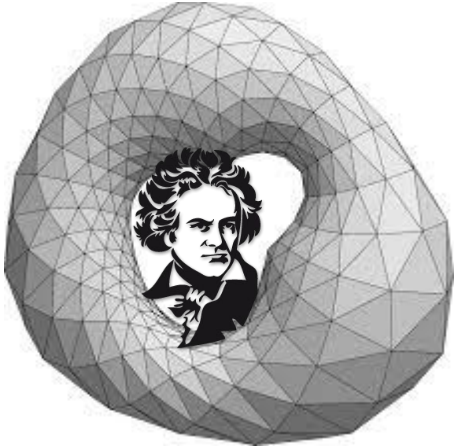
NEURAL INFORMATION PROCESSING SYSTEMS

34th Conference on Neural Information Processing Systems
NeurIPS 2020



Ode to an ODE

Krzysztof Choromanski, Jared Quincy Davis, Valerii Likhosherstov, Xingyou Song, Jean-Jacques Slotine, Jacob Varley, Honglak Lee, Adrian Weller, Vikas Sindhwani

Google AI

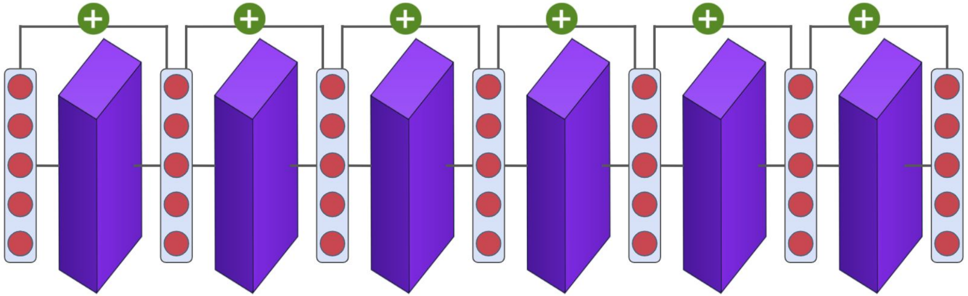
The Alan Turing Institute

Neural ODEs:

- Continuous variants of standard ResNet networks:

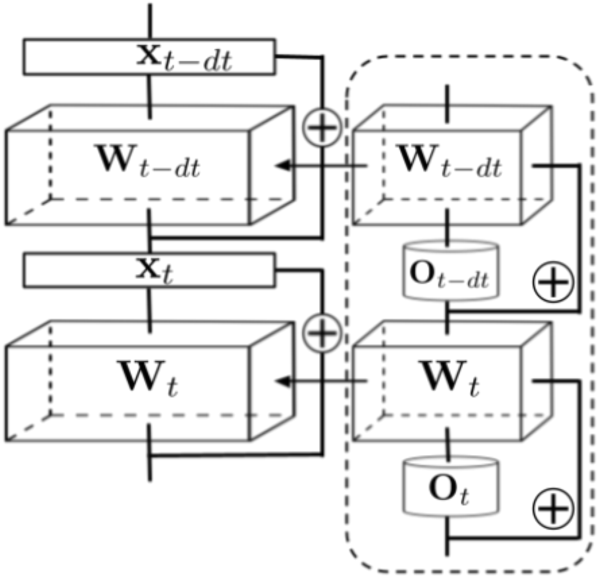
$$\frac{d\mathbf{x}(t)}{dt} = f(\mathbf{x}_t, t, \theta) \qquad \mathbf{x}_t = \mathbf{x}_{t_0} + \int_{t_0}^{t} f(\mathbf{x}_s, s, \theta) ds \tag{1}$$

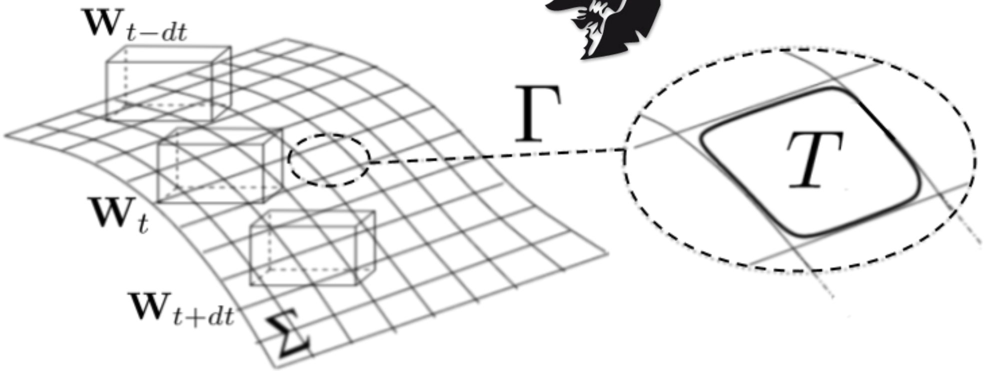
- Emulate deep discrete neural networks with **compact** number of parameters.
- Parameters of the Neural ODEs encapsulated in the mapping $\theta(t)$.
  How to design it ?
- As every deep neural network system, suffer from exploding/vanishing gradients which makes training challenging. **Can we robustify Neural ODEs ?**



