Universal approximation and convexified training in neural networks

Kang Liu

Université Bourgogne Europe

DeustoCCM Seminar 01/07/2025

CHAIR OF COMPUTATIONAL MATHEMATICS

INSTITUT DE MATHÉMATIQUES DE BOURGOGNE

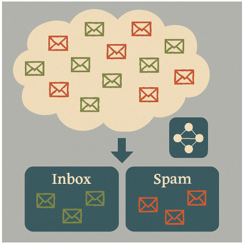
cnrs

UNIVERSITÉ BOURGOGNE EUROPE

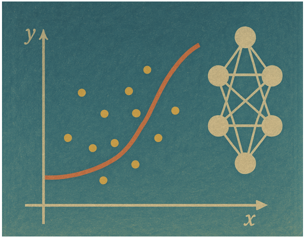
Table of Contents

Examples of Machine Learning

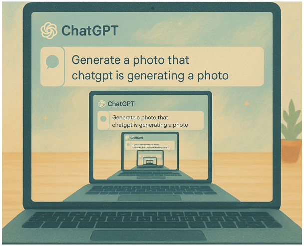
Classification



Regression



Generation



- **Why it works** : Universal approximation property (UAP),

$$f(x) \approx f_\Theta(x);$$

- **How it works** : Optimization (training),

$$\inf_{\Theta} \sum_{i=1}^{N} \text{Loss}\left(f_\Theta(x_i), f(x_i)\right) + r(\Theta).$$

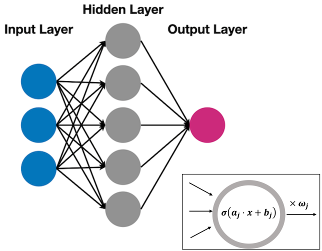
Shallow (One-hidden-layer) Neural Network

Formulation:

$$f(x;\Theta) = \sum_{i=1}^{P} w_i \, \sigma\big(\langle a_i, x\rangle + b_i\big),$$

where

- $\Theta = \{(w_i, a_i, b_i) \in \mathbb{R}^{d+2}\}_{i=1}^{P}$;
- σ is an activation function.



UAP: A Historical Overview

Qualitative Results

- Wiener (1932)
- Cybenko (1989)
- Hornik (1991)
- ...
- Pinkus (1999)

Quantitative Bounds

- Barron (1993)
- Bach (2017)
- Klusowski-Barron (2018)
- E–Ma–Wu (2022)
- Siegel-Xu (2024)...

Qualitative UAP

Universal approximation property [Pinkus 1999, Acta Numer.]

Fix any compact set $X \subseteq \mathbb{R}^d$. Let σ be a non-polynomial continuous function. For any function $f \in \mathcal{C}(X)$ and $\epsilon > 0$, there exists $P \in \mathbb{N}_+$ and parameters $\Theta = (\omega_i, a_i, b_i)_{i=1}^{P}$ such that

$$\|f - f_{\text{shallow}}(\cdot, \Theta)\|_{\mathcal{C}(X)} \leq \epsilon.$$

Quantitative UAP I: Barron Space

Fix the domain $X = [-1, 1]^d$, and let σ denote the ReLU activation function.

Definition: Barron Space $\mathcal{S}_B(X)$

A function $f \in \mathcal{C}(X)$ belongs to the **Barron space** $\mathcal{S}_B(X)$ if there exists a probability measure $\mu \in \mathcal{P}(\mathbb{R}^{d+2})$ such that

$$f(x) = \int_{\mathbb{R}^{d+2}} w\, \sigma(\langle a, x\rangle + b)\, d\mu(w, a, b), \quad \forall x \in X.$$

Sufficient Condition [Klusowski–Barron 2018, IEEE Trans. Inf. Theory]

If $f \in \mathcal{C}(X)$ admits an extension $\widetilde{f} \in \mathcal{C}(\mathbb{R}^d)$ whose Fourier transform satisfies

$$v_{f,2} := \int_{\mathbb{R}^d} \|\omega\|^2 \left|\mathcal{F}(\widetilde{f})(\omega)\right| d\omega < \infty,$$

then $f \in \mathcal{S}_B(X)$.

Sobolev embedding into Barron Space: $H^k(X) \subseteq \mathcal{S}_B(X) \quad$ if $k > \frac{d}{2} + 2$.

