Kinetic Description of Neural Differential Equations

Master Thesis

Flora Valerio VR481426

October 9, 2023

University of Verona

UNIVERSITÀ di VERONA

Presentation

- **Goal**: analyze the connection between neural networks and partial differtial equations
- **General approach**: approximate PDE solutions through kinetic propagation of data
- **Why** Neural Differential Equations

	Advantage
Differential Equations	Unparalleled modeling capacity
Neural Networks	Solid theoretical foundation

.

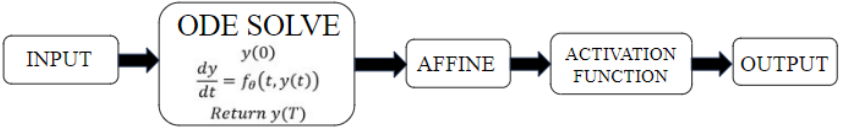
Neural Differential Equations

Formal explanation:

$$y(0) = y_0 \qquad \frac{dy}{dt}(t) = f_\theta(t, y(t))$$

where

- θ parameters;
- $f_\theta : \mathbb{R} \times \mathbb{R}^{d_1 \times d_2 \cdots \times d_k} \to \mathbb{R}^{d_1 \times d_2 \cdots \times d_k}$ is any standard neural architecture
- $y : [0, T] \to \mathbb{R}^{d_1 \times d_2 \cdots \times d_k}$ is the solution



Link between NDEs and Residual Neural Networks

ResNet

$$y^{(k+1)} = y^{(k)} + f_\theta(k, y^{(k)})$$

where $f_\theta(k, \cdot)$ is the $k-th$ residual block.

NDE

$$\frac{dy}{dt}(t) = f_\theta(t, y(t)) \qquad (1)$$

- Discretizing (1) via explicit Euler method at times t_k uniformly separated by Δt:

$$\frac{y(t_{k+1}) - y(t_k)}{\Delta t} \approx \frac{dy}{dt}(t_k) = f_\theta(t_k, y(t_k))$$

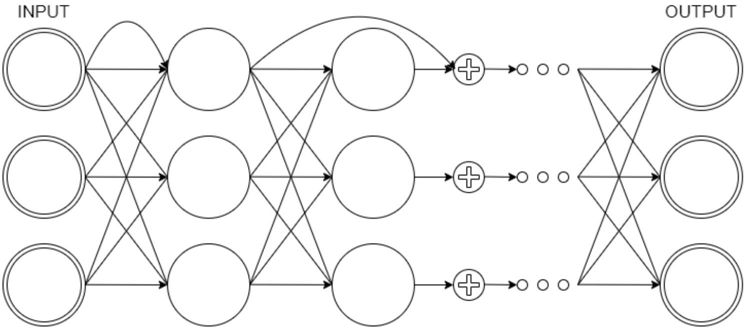
$$y(t_{k+1}) = y(t_k) + \Delta t f_\theta(t_k, y(t_k))$$

- ResNets are optimal for addressing optimal control problems by mitigating the challenges associated with training deep neural networks.

Simplified Resisual Neural Network

Assumption (SimResNet)

The number of neurons in each layer is fixed and determined by the dimension of the input data.



Simplified Resisual Neural Network

Properties

1. Stable gradient flow
2. Control over network complexity
3. Analytical tractability:
4. Relationship to traditional differential equations

- Well-suited for deriving neural differential equations and studying the dynamics of complex systems in a more analytically tractable and interpretable manner.
- Satisfy the universal approximating theorem

Mean-field formulation of SimResNet

- SimResNet defines precise microscopic dynamics:

$$\begin{cases} \mathbf{x}_i^{(\ell+1)} & = \mathbf{A}^{(\ell)}\mathbf{x}_i^{(\ell)} + \Delta t \sigma \left(\boldsymbol{\omega}^{(\ell)} \mathbf{x}_i^{(\ell)} + \mathbf{b}^{(\ell)} \right), \quad \ell = 0, \ldots, L \\ \mathbf{x}_i^{(0)} & = \mathbf{x}_i^0 \end{cases} \tag{2}$$

- L number of layers
- $\mathbf{x}_i^{(\ell)} \in \mathbb{R}^{\bar{d}}$
- $\mathbf{A}^{(\ell)} \in \mathbb{R}^{\bar{d}\times\bar{d}}$ is a deterministic matrix
- $\sigma(\cdot) : \mathbb{R} \to \mathbb{R}$ activation function
- $\boldsymbol{\omega}^{(\ell)} \in \mathbb{R}^{\bar{d}\times\bar{d}}$ weights
- $\mathbf{b}(\ell) \in \mathbb{R}^{\bar{d}}$ bias

Continuous SimResNet

- Considering $\Delta t \to 0^+$ and $L \to \infty$, we obtain the following *continuous structure of SimResNet*:

$$\begin{cases} \frac{d}{dt}\mathbf{x}_i(t) & = \sigma\left(\boldsymbol{\omega}(t)\mathbf{x}_i(t) + \mathbf{b}(t)\right), \quad t > 0 \\ \mathbf{x}_i^{(0)} & = \mathbf{x}_i^0 \end{cases} \tag{3}$$

