Neural ODEs: A Representer Theorem

Daniel López Montero

Joint work with Zhengping Ji

September 2, 2026

FAU Erlangen-Nürnberg · Chair for Dynamics, Control, Machine Learning and Numerics

Outline

1. Neural ODEs: Representer theorem

2. Regularity in time

3. Strong regularization

4. Extra: Optimal Transport

Neural ODEs

Neural ODEs model dynamics on $X := \mathbb{R}^d$ with a neural-network drift.

$$\dot{x}(t) = f_\theta(x, t), \qquad t \in [0, 1],$$
$$x(0) = x_0.$$

Consider the time-varying neural network of width m,

$$f_\theta(x, t) := \sum_{j=1}^{m} \sigma\big(x, \xi_j(t)\big)\, w_j(t),$$

where $\theta(t) := \big(w(t), \xi(t)\big) \in X^m \times \Omega^m$, σ is a scalar activation, and $\Omega \subset \mathbb{R}^{d_\xi}$ is compact.

Example

Take $\sigma(x, \xi) := \tanh(\langle a, x\rangle + b)$ with $\xi = (a, b) \in \Omega$, where $\Omega \subset \mathbb{R}^{d+1}$ is compact.

We denote by $\mathcal{N}_{\leq m}$ the set of such time-varying neural networks of width at most m.

Neural ODEs: the training problem

Given n samples $x_{1,0}, \dots, x_{n,0} \in X$ and losses $\varphi_i \geq 0$ (running), $\phi_i \geq 0$ (terminal):

$$\inf_{\theta \in \mathcal{N}_{\leq m}} \mathcal{J}_{\mathrm{NN}}(\theta) := \sum_{i=1}^{n} \left(\int_0^1 \varphi_i\big(t, x_i(t)\big)\, \mathrm{d}t + \phi_i\big(x_i(1)\big) \right) + \lambda \|w\|^q_{L^q(0,1;\ell^1(X))},$$
$$\text{s.t.} \quad \dot{x}_i = f_\theta(x_i, t), \quad x_i(0) = x_{i,0}, \quad i = 1, \dots, n,$$

with $\lambda > 0$, $q \in [1, \infty)$.

- The ℓ^1 (Lasso) penalty *promotes* neuron sparsity.
- We use an L^q penalty in time.

Example (trajectory fitting)

For observed $y_i \in C([0,1]; X)$, take $x_{i,0} := y_i(0)$, $\varphi_i(t, x) := \|x - y_i(t)\|_X^2$ and $\phi_i(x) := \|x - y_i(1)\|_X^2$.

Neural ODEs: the training problem

Given n samples $x_{1,0}, \dots, x_{n,0} \in X$ and losses $\varphi_i \geq 0$ (running), $\phi_i \geq 0$ (terminal):

$$\inf_{\theta \in \mathcal{N}_{\leq m}} \mathcal{J}_{\mathrm{NN}}(\theta) := \sum_{i=1}^{n} \left(\int_0^1 \varphi_i\big(t, x_i(t)\big)\, \mathrm{d}t + \phi_i\big(x_i(1)\big) \right) + \lambda \|w\|^q_{L^q(0,1;\ell^1(X))},$$
$$\text{s.t.} \quad \dot{x}_i = f_\theta(x_i, t), \quad x_i(0) = x_{i,0}, \quad i = 1, \dots, n,$$

with $\lambda > 0$, $q \in [1, \infty)$.

Questions

1. How many neurons do we need?
2. What is the regularity in time of an optimal θ?
3. How does λ affect the optimal solution?

Finite width versus mean field

Network of width m

$$f_\theta(x,t) = \sum_{j=1}^m \sigma(x,\xi_j(t))\, w_j(t)$$

- control $\theta = (w(t), \xi(t)) \in X^m \times \Omega^m$;
- nonlinear in the control;
- penalty $\|w(t)\|_{\ell^1(X)}$.

