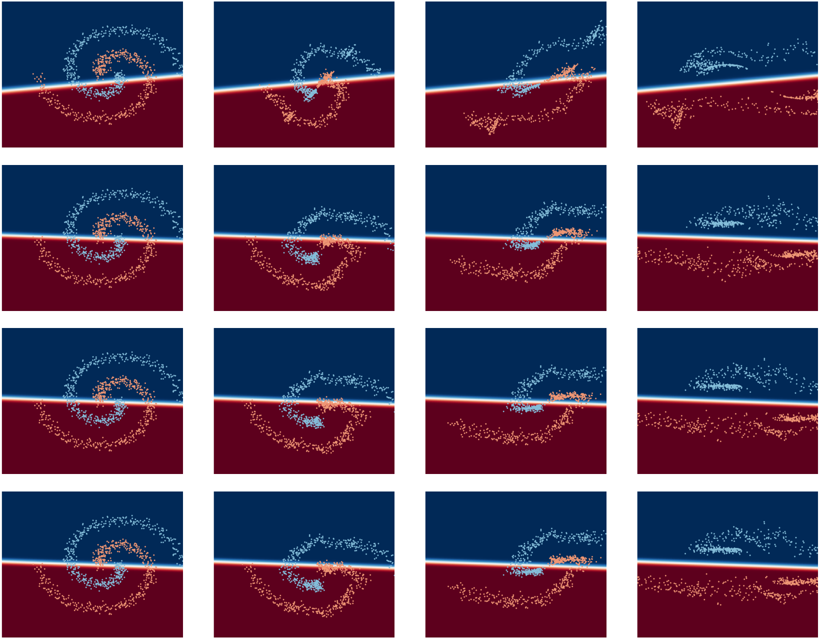
1. DTO. Discretise the forward ODE ($\dot{z} = f(z;\theta)$) by some numerical method ϕ. Then solve the discrete optimisation problem, based on the gradients $\nabla_{\theta^k} E(z^K;\theta^K)$.
2. OTD. Solve the Hamiltonian boundary value problem by a numerical method $\bar{\phi} : (z^k, p^k) \mapsto (\phi(z^k), p^{k+1})$ and compute $\frac{\partial \phi}{\partial \theta}(z^k, \theta^k)^T p^{k+1}$ for each k.

Theorem (Benning et al 2020, Sanz-Serna 2015)

DTO and OTD are equivalent if the overall method $\bar{\phi}$ for the Hamiltonian boundary value problem preserves quadratic invariants (a.k.a. symplectic). That is,

$$\nabla_{\theta^k} E(z^K;\theta^K) = \frac{\partial \phi}{\partial \theta}(z^k, \theta^k)^T p^{k+1}$$

An illustration



Once the network has been trained, the parameters $\theta(t)$ are known. Generalisation (the forward problem) becomes a non-autonomous initial value problem

$$\dot{z} = \bar{f}(t, z) := f(z; \theta(t)), \quad z(0) = x.$$

- Arguably, one may ask for good "stability properties" for the forward problem. Haber & Ruthotto (2017), Zhang & Schaeffer (2020).
- Stability may also be desired in "backward time", Chang et al. (2018).

What is our freedom in choosing good models?

- Restrict parameter space Θ (A skew-symmetric, negative definite, manifold-valued,. . .)
- Alter the structure of the vector field f (Hamiltonian, dissipative, measure preserving,. . .)
- Apply integrator with good stability properties

- Linear stability analysis (Haber and Ruthotto). Nonlinear vector field $f(t,z)$ look at spectrum of

$$J(t,z) := \frac{\partial f}{\partial z}(t,z), \quad \mathrm{Re}\,\lambda_i \leq 0$$

Works only locally and only with autonomous vector fields.