- *neural differential equation* associated to the SimResNet
- *Picard-Lindelőf Theorem* ensures the existence and uniqueness of solution
- **Goal**: pass to a statistical interpretation of (3)

Mean-field limit

- Mean-field limit $(N \to \infty) \implies$ **Hyperbolic Vlasov-type PDE**

$$\begin{cases} \partial_t f(t,\mathbf{x}) + \nabla_{\mathbf{x}} \cdot \big(\sigma(\boldsymbol{\omega}(t)\mathbf{x} + \mathbf{b}(t))f(t,\mathbf{x})\big) = 0, \quad t > 0 \\ f(0,\mathbf{x}) = f_0(\mathbf{x}), \qquad \int_{\mathbb{R}^d} f_0(\mathbf{x})d\mathbf{x} = 1 \end{cases} \tag{4}$$

- $f(t,\mathbf{x}) : \mathbb{R}_0^+ \times \mathbb{R}^d \to \mathbb{R}_0^+$ is the **probability distribution function**
- statistical interpretation of the neural network

1. Well-posedness of (4)
2. existence and uniqueness of weak solution
3. continuous dependence on the initial condition and on the paramaters
4. convergence of the continuous SimResNet (3) to (4) as $N \to \infty$.

Formal derivation

- Empirical distribution measure: $f^N(x,t) := \frac{1}{N}\sum_{i=1}^{N} \delta(x, x_i(t))$
- By Liouville's theorem, for $\phi \in C_0^1(\mathbb{R})$ *test function*:

$$\int_{\mathbb{R}^d} \phi(x(t)) f^N(x,t) dx = \frac{1}{N}\sum_{i=1}^{N} \phi(x_i(t))$$

- Applying derivative over time and integrating by parts the extension of the right term:

$$\partial_t f^N(x,t) + \nabla_x \cdot (\sigma(w(t)x(t) + b(t)) \cdot f^N(x,t)) = 0$$

Rigourous derivation - previous tools

Definition (1-Wasserstein distance)

Let μ and ν two probability measures on $\mathbb{R}^d$. Then the 1-Wasserstein distance is defined by

$$W(\mu,\nu) := \inf_{\pi \in \mathcal{P}^*(\mu,\nu)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\xi - \eta| d\pi(\xi,\eta)$$

where $\mathcal{P}^*$ is the space of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ such that the marginals are μ and ν, i.e.

$$\int_{\mathbb{R}^d} d\pi(\cdot,\eta) = d\mu(\cdot), \qquad \int_{\mathbb{R}^d} d\pi(\xi,\cdot) = d\nu(\cdot).$$

Rigourous derivation - previous tools

Definition (Weak solution)

Let $T > 0$ be fixed. Assume that $F_0 \in \mathcal{P}_1(\mathbb{R}^{d+1})$. We say that the time dependent measure $F_t \in C([0,T];\mathcal{P}_1(\mathbb{R}^{d+1}))$ is a weak solution to the mean-field equation

$$\partial_t F_t + \nabla_x \cdot \big(\sigma(\omega(\tau)x + b(\tau))F_t\big) + \partial_\tau F_t = 0 \tag{5}$$

with initial condition F_0 if for all $\phi = \phi(x,\tau) \in C_0^\infty(\mathbb{R}^{d+1})$ and for all $t \in [0,T]$ the following equality holds:

$$\begin{aligned}\int_{\mathbb{R}^{d+1}} \phi(x,\tau)dF_t(x,\tau) = &\int_{\mathbb{R}^{d+1}} \phi(x,\tau)dF_0(x,\tau) + \\ &+ \int_0^t \int_{\mathbb{R}^{d+1}} \nabla_{(x,\tau)}\phi(x,\tau) \cdot G(x,\tau)dF_s(x,\tau)ds\end{aligned}$$

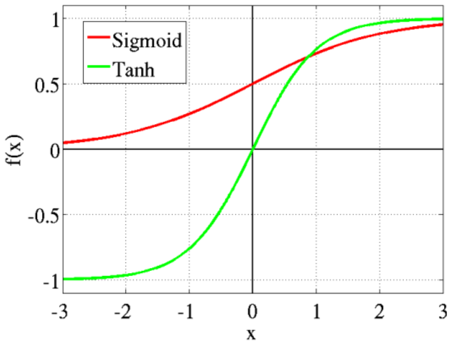
Rigourous derivation - assumptions

- *existence* and *uniqueness* of weak solution F_t of the mean-field equation (4) is obtained under the following assumptions

(A1) $\sigma \in C^{0,1}(\mathbb{R}^d), \quad \omega, b \in C^{0,1}(\mathbb{R});$

(A2) $|\sigma(x)| \leq C_0, \quad \forall x \in \mathbb{R}^d$

- To prove the mean-field limit, it is necessary to go through an auxiliary system that satisfies the following proposition



Convergence and continuous dependence on initial data

Proposition

Let $F_0 \in \mathcal{P}_1(\mathbb{R}^{d+1})$ be given and let $T > 0$.
Then, under the assumption (A1) and (A2), there exists a unique solution $F_t \in C([0, T]; \mathcal{P}_1(\mathbb{R}^{d+1}))$ of the mean-field equation (5)*.*
In particular $F_t = \Phi_t \# F_0$ and F_t is continuously dependent on the initial data F_0 with respect to the 1-Wasserstain distance. Furthermore, the solution of the ausiliary dynamical system converges to F_t in Wasserstain for $N \to \infty$.