Mean field (infinite width)

$$f_\mu(x,t) = \int_\Omega \sigma(x,\xi)\, \mathrm{d}\mu_t(\xi)$$

- control $\mu_t \in \mathcal{M}(\Omega; X)$;
- linear in the control;
- penalty $\|\mu_t\|_{TV}$, a convex norm.

We can identify networks with atomic measures in the mean-field formulation:

$$\mu_t^\theta := \sum_{j=1}^m w_j(t)\, \delta_{\xi_j(t)} \implies f_{\mu^\theta}(\cdot,t) = f_\theta(\cdot,t).$$

The regularization is consistent as well:

$$\|\mu_t^\theta\|_{TV} = \sum_{j=1}^m \|w_j(t)\| = \|w(t)\|_{\ell^1(X)}.$$

The mean-field training problem

$$\inf_{\mu \in \mathcal{U}_q} \mathcal{J}(\mu) := \sum_{i=1}^{n} \left(\int_0^1 \varphi_i(t, x_i(t))\, \mathrm{d}t + \phi_i(x_i(1)) \right) + \lambda \|\mu\|^q_{L^q(0,1;\mathrm{TV})},$$
$$\text{s.t.} \quad \dot{x}_i = f_\mu(x_i, t), \quad x_i(0) = x_{i,0}, \quad i = 1, \dots, n,$$

Admissible controls (Trautmann et al. 2018)

$$\mathcal{U}_q := \{\mu : [0,1] \to \mathcal{M}(\Omega; X) \text{ weak-}* \text{ measurable}, \ \|\mu\|_{L^q} < \infty\}$$

Under minimal regularity conditions on σ and the losses:

- **Well-posedness.** The state and the adjoint depend on μ *only* through $t \mapsto \|\mu_t\|_{TV} \in L^1(0,1)$.
- **Existence.** For $q > 1$, the problem attains its minimum on $\mathcal{U}_q$.

Adjoint equation

The adjoint carries the loss gradients backwards along the trajectories:

$$-\dot{p}_i(t) = \nabla_x \varphi_i\big(t, x_i(t)\big) + \big(D_x f_\mu(x_i(t), t)\big)^\top p_i(t), \qquad p_i(1) = \nabla_x \phi_i\big(x_i(1)\big).$$

It defines the discriminator, the level and the saturation set:

$$h_t(\xi) := \sum_{i=1}^{n} \sigma\big(x_i(t), \xi\big)\, p_i(t), \quad \Lambda(t) := \max_{\xi \in \Omega} \|h_t(\xi)\|, \quad \Omega^\star(t) := \big\{\xi : \|h_t(\xi)\| = \Lambda(t)\big\}.$$

Let $\mu^\star$ be a minimizer.

1. Support inclusion

For a.e. t,

$$\operatorname{supp} \mu_t^\star \subseteq \Omega^\star(t).$$

2. L^1 sparsity

If $q = 1$, then

$$\Lambda(t) \leq \lambda,$$

and $\mu_t^\star = 0$ if $\Lambda(t) < \lambda$.

