Weierstraß-Institut für
Angewandte Analysis und Stochastik

Leibniz Gemeinschaft

Estimation and Inference for smooth DNNs. Blessing of dimension

Vladimir Spokoiny ,
HSE University and IITP RAS Moscow,
WIAS Berlin

4. Juni 2025

Given data $(Y_i, \boldsymbol{X}_i)$ with response Y_i and regressors $\boldsymbol{X}_i \in \mathbb{R}^q$.

Aim: to build a network with a small prediction error $Y_i - m(\boldsymbol{X}_i)$.

Shallow network: $\mathbb{X} = \sigma(W\boldsymbol{X})$,

DNN: $\mathbb{X} = \mathbb{X}^{(K)} = \mathrm{DNN}(\boldsymbol{X}) \in \mathbb{R}^M$,

$$\mathrm{DNN}(\boldsymbol{X}) = \sigma(W^{(K)}\sigma(W^{(K-1)} \dots \sigma(W^{(1)}\boldsymbol{X}))),$$

where M is the number of nodes in the last hidden layer,
$W^{(k)} = (w_{mj})$ is a weighting $M_k \times M_{k-1}$-matrix with rows $W_m^{(k)}$,
and σ is an entry-wise activating function.

Prediction by linear regression on $\mathbb{X}$:

$$m(\boldsymbol{X}_i) = \mathbb{X}_i^\top \boldsymbol{\beta} \quad = \sigma(W\boldsymbol{X}_i)^\top \boldsymbol{\beta}.$$

Model:

$$L(\boldsymbol{Y}, W, \boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^{n} \left| Y_i - \sigma(W \boldsymbol{X}_i)^\top \boldsymbol{\beta} \right|^2$$

MLE:

$$(\widehat{W}, \widehat{\boldsymbol{\beta}}) = \operatorname*{argmin}_{W, \boldsymbol{\beta}} L(W, \boldsymbol{\beta}).$$

Issues: Non-convexity, lack of identifiability, overparametrization.

Some literature on DNN regression

- [Schmidt-Hieber, 2020]. Nonparametric regression using deep neural networks with ReLU activation function, Kolmogorov-Arnold representation,
- [Zuowei Shen et al., 2020], Deep Network Approximation Characterized by Number of Neurons,
- [Fan et al., 2024], How do noise tails impact on deep ReLU networks? polynomial noise, tails, Huber, DNN approximation,
- [Kohler and Krzyzak, 2022], Analysis of the rate of convergence of an over-parametrized deep neural network estimate learned by gradient descent,
- [Shen et al., 2022], Approximation with CNNs in Sobolev Space: with Applications to Classification,
- [Liu et al., 2022], Benefits of Overparameterized Convolutional Residual Networks: Function Approximation under Smoothness Constraint,
- [Simionescu-Badea, 2022], Analysis of the rate of convergence of fully connected deep neural network regression estimates with smooth activation,

Manifold regression

- [Chen et al., 2019], Efficient Approximation of Deep ReLU Networks for Functions on Low Dimensional Manifolds,
- [Kohler et al., 2019], Estimation of a Function of Low Local Dimensionality by DNN,
- [Jiao et al., 2021], Deep Nonparametric Regression on Approximately Low-dim Manifolds,
- [Liu et al., 2021], Besov Function Approximation . . . on Low-Dim Manifolds . . . ,
- [Cloninger and Klock, 2021], A deep network construction that adapts to intrinsic dimensionality beyond the domain,
- [Zhang et al., 2023], Effective Minkowski Dimension of Deep Nonparametric Regression: Function Approximation and Statistical Theories,
- [Shen et al., 2023], RePU DNN, manifolds, Differentiable Neural Networks with RePU Activation: with Applications . . .
- [Jiao et al., 2023], Deep nonparametric regression on approximate manifolds

Asymptotic minimax rate results

- are very poor (small power of n, hidden constants, . . .)
- are useless (do not help in practical applications because involve unknown constants)
- are sometimes even misleading (rate optimality does not lead to a practically efficient method)
- are too concervative (functional classes are too big)
- depend on unknown parameters of a functional class
- do not properly explain the important notions like effective and critical dimension or curse of dimension (rather curse of minimax rates)

- Parameter dimension p large or infinite;
- Moderate sample size n ;
- Finite sample “oracle type” theoretical guaranties;
- Explicit dependence on the “effective dimension”;
- mild and realistic model assumptions
- Efficient numerical procedures

Blessing of dimension:

When and why overparametrization can be useful?

Extending the parameter space can improve geometric properties of the loss function

without increasing the effective dimension.

Let $L(\upsilon)$ be a random function (loss, empirical risk, negative log-likelihood). Consider

$$\widetilde{\upsilon} = \operatorname*{argmin}_{\upsilon} L(\upsilon); \quad \upsilon^* = \operatorname*{argmin}_{\upsilon} \mathbb{E}\, L(\upsilon);$$

Includes MLE, LSE, LAD, MC, and many others procedures.

Aim: describe $\widetilde{\upsilon} - \upsilon^*$ (estimation loss) and $L(\widetilde{\upsilon}) - L(\upsilon^*)$ (excess).