Quantitative UAP II: L^∞ Approximation Rate

L^∞-Approximation Rate [Klusowski-Barron 2018, IEEE Trans. Inf. Theory]

If $f \in \mathcal{C}(X)$ has an extension $\widetilde{f}$ to $\mathbb{R}^d$ such that $v_{f,2} < \infty$. Then, for every integer $P \geq 3$ there exist $(w_i, a_i, b_i) \in \mathbb{R}^{n+2}$, for $i = 1, \dots, P$, such that

$$\left\| f - \sum_{i=1}^{P} w_i \, \sigma\big(\langle a_i, \cdot\rangle + b_i\big) \right\|_{L^\infty(X)} \leq \frac{C_d \, v_{f,2}}{\sqrt{P}}, \quad \text{and}$$

$$\mathrm{Lip}\left(\sum_{i=1}^{P} w_i \, \sigma\big(\langle a_i, \cdot\rangle + b_i\big) \right) \leq \|\nabla f(0)\| + 2 \, v_{f,2},$$

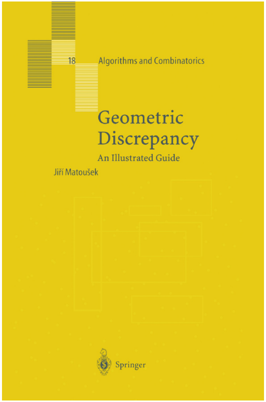
where $C_d > 0$ depends only on the dimension d.

Key elements in the proof:

1. $f \in \mathcal{S}_B(X) \Rightarrow \exists\, \mu \in \mathcal{P}(\mathbb{R}^{d+2})$ such that $f(x) = \int_{\mathbb{R}^{d+2}} w\, \sigma(\langle a, x\rangle + b)\, d\mu$;
2. Let $\mathcal{P}_P(\mathbb{R}^{d+2})$ be the set of empirical measures supported on at most P points. Find an upper bound of the following Minimax problem:

$$\inf_{\mu_P \in \mathcal{P}_P(\mathbb{R}^{d+2})} \sup_{x \in X} \left| \int_{\mathbb{R}^{d+2}} w\, \sigma(\langle a, x\rangle + b)\, d(\mu - \mu_P) \right| \leq \mathcal{O}\left(\frac{1}{P^r} \right)$$

Core estimate: $r \geq 1/2$. The analysis is based on techniques from *geometric discrepancy theory*.



Acta Math., 177 (1996), 55–73

Improved upper bounds for approximation by zonotopes

by

JIŘÍ MATOUŠEK

Charles University
Prague, Czech Republic

1. Introduction

A *zonotope* in $\mathbf{R}^n$ is a special type of a convex polytope; it is defined as a Minkowski sum of finitely many segments. That is, a zonotope is a set in $\mathbf{R}^n$ of the form $\{x_1+x_2+...+x_m: x_1 \in I_1, ..., x_m \in I_m\}$, where $I_1, ..., I_m$ are segments in $\mathbf{R}^n$. A convex body that can be approximated by zonotopes arbitrarily closely is called a *zonoid*. Several authors have recently studied the following question: what is the minimum number, N, of summands of a zonotope needed to approximate a given zonoid Z in $\mathbf{R}^n$ with error at most ε (this means that $Z \subseteq A \subseteq (1+\varepsilon)Z$, where A is the approximating zonotope, and we assume that the center of symmetry of Z is the origin). Here we consider the dimension n fixed, and we investigate the dependence of N on ε (we assume that $n \geqslant 3$, as the case $n=2$ is simple—see [4]).

Other quantitative UAP results using this technique:

- Bach (2017), *Journal of Machine Learning Research*;
- Siegel (2025), *Constructive Approximation*.

Outline

Semi-autonomous Neural ODE

- Consider an ODE system in $\mathbb{R}^d$ with an unknown vector field f:

$$\begin{cases} \dot{\boldsymbol{z}}_{z_0} = f(\boldsymbol{z}_{z_0}, t), \quad t \in (0, T), \\ \boldsymbol{z}_{z_0}(0) = z_0. \end{cases}$$

- We propose the following **SA-NODE** to approximate it:

$$\begin{cases} \dot{\boldsymbol{x}}_{z_0} = \sum_{i=1}^{P} W_i \circ \boldsymbol{\sigma}(A_i^1 \boldsymbol{x}_{z_0} + A_i^2 t + B_i), \quad t \in (0, T), \\ \boldsymbol{x}_{z_0}(0) = z_0. \end{cases}$$

with

- $\circ$ the Hadamard product,
- $W_i \in \mathbb{R}^d$,
- $A_i^1 \in \mathbb{R}^{d \times d}$, $A_i^2 \in \mathbb{R}^d$
- $B_i \in \mathbb{R}^d$.