Continuos dependence on parameters

Proposition

Let $F_0 \in \mathcal{P}_1(\mathbb{R}^{d+1})$ be given and let $T > 0$. Then, under the assumptions (A1) and (A2), the unique solution $F_t \in C([0,T];\mathcal{P}_1(\mathbb{R}^{d+1}))$ of the mean-field equation (5) is continuosly dependent on (ω, b).

Implications:

- Robustness
- Sensitivity Analysis
- Optimization

Steady-state solution analysis

Proposition

Let $f : \mathbb{R}_o^+ \times \mathbb{R}^d \to \mathbb{R}_o^+$ *be the compactly supported weak solution of the mean-field equation* (4). *Assume that the activation function* $\sigma : \mathbb{R} \to \mathbb{R}$ *has disjunct zeros* z_k, $k = 1, \ldots, K$ *for some* $K > 0$. *Let* $b^\infty = lim_{t\to\infty} b(t)$ *and* $\omega^\infty = lim_{t\to\infty}\omega(t)$ *exists and finite.*

Then

$$f^\infty(x) = \sum_{i=1}^{n^d} \rho_i \delta(x - y_i)$$

is a steady state solution of (4) *in sense of distributions provided that* $\mathbf{y}_i$ *for* $i = 1, \ldots, n^d$ *are disjunct solutions of* $\omega^\infty y_i + b^\infty = z_k$ *for some* k *and* $\rho_i \in [0, 1]$ *such that* $\sum_{i=1}^{n^d} \rho_i = 1$.

Moment Analysis

Moment Analysis - Criteria

Definition (k-th moment & variance)

Given $k \geq 0$, the k-th moment of the probability distribution $f(t,x)$ is defined as

$$m_k(t) := \int_{\mathbb{R}} x^k f(t,x) dx$$

The variance of the probability distribution $f(t,x)$ is defined as

$$\mathbb{V}(t) = m_2(t) - (m_1(t))^2$$

(i) **local energy bound** if

$$m_2(0) > m_2(t),$$

holds at a fixed time t;

(ii) **energy decay** if

$$m_2(t_1) > m_2(t_2),$$

holds for any $t_1 < t_2$;

(ii) **concentration** or **clustering** if

$$\lim_{t \to \infty} \mathbb{V}(t) = 0$$

Moment Analysis

$$\frac{d}{dt} m_k(t) = k\big(\omega(t) m_k(t) + b(t) m_{k-1}(t)\big), \qquad m_k(0) = m_k^0 \tag{6}$$

$$\implies m_k(t) = e^{\Phi_k(t)}\left(m_k(0) + k\int_0^t e^{-\Phi_k(s)} b(s) m_{k-1}(s) ds\right), \qquad \Phi_k(t) := k\int_0^t \omega(s) ds \tag{7}$$

Moment Analysis - Characterization

$$b(t) \equiv 0$$

(a) **local energy bound** if $\Phi_1(t) < 0$ at a fixed time t;

(a) **energy decay** if and only if $\omega(t) < 0$ for all time $t > 0$;

(a) **clustering** if and only if $\lim_{t\to\infty} \Phi_1(t) = -\infty$. In particular, the steady state is distributed as a Dirac delta

$$b(t) := -\omega(t)m_1(t)$$

(a) local energy bound if $\Phi_2(t) < 0$ at a fixed time t;

(b) clustering phenomenon if $\omega(t) < 0$ holds for all $t \geq 0$. In particular, the steady state is distributed as a Dirac delta centered at $x = m_1(0)$.

Forward re-training algorithm

Optimization algorithm

- Proposition (3) establishes the existence of steady solutions for the mean-field equation, which are distinguished by their **constant weights** and **biases**.

$$\Downarrow$$

Forward re-training algorithm

$$\Downarrow$$

Optimal control problem

- based on the motion of each single particle

3 OCP

1. 1D Classification
2. 2D Regression
3. Multivariate Regression

OCP - 1D Classification

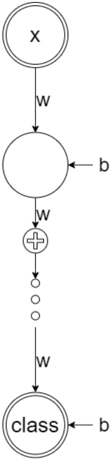
- **Goal**: classify data between car and truck
- $y_j \sim \xi(x), \quad j = 1.\dots, N$ and N_T the last time step

$$\arg\min_{\omega,b} \frac{1}{N}\sum_{j=1}^{N}\frac{1}{2}\|x_j^{(N_T)} - y_j\|_2^2$$

$$s.t. \quad x_j^{(n+1)} = x_j^{(n)} + h\cdot\sigma(\omega^{(n)}x_j^{(n)} + b^{(n)})$$

$$x_j^{(o)} \in \mathbb{R}^d$$

- through the Lagrangian and implementing optimality conditions we get a forward retraining algorithm.