Perturbed optimization: $L(\upsilon)$ is a perturbation of $f(\upsilon) = \mathbb{E}L(\upsilon)$ by a stochastic component

$$\zeta(\upsilon) = L(\upsilon) - \mathbb{E}L(\upsilon).$$

- **Truth** $\boldsymbol{\upsilon}^* \stackrel{\text{def}}{=} \operatorname{argmin}_{\boldsymbol{\upsilon}} \mathbb{E} L(\boldsymbol{\upsilon})$;
- **MLE** $\widetilde{\boldsymbol{\upsilon}} \stackrel{\text{def}}{=} \operatorname{argmin}_{\boldsymbol{\upsilon}} L(\boldsymbol{\upsilon})$;
- **Fisher information** $\mathbb{F}(\boldsymbol{\upsilon}) = \nabla^2 \mathbb{E} L(\boldsymbol{\upsilon})\,, \ \mathbb{F} = \mathbb{F}(\boldsymbol{\upsilon}^*)$;
- **Effective sample size** $\mathbb{N} \stackrel{\text{def}}{=} \lambda_{\min}(\mathbb{F})$;
- **Score** $\zeta(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} L(\boldsymbol{\upsilon}) - \mathbb{E} L(\boldsymbol{\upsilon})\,, \ \nabla\zeta = \nabla\zeta(\boldsymbol{\upsilon}^*)$;
- **Effective dimension** $\mathbb{p} \stackrel{\text{def}}{=} \operatorname{tr}\bigl(\mathbb{F}^{-1} \operatorname{Var}(\nabla\zeta)\bigr)$;

- **Concentration** (requires $\mathbb{p} \ll \mathbb{N}$)

$$\|\mathbb{F}^{1/2}(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*)\| \lesssim \sqrt{\mathbb{p}}\,.$$

- **Fisher expansion** (requires $\mathbb{p} \ll \mathbb{N}$)

$$\|\mathbb{F}^{1/2}(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^* + \mathbb{F}^{-1}\nabla\zeta)\| \lesssim \frac{\|\mathbb{F}^{-1}\nabla\zeta\|^2}{\sqrt{\mathbb{N}}}\,.$$

- **Wilks expansion** (requires $\mathbb{p} \ll \mathbb{N}$)

$$|L(\widetilde{\boldsymbol{\upsilon}}) - L(\boldsymbol{\upsilon}^*) + \tfrac{1}{2}\|\mathbb{F}^{-1/2}\nabla\zeta\|^2| \lesssim \frac{\|\mathbb{F}^{-1}\nabla\zeta\|^{3/2}}{\sqrt{\mathbb{N}}}\,.$$

Consider a penalized MLE for a penalty function $\mathrm{pen}_G(\boldsymbol{\upsilon})$

$$\widetilde{\boldsymbol{\upsilon}}_G = \operatorname*{argmin}_{\boldsymbol{\upsilon}} L_G(\boldsymbol{\upsilon}) = \operatorname*{argmin}_{\boldsymbol{\upsilon}} \bigl\{ L(\boldsymbol{\upsilon}) + \mathrm{pen}_G(\boldsymbol{\upsilon}) \bigr\}$$

Typical examples: $\mathrm{pen}_G(\boldsymbol{\upsilon}) = \|G\boldsymbol{\upsilon}\|^2/2$ and $\mathrm{pen}_\lambda(\boldsymbol{\upsilon}) = \lambda\|\boldsymbol{\upsilon}\|_1$.

Consider three optimization problems

$$\widetilde{\boldsymbol{\upsilon}}_G = \operatorname*{argmin}_{\boldsymbol{\upsilon}} L_G(\boldsymbol{\upsilon}),$$

$$\boldsymbol{\upsilon}^*_G = \operatorname*{argmin}_{\boldsymbol{\upsilon}} \mathbb{E} L_G(\boldsymbol{\upsilon}),$$

$$\boldsymbol{\upsilon}^* = \operatorname*{argmin}_{\boldsymbol{\upsilon}} \mathbb{E} L(\boldsymbol{\upsilon}).$$

The penalized MLE $\widetilde{\boldsymbol{\upsilon}}_G$ estimates rather $\boldsymbol{\upsilon}^*_G$ than $\boldsymbol{\upsilon}^*$.

$$\boldsymbol{\upsilon}_G^* = \operatorname*{argmin}_{\boldsymbol{\upsilon}} \mathbb{E} L_G(\boldsymbol{\upsilon}) = \operatorname*{argmin}_{\boldsymbol{\upsilon}} \{ \mathbb{E} L(\boldsymbol{\upsilon}) + \mathrm{pen}_G(\boldsymbol{\upsilon}) \},$$

$$\boldsymbol{\upsilon}^* = \operatorname*{argmin}_{\boldsymbol{\upsilon}} \mathbb{E} L(\boldsymbol{\upsilon}).$$

Proposition

Define

$$\mathbb{F}_G \stackrel{\mathrm{def}}{=} \nabla^2 \mathbb{E} L_G(\boldsymbol{\upsilon}_G^*) = \mathbb{F} + \nabla^2 \mathrm{pen}_G(\boldsymbol{\upsilon}_G^*)\,,$$

$$\boldsymbol{M}_G \stackrel{\mathrm{def}}{=} \nabla \mathrm{pen}_G(\boldsymbol{\upsilon}^*)\,.$$

Then

$$\| \mathbb{F}_G^{1/2} (\boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^* + \mathbb{F}_G^{-1} \boldsymbol{M}_G) \| \lesssim \frac{\| \mathbb{F}_G^{-1/2} \boldsymbol{M}_G \|^2}{\sqrt{\mathbb{N}}}\,.$$