UAP of SA-NODE I: ODE

Assume that

$$f \in \mathcal{H}^k_{\text{loc}}(\mathbb{R}^d \times [0,T]; \mathbb{R}^d), \quad \text{with } k > (d+1)/2 + 2. \tag{2.1}$$

Theorem 1 [Li-L.-Liverani-Zuazua 2024]

Let (2.1) hold true. Fix any compact set $K \subseteq \mathbb{R}^d$. For any $P \geq 3$, there exist parameters $(W_i, A_i^1, A_i^2, B_i)_{i=1}^P$ such that

$$\|\boldsymbol{z}_{z_0}(t) - \boldsymbol{x}_{z_0}(t)\|_{L^\infty([0,T];\mathbb{R}^d)} \leq \frac{C_{T,K,f}}{\sqrt{P}}, \quad \forall z_0 \in K,$$

where $C_{T,K,f}$ is a constant independent of P.

Key proof steps:

- A priori estimate on the SA-NODE domain via bootstrapping;
- Apply L^∞ approximation of f [Klusowski–Barron 2018];
- Use Grönwall's inequality.

UAP of SA-NODE II: Continuity Equation

- Continuity equation

$$\begin{cases} \partial_t \rho(x,t) + \mathrm{div}_x(f(x,t)\rho(x,t)) = 0, & (x,t) \in \mathbb{R}^d \times [0,T], \\ \rho(\cdot, 0) = \rho_0 \in \mathcal{P}(\mathbb{R}^d). \end{cases} \tag{2.2}$$

- Neural counterpart: ρ_Θ satisfies (2.2) with f replaced by

$$f_\Theta(x,t) = \sum_{i=1}^{P} W_i \circ \boldsymbol{\sigma}(A_i^1 x + A_i^2 t + B_i).$$

Theorem 2 [Li-L.-Liverani-Zuazua 2024]

Let (2.1) hold true. Assume that ρ_0 has compact support set. Then, for any $P \geq 3$, there exist parameters $\Theta = \{(W_i, A_i^1, A_i^2, B_i)\}_{i=1}^P$ such that

$$\sup_{t\in[0,T]} \mathbb{W}_1(\rho(\cdot,t), \rho_\Theta(\cdot,t)) \leq \frac{C_{T,f,\rho_0}}{\sqrt{P}},$$

where C_{T,f,ρ_0} is a constant independent of P and $\mathbb{W}_1$ is the Wasserstein-1 distance.

Key proof steps: Theorem 1 + Superposition principle of the continuity equation.

Outline

1. Introduction to UAP of NNs
2. UAP in dynamical systems: SA-NODE
   - Theory
   - Training, optimization, and numerical simulations
3. Sparse optimization and mean-field relaxation
   - Mean-filed (convex) relaxation
   - Algorithms and numerical simulations

True dynamic system

$$\begin{cases} \dot{\boldsymbol{z}}_{z_0} = f(\boldsymbol{z}_{z_0}, t), \quad t \in (0, T), \\ \boldsymbol{z}_{z_0}(0) = z_0. \end{cases}$$

SA-NODE

$$\begin{cases} \dot{\boldsymbol{x}}_{z_0} = \sum_{i=1}^{P} W_i \circ \boldsymbol{\sigma}(A_i^1 \boldsymbol{x}_{z_0} + A_i^2 t + B_i), \quad t \in (0, T), \\ \boldsymbol{x}_{z_0}(0) = z_0. \end{cases}$$

Optimal control problem for learning (continuous)

$$\inf_{\Theta} L(\Theta) = \int_0^T \int_K \|\boldsymbol{z}_{z_0}(t) - \boldsymbol{x}_{z_0}(t)\|^2 d\boldsymbol{z}_0 dt + \lambda \underbrace{\left\| \sum_{i=1}^{P} |W_i| \circ \|A_i^1\|_{\ell^2} \right\|}_{\text{Lipschitz constant of SA-NODE}} .$$

In practice, we observe N trajectories, $\mathbf{z}^k$, of $\mathbf{z}$ from N different initial positions.