Forward Retraining Algorithm

Initialize $\omega^{(0)}$ and $b^{(0)}$ randomly

for *iter* $\leftarrow 1$ **to** *n_iter* **do**

  Propagate training data with the current parameters

  Compute the loss

  Compute the new updates following

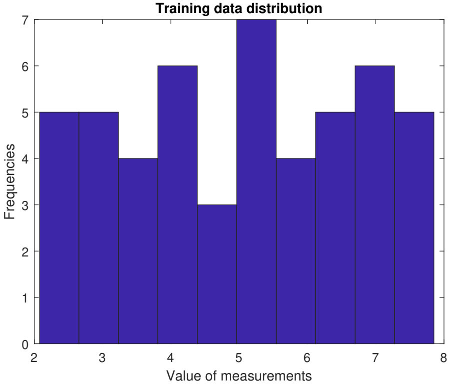
$$\lambda_j^{(n)} = \lambda_j^{(n+1)} + \lambda_j^{(n+1)} \cdot h \cdot \sigma'(\omega^{(n)} x_j^{(n)} + b^{(n)}) \omega^{(n)}$$

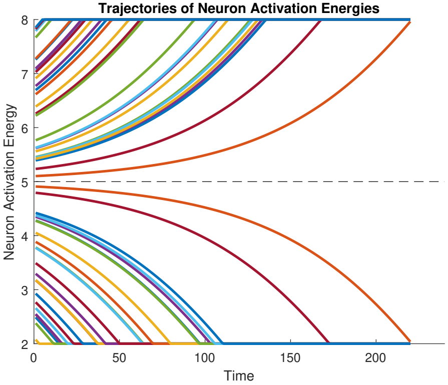
$$\omega^{(n+1)} = \omega^{(n)} - \gamma \frac{h}{N} \sum_{j=1}^{N} \lambda_j^{(n+1)} \cdot h \cdot \sigma'(\omega^{(n)} x_j^{(n)} + b^{(n)}) \cdot x_j^{(n)}$$

$$b^{(n+1)} = b^{(n)} - \gamma \frac{h}{N} \sum_{j=1}^{N} \lambda_j^{(n+1)} \cdot h \cdot \sigma'(\omega^{(n)} x_j^{(n)} + b^{(n)})$$

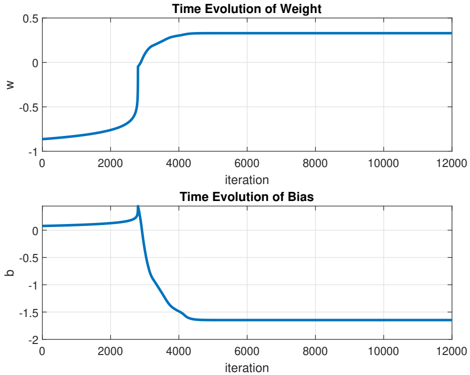
end

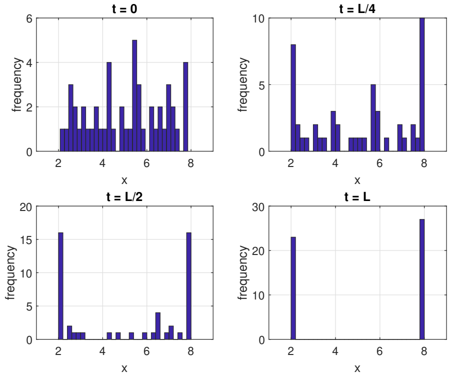
1D Classification - Numerical





1D Classification - Numerical





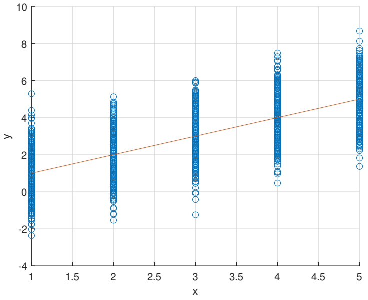
OCP - Regression 2D

- **Goal**: learn the slope and intercept of the regression line
- $m \rightarrow$ slope, $q \rightarrow$ *intercept* $\implies y = mx + q$

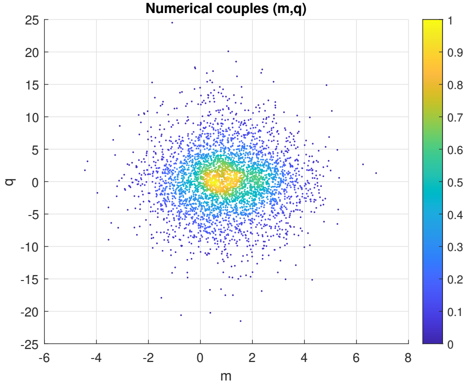
$$\underset{\omega,b}{\arg\min} \frac{1}{n*} \sum_{i=1}^{n^*} \frac{1}{2} ||y_i - m^{(N_T)} x_i - q^{(N_T)}||^2$$

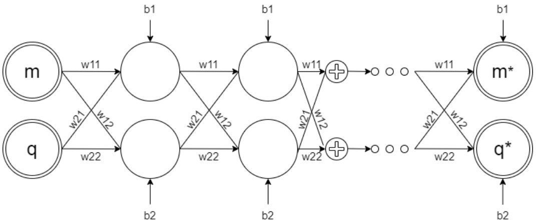
$$s.t. \quad m^{(n+1)} = m^{(n)} + h \cdot \sigma(\omega_{11}^{(n)} m^{(n)} + \omega_{21}^{(n)} q^{(n)} + b_1^{(n)})$$

$$q^{(n+1)} = q^{(n)} + h \cdot \sigma(\omega_{12}^{(n)} m^{(n)} + \omega_{22}^{(n)} q^{(n)} + b_2^{(n)})$$