Fix $Q : \mathbb{R}^p \to \mathbb{R}^q$

$$\mathbb{p}_Q \stackrel{\text{def}}{=} \operatorname{tr} \operatorname{Var}(Q I\!\!F_G^{-1} \nabla\zeta) , \qquad \text{variance}$$

$$\mathrm{b}_Q = \| Q I\!\!F_G^{-1} \boldsymbol{M}_G \| , \qquad \text{bias}$$

$$\mathscr{R}_Q \stackrel{\text{def}}{=} I\!\!E\{ \| Q I\!\!F_G^{-1} (\nabla\zeta + \boldsymbol{M}_G) \|^2 \, 1\!\!1_{\Omega(\mathtt{x})} \} \qquad \text{squared risk}$$

$$\leq \mathbb{p}_Q + \mathrm{b}_Q^2 .$$

Then $\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^* \approx I\!\!F_G^{-1} (\nabla\zeta + \boldsymbol{M}_G)$ and

$$(1 - \alpha_Q)^2 \mathscr{R}_Q \leq I\!\!E\big\{ \| Q \, (\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*) \|^2 \, 1\!\!1_{\Omega(\mathtt{x})} \big\} \leq (1 + \alpha_Q)^2 \mathscr{R}_Q .$$

- Stochastic linearity: the stochastic component $\zeta(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - \mathbb{E}L(\boldsymbol{\upsilon})$ is linear in $\boldsymbol{\upsilon}$ or $\nabla\zeta(\boldsymbol{\upsilon}) \equiv \nabla\zeta$;
- Sub-gaussian tails for $\nabla\zeta$;
- Convexity: $\mathbb{E}L_G(\boldsymbol{\upsilon})$ is strongly convex in $\boldsymbol{\upsilon}$;
- Smoothness: $\mathbb{E}L_G(\boldsymbol{\upsilon})$ is third order smooth in uniform gateaux sense (self-concordance);

- The results are dimension- and coordinate-free. The ambient dimension $p \leq \infty$ is unimportant; only the effective dimension $\mathbb{p}$ matters. Can be controlled by a choice of penalization, replacing model reduction. The results require $\mathbb{p} \ll \boldsymbol{N}$.
- All the remainder are explicit and can be handled. Helpful for further analysis of iterative procedures like stochastic optimization, EM-procedures, etc.
- Conditions are restrictive. Convexity is the most critical, but can be ensured by extending the parameter set (blessing of dimension) and localization.
- The main tool is perturbed optimization, [Spokoiny, 2025c, Spokoiny, 2025b].

1 Introduction

2 Regression using DNN

- Training procedure
- DNN regression. Theoretical study
- A multilayer network

Model:

$$L(\boldsymbol{Y}, W, \boldsymbol{\beta}) = -\frac{1}{2}\sum_{i=1}^{n}\left|Y_i - \sigma(W\boldsymbol{X}_i)^\top\boldsymbol{\beta}\right|^2$$

MLE:

$$(\widehat{W}, \widehat{\boldsymbol{\beta}}) = \operatorname*{argmin}_{W,\boldsymbol{\beta}} L(W, \boldsymbol{\beta}).$$

Issue: The major assumptions of stochastic linearity and convexity are not fulfilled.

Decouple the structural relations $\mathbb{X} = \sigma(W\boldsymbol{X})$ and $m = \mathbb{X}^\top \boldsymbol{\beta}$ and replace them by structural penalty $\lambda^* \|\mathbb{X} - \sigma(W\boldsymbol{X})\|^2/2$ and $\lambda^* \|m - \mathbb{X}^\top \boldsymbol{\beta}\|^2/2$ (later $\lambda^* = 1$): for $\boldsymbol{\upsilon} = (W, \mathbb{X}, \boldsymbol{\beta}, m)$

$$\mathscr{L}(\boldsymbol{\upsilon}) = \frac{1}{2}\|\boldsymbol{Y} - m\|^2 + \frac{1}{2}\|\mathbb{X} - \sigma(W\boldsymbol{X})\|^2 + \frac{1}{2}\|m - \mathbb{X}^\top \boldsymbol{\beta}\|^2 .$$

Noise reduction by presmoothing by $\mathcal{S} : \boldsymbol{Y} \to \boldsymbol{Z} \stackrel{\text{def}}{=} \mathcal{S}\boldsymbol{Y}$:

$$\mathscr{L}(\boldsymbol{\upsilon}) = \frac{\mu^2}{2}\|\boldsymbol{Z} - \boldsymbol{z}\|^2 + \frac{1}{2}\|\mathbb{X} - \sigma(W\boldsymbol{X})\|^2 + \frac{1}{2}\|\boldsymbol{z} - \mathcal{S}\mathbb{X}^\top \boldsymbol{\beta}\|^2 .$$

Identifiability issue: partially resolved by ridge penalties:

$$\mathscr{L}(\boldsymbol{\upsilon}) = \frac{\mu^2}{2}\|\boldsymbol{Z} - \boldsymbol{z}\|^2 + \frac{1}{2}\|\boldsymbol{z} - \mathcal{S}\mathbb{X}^\top\boldsymbol{\beta}\|^2 + \frac{1}{2}\|\mathbb{X} - \sigma(W\boldsymbol{X})\|^2$$
$$+ \frac{\lambda_{\mathbb{W}}}{2}\|W\|^2 + \frac{\lambda_{\mathbb{X}}}{2}\|\mathbb{X}\|^2 + \frac{\lambda_{\boldsymbol{\beta}}}{2}\|\boldsymbol{\beta}\|^2 + \frac{\lambda_{\boldsymbol{z}}}{2}\|\boldsymbol{z}\|^2 .$$

The estimator $\widetilde{\boldsymbol{\upsilon}} = (\widetilde{W}, \widetilde{\mathbb{X}}, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{z}})$ and the background truth $\boldsymbol{\upsilon}^* = (W^*, \mathbb{X}^*, \boldsymbol{\beta}^*, \boldsymbol{z}^*)$:

$$\widetilde{\boldsymbol{\upsilon}} = \operatorname*{argmin}_{\boldsymbol{\upsilon}} \mathscr{L}(\boldsymbol{\upsilon}), \qquad \boldsymbol{\upsilon}^* = \operatorname*{argmin}_{\boldsymbol{\upsilon}} \mathbb{E}\mathscr{L}(\boldsymbol{\upsilon}).$$

The main results describe the error of estimation $\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*$.