Discretized problem

$$\inf_{\Theta} \hat{L}(\Theta) = \frac{\Delta t}{N} \sum_{k=1}^{N} \sum_{l=1}^{M} \left(\boldsymbol{z}^k(t_l) - \boldsymbol{x}^k(t_l, \Theta) \right)^2 + \lambda \left\| \sum_{i=1}^{P} |W_i| \circ \|A_i^1\|_{\ell^2} \right\|,$$

where $\boldsymbol{x}^k$ is the solution of SA-NODE with the same initial position as $\boldsymbol{z}^k$.

Adjoint method, backpropagation, and SGD

Theorem 3 (Gradient of L)

Let $\widetilde{f}(\Theta, x, t) = f_\Theta(x, t)$ and let $g(\Theta) = \left\| \sum_{i=1}^{P} |W_i| \circ \|A_i^1\|_{\ell^2} \right\|$. It holds that

$$\nabla L(\Theta) = \int_K \int_0^T \frac{\partial \widetilde{f}}{\partial \Theta}(\Theta, \boldsymbol{x}_{z_0}(t), t)^\top \boldsymbol{a}_{z_0}(t) dt dz_0 + \lambda \nabla g(\Theta), \quad \text{for } \Theta \text{ a.e.},$$

where $\boldsymbol{a}_{z_0}$ satisfies the adjoint equation (backpropagation)

$$\begin{cases} -\dot{\boldsymbol{a}}_{z_0}(t) = \frac{\partial \widetilde{f}}{\partial x}(\Theta, \boldsymbol{x}_{z_0}(t), t)^\top \boldsymbol{a}_{z_0}(t) + 2(\boldsymbol{x}_{z_0}(t) - \boldsymbol{z}_{z_0}(t)), & t \in [0, T], \\ \boldsymbol{a}_{z_0}(T) = 0, & z_0 \in K. \end{cases}$$

Discrete version:

$$\nabla \hat{L}(\Theta) = \frac{1}{N} \sum_{k=1}^{N} \underbrace{\left(\Delta t \sum_{l=1}^{M} \frac{\partial \widetilde{f}}{\partial \Theta}(\Theta, \boldsymbol{x}^k(t_l), t_l)^\top \boldsymbol{a}^k(t_l) \right)}_{\nabla \hat{L}_k(\Theta)} + \lambda \nabla g(\Theta).$$

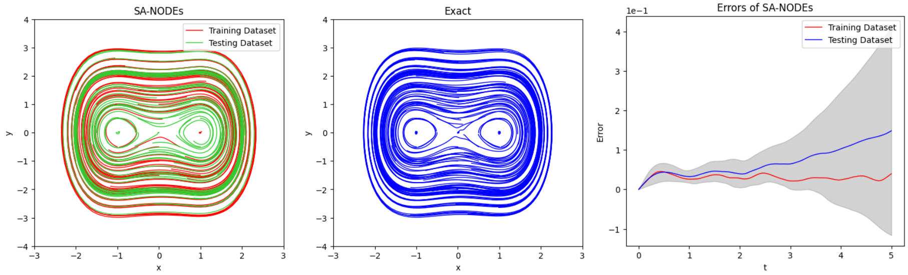
SGD:

$$\Theta^{m+1} = \Theta^m - \tau^m \left(\nabla \hat{L}_k(\Theta^m) + \lambda \nabla g(\Theta^m) \right)_{k \sim \text{Uni}\{1,\dots,N\}}.$$

Numerical Example I

Forced Duffing oscillator:

$$\begin{cases} \dot{z}_1 = z_2, \\ \dot{z}_2 = z_1 - z_1^3 + \delta \cos(\omega t). \end{cases}$$



