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

3. September 2025

Which dimension?

Which dimension?

Number of features d?

Which dimension?

Number of features d?

or

Which dimension?

Number of features d?

or

number of parameters p of the model?

Coincide for linear models.

- Full parameter dimension p does not enter the accuracy guarantees;
- what matters?
  - effective dimension $\mathbb{p}$;
  - geometric properties (like strong convexity) of the objective function;
- A proper extension of the parameter space can be useful and helpful to improve geometric properties of the objective function without increasing the effective dimension.

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

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

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

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

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

where $M = M^{(K)}$ 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 $\mathbb{X}$:

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

Model (shallow): $\mathrm{DNN}(\boldsymbol{X}_i \,|\, W) = \sigma(W\boldsymbol{X}_i)$.

Model (deep): $W = (W^{(1)}, \ldots, W^{(K)})$,

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

Method (MLE):

$$m(\boldsymbol{X}_i \,|\, W, \boldsymbol{\beta}) = \mathrm{DNN}(X_i \,|\, W)^\top \boldsymbol{\beta},$$

$$L(W, \boldsymbol{\beta}) = \sum_{i=1}^{n} \bigl| Y_i - m(\boldsymbol{X}_i \,|\, W, \boldsymbol{\beta}) \bigr|^2,$$

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

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

- [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,

- [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

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

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

Includes MLE, LSE, LAD, Minimum Contrast, and many others procedures.

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

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

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

- **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** [Wilks, 1938] (requires $\mathbb{p} \ll \mathbb{N}$)

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

[Spokoiny, 2024].

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

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

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

Consider three optimization problems

$$\begin{aligned} \widetilde{\boldsymbol{v}}_G &= \operatorname*{argmin}_{\boldsymbol{v}} L_G(\boldsymbol{v}), \\ \boldsymbol{v}_G^* &= \operatorname*{argmin}_{\boldsymbol{v}} \mathbb{E} L_G(\boldsymbol{v}), \\ \boldsymbol{v}^* &= \operatorname*{argmin}_{\boldsymbol{v}} \mathbb{E} L(\boldsymbol{v}). \end{aligned}$$

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

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

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

Proposition

Define

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

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

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 $Q \colon \mathbb{R}^p \to \mathbb{R}^q$

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

$$\mathsf{b}_Q = \| Q \mathbb{F}_G^{-1} \boldsymbol{M}_G \| = \| Q \mathbb{F}_G^{-1} \nabla \operatorname{pen}_G(\boldsymbol{\upsilon}^*) \| \qquad \text{bias}$$

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

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

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

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

[Spokoiny, 2024].

- Stochastic linearity: the stochastic component $\zeta(\boldsymbol{v}) = L(\boldsymbol{v}) - \mathbb{E}L(\boldsymbol{v})$ is linear in $\boldsymbol{v}$ or $\nabla\zeta(\boldsymbol{v}) \equiv \nabla\zeta$;
- Sub-gaussian tails for $\nabla\zeta$;
- Convexity: $\mathbb{E}L_G(\boldsymbol{v})$ is strongly convex in $\boldsymbol{v}$;
- Smoothness: $\mathbb{E}L_G(\boldsymbol{v})$ 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 $\mathbb{p} \ll \mathbb{N}$.
- 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, 2025b, Spokoiny, 2025a].

Let $f(\boldsymbol{v})$ be a smooth convex function,

$$\boldsymbol{v}^* = \operatorname*{argmin}_{\boldsymbol{v}} f(\boldsymbol{v}), \quad \mathbb{F} = \nabla^2 f(\boldsymbol{v}^*).$$

Let another function $g(\boldsymbol{v})$ satisfy for some vector $\boldsymbol{A}$

$$g(\boldsymbol{v}) - g(\boldsymbol{v}^*) = \bigl\langle \boldsymbol{v} - \boldsymbol{v}^*, \boldsymbol{A} \bigr\rangle + f(\boldsymbol{v}) - f(\boldsymbol{v}^*).$$

Define

$$\boldsymbol{v}^\circ \stackrel{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{v}} g(\boldsymbol{v}), \qquad g(\boldsymbol{v}^\circ) = \min_{\boldsymbol{v}} g(\boldsymbol{v}).$$

Aim: evaluate the quantities $\boldsymbol{v}^\circ - \boldsymbol{v}^*$ and $g(\boldsymbol{v}^\circ) - g(\boldsymbol{v}^*)$.

Consider $\boldsymbol{\upsilon}^{*} \stackrel{\text{def}}{=} \operatorname{argmin} f(\boldsymbol{\upsilon})\,, \boldsymbol{\upsilon}^{\circ} \stackrel{\text{def}}{=} \operatorname{argmin}\{f(\boldsymbol{\upsilon}) + \langle \boldsymbol{\upsilon}, \boldsymbol{A} \rangle\}$.