For $\boldsymbol{\upsilon} = (W, \mathbb{X}, \boldsymbol{\beta}, \boldsymbol{z})$, consider

$$\mathscr{L}(\boldsymbol{\upsilon}) = \frac{\mu^2}{2}\|\boldsymbol{Z} - \boldsymbol{z}\|^2 + \frac{1}{2}\|\boldsymbol{z} - \mathcal{S}\mathbb{X}^\top\boldsymbol{\beta}\|^2 + \frac{1}{2}\|\mathbb{X} - \sigma(W\boldsymbol{X})\|^2$$
$$+ \frac{\lambda_{\mathbb{W}}}{2}\|W\|^2 + \frac{\lambda_{\mathbb{X}}}{2}\|\mathbb{X}\|^2 + \frac{\lambda_{\boldsymbol{\beta}}}{2}\|\boldsymbol{\beta}\|^2 + \frac{\lambda_{\boldsymbol{z}}}{2}\|\boldsymbol{z}\|^2 .$$

Sufficient condition for local convexity: for $\boldsymbol{\upsilon} \in U(\boldsymbol{\upsilon}^*)$:

$$\mathscr{F}(\boldsymbol{\upsilon}) = \nabla^2\mathscr{L}(\boldsymbol{\upsilon}) > 0.$$

$$\mathscr{L}(\boldsymbol{\upsilon}) = \frac{\mu^2}{2}\|\boldsymbol{Z} - \boldsymbol{z}\|^2 + \frac{1}{2}\|\boldsymbol{z} - \mathcal{S}\mathbb{X}^\top\boldsymbol{\beta}\|^2 + \frac{1}{2}\|\mathbb{X} - \sigma(W\boldsymbol{X})\|^2$$
$$+ \frac{\lambda_{\mathbb{W}}}{2}\|W\|^2 + \frac{\lambda_{\mathbb{X}}}{2}\|\mathbb{X}\|^2 + \frac{\lambda_{\boldsymbol{\beta}}}{2}\|\boldsymbol{\beta}\|^2 + \frac{\lambda_{\boldsymbol{z}}}{2}\|\boldsymbol{z}\|^2 .$$

Lemma

Fix $\boldsymbol{\upsilon} = (W, \mathbb{X}, \boldsymbol{\beta}, \boldsymbol{z})$ *with* $\mathbb{X} = \sigma(W\boldsymbol{X})$ *and* $\boldsymbol{z} = \mathcal{S}\mathbb{X}^\top\boldsymbol{\beta}$. *Then for any* $\boldsymbol{u} = (\mathbb{w}, \mathbb{x}, \boldsymbol{b}, \boldsymbol{h})$

$$\frac{d^2}{dt^2}\mathscr{L}(\boldsymbol{\upsilon} + t\boldsymbol{u})\Big|_{t=0} = (\mu^2 + \lambda_{\boldsymbol{z}})\|\boldsymbol{h}\|^2 + \lambda_{\mathbb{X}}\|\mathbb{x}\|^2 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{b}\|^2 + \lambda_{\mathbb{W}}\|\mathbb{w}\|^2$$
$$+ \sum_{i=1}^{n}\sum_{m=1}^{M}\left\{\sigma'(W_m\boldsymbol{X}_i)\,\mathbb{w}_m\boldsymbol{X}_i - \mathbb{x}_{mi}\right\}^2 + \|\boldsymbol{h} - \mathcal{S}(\mathbb{X}^\top\boldsymbol{b} + \mathbb{x}^\top\boldsymbol{\beta})\|^2 .$$

Define

$$I\!F_m(\boldsymbol{\upsilon}) \;=\; I\!F_m(W_m) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \sigma'\big(W_m \boldsymbol{X}_i\big)^2 \, \boldsymbol{X}_i \, \boldsymbol{X}_i^\top + \frac{4}{3}\lambda_{\mathbb{w}} I\!I_d \,,$$

$$\mathscr{F}_{\mathbb{ww}}(\boldsymbol{\upsilon}) = \mathscr{F}_{\mathbb{ww}}(W) = \mathrm{block}\{I\!F_1(W_1), \ldots, I\!F_M(W_M)\} \,.$$

Lemma

Let $\mu^2 \geq 3$, $\lambda_{\mathbb{x}} \geq 4 + \frac{3b_0^2}{2}$. *For* $\boldsymbol{\upsilon} = (W, \mathbb{X}, \boldsymbol{\beta}, \boldsymbol{z})$ *with* $\mathbb{X} = \sigma(W\boldsymbol{X})$, $\boldsymbol{z} = \mathcal{S}\mathbb{X}^\top\boldsymbol{\beta}$, *and* $\|\boldsymbol{\beta}\| \leq b_0$

$$\mathscr{F}(\boldsymbol{\upsilon}) \geq \mathrm{block}\Big\{ \frac{3}{4}\mathscr{F}_{\mathbb{ww}} \,,\; \Big(\lambda_{\mathbb{x}} - 3 - \frac{3b_0^2}{2}\Big) I\!I_{\mathbb{x}} \,, \\ \frac{1}{2}\mathbb{X}\boldsymbol{A}\mathbb{X}^\top + \lambda_{\boldsymbol{\beta}} I\!I_{\boldsymbol{\beta}} \,,\; (\mu^2 - 3) I\!I_{\boldsymbol{z}} \Big\} \,.$$