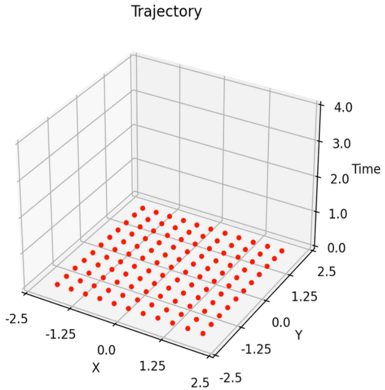
Numerical Example II

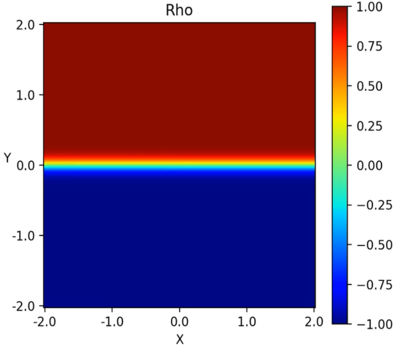
Doswell frontogenesis [Doswell 1984]:

$$\begin{cases} \partial_t \rho + \operatorname{div}\left(\left(-yg(r), xg(r)\right)\rho\right) = 0, & (x, y, t) \in \mathbb{R}^2 \times [0, T], \\ \rho(\cdot, 0) = \rho_0, \end{cases}$$

where

$$g(r) = \frac{1}{r}\ \overline{v}\ \mathrm{sech}^2(r) \tanh(r), \quad r = \sqrt{x^2 + y^2}.$$





t = 0.00

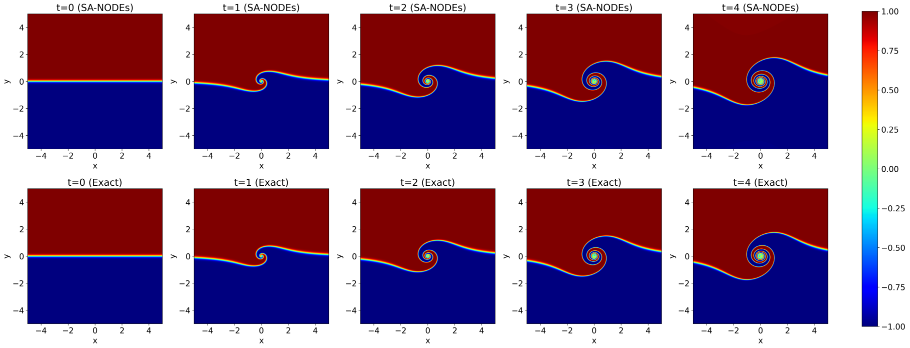
Numerical Example II

Doswell frontogenesis [Doswell 1984]:

$$\begin{cases} \partial_t \rho + \mathrm{div}\left(\left(-yg(r), xg(r)\right)\rho\right) = 0, \quad (x, y, t) \in \mathbb{R}^2 \times [0, T], \\ \rho(\cdot, 0) = \rho_0, \end{cases}$$

where

$$g(r) = \frac{1}{r}\ \overline{v}\ \mathrm{sech}^2(r)\tanh(r), \quad r = \sqrt{x^2 + y^2}.$$



Outline

Training problems for shallow NNs

- **Data**: $\{(x_i, y_i) \in \mathbb{R}^{d+1}\}_{i=1}^{N}$.
- **NNs architecture**:
  - feature (input): $x \in \mathbb{R}^d$;
  - parameter (control): $\Theta = (\omega, a, b) \in \mathbb{R}^{P\times(d+2)}$;
  - prediction (output):

$$f_{\text{shallow}}(x, \Theta) = \sum_{p=1}^{P} \omega_p \sigma(\langle a_p, x\rangle + b_p).$$

Training problem for shallow NNs

Let ℓ be a proper, convex and l.s.c. function. Consider the following optimization:

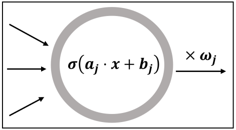
$$\inf_{(\omega,a,b)} \frac{1}{N}\sum_{i=1}^{N} \ell\left(y_i - \sum_{p=1}^{P} \omega_p \sigma(\langle a_p, x_i\rangle + b_p)\right) + \underbrace{R(\omega, a, b)}_{\text{Regularization}}.$$

Design of the regularization term

- A well-known principle [1] in machine learning is the following:

"sparsity" mitigates "overfitting".

- In shallow NNs, the number of activated neurons is $\|\omega\|_{\ell^0}$.



- The function $\|\omega\|_{\ell^0}$ is non-convex. A practical replacement from compressed sensing [2]:

$$\|\omega\|_{\ell^0} \mapsto \|\omega\|_{\ell^1}.$$

[1] Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov. "Dropout: A simple way to prevent Neural Networks from overfitting". In JMLR, 2014.