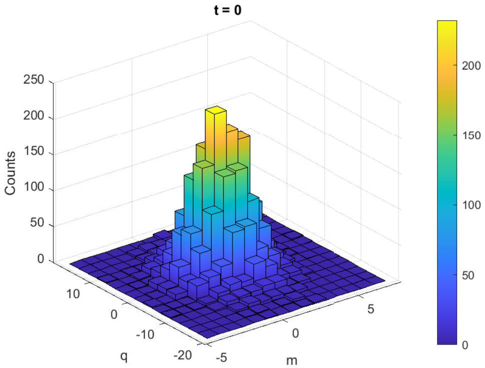


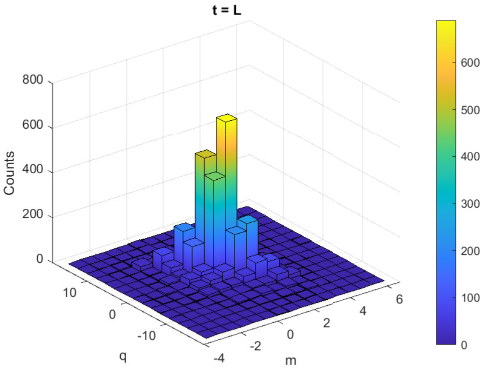
2D Regression - Numerical



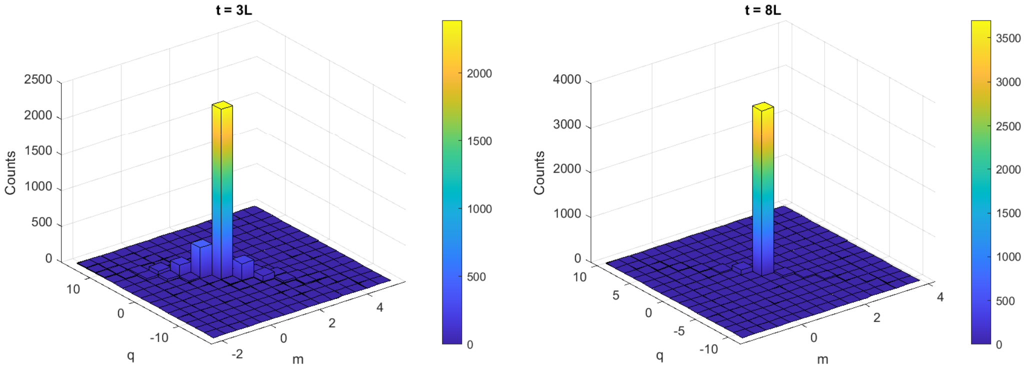


2D Regression - Numerical





2D Regression - Numerical



OCP - Multivariate Regression

- **Goal**: fit hyperplanes $A^T x = c$ in dimension d.

$$A = [a_1, a_2, \dots a_d, a_{d+1}] \in \mathbb{R}^{d+1}, \qquad x = [x_1, x_2, \dots x_d, -1] \in \mathbb{R}^{d+1}$$

where a_{d+1} corresponds to the coefficient c.

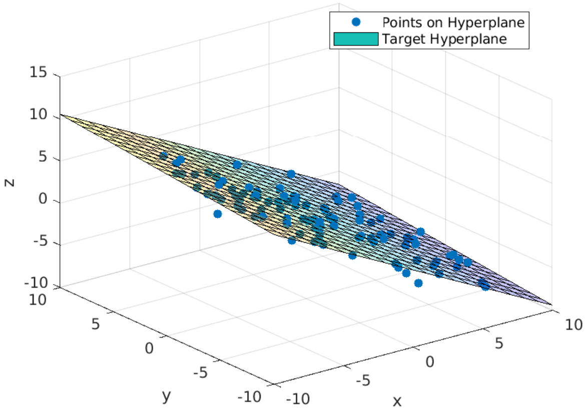
$$\arg\min_{\omega,b} \frac{1}{n^*} \sum_{k=1}^{n^*} \frac{1}{2} \|A^{(N_T)} x_k\|_2^2$$

$$s.t \quad A^{(n+1)} = A^{(n)} + h \cdot \sigma(W^{(n)} * A^{(n)} + b^{(n)})$$

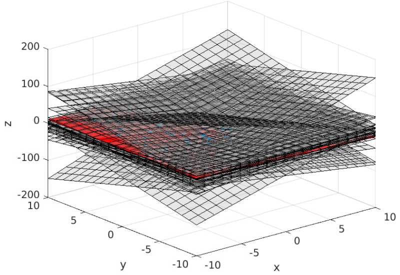
$$A^o \in \mathbb{R}^{d+1}$$

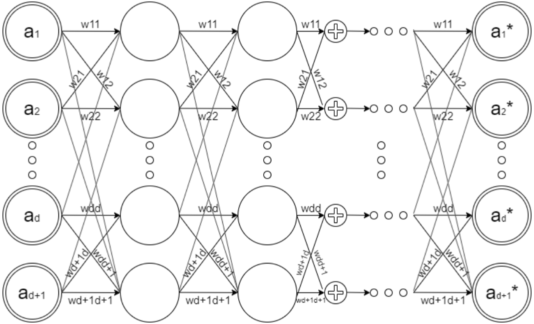
- $W \in \mathbb{R}^{d+1 \times d+1}$ weights' matrix
- $b \in \mathbb{R}^{d+1}$ is the bias vector.