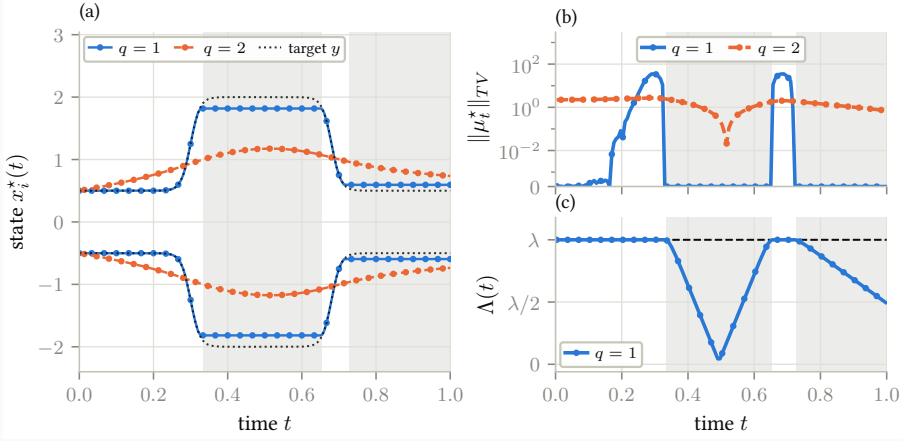
3. Hamiltonian

For a.e. t, $\mu_t^\star$ minimizes, over $\nu \in \mathcal{M}(\Omega; X)$,

$$\mathcal{H}^t(\nu) = \langle \nu, h_t \rangle + \lambda \|\nu\|_{TV}^q.$$

Sparsity in the solution: $q = 1$ versus $q = 2$

$$\min_{\mu \in \mathcal{U}_q} \sum_{i=1,2} \Big(\int_0^1 |x_i(t) - y_i(t)|^2 \, \mathrm{d}t + \tfrac{1}{2} |x_i(1) - y_i(1)|^2 \Big) + \lambda \|\mu\|^q_{L^q(0,1;\mathrm{TV})},$$



1. Neural ODEs: Representer theorem

The sparsity tool: an extreme-point theorem

Theorem (Bredies and Carioni 2020, Thm. 3.3)

Let E be a locally convex space, $\mathcal{R} : E \to [0, +\infty]$ a lower semicontinuous *seminorm* with trivial null space, let $\mathcal{A} : E \to H$ be linear and continuous into a finite-dimensional Hilbert space H, and let $F : H \to (-\infty, +\infty]$ be proper, convex, lower semicontinuous and coercive. Assume $\mathcal{A}(\operatorname{dom} \mathcal{R}) = H$ and that $\{\mathcal{R} \leq c\}$ is compact in E for every $c > 0$. Then

$$\min_{u \in E} \; F(\mathcal{A}u) + \mathcal{R}(u)$$

admits a minimizer of the form

$$\bar{u} = \sum_{k=1}^{p} \gamma_k \, u_k,$$

where $p \leq \dim H$ and the u_k are extreme points of the unit ball $\{\mathcal{R} \leq 1\}$.

Let $\mu^\star$ be the optimal control and $x^\star$ its state. At a fixed t, let $E := \mathcal{M}(\Omega^\star(t); X)$.

$$\mathcal{A}_t : E \to X^n \cong \mathbb{R}^{nd}, \quad (\mathcal{A}_t \nu)_i := \int_{\Omega^\star(t)} \sigma\big(x_i^\star(t), \xi\big) \, \mathrm{d}\nu(\xi),$$

$$\mathcal{R} : E \to [0, +\infty], \qquad \mathcal{R}(\nu) := \|\nu\|_{TV}.$$

$$F(b) := \begin{cases} 0, & b = f_{\mu^\star}(x^\star(t), t), \\ +\infty, & \text{otherwise.} \end{cases}$$

Neural ODEs: the representer theorem

$$\inf_{\mu \in \mathcal{U}_q} \mathcal{J}(\mu) := \sum_{i=1}^{n} \left(\int_0^1 \varphi_i(t, x_i(t))\, \mathrm{d}t + \phi_i(x_i(1)) \right) + \lambda \|\mu\|^q_{L^q(0,1;\mathrm{TV})},$$
$$\text{s.t.} \quad \dot{x}_i = f_\mu(x_i, t), \quad x_i(0) = x_{i,0}, \quad i = 1, \dots, n,$$

Theorem

Let $\mu^\star \in \mathcal{U}_q$ be a minimizer. Then there is minimizer $\bar{\mu} \in \mathcal{U}_q$ whose drift takes the form