$(\boldsymbol{\mathcal{T}_3^{*}})$ $f(\boldsymbol{\upsilon})$ *is strongly convex,* $D^2 \leq \mathbb{F} \stackrel{\text{def}}{=} \nabla^2 f(\boldsymbol{\upsilon}^{*})$ *, and*

$$\sup_{\boldsymbol{u}\colon \|D\boldsymbol{u}\| \leq \mathtt{r}} \sup_{\boldsymbol{z} \in \mathbb{R}^p} \frac{\bigl|\langle \nabla^3 f(\boldsymbol{\upsilon}^{*} + \boldsymbol{u}), \boldsymbol{z}^{\otimes k} \rangle\bigr|}{\|D\boldsymbol{z}\|^3} \leq \tau_3\,.$$

Proposition ([Spokoiny, 2025b, Spokoiny, 2025a])

Assume $(\boldsymbol{\mathcal{T}_3^{*}})$ *at* $\boldsymbol{\upsilon}^{*}$ *with* D^2 *,* $\mathtt{r}$ *, and* τ_3 *s.t.*

$$\mathtt{r} \geq \frac{3}{2}\|D\,\mathbb{F}^{-1}\boldsymbol{A}\|\,, \quad \tau_3\|D\,\mathbb{F}^{-1}\boldsymbol{A}\| < \frac{4}{9}\,.$$

Then $\|D(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*})\| \leq (3/2)\|D\,\mathbb{F}^{-1}\boldsymbol{A}\|$ *and moreover,*

$$\|D^{-1}\mathbb{F}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*} - \mathbb{F}^{-1}\boldsymbol{A})\| \leq \frac{3}{4}\,\tau_3\|D\,\mathbb{F}^{-1}\boldsymbol{A}\|^2\,.$$

Model:

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

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.

$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top$ – observations,
$m(\boldsymbol{v}) = \big(m(\boldsymbol{X}_1, \boldsymbol{v}), \ldots, m(\boldsymbol{X}_n, \boldsymbol{v})\big)^\top$, regression function.

Nonlinear least squares (Gauss, Legendre):

$$\widehat{\boldsymbol{v}} = \operatorname*{argmin}_{\boldsymbol{v}} L(\boldsymbol{v}) = \operatorname*{argmin}_{\boldsymbol{v}} \|\boldsymbol{Y} - m(\boldsymbol{v})\|^2 .$$

Hessian

$$\nabla^2 L(\boldsymbol{v}) = \nabla m(\boldsymbol{v}) \nabla m(\boldsymbol{v})^\top - \{\boldsymbol{Y} - m(\boldsymbol{v})\}^\top \nabla^2 m(\boldsymbol{v}) .$$

Noiseless case: $\boldsymbol{Y} = m(\boldsymbol{v}^*)$, $\boldsymbol{v} \approx \boldsymbol{v}^*$

$$\nabla^2 L(\boldsymbol{v}) \approx \nabla m(\boldsymbol{v}) \nabla m(\boldsymbol{v})^\top .$$

Gauss–Newton iterations; [Gratton et al., 2007]. Requires “small noise”.

Original problem

$$L(W,\boldsymbol{\beta}) = \frac{1}{2}\big\|\boldsymbol{Y} - \sigma(W\boldsymbol{X})^{\top}\boldsymbol{\beta}\big\|^{2}.$$

- Introduce $\mathbb{X} = \sigma(W\boldsymbol{X})$ and $m = \mathbb{X}^{\top}\boldsymbol{\beta}$;
- Replace the structural relations 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 $\boldsymbol{\upsilon} = (W, \mathbb{X}, \boldsymbol{\beta}, m)$, results in

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

Dimension before calming:

$$p = \dim(W) + \dim(\boldsymbol{\beta}) =$$

Dimension before calming:

$$p = \dim(W) + \dim(\boldsymbol{\beta}) = dM + M \,.$$

Dimension after calming:

$$\begin{aligned} p &= \dim(W) + \dim(\mathbb{X}) + \dim(m) + \dim(\boldsymbol{\beta}) \\ &= \end{aligned}$$

Dimension before calming:

$$p = \dim(W) + \dim(\boldsymbol{\beta}) = dM + M\,.$$

Dimension after calming:

$$\begin{aligned} p &= \dim(W) + \dim(\mathbb{X}) + \dim(m) + \dim(\boldsymbol{\beta}) \\ &= dM + nM + n + M \gg n\,d + n\,. \end{aligned}$$

DNN require very good (purified) data satisfying the small noise condition.