Multivariate Regression - Numerical

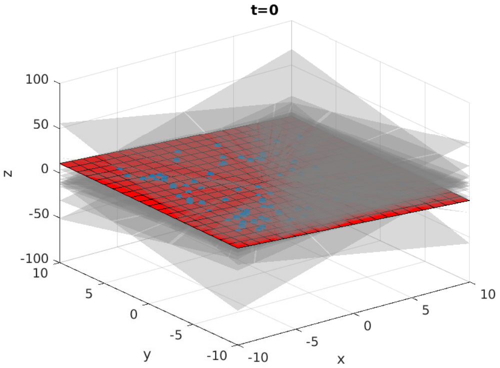


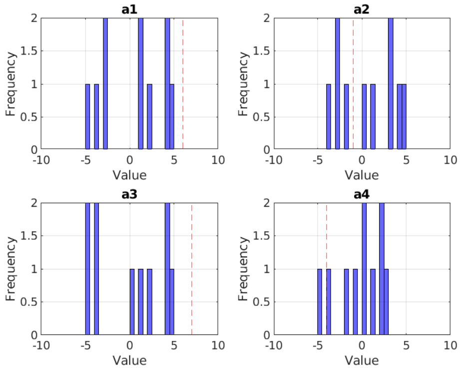
Multivariate Regression - Numerical



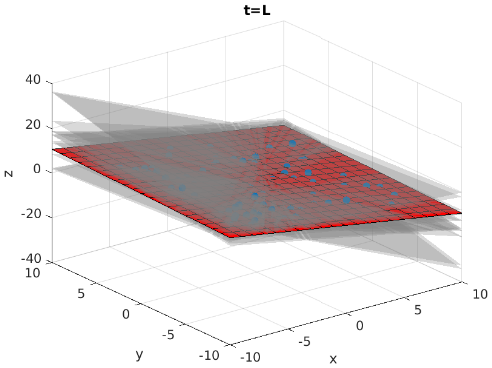


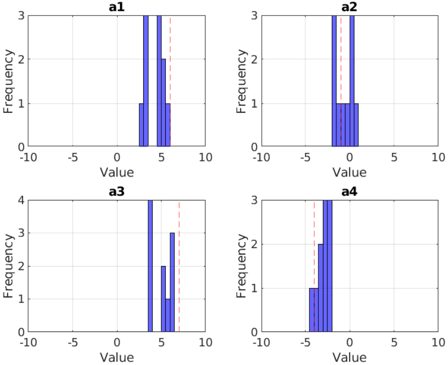
Multivariate Regression - Numerical



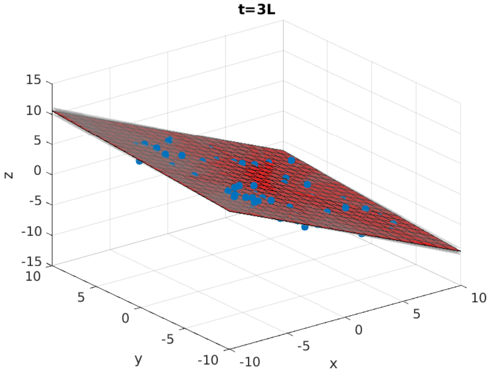


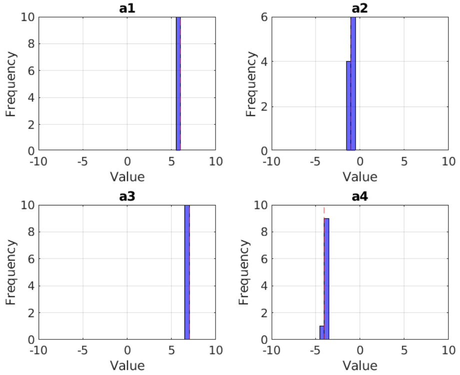
Multivariate Regression - Numerical



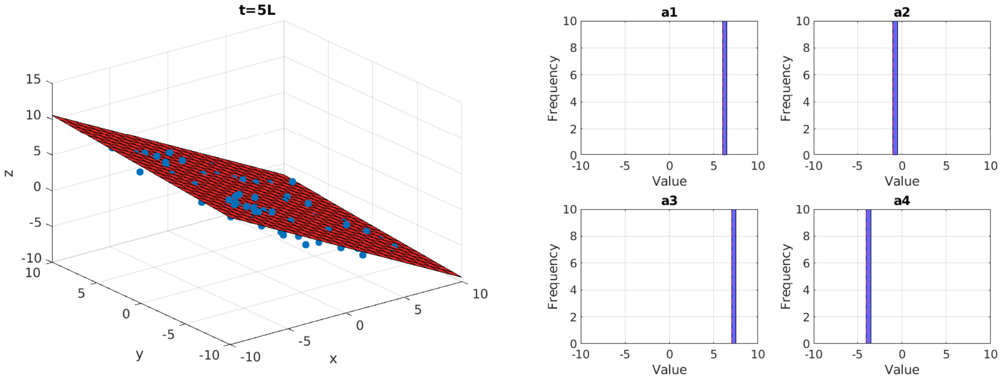


Multivariate Regression - Numerical





Multivariate Regression - Numerical



Boltzmann-type formulation of SimResNet

Boltzmann-type formulation of SimResNet

- Goal: reach other kind of stationary solution
- Add noise to the dynamic + grazing limit ($\epsilon \to 0$)