Introduce a full dimensional metric tensor $\mathcal{D}$ by

$$\mathcal{D}^2 \stackrel{\text{def}}{=} \operatorname{block}\bigl\{ D^2_{\mathbb{w}} \,,\; \mathbb{I}_{\mathbb{x}} \,,\; D^2_{\boldsymbol{\beta}} \,,\; \mathbb{I}_{\boldsymbol{z}} \bigr\},$$

$$D^2_{\mathbb{w}} \stackrel{\text{def}}{=} \frac{3}{4} \operatorname{block}\{\mathbb{F}_1, \ldots, \mathbb{F}_M\}\,, \quad D^2_{\boldsymbol{\beta}} \stackrel{\text{def}}{=} \frac{1}{2} \mathbb{X}^* \boldsymbol{A}\, \mathbb{X}^{*\top} + \lambda_{\boldsymbol{\beta}} \mathbb{I}_{\boldsymbol{\beta}}\,,$$

$$\mathbb{F}_m(W_m) = \sum_{i=1}^{n} \sigma'\bigl(W_m \boldsymbol{X}_i\bigr)^2 \boldsymbol{X}_i\, \boldsymbol{X}_i^\top + \frac{4}{3} \lambda_{\mathbb{w}} \mathbb{I}_d.$$

With $\sigma(t) = \alpha^{-1} \log(1 + \mathrm{e}^{\alpha t})$, define

$$N^{-1/2} = \max_{m \le M,\, i \le n} \left\{ \frac{\sigma''(W^*_m \boldsymbol{X}_i)}{\sigma'(W^*_m \boldsymbol{X}_i)} \, \| \mathbb{F}_m^{-1/2} \boldsymbol{X}_i \| \right\}, \quad N_1^{-1/2} = \| D_{\boldsymbol{\beta}}^{-1} \|\,.$$

Under $\|\boldsymbol{\beta}\| \le b_0$, $\mathbb{X} \boldsymbol{A} \mathbb{X}^\top \le 2 \mathbb{X}^* \boldsymbol{A} \mathbb{X}^{*\top}$, smoothness cond. holds with

$$\tau_3 \stackrel{\text{def}}{=} 6\sqrt{\mathrm{e}} N^{-1/2} + 6(b_0 + 1) N_1^{-1/2}\,.$$

For the score vector $\nabla\zeta$, it holds $\nabla\zeta = (0, 0, 0, \nabla_{\boldsymbol{z}}\zeta)$ with

$$\nabla_{\boldsymbol{z}}\zeta = \boldsymbol{Z} - I\!\!E\boldsymbol{Z}\,, \quad \mathrm{Var}(\nabla_{\boldsymbol{z}}\zeta) = \mathrm{Var}(\mathcal{S}\boldsymbol{Y}) = \mathcal{S}\,\mathrm{Var}(\boldsymbol{Y})\mathcal{S}^{\top}\,.$$

With $\mathscr{F} = \mathscr{F}(\boldsymbol{\upsilon}^*) = \nabla^2\mathscr{L}(\boldsymbol{\upsilon}^*)$, the effective dimension

$$\mathbb{p} \stackrel{\mathrm{def}}{=} \mathrm{tr}\bigl\{\mathscr{F}^{-1}\,\mathrm{Var}(\nabla\zeta)\bigr\}\,.$$

Then $\mathscr{F}^{-1} \leq \mathcal{D}^{-2}$ and

$$\|\mathcal{D}\mathscr{F}^{-1}\nabla\zeta\| \leq \|\mathcal{D}^{-1}\nabla\zeta\| = \|\mathcal{S}\boldsymbol{\varepsilon}\|\,,$$

$$\mathbb{p} \leq \mathrm{tr}\bigl\{\mathcal{D}^{-2}\,\mathrm{Var}(\nabla\zeta)\bigr\} \leq I\!\!E\|\mathcal{S}\boldsymbol{\varepsilon}\|^2\,.$$

The effective radius $\mathtt{r}_{\mathcal{D}}$ fulfills with $\mathbb{V}^2 \stackrel{\mathrm{def}}{=} \mathrm{Var}(\mathcal{S}\boldsymbol{\varepsilon})$

$$\mathtt{r}_{\mathcal{D}} \leq z(\mathbb{V}^2, \mathtt{x}) \leq \sqrt{\mathrm{tr}\,\mathbb{V}^2} + \sqrt{2\mathtt{x}\|\mathbb{V}^2\|} \approx \sqrt{\mathbb{p}}\,.$$

Identifiability of the product $\mathbb{X}^\top \boldsymbol{\beta}$ after restricting to a bounded set

$$\Upsilon^\circ \stackrel{\text{def}}{=} \{\boldsymbol{\upsilon} = (W, \mathbb{X}, \boldsymbol{\beta}, \boldsymbol{z}) \colon \|\boldsymbol{\beta}\| \le b_0\,,\ \mathbb{X}\boldsymbol{A}\mathbb{X}^\top \le 2\mathbb{X}^*\boldsymbol{A}\mathbb{X}^{*\top}\}\,.$$

Here $\boldsymbol{A} = \mathcal{S}^\top\mathcal{S}$ and b_0 is a fixed constant, e.g. $b_0 = 2$.