[2] Candes and Romberg. "Quantitative robust uncertainty principles and optimally sparse decompositions". In FOCM, 2006.

Optimization via Mean-Field Relaxation

Primal Problem

Optimization problem based on sparsity:

$$\inf_{(\omega,a,b)} \frac{1}{N} \sum_{i=1}^{N} \ell\left(y_i - \sum_{p=1}^{P} \omega_p \sigma(\langle a_p, x_i \rangle + b_p)\right) + \underbrace{\lambda \sum_{p=1}^{P} |\omega_p|}_{\text{Regularization for sparsity}} . \qquad \text{(P)}$$

Observation: the non-convexity arises from the non-linearity of neural networks.

$$\sum_{p=1}^{P} \omega_p \sigma(\langle a_p, x \rangle + b_p) \xrightarrow{\text{Mean-field relaxation}} \int_{\mathbb{R}^{d+1}} \sigma(\langle a, x \rangle + b)\, d\mu(a, b).$$

Relaxed Problem

$$\inf_{\mu} \frac{1}{N} \sum_{i=1}^{N} \ell\left(y_i - \int_{\mathbb{R}^{d+1}} \sigma(\langle a, x_i \rangle + b)\, d\mu\right) + \lambda \|\mu\|_{\mathrm{TV}}. \qquad \text{(PR)}$$

Free of relaxation gap

Theorem (L.-Zuazua, 2025)

Under mild assumptions [1] *on* σ *and the constraint domain* Ω *for* (a_p, b_p), *if* $P \geq N$, *then*

$$\text{val(P)} = \text{val(PR)}.$$

Moreover, the extreme points of the solution sets of relaxed problems have the following form:

$$\mu^* = \sum_{j=1}^{N} \omega_j^* \delta_{(a_j^*, b_j^*)}.$$

Key proof steps:

- Existence of solutions: finite-sample representation property from [Pinkus, 1999].
- "Representer Theorem" from [Fisher-Jerome, 1975].

[1] An example of (σ, Ω): σ is the ReLU function and Ω is the unit ball.

A generalization bound

- **Training/Testing** dataset: $\{(x_i, y_i)\}_{i=1}^{N}$ / $\{(x_i', y_i')\}_{i=1}^{N'}$.
- **Predictions** on testing set by the shallow NN with parameter Θ:

$$\{(x_i', f_{\text{shallow}}(x_i', \Theta))\}_{i=1}^{N'}$$

- **Empirical measures**:

$$m_{\text{train}} = \frac{1}{N}\sum_{i=1}^{N}\delta_{(x_i,y_i)}, \quad m_{\text{test}} = \frac{1}{N'}\sum_{i=1}^{N'}\delta_{(x_i',y_i')}, \quad m_{\text{pred}}(\Theta) = \frac{1}{N'}\sum_{i=1}^{N'}\delta_{(x_i',f_{\text{shallow}}(x_i',\Theta))}.$$

A generalization bound

- **Training/Testing** dataset: $\{(x_i, y_i)\}_{i=1}^{N}$ / $\{(x_i', y_i')\}_{i=1}^{N'}$.
- **Predictions** on testing set by the shallow NN with parameter Θ:

$$\{(x_i', f_{\text{shallow}}(x_i', \Theta))\}_{i=1}^{N'}$$

- **Empirical measures**:

$$m_{\text{train}} = \frac{1}{N}\sum_{i=1}^{N}\delta_{(x_i, y_i)}, \quad m_{\text{test}} = \frac{1}{N'}\sum_{i=1}^{N'}\delta_{(x_i', y_i')}, \quad m_{\text{pred}}(\Theta) = \frac{1}{N'}\sum_{i=1}^{N'}\delta_{(x_i', f_{\text{shallow}}(x_i', \Theta))}.$$

Theorem (L.-Zuazua,2025)

Let $W_1(\cdot,\cdot)$ denote the Wasserstein-1 distance. If σ is 1-Lipschitz, then for any Θ,

$$W_1(m_{\text{test}}, m_{\text{pred}}(\Theta)) \leq \underbrace{2W_1(m_{\text{train}}, m_{\text{test}})}_{\text{Bias from datasets}} + r(\Theta), \quad \text{where}$$

$$r(\Theta) = \underbrace{\frac{1}{N}\sum_{i=1}^{N}|f_{\text{shallow}}(x_i, \Theta) - y_i|}_{\text{Bias from training}} + \underbrace{W_1(m_{\text{train}}, m_{\text{test}})\sum_{j=1}^{P}|\omega_j|\|a_j\|}_{\text{"Standard deviation"}}.$$

Outline

Guideline for numerical algorithms

The relaxed problem is convex, but in an infinite-dimensional space.