$$\Downarrow$$

Fokker-Planck equation

- $\eta \sim \mathcal{N}(0, \nu^2)$, $K(x)$ diffusion function, ϵ interactions's weight

New dynamic

$$x^* = x + \epsilon\sigma(\omega(t)x + b(t)) + \sqrt{\epsilon}K(x)\eta$$

Statistical description

$$\frac{d}{dt}\int_{\mathbb{R}} \Phi(x) f(t,x)dx = \mathbb{E}\left[\frac{1}{\epsilon}\int_{\mathbb{R}} \Big(\Phi(x^*) - \Phi(x)\Big) f(t,x)dx\right]$$

where

$$\Phi(x^*) \approx \Phi(x) + (x^* - x)\Phi'(x) + \frac{(x^* - x)^2}{2}\Phi''(x) + \mathcal{R}(x)$$

Fokker-Planck equation

- Extending the right term and applying grazing limit ($\epsilon \to 0$):

$$\partial_t f(t,x) + \partial_x[\mathcal{B} f(t,x) - \mathcal{D}\partial_x f(t,x)] = 0$$

where

$$\mathcal{B} = \sigma(\omega(t)x + b(t)) - \frac{\nu^2}{2}\partial_x K^2(x)$$

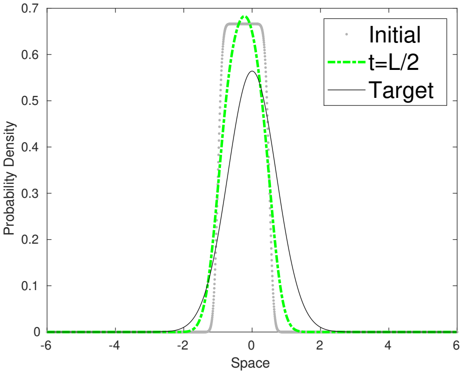
$$\mathcal{D} = \frac{\nu^2}{2} K^2(x)$$

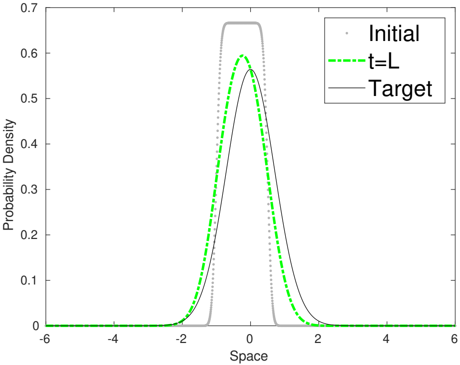
Steady-state solution

$$f^\infty(x) = \frac{C}{K^2(x)} exp\left(\int \frac{2\sigma(\omega^\infty x + b^\infty)}{\nu^2 K^2(x)} dx\right)$$

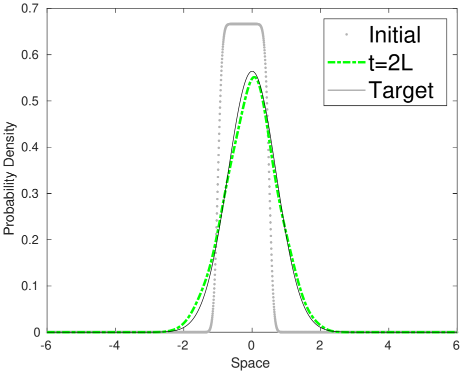
Fokker-Planck - Numerical

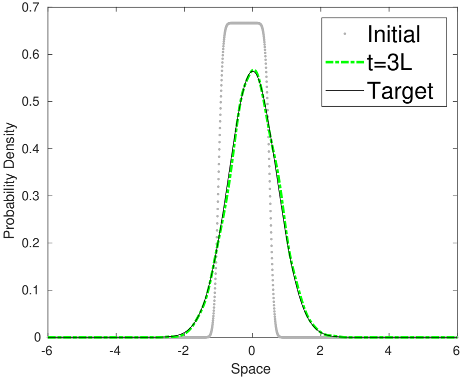
$$\omega = -1, \quad b = 0, \quad K(x) = 1 \implies f^{\infty} = \frac{\sqrt{\nu^{-2}}}{\sqrt{\pi}} exp\left(-x^2 \cdot \nu^{-2}\right)$$





Fokker-Planck - Numerical





References I

- Michael Herty, Torsten Trimborn. *Mean-field and kinetic descriptions of neural differential equations*. Foundation of Data Science Vol. 4 No. 2, June 2022.
- Patrick Kidger. *On Neural Differential Equations*. Thesis for the degree of Doctor of Philosophy, 4 February 2022.
- Francois Golse. *On the Dynamics of Large Particle Systems in the Mean Field Limit* . Centre de Mathematiques Laurent Schwartz, 23 January 2013.
- Francois Golse. *On the dynamics of large particle systems in the Mean Field Limit.* In *Macroscopic and large scale phenomena: coarse graining, mean field limits and ergodicity*, pages 1-144. Springer, 2016.
- Michael Herty, Anna Thunen, Torsten Trimborn, Giuseppe Visconti. *Continuous limits of residual neural networks in case of large input data*. 11 May 2022.