- Nonlinear stability analysis, look at norm contractivity/growth

$$\|z_2(t) - z_1(t)\| \leq C(t)\|z_2(0) - z_1(0)\|$$

Such conditions can be ensured by imposing Lipschitz type conditions. E.g. for inner product spaces $\nu \in \mathbb{R}$

$$\langle f(t,z_2) - f(t,z_1), z_2 - z_1\rangle \leq \nu\|z_2 - z_1\|_2^2, \ \forall z_1, z_2, \ t \in [0,T]$$

$$\Rightarrow \|z_2(t) - z_1(t)\| \leq e^{\nu t}\|z_2(0) - z_1(0)\|$$

Example of a stability result (Celledoni et al. (2020))

We consider for simplicity the ODE model

$$\dot{z} = -A(t)^T \sigma(A(t)z + b(t)) = f(t, z),$$

Here $\dot{z} = -\nabla_z V$ with $V = \gamma(A(t)z + b(t))\mathbf{1}$ where $\gamma' = \sigma$

Theorem

1. *Let $V(t, z)$ be twice differentiable and convex in the second argument. Then the vector field $f(t, z) = -\nabla V(t, z)$ satisfies a one-sided Lipschitz condition with $\nu \leq 0$.*
2. *Suppose that $\sigma(s)$ is absolutely continuous and $0 \leq \sigma'(s) \leq 1$ a.e. in $\mathbb{R}$. Then the one-sided Lipschitz condition holds for any $A(t)$ and $b(t)$ with*

$$-\mu_*^2 \leq \nu_\sigma \leq 0$$

where $\mu_* = \min_t \mu(t)$ and where $\mu(t)$ is the smallest singular value of $A(t)$. In particular $\nu_\sigma = -\mu_*^2$ is obtained when $\sigma(s) = s$.

Let

$$H(t, z, p) = T(t, p) + V(t, z)$$

Let $\gamma_i : \mathbb{R} \to \mathbb{R}$ be such that $\gamma_i'(t) = \sigma_i(t),\ i = 1, 2$ and set

$$T(t, p) = \gamma_1(A_1(t)p + b_1(t))\mathbf{1}, \quad V(t, z) = \gamma_2(A_2(t)z + b_2(t))\mathbf{1}$$

where $\mathbf{1} = (1, \ldots, 1)^T$.

This leads to models of the form

$$\dot{z} = \partial_p H = A_1(t)^T \sigma_1(A_1(t)p + b_1(t))$$
$$\dot{p} = -\partial_z H = -A_2(t)^T \sigma_2(A_2(t)z + b_2(t))$$

1. A simple case is obtained by choosing $\sigma_1(s) := s$, $A_1(t) \equiv I$, $b_1(t) \equiv 0$ and $\sigma_2(s) := \sigma(s)$ which after eliminating p yields the second order ODE

$$\ddot{z} = -\partial_z V = -A(t)^T \sigma(A(t)z + b(t))$$

2. A second example

$$\dot{z} = A(t)^T \sigma(A(t)p + b(t))$$
$$\dot{p} = -A(t)^T \sigma(A(t)z + b(t))$$

Autonomous problems

- Two important geometric properties
  - The flow preserves the Hamiltonian
  - The flow is symplectic
- Numerical schemes can be symplectic or energy preserving, excellent long time behaviour

Non-autonomous Hamiltonian problems

- The situation is less clear, at least two ways to interpret the dynamics
  1. 'Autonomise' by adding time as dependent variable (contact manifold). A preserved two-form can be introduced

     $$\omega = dp \wedge dq - dH \wedge dt$$

     but the Hamiltonian is not preseved along the flow
  2. Extend system by adding time and a conjugate momentum variable p_t. Define extended Hamiltonian $K(q, p, t, p_t) = H(q, p, t) + p_t$ and symplectic form

     $$\Omega = dp \wedge dq + dp_t \wedge dt$$

$$\dot{z} = \partial_p H, \ \dot{p} = -\partial_z H, \ \dot{t} = 1, \ \dot{p}_t = -\partial_t H$$

- An obvious strategy would be to study the dynamics of the extended autonomous Hamiltonian system.
- Unfortunately, it does not give a lot of information
- Any level set of K is unbounded
- Chang et al (2018) report good numerical results with this type of model, I am not aware of any theoretical justification
- Asorey et al. (1983) contains a number of results for the relations between the dynamics on the contact manifold and the extended manifold, [more work to be done in this direction]
- LO Jay (2020), Marthinsen & O (2016) provide conditions on numerical integrators to be canonical in the non-autonomous case

Without regularisaton, the learned parameters become irregular in time [see figure].
In the continuous model one may add a regularisation e.g.