Ode to an ODE System:

- **IDEA:** Design $\theta(t)$, so that when integrated, Neural ODE emulates deep ResNet with **orthogonal** connection matrices.



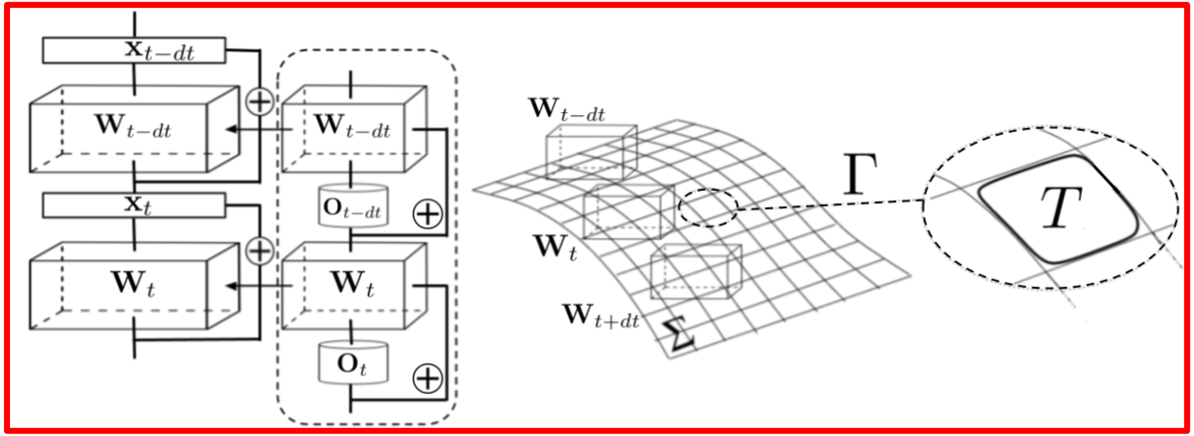


- This leads to the matrix-flow on the **orthogonal group** and effectively: to a **nested system of flows**, where the orthogonal flow encoding $\theta(t)$ determines main flow.

How to design learnable orthogonal flows and why are they good ?

Orthogonal Flows:

(2) $$\begin{cases} \dot{\mathbf{x}}_t = f(\mathbf{W}_t \mathbf{x}_t) \\ \dot{\mathbf{W}}_t = \mathbf{W}_t \underbrace{b_\psi(t, \mathbf{W}_t)}_{\text{mapping to skew-symmetric matrices}} \end{cases}$$