References II

- Paulo Tabuada, Bahman Gharesifard. *Universal Approximation Power of Deep Residual Neural Networks via nonlinear control theory*. 16 Dec 2020.
- Patrick Kidger, Terry Lyons. *Universal approximation with deep narrow networks*, In *Conference on Learning Theory*, 2020.
- Hongzhou Lin, Stefanie Jegelka. *Resent with one-neuron hidden layers is a universal approximator*, NIPS'18, Red Hook, NY, USA, Curran Associates Inc, pages 6172-6181. 2018.
- Yulong Lu, Jianfeng Lu. *A universal approximation theorem of deep neural networks for expressing probability distribution*Advances in Neural Information Processing Systems, Curran Associates, Inc., 33. pages 3094-3105 2020.

References III

- L. Pareschi, G.Toscani. *Interacting Multiagent Systems. Kinetic equations and Monte Carlo methods*, Oxford University Press, 2013.
- B. Bonnet, C. Cipriani, M. Fornasier, H. Huang. *A Measure Theoretical Approach to the Mean-field Maximum Principle for Training NeurODEs*, 2022.

Kinetic Description of Neural Differential Equations

Master Thesis

Flora Valerio VR481426

October 9, 2023

University of Verona

Appendix - Moment Analysis - Criteria

(i) **local energy bound** if

$$m_2(0) > m_2(t),$$

holds at a fixed time t;

(ii) **energy decay** if

$$m_2(t_1) > m_2(t_2),$$

holds for any $t_1 < t_2$;

(iii) **local aggregation** if

$$\mathbb{V}(0) > \mathbb{V}(t),$$

(iv) **aggregation** if

$$\mathbb{V}(t_1) > \mathbb{V}(t_2),$$

holds for any $t_1 < t_2$;

(v) **concentration** or **clustering** if

$$\lim_{t\to\infty} \mathbb{V}(t) = 0$$

Appendix - Formal derivation

By *Liouville's theorem*:

- *test function* $\phi \in C_0^1(\mathbb{R})$
- $$\int_{\mathbb{R}^d} \phi(x(t)) f^N(x,t) dx = \frac{1}{N} \sum_{i=1}^{N} \phi(x_i(t))$$
- Applying the time derivative and expanding the right term:

Appendix - Formal derivation

$$
\begin{aligned}
\frac{d}{dt}\int_{\mathbb{R}^d}\phi(x(t))f^N(x,t)dx &= \frac{d}{dt}\left(\frac{1}{N}\sum_{i=1}^{N}\phi(x_i(t))\right)\\
&= \frac{1}{N}\sum_{i=1}^{N}\nabla_x\phi(x_i(t))\cdot\dot{x}_i(t)\\
&= \frac{1}{N}\sum_{i=1}^{N}\nabla_x\phi(x_i(t))\cdot\sigma\big(\omega(t)x_i(t)+b(t)\big)\\
&= \int_{\mathbb{R}^d}\nabla_x\phi(x(t))\cdot\sigma\big(\omega(t)x(t)+b(t)\big)\cdot f^N(x,t)dx
\end{aligned}
$$

Appendix - Formal derivation

$$\int_{\mathbb{R}^d} \nabla_x \phi(x(t)) \cdot \sigma\big(\omega(t)x(t) + b(t)\big) \cdot f^N(x,t)dx =$$

$$= \big[\phi(x(t)) \cdot \sigma(\omega(t)x(t) + b(t)) \cdot f^N(x,t)\big] - \int_{\mathbb{R}^d} \phi(x(t))\nabla_x \cdot \big(\sigma(w(t)x(t) + b(t)) \cdot f^N(x,t)$$

$$= 0 - \int_{\mathbb{R}^d} \phi(x(t))\nabla_x \cdot \big(\sigma(w(t)x(t) + b(t)) \cdot f^N(x,t)\big)dx$$

Appendix - Formal derivation

$$\frac{d}{dt}\int_{\mathbb{R}^d}\phi(x(t))f^N(x,t)dx = -\int_{\mathbb{R}^d}\phi(x(t))\nabla_x\cdot(\sigma(w(t)x(t)+b(t))\cdot f^N(x,t))dx$$

$$\implies \int_{\mathbb{R}^d}\phi(x(t))\big[\partial_t f^N(x,t)+\nabla_x\cdot(\sigma(w(t)x(t)+b(t))\cdot f^N(x,t))\big]dx = 0$$

$$\implies \partial_t f^N(x,t)+\nabla_x\cdot(\sigma(w(t)x(t)+b(t))\cdot f^N(x,t)) = 0$$