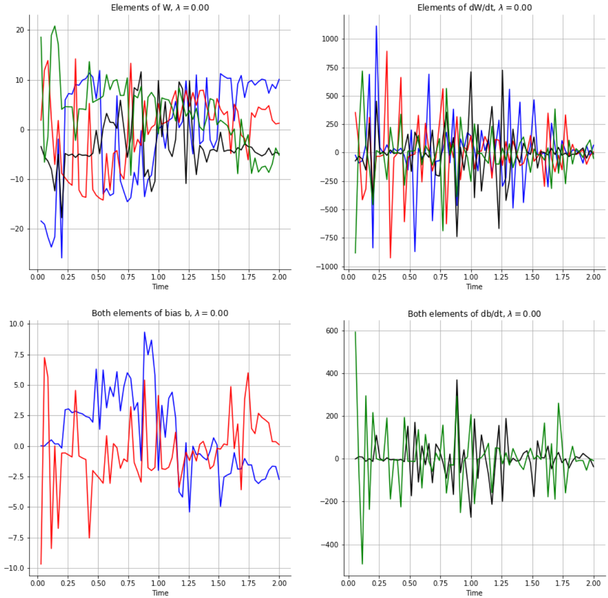
$$R(\theta) = \lambda \int_0^T \|\dot{\theta}\|^2 \,\mathrm{d}t$$

discretised, say, as

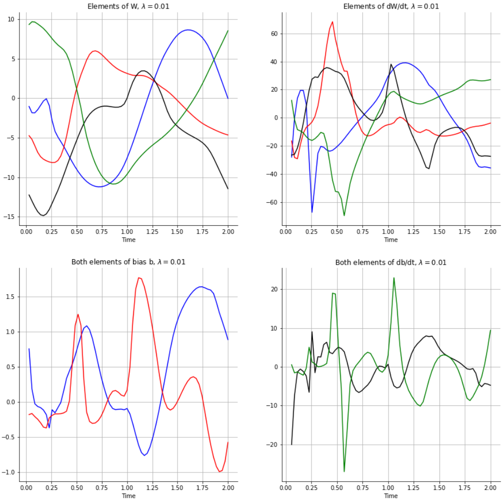
$$R_h(\theta) = \lambda\, h \sum_k \left(\frac{\|\theta(t_{k+1} - \theta(t_k)\|}{h} \right)^2$$

We tried $\lambda \in \{0.0,\ 0.1,\ 1.0\}$

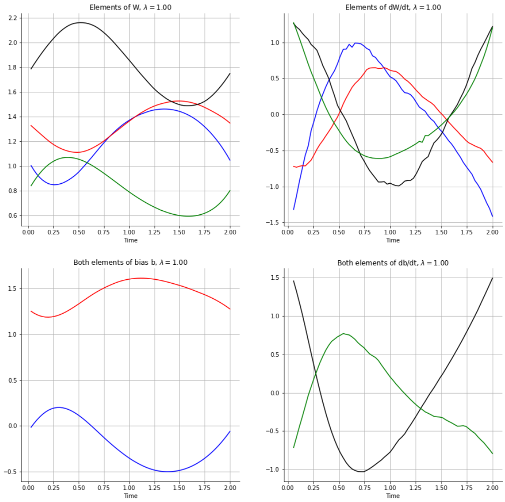
A test on the spiral problem, $\lambda = 0$