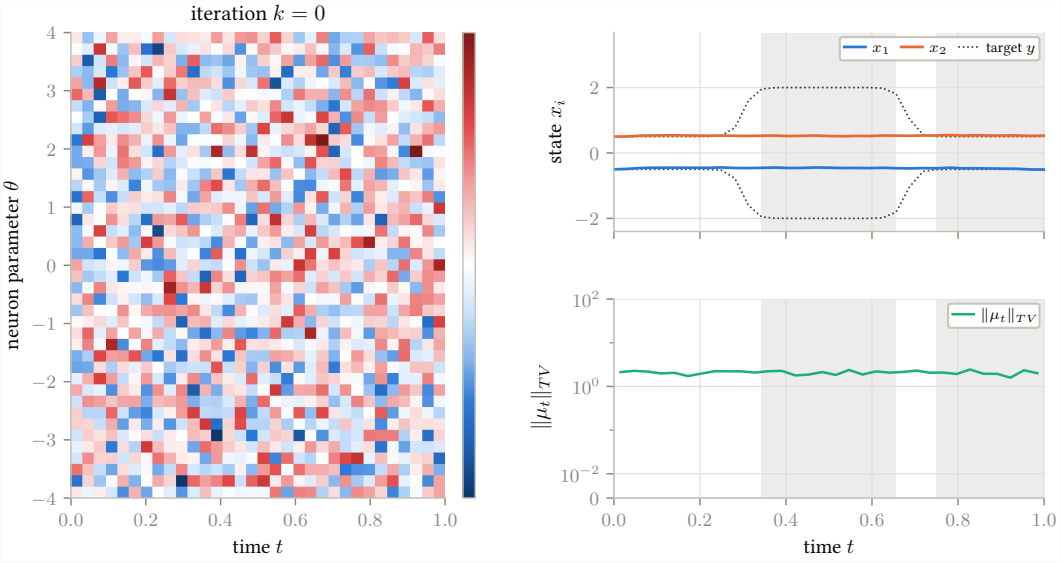
$$f_{\bar{\mu}}(x, t) = \sum_{j=1}^{nd} \sigma\big(x, \xi_j(t)\big)\, w_j(t),$$

with w_j, ξ_j measurable and $\xi_j(t) \in \Omega^\star(t)$ for a.e. t.

Corollary

$$\min_{\mu \in \mathcal{U}_q} \mathcal{J}(\mu) = \min_{\theta \in \mathcal{N}_\infty} \mathcal{J}_{\mathrm{NN}}(\theta) = \min_{\theta \in \mathcal{N}_{\leq nd}} \mathcal{J}_{\mathrm{NN}}(\theta).$$

Sparsity emerges during training ($\Omega = [-4, 4]$)



2. Regularity in time

What is the temporal structure of the sparse network?

Proposition (branches of active neurons)

Set $H_t := \frac{1}{2}\|h_t\|^2$. If at t_0 every $\zeta \in \Omega^\star(t_0)$ lies in $\operatorname{int}\Omega$ with nonsingular Hessian $\nabla^2_\xi H_{t_0}(\zeta)$, then $\Omega^\star(t_0)$ is finite, say M points, and near t_0

$$\mu^\star_t = \sum_{j=1}^{M} w_j(t)\,\delta_{\xi_j(t)}, \qquad \xi_j \in W^{1,q}, \qquad w_j \in L^q.$$

If moreover the data are C^{k+1} and the bracket matrix $\mathcal{B}(t_0)$ is nonsingular[1], then, near t_0,

$$\mathrm{x}^\star, \mathrm{p}^\star \in C^{k+1}, \qquad w_j,\ \xi_j \in C^k.$$

[1] $\mathcal{B}_{jl}(t) := \sum_i p_i^\star(t)^\top [X_{t,\xi_j}, X_{t,\xi_l}](x_i^\star(t))$, built from the directed neuron fields $X_{t,\xi} := -\sigma(\cdot,\xi)\,h_t(\xi)/\Lambda(t)$.