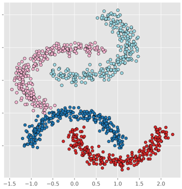
$$\inf_{\mu \in \mathcal{M}(\Omega)} \|\mu\|_{\mathrm{TV}} + \frac{\lambda}{N} \sum_{i=1}^{N} |\phi_i \, \mu - y_i|,$$

where $\phi_i \, \mu = \int_\Omega \sigma(\langle a, x_i \rangle + b) d\mu(a, b)$.

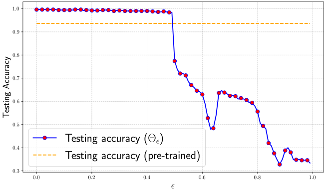
Two numerical scenarios

1. When $\dim(\Omega) = d + 1$ is small: Discretization, then Optimization.
   - Discretize Ω by a mesh, then optimize by the simplex method.
2. When $\dim(\Omega) = d + 1$ is great: Optimization, Discretization, then Sparsification.
   - Write the gradient flow associated with (PR);
   - The SGD algorithm on an overparameterized (P) (a large P) is seen as a discretization of the gradient flow;
   - Filter the overparameterized result by our Sparsification algorithm.

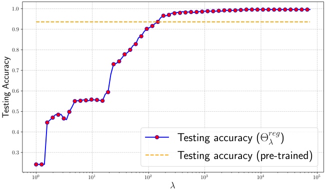
Classification in 2-D



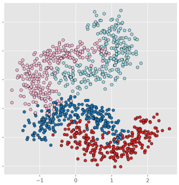
(a) Datasets.



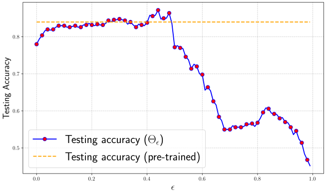
(b) Testing accuracy w.r.t. ϵ.



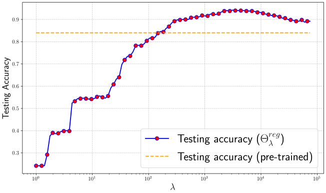
(c) Testing accuracy w.r.t. λ.



(a) Datasets.

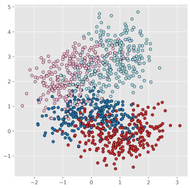


(b) Testing accuracy w.r.t. ϵ.

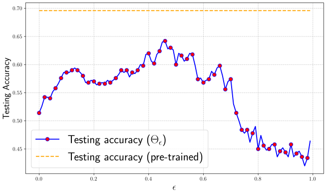


(c) Testing accuracy w.r.t. λ.

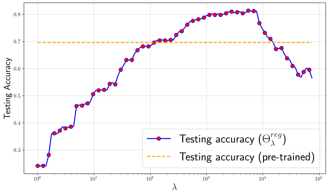
Classification in 2-D



(a) Datasets.



(b) Testing accuracy w.r.t. ϵ.

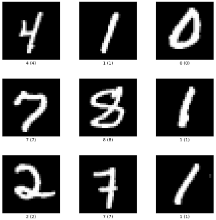


(c) Testing accuracy w.r.t. λ.

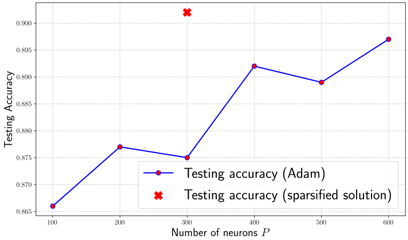
Conclusion:

- If the datasets have clear separable boundaries, take $\lambda \to \infty$;
- If the datasets have heavily overlapping areas, consider a particular range of $\lambda \sim W_1^{-1}(m_{\text{train}}, m_{\text{test}})$.