A test on the spiral problem, $\lambda = 0.1$



A test on the spiral problem, $\lambda = 1.0$

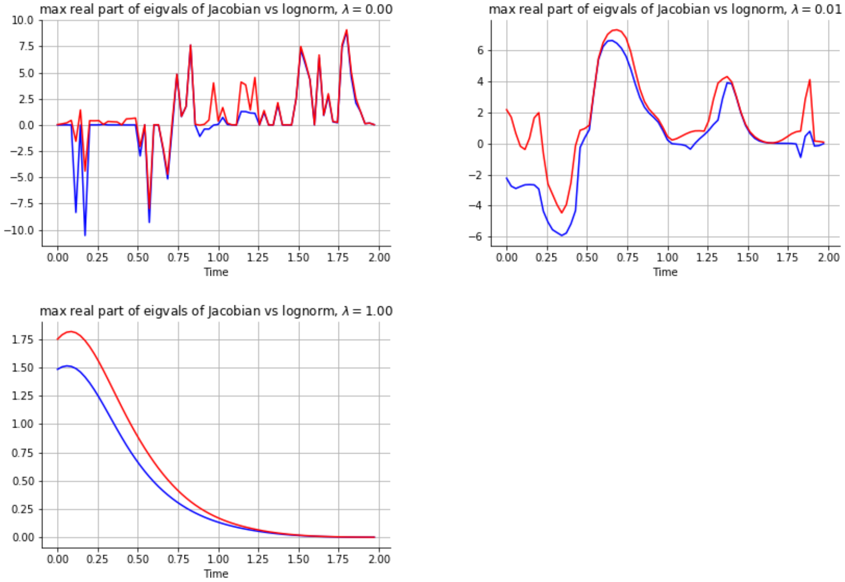


Making the parameters more regular may intuitively make the system “more autonomous".

Can we then use eigenvalue analysis for stability?
In the next plot we show

- the largest real part of the Jacobian eigenvalues (blue)
- the one-sided Lipschitz constant (red)

Eigenvalues (real part) vs one-sided Lipschitz constants

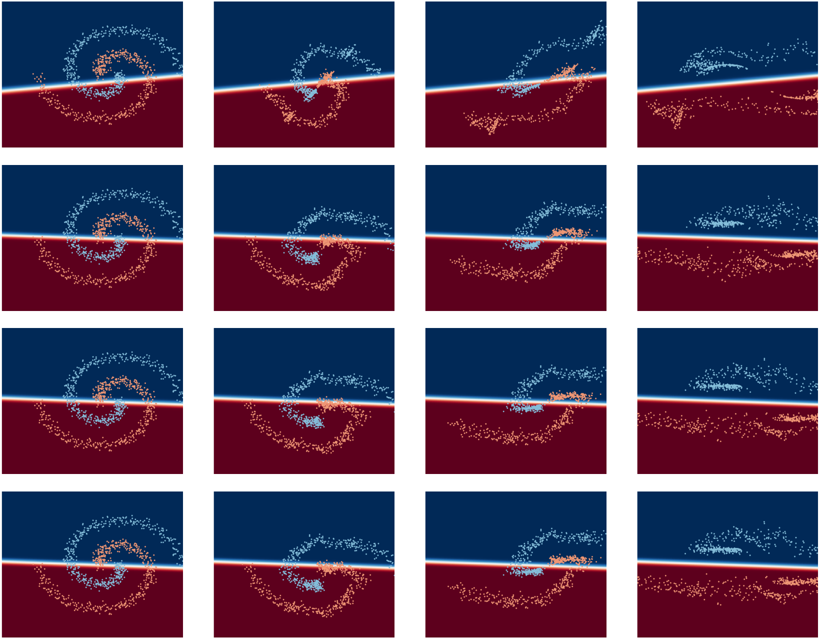


- Deep limits – convergence as $K \to \infty$
- Invertible networks (similar to ODE-based networks)
- Features evolving on homogeneous manifolds
- Equivariance in Convolutional networks
- Algorithms for optimisation
  - Descent methods accelerated by momentum, and ADAM-like methods
  - Hamiltonian descent methods
  - Learning in Riemannian metric spaces
  - Parameters evolving on manifolds