3. Strong regularization

Strong regularization drives the network to rest

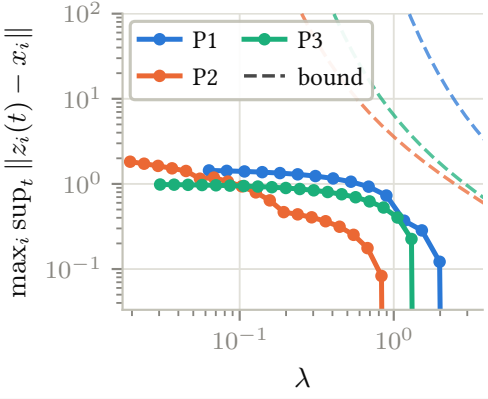
Proposition (decay and extinction)

For every $q \geq 1$ and $\lambda > 0$,

$$\max_i \sup_t \left\| x_i^{\star,\lambda}(t) - x_{i,0} \right\| \leq C \left(\frac{\mathcal{J}(0)}{\lambda} \right)^{1/q}.$$

Moreover, there is a threshold $\Lambda_0 < \infty$, *explicit and independent of* λ, such that

- for $q = 1$: $\mu^{\star,\lambda} \equiv 0$ and $x^{\star,\lambda} \equiv x_0$ for every $\lambda > \Lambda_0$.



4. Extra: Optimal Transport

Optimal transport

Consider the following optimization problem for n particles, with a prescribed target ρ_1:

$$\inf_{\mu \in \mathcal{U}_q} \sum_{i=1}^{n} \int_0^1 L\big(x_i(t),\, \dot{x}_i(t),\, t\big)\, \mathrm{d}t \;+\; \lambda \, \|\mu\|^q_{L^q(0,1;\mathrm{TV})},$$
$$\text{s.t.} \;\; \dot{x}_i = f_\mu(x_i, t), \quad x_i(0) = x_{i,0}, \qquad \textstyle\sum_i \delta_{x_i(1)} = \rho_1.$$

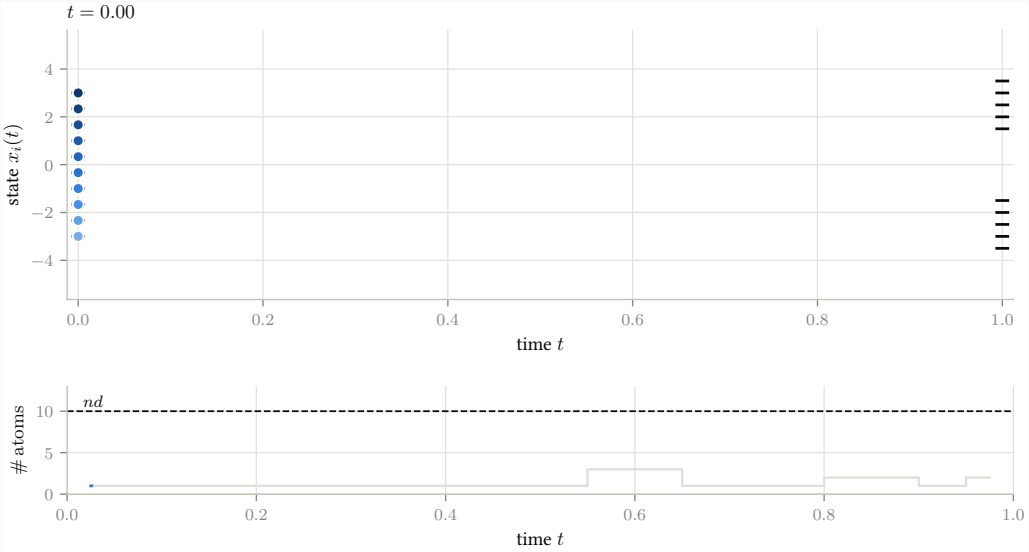
- The running cost sees the velocity $\dot{x}_i(t)$.
- The terminal cost is a hard constraint: neither smooth nor convex.
- The regularizer penalizes the control measure μ.

Theorem

Let $\mu^\star$ be a minimizer. Then there is a minimizer $\bar{\mu}$ whose drift takes the form

$$f_{\bar{\mu}}(x, t) = \sum_{j=1}^{nd} \sigma\big(x, \xi_j(t)\big)\, w_j(t).$$

Experiment ($d = 1$)



5. Conclusions and Outlook

Contributions

1. **Representer theorem.** An optimal drift is realized by *nd* neurons at a.e. time.
2. **Regularity in time.** Nondegenerate saturation points trace C^k branches.
3. **Turnpike.** At $q = 1$, the bound $\Lambda \leq \lambda$ creates exact rest phases.
4. **Optimal transport.** The representer theorem extends to a hard terminal constraint.