Theorem

Let $\|\mathcal{S}^\top\mathcal{S}\|_{\mathrm{op}} \le 1$. *If* $\lambda_{\boldsymbol{z}} \ge 3$ *and* $\lambda_{\mathbb{X}} \ge 4 + 3b_0^2/2$, *let* $\mathtt{r}_{\mathcal{D}}\,\tau_3 \le 4/9$. *For any* $\mathcal{Q}\colon \mathbb{R}^{\bar{p}} \to \mathbb{R}^q$, *on a random set* $\Omega(\mathtt{x})$ *with* $\mathbb{P}(\Omega(\mathtt{x})) \ge 1 - \mathrm{e}^{-\mathtt{x}}$

$$\|\mathcal{Q}(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^* + \mathscr{F}^{-1}\nabla\zeta)\| \le \|\mathcal{Q}\,\mathcal{D}^{-1}\|\frac{3\tau_3}{4}\|\mathcal{S}\boldsymbol{\varepsilon}\|^2\,.$$

where $\mathcal{D}^2 = \mathrm{block}\{D^2_{\mathtt{w}}\,,\ \mathbb{I}_{\mathbb{X}}\,,\ D^2_{\boldsymbol{\beta}}\,,\ \mathbb{I}_{\boldsymbol{z}}\}\,,$

$$D^2_{\mathtt{w}} \stackrel{\text{def}}{=} \frac{3}{4}\,\mathrm{block}\{\mathbb{F}_1, \ldots, \mathbb{F}_M\}\,, \quad D^2_{\boldsymbol{\beta}} \stackrel{\text{def}}{=} \frac{1}{2}\mathbb{X}^*\boldsymbol{A}\,\mathbb{X}^{*\top} + \lambda_{\boldsymbol{\beta}}\mathbb{I}_{\boldsymbol{\beta}}\,.$$

Theorem

Introduce

$$\mathbb{p}_{\mathcal{Q}} \stackrel{\mathrm{def}}{=} \mathbb{E}\{\|\mathcal{Q}\mathscr{F}^{-1}\nabla\zeta\|^2\,\mathbb{1}_{\Omega(\mathtt{x})}\}\,,$$

assume $\mathbb{E}\{\|\mathcal{S}\boldsymbol{\varepsilon}\|^4\,\mathbb{1}_{\Omega(\mathtt{x})}\} \leq \mathtt{C}_4^2\,\mathbb{p}^2$ *, and define*

$$\alpha_{\mathcal{Q}} \stackrel{\mathrm{def}}{=} \frac{\|\mathcal{Q}\mathcal{D}^{-1}\|\,(3/4)\tau_3\,\mathtt{C}_4\,\mathbb{p}}{\sqrt{\mathbb{p}_{\mathcal{Q}}}}\,.$$

If $\alpha_{\mathcal{Q}} < 1$ *then*

$$(1-\alpha_{\mathcal{Q}})^2\mathbb{p}_{\mathcal{Q}} \leq \mathbb{E}\{\|\mathcal{Q}\,(\widetilde{\boldsymbol{v}}-\boldsymbol{v}^*)\|^2\,\mathbb{1}_{\Omega(\mathtt{x})}\} \leq (1+\alpha_{\mathcal{Q}})^2\mathbb{p}_{\mathcal{Q}}\,.$$

Denote $\mathcal{I} = \mathscr{F}^{-1}$. Then

$$\mathscr{F}^{-1} \nabla \zeta = (\mathcal{I}_{\mathbb{w}z} \mathcal{S}\varepsilon, \mathcal{I}_{\mathbb{x}z} \mathcal{S}\varepsilon, \mathcal{I}_{\boldsymbol{\beta} z} \mathcal{S}\varepsilon, \mathcal{I}_{zz} \mathcal{S}\varepsilon) \,,$$

$$D^2_{\mathbb{w}} \stackrel{\text{def}}{=} \frac{3}{4} \operatorname{block}\{\mathbb{F}_1, \ldots, \mathbb{F}_M\} \,, \quad D^2_{\boldsymbol{\beta}} \stackrel{\text{def}}{=} \frac{1}{2} \mathbb{X}^* \boldsymbol{A} \, \mathbb{X}^{*\top} + \lambda_{\boldsymbol{\beta}} \mathbb{I}_{\boldsymbol{\beta}} \,.$$

Corollary

On $\Omega(\mathrm{x})$, *with* $\mathcal{I} = \mathscr{F}^{-1}$,

$$\begin{aligned}
\| D_{\mathbb{w}} (\widetilde{W} - W^* + \mathcal{I}_{\mathbb{w}z} \mathcal{S}\varepsilon) \| &\leq (3/4) \tau_3 \, \| \mathcal{S}\varepsilon \|^2 \,, \\
\| \widetilde{\mathbb{X}} - \mathbb{X}^* + \mathcal{I}_{\mathbb{x}z} \mathcal{S}\varepsilon) \| &\leq (3/4) \tau_3 \, \| \mathcal{S}\varepsilon \|^2 \,, \\
\| D_{\boldsymbol{\beta}} (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* + \mathcal{I}_{\boldsymbol{\beta} z} \mathcal{S}\varepsilon) \| &\leq (3/4) \tau_3 \, \| \mathcal{S}\varepsilon \|^2 \,, \\
\| \widetilde{\boldsymbol{z}} - \boldsymbol{z}^* + \mathcal{I}_{zz} \mathcal{S}\varepsilon) \| &\leq (3/4) \tau_3 \, \| \mathcal{S}\varepsilon \|^2
\end{aligned}$$