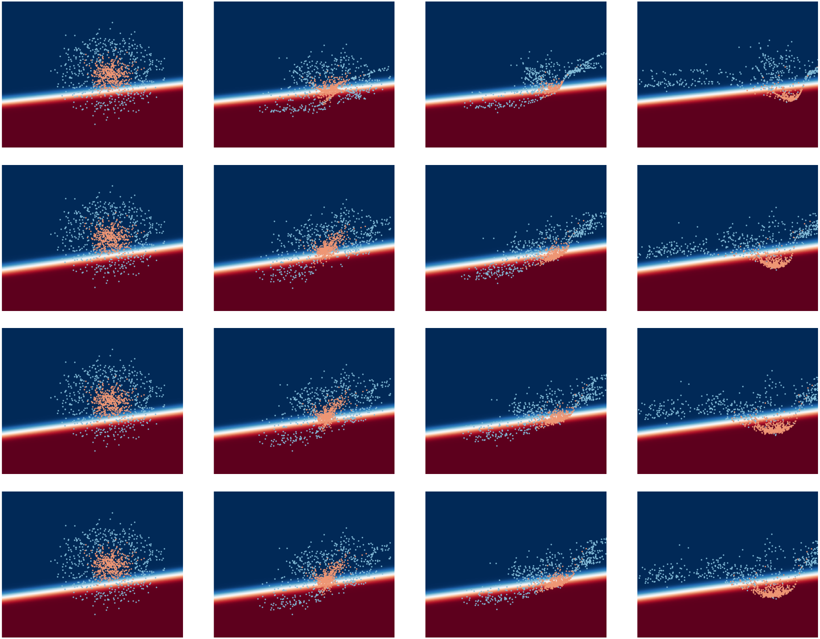
Thank you!

Additional plots

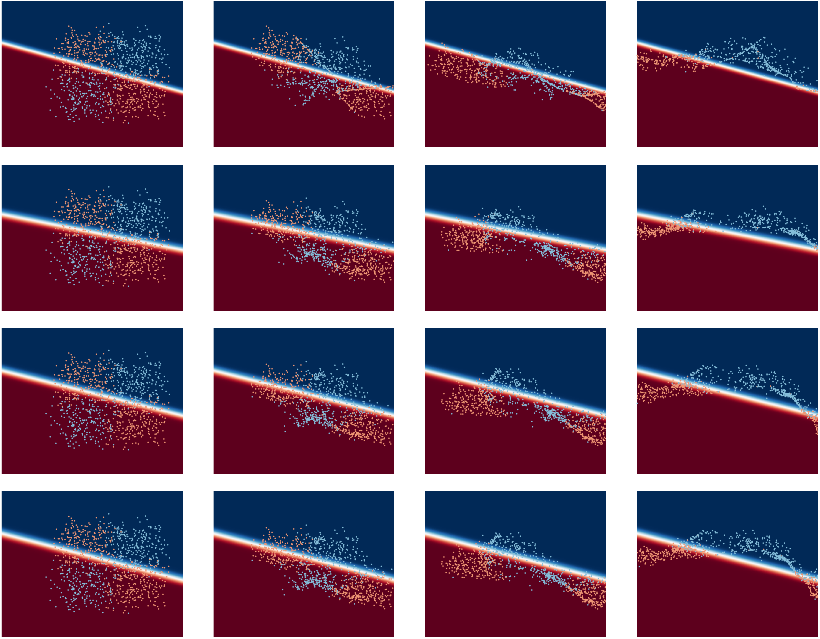
Transitions in Runge–Kutta methods – spiral



Transitions in Runge–Kutta methods – donut2d



Transitions in Runge–Kutta methods – squares



1. Bo Chang, Lili Meng, Eldad Haber, Lars Ruthotto, David Begert, and Elliot Holtham. *Reversible architectures for arbitrarily deep residual neural networks*. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
2. Eldad Haber and Lars Ruthotto. *Stable architectures for deep neural networks.* Inverse Problems, 34(1):014004, 2017.
3. J. M. Sanz-Serna. *Symplectic Runge-Kutta schemes for adjoint equations automatic differentiation, optimal control and more*. SIAM Review, 58:3–33, 2015.
4. Yann LeCun. *A theoretical framework for back-propagation*. In Proceedings of the 1988 connectionist models summer school, volume 1, pages 21–28. CMU, Pittsburgh, Pa: Morgan Kaufmann, 1988.
5. Qianxiao Li and Shuji Hao. *An Optimal Control Approach to Deep Learning and Applications to Discrete-Weight Neural Networks.* arXiv:1803.01299v2, 2018