References i

Arens, R. F. and J. L. Kelley (1947). "Characterization of the Space of Continuous Functions over a Compact Hausdorff Space". In: *Transactions of the American Mathematical Society* 62.3, pp. 499–508. ISSN: 1088-6850, 0002-9947. DOI: 10.1090/S0002-9947-1947-0022999-0.

Bartolucci, Francesca et al. (2024). *Neural Reproducing Kernel Banach Spaces and Representer Theorems for Deep Networks*. Version 2. DOI: 10.48550/ARXIV.2403.08750. Pre-published.

Beck, Amir and Marc Teboulle (Jan. 1, 2009). "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems". In: *SIAM Journal on Imaging Sciences* 2.1, pp. 183–202. ISSN: 1936-4954. DOI: 10.1137/080716542.

Bonnet, Benoît et al. (Apr. 8, 2022). *A Measure Theoretical Approach to the Mean-field Maximum Principle for Training NeurODEs*. DOI: 10.48550/arXiv.2107.08707. arXiv: 2107.08707 [math.OC]. Pre-published.

Bredies, Kristian and Marcello Carioni (Feb. 2020). "Sparsity of Solutions for Variational Inverse Problems with Finite-Dimensional Data". In: *Calculus of Variations and Partial Differential Equations* 59.1, p. 14. ISSN: 0944-2669, 1432-0835. DOI: 10.1007/s00526-019-1658-1. arXiv: 1809.05045 [math].

Chen, Ricky T. Q. et al. (2018). *Neural Ordinary Differential Equations*. Version 5. DOI: 10.48550/ARXIV.1806.07366. Pre-published.

Chizat, Lenaic and Francis Bach (Oct. 29, 2018). *On the Global Convergence of Gradient Descent for Over-parameterized Models Using Optimal Transport*. DOI: 10.48550/arXiv.1805.09545. arXiv: 1805.09545 [math]. Pre-published.

Esteve-Yagüe, Carlos and Borjan Geshkovski (Sept. 8, 2022). *Sparsity in Long-Time Control of Neural ODEs*. DOI: 10.48550/arXiv.2102.13566. arXiv: 2102.13566 [cs.LG]. Pre-published.

References ii

Goh, B. S. (Nov. 1966). "Necessary Conditions for Singular Extremals Involving Multiple Control Variables". In: *SIAM Journal on Control* 4.4, pp. 716–731. ISSN: 0036-1402. DOI: 10.1137/0304052.

Gugat, Martin et al. (2021). "The Finite-Time Turnpike Phenomenon for Optimal Control Problems: Stabilization by Non-smooth Tracking Terms". In: *Stabilization of Distributed Parameter Systems: Design Methods and Applications*. Ed. by Grigory Sklyar and Alexander Zuyev. Cham: Springer International Publishing, pp. 17–41. ISBN: 978-3-030-61742-4. DOI: 10.1007/978-3-030-61742-4_2.

Jabir, Jean-François et al. (Mar. 16, 2021). *Mean-Field Neural ODEs via Relaxed Optimal Control*. DOI: 10.48550/arXiv.1912.05475. arXiv: 1912.05475 [math.PR]. Pre-published.

Trautmann, Philip et al. (2018). "Finite Element Error Analysis for Measure-Valued Optimal Control Problems Governed by a 1D Wave Equation with Variable Coefficients". In: *Mathematical Control & Related Fields* 8.2, pp. 411–449. ISSN: 2156-8499. DOI: 10.3934/mcrf.2018017. arXiv: 1702.00362 [math.NA].

Zaslavski, Alexander J. (2014). *Turnpike Phenomenon and Infinite Horizon Optimal Control*. Vol. 99. Springer Optimization and Its Applications. Cham: Springer International Publishing. ISBN: 978-3-319-08827-3 978-3-319-08828-0. DOI: 10.1007/978-3-319-08828-0.