Construction can be extended to a K-layer network using recurrence

$$\mathbb{X}^{(k)} = \sigma(W^{(k)}\mathbb{X}^{(k-1)})$$

for $k = 1, \ldots, K$ and $\mathbb{X}^{(0)} = \boldsymbol{X}$. With $\mathbb{X} = (\mathbb{X}^{(1)}, \ldots, \mathbb{X}^{(K)})$ and $W = (W^{(1)}, \ldots, W^{(K)})$, this leads to the log-likelihood

$$\begin{aligned}\mathscr{L}(\boldsymbol{\upsilon}) = \mathscr{L}(W, \mathbb{X}, \boldsymbol{\beta}, \boldsymbol{z}) = {} & \frac{\mu^2}{2}\|\boldsymbol{Z} - \boldsymbol{z}\|^2 \\ & + \frac{1}{2}\|\boldsymbol{z} - \mathcal{S}\mathbb{X}^{(K)\top}\boldsymbol{\beta}\|^2 + \frac{1}{2}\sum_{k=1}^{K}\|\mathbb{X}^{(k)} - \sigma(W^{(k)}\mathbb{X}^{(k-1)})\|^2 \\ & + \frac{\lambda_{\mathbb{X},K}}{2}\|\mathbb{x}^{(K)}\|^2 + \frac{\lambda_{\boldsymbol{\beta}}}{2}\|\boldsymbol{\beta}\|^2 + \frac{\lambda_{\boldsymbol{z}}}{2}\|\boldsymbol{z}\|^2 + \frac{\lambda_{\mathbb{W}}}{2}\|\mathbb{W}\|^2 + \frac{\lambda_{\mathbb{X}}}{2}\|\mathbb{X}\|^2 .\end{aligned}$$

For any $\boldsymbol{u} = (\mathbb{w}, \mathbb{x}, \boldsymbol{b}, \boldsymbol{h})$ with $\mathbb{w} = (\mathbb{w}^{(k)}) \in \mathbb{R}^{n_{\mathbb{w}}}$, $\mathbb{x} = (\mathbb{x}^{(k)}) \in \mathbb{R}^{n \times d}$, $\boldsymbol{b} \in \mathbb{R}^{M_K}$, and $\boldsymbol{h} \in \mathbb{R}^q$, it holds

$$\frac{d^2}{dt^2} \mathscr{L}(\boldsymbol{v} + t\boldsymbol{u})\Big|_{t=0} = \lambda_{\mathbb{x}} \|\mathbb{x}\|^2 + \lambda_{\mathbb{w}} \|\mathbb{w}\|^2 + (1 + \lambda_{\boldsymbol{h}}) \|\boldsymbol{h}\|^2 + \lambda_{\boldsymbol{\beta}} \|\boldsymbol{b}\|^2$$

$$+ \sum_{k=1}^{K} \sum_{m=1}^{M_k} \sum_{i=1}^{n} \left\{ \sigma'(W_m^{(k)} \mathbb{X}_i^{(k-1)}) \left(\mathbb{w}_m^{(k)} \mathbb{X}_i^{(k-1)} + W_m^{(k)} \mathbb{x}_i^{(k-1)} \right) - \mathbb{x}_{mi}^{(k)} \right\}^2$$

$$+ \left\| \boldsymbol{h} - \mathcal{S}(\mathbb{X}^{(K)\top} \boldsymbol{b} + \mathbb{x}^{(K)\top} \boldsymbol{\beta}) \right\|^2 + \lambda_{\mathbb{x},K} \|\mathbb{x}^{(K)}\|^2 .$$

With $\upsilon = (W^{(1)}, \mathbb{X}^{(1)}, \ldots, W^{(K)}, \mathbb{X}^{(K)})$ and $\upsilon^{(k)} = (W^{(k)}, \mathbb{X}^{(k-1)})$, define for $k \le K$ and $m \le M_k$

$$\mathbb{F}_m^{(k)}(\upsilon^{(k)}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \sigma'\bigl(W_m^{(k)} \mathbb{X}_i^{(k-1)}\bigr)^2 \, \mathbb{X}_i^{(k-1)} \, \mathbb{X}_i^{(k-1)\top} + \lambda_k \mathbb{I}_{M_k} ,$$

$$\mathbb{F}^{(k)}(\upsilon^{(k)}) \stackrel{\text{def}}{=} \operatorname{block}\bigl\{ \mathbb{F}_1^{(k)}(\upsilon^{(k)}), \ldots, \mathbb{F}_{M_k}^{(k)}(\upsilon^{(k)}) \bigr\} .$$

Similarly define for $k = 2, \ldots, K$ and $i = 1, \ldots, n$

$$\mathbb{G}_i^{(k)}(\upsilon^{(k)}) \stackrel{\text{def}}{=} \sum_{m=1}^{M_k} \sigma'\bigl(W_m^{(k)} \mathbb{X}_i^{(k-1)}\bigr)^2 \, W_m^{(k)\top} W_m^{(k)} ,$$

$$\mathbb{G}^{(k)}(\upsilon^{(k)}) \stackrel{\text{def}}{=} \operatorname{block}\bigl\{ \mathbb{G}_1^{(k)}(\upsilon^{(k)}), \ldots, \mathbb{G}_n^{(k)}(\upsilon^{(k)}) \bigr\} .$$

Also set $\mathbb{G}^{(1)}(\upsilon^{(1)}) \equiv 0$.