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

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

Example: convolutional networks.

Identifiability issue: partially resolved by ridge penalties:

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

$$+ \frac{\lambda_{\mathbb{W}}}{2}\|W - W_0\|^2 + \frac{\lambda_{\mathbb{X}}}{2}\|\mathbb{X} - \mathbb{X}_0\|^2 + \frac{\lambda_{\boldsymbol{\beta}}}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + \frac{\lambda_{\boldsymbol{z}}}{2}\|\boldsymbol{z} - \boldsymbol{z}_0\|^2 .$$

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

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

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

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

$$\mathscr{L}(\boldsymbol{v}) = \mathscr{L}(W, \mathbb{X}, \boldsymbol{\beta}, \boldsymbol{z}) = \frac{\mu^2}{2} \| \boldsymbol{Z} - \boldsymbol{z} \|^2$$

$$+ \frac{1}{2} \| \boldsymbol{z} - \mathcal{S} \mathbb{X}^{(K)^\top} \boldsymbol{\beta} \|^2 + \frac{1}{2} \sum_{k=1}^{K} \| \mathbb{X}^{(k)} - \sigma(W^{(k)} \mathbb{X}^{(k-1)}) \|^2$$

$$+ \frac{\lambda_{\mathbb{W}}}{2} \| W - W_0 \|^2 + \frac{\lambda_{\mathbb{X}}}{2} \| \mathbb{X} - \mathbb{X}_0 \|^2 + \frac{\lambda_{\boldsymbol{\beta}}}{2} \| \boldsymbol{\beta} - \boldsymbol{\beta}_0 \|^2 + \frac{\lambda_{\boldsymbol{z}}}{2} \| \boldsymbol{z} - \boldsymbol{z}_0 \|^2 .$$

Parameter dimension before calming: with $M^{(0)} = d$

$$p = \sum_{k=1}^{K} \dim(W^{(k)}) + \dim(\boldsymbol{\beta}) =$$

Parameter dimension before calming: with $M^{(0)} = d$

$$p = \sum_{k=1}^{K} \dim(W^{(k)}) + \dim(\boldsymbol{\beta}) = \sum_{k=1}^{K} M^{(k-1)} M^{k} + M^{(K)} .$$

Dimension after calming: $\boldsymbol{v} = (W, \mathbb{X}, \boldsymbol{\beta}, \boldsymbol{z})$

$$p = \sum_{k=1}^{K} \bigl\{ \dim(W^{(k)}) + \dim(\mathbb{X}^{(k)}) \bigr\} + \dim(\boldsymbol{\beta}) + \dim(\boldsymbol{z})$$

$$=$$

Parameter dimension before calming: with $M^{(0)} = d$

$$p = \sum_{k=1}^{K} \dim(W^{(k)}) + \dim(\boldsymbol{\beta}) = \sum_{k=1}^{K} M^{(k-1)} M^{k} + M^{(K)} .$$

Dimension after calming: $\boldsymbol{\upsilon} = (W, \mathbb{X}, \boldsymbol{\beta}, \boldsymbol{z})$

$$p = \sum_{k=1}^{K} \bigl\{ \dim(W^{(k)}) + \dim(\mathbb{X}^{(k)}) \bigr\} + \dim(\boldsymbol{\beta}) + \dim(\boldsymbol{z})$$

$$= \sum_{k=1}^{K} \bigl\{ M^{(k-1)} M^{(k)} + n M^{(k)} \bigr\} + M^{(k)} + M^{(K)} + q .$$

For $\boldsymbol{v} = (W, \mathbb{X}, \boldsymbol{\beta}, \boldsymbol{z})$, and a starting point $\boldsymbol{v} = (W_0, \mathbb{X}_0, \boldsymbol{\beta}_0, \boldsymbol{z}_0)$ consider

$$\mathscr{L}(\boldsymbol{v}) = \frac{\mu^2}{2}\|\boldsymbol{Z} - \boldsymbol{z}\|^2 + \frac{1}{2}\|\boldsymbol{z} - \mathcal{S}\mathbb{X}^\top \boldsymbol{\beta}\|^2 + \frac{1}{2}\|\mathbb{X} - \sigma(W\boldsymbol{X})\|^2$$
$$+ \frac{\lambda_{\mathbb{W}}}{2}\|W - W_0\|^2 + \frac{\lambda_{\mathbb{X}}}{2}\|\mathbb{X} - \mathbb{X}_0\|^2 + \frac{\lambda_{\boldsymbol{\beta}}}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + \frac{\lambda_{\boldsymbol{z}}}{2}\|\boldsymbol{z} - \boldsymbol{z}_0\|^2 .$$

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

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

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

$$+ \frac{\lambda_{\mathbb{W}}}{2}\|W - W_0\|^2 + \frac{\lambda_{\mathbb{X}}}{2}\|\mathbb{X} - \mathbb{X}_0\|^2 + \frac{\lambda_{\boldsymbol{\beta}}}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + \frac{\lambda_{\boldsymbol{z}}}{2}\|\boldsymbol{z} - \boldsymbol{z}_0\|^2 .$$

Later $\lambda_{\mathbb{W}} = \lambda_{\boldsymbol{\beta}} = \lambda_{\boldsymbol{z}} = 0$ provided $\mu^2 > 3$.