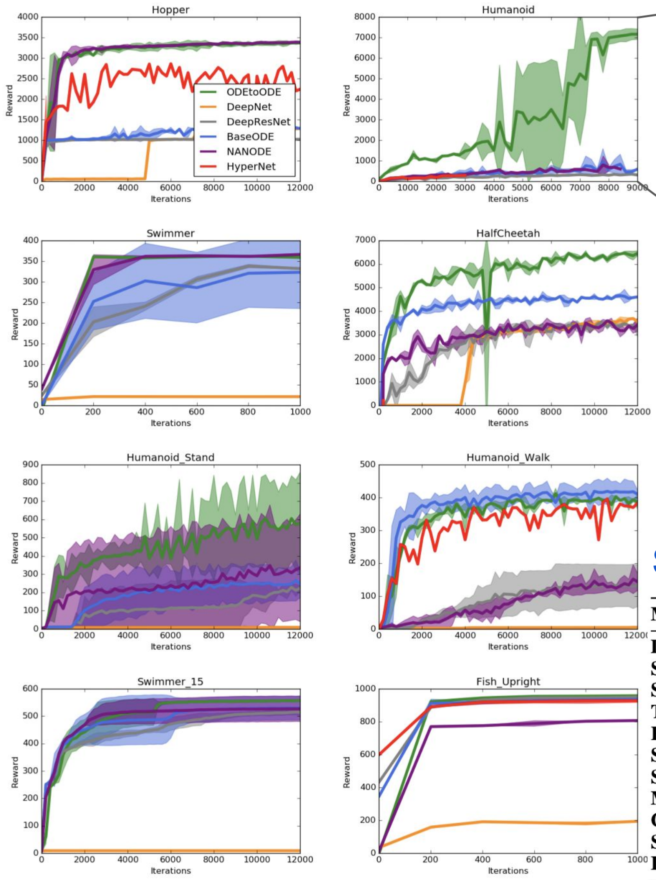


- b_ψ can be modeled by a **neural network producing skew-symmetric matrices** or via parameterized **isospectral flows** (e.g. double-bracket flows)

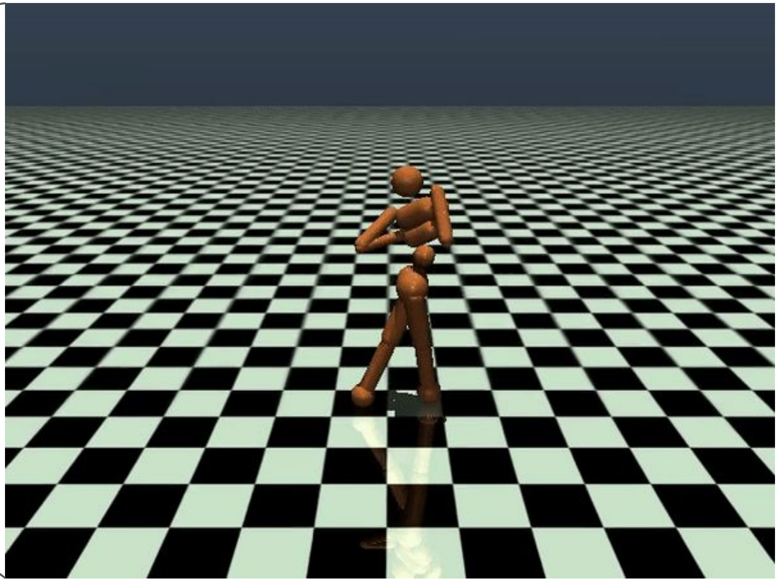
Lemma 4.1 (ODEtoODES for gradient stabilization). *Consider a Neural ODE on time interval* $[0, T]$ *and given by Formula 2. Let* $\mathcal{L} = \mathcal{L}(\mathbf{x}_T)$ *be a differentiable loss function. The following holds for any* $t \in [0, 1]$, *where* $e = 2.71828\ldots$ *is Euler constant:*

$$\boxed{\frac{1}{e}\|\frac{\partial \mathcal{L}}{\partial \mathbf{x}_T}\|_2 \leq \|\frac{\partial \mathcal{L}}{\partial \mathbf{x}_t}\|_2 \leq e\|\frac{\partial \mathcal{L}}{\partial \mathbf{x}_T}\|_2.}$$

Ode to an ODE System in Action:



RL: comparison with Deep(Res)Nets, NANODE, Base ODEs and ANODEV2-hypernets



Supervised: MNIST-Corrupted

Models	Dense-1	Dense-10	NODE	NANODE-1	NANODE-10	HyperNet-1	HyperNet-10	ODEtoODE-1	ODEtoODE-10
Dotted Lines	92.99	88.22	91.54	92.42	92.74	91.88	91.9	**95.42**	95.22
Spatter	93.54	89.52	92.49	93.13	93.15	93.19	93.28	**94.9**	**94.9**
Stripe	30.55	20.57	36.76	16.4	21.37	19.71	18.69	**44.51**	44.37
Translate	25.8	24.82	27.09	28.97	29.31	**29.42**	29.3	26.82	26.63
Rotate	82.8	80.38	82.76	83.19	83.65	83.5	83.56	83.9	**84.1**
Scale	58.68	62.73	62.05	66.43	66.63	67.84	**68.11**	66.68	66.76
Shear	91.73	89.52	91.82	92.18	93.11	92.33	92.48	**93.37**	92.93
Motion Blur	78.16	67.25	75.18	76.53	**79.22**	78.82	78.33	78.63	77.58
Glass Blur	91.62	84.89	87.94	90.18	91.07	91.3	91.17	**93.91**	93.29
Shot Noise	96.16	91.63	94.24	94.97	94.73	94.76	94.81	**96.91**	96.71
Identity	97.55	95.73	97.61	97.65	97.69	97.72	97.54	**97.94**	97.88

Thank you for your Attention !

https://arxiv.org/abs/2006.11421