Lemma

Let $\mathbb{X}^{(k)} \equiv \sigma\big(W^{(k)}\mathbb{X}^{(k-1)}\big)\,,\ k = 1, \ldots, K$ *. Also, assume* $\|\boldsymbol{\beta}\| \leq b_0$ *and*

$$\max_{i=1,\ldots,n} \max_{k=1,\ldots,K} \frac{3}{2} \|\mathbb{G}_i^{(k)}(\boldsymbol{v}^{(k)})\|_{\mathrm{op}} \leq G_0 \leq \lambda_{\mathbb{X}} - 4\,.$$

Then it holds

$$\mathscr{F}(\boldsymbol{v}) \geq \mathrm{block}\Big\{\frac{1}{2}\mathbb{F}^{(1)}(\boldsymbol{v}^{(1)}),\ \mathbb{I}_{\mathbb{X}^{(1)}}, \ldots, \frac{1}{2}\mathbb{F}^{(K)}(\boldsymbol{v}^{(K)}),\ \mathbb{I}_{\mathbb{X}^{(K)}},$$
$$\frac{1}{2}\,\mathbb{X}^{(K)}\boldsymbol{A}\mathbb{X}^{(K)^\top} + \lambda_{\boldsymbol{\beta}}\mathbb{I}_{\boldsymbol{\beta}}\,,\ (\lambda_{\boldsymbol{z}} - 2)\mathbb{I}_{\boldsymbol{z}}\Big\}\,.$$

- (Parametric) lower bounds. Please, ask about
- LASSO-type procedures. Issue $\|\boldsymbol{v}\|_1$ is non-smooth.
  Idea: replace by a smooth penalty ensuring $\mathbb{p} \asymp \#(\mathrm{active\,set})$.
- Other models including
  - dimension reduction, manifold learning
  - density estimation
  - error-in-operator models
  - matrix completion
  - inference for stochastic processes
  - statistical nonlinear inverse problems, PDEs
  - ...
- adaptation, parameter/penalty choice
- higher-order expansions, estimation of smooth functionals.

References

Banach, S. (1938).

Über homogene Polynome in (L^2).

Studia Mathematica, 7(1):36–44.

Chen, M., Jiang, H., Liao, W., and Zhao, T. (2019).

Efficient approximation of deep relu networks for functions on low dimensional manifolds.

ArXiv, abs/1908.01842:null.

Cloninger, A. and Klock, T. (2021).

A deep network construction that adapts to intrinsic dimensionality beyond the domain.

Neural Networks, 141:404–419.

Fan, J., Gu, Y., and Zhou, W.-X. (2024).

How do noise tails impact on deep ReLU networks?

The Annals of Statistics, 52(4):1845–1871.

Jiao, Y., Shen, G., Lin, Y., and Huang, J. (2021).

Deep nonparametric regression on approximately low-dimensional manifolds.

arXiv: Statistics Theory, page null.

Jiao, Y., Shen, G., Lin, Y., and Huang, J. (2023).

Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors.

The Annals of Statistics, 51:691–716.

Kohler, M. and Krzyzak, A. (2022).

Analysis of the rate of convergence of an over-parametrized deep neural network estimate learned by gradient descent.

References

Kohler, M., Krzyżak, A., and Langer, S. (2019).
Estimation of a function of low local dimensionality by deep neural networks.
IEEE Transactions on Information Theory, 68:4032–4042.

Liu, H., Chen, M., Er, S., Liao, W., Zhang, T.-M., and Zhao, T. (2022).
Benefits of overparameterized convolutional residual networks: Function approximation under smoothness constraint.
ArXiv, abs/2206.04569:null.

Liu, H., Chen, M., Zhao, T., and Liao, W. (2021).
Besov function approximation and binary classification on low-dimensional manifolds using convolutional residual networks.

Puchkin, N., Noskov, F., and Spokoiny, V. (2025).
Sharper dimension-free bounds on the Frobenius distance between sample covariance and its expectation.
Bernoulli, 31(2):1664–1691.
https://arxiv.org/abs/2308.14739.

Schmidt-Hieber, J. (2020).
Nonparametric regression using deep neural networks with ReLU activation function.
The Annals of Statistics, 48(4):1875–1897.

Shen, G., Jiao, Y., Lin, Y., and Huang, J. (2022).
Approximation with cnns in sobolev space: with applications to classification.

Shen, G., Jiao, Y., Lin, Y., and Huang, J. (2023).
Differentiable neural networks with repu activation: with applications to score estimation and isotonic regression.
ArXiv, abs/2305.00608:null.

References

Simionescu-Badea, C. (2022).
Analysis of the rate of convergence of fully connected deep neural network regression estimates with smooth activation function.

Spokoiny, V. (2024).
Sharp deviation bounds and concentration phenomenon for the squared norm of a sub-gaussian vector.
https://arxiv.org/abs/2305.07885.

Spokoiny, V. (2025a).
Deviation bounds for the norm of a random vector under exponential moment conditions with applications.
Probability, Uncertainty and Quantitative Risk, 10(1):135–158.
https://arxiv.org/abs/2309.02302v1.

Spokoiny, V. (2025b).
Marginal minimization and sup-norm expansions in perturbed optimization.
arxiv.org/2505.02562.

Spokoiny, V. (2025c).
Sharp bounds in perturbed smooth optimization.
arxiv.org/2504.11834.

Zhang, Z., Chen, M., Wang, M., Liao, W., and Zhao, T. (2023).
Effective minkowski dimension of deep nonparametric regression: Function approximation and statistical theories.
ArXiv, abs/2306.14859:null.

Zuowei Shen, Z. S., Haizhao Yang, H. Y., and Shijun Zhang, S. Z. (2020).
Deep network approximation characterized by number of neurons.
Communications in Computational Physics, 28(5):1768–1811.