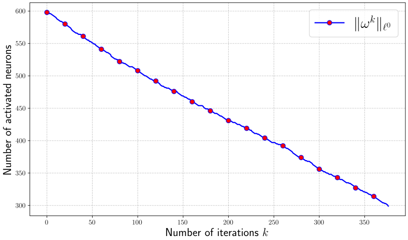
Classification in a high-dimensional space



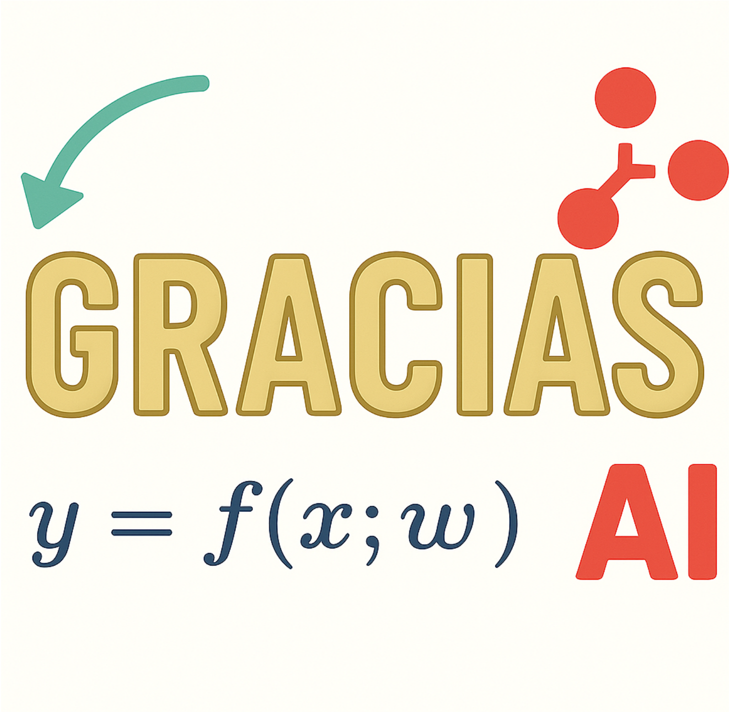
- The Mnist dataset, vectors in $\mathbb{R}^{28\times 28}$.
- Training data: 300 samples of numbers 0, 1, and 2.
- Testing data: 1000 samples of numbers 0, 1, and 2.



(a) Testing accuracy w.r.t. P.



(b) $\|\omega\|_{\ell^0}$ w.r.t. the iteration number.


GRACIAS
y = f(x; w)
AI

Low-dimensional scenario

- Discretization of the domain:

$$\Omega \to \Omega_h = \{(a_j, b_j)\}_{j=1}^{M}.$$

- Discretized problems:

$$\inf_{\omega \in \mathbb{R}^M} \|\omega\|_{\ell^1} + \frac{\lambda}{N}\|A\omega - Y\|_{\ell^1}, \qquad \text{(PD)}$$

where $A \in \mathbb{R}^{N \times M}$ with $A_{ij} = \sigma(\langle a_j, x_i \rangle + b_j)$.

- Error estimates:

$$|\mathrm{val(PD)} - \mathrm{val(PR)}| = \mathcal{O}(d_{\mathrm{Hausdorff}}(\Omega, \Omega_h)).$$

- Equivalent to linear programming problems, solvable using the simplex method.
  - **Advantage**: Terminates at an extreme point of the solution set, which corresponds to a solution of the primal problems.
  - **Limitation**: Suffer from the curse of dimensionality.

High-dimensional scenario

- Apply the SGD algorithm to the following overparameterized problem:

$$\inf_{\Theta\in(\mathbb{R}\times\Omega)^{\bar{P}}} \|\omega\|_{\ell^1} + \frac{\lambda}{N}\sum_{i=1}^{N}\ell\left(\sum_{j=1}^{\bar{P}}\omega_j\sigma(\langle a_j, x_i\rangle + b_j) - y_i\right),$$

  where $\bar{P}$ is large [1].

- Use the sparsification method developed in [L.-Zuazua, 2024] to filter the previous solution, obtaining one with fewer than N activated neurons.

This approach is free from the curse of dimensionality but lacks rigorous convergence analysis.

[1] The convergence properties of SGD for the training of overparameterized NNs have been extensively studied recently, including [Chitzat-Bach, 2018], [Zhu-Li-Song, 2019], [Bach, 2024, Chp.12], etc.