Lemma

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

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

$$+ \sum_{i=1}^{n}\sum_{m=1}^{M}\bigl\{\sigma'(W_m\boldsymbol{X}_i)\,\mathbb{w}_m\boldsymbol{X}_i - \mathbb{x}_{mi}\bigr\}^2 + \|\boldsymbol{h} - \mathcal{S}(\mathbb{X}^\top\boldsymbol{b} + \mathbb{x}^\top\boldsymbol{\beta})\|^2 .$$

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

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

$$\mathscr{F}_{\mathbb{W}\mathbb{W}}(\boldsymbol{\upsilon}) = \mathscr{F}_{\mathbb{W}\mathbb{W}}(W) = \operatorname{block}\{\mathbb{F}_1(W_1), \ldots, \mathbb{F}_M(W_M)\} \,.$$

Lemma

Let $\mathbb{A} = \mathcal{S}^\top \mathcal{S}$, $\mu^2 \geq 3$, $\lambda_{\mathbb{X}} \geq \frac{1}{3} + \frac{3 b_0^2 \|\mathbb{A}\|_{\mathrm{op}}}{2}$. *Let also* $\mathbb{X} = \sigma(W\boldsymbol{X})$, $\boldsymbol{z} = \mathcal{S}\mathbb{X}^\top \boldsymbol{\beta}$, *and* $\|\boldsymbol{\beta}\| \leq b_0$

$$\mathscr{F}(\boldsymbol{\upsilon}) \geq \operatorname{block}\Bigl\{\frac{3}{4}\mathscr{F}_{\mathbb{W}\mathbb{W}} \,,\ \Bigl(\lambda_{\mathbb{X}} - \frac{1}{3} - \frac{3 b_0^2 \|\mathbb{A}\|_{\mathrm{op}}}{2}\Bigr)\mathbb{I}_{\mathbb{X}} \,, \\ \frac{1}{2}\mathbb{X}\mathbb{A}\mathbb{X}^\top + \lambda_{\boldsymbol{\beta}}\mathbb{I}_{\boldsymbol{\beta}} \,,\ (\mu^2 - 3)\mathbb{I}_{\boldsymbol{z}}\Bigr\} \,.$$

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

$$\mathcal{D}^2 \stackrel{\text{def}}{=} \operatorname{block}\{D_{\mathbb{w}}^2\,,\ I\!I_{\mathbb{x}}\,,\ D_{\boldsymbol{\beta}}^2\,,\ \mu^2 I\!I_{\boldsymbol{z}}\},$$

$$D_{\mathbb{w}}^2 \stackrel{\text{def}}{=} \frac{1}{4}\operatorname{block}\{I\!F_1,\ldots,I\!F_M\}\,,\quad D_{\boldsymbol{\beta}}^2 \stackrel{\text{def}}{=} \frac{1}{2}\mathbb{X}^*\boldsymbol{A}\,\mathbb{X}^{*\top} + \lambda_{\boldsymbol{\beta}} I\!I_{\boldsymbol{\beta}}\,,$$

$$I\!F_m(W_m) = \sum_{i=1}^{n} \sigma'\bigl(W_m \boldsymbol{X}_i\bigr)^2\, \boldsymbol{X}_i\, \boldsymbol{X}_i^\top + 4\lambda_{\mathbb{w}} I\!I_d.$$

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

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

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

$$\nabla_{\boldsymbol{z}}\zeta = \mu^2(\boldsymbol{Z} - \mathbb{E}\boldsymbol{Z}), \quad \operatorname{Var}(\nabla_{\boldsymbol{z}}\zeta) = \operatorname{Var}(\mathcal{S}\boldsymbol{Y}) = \mu^2 \operatorname{Var}(\mu\mathcal{S}\boldsymbol{\varepsilon}).$$

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

$$\mathbb{p} \stackrel{\text{def}}{=} \operatorname{tr}\{\mathscr{F}^{-1}\operatorname{Var}(\nabla\zeta)\}.$$

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

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

$$\mathbb{p} \le \operatorname{tr}\{\mathcal{D}^{-2}\operatorname{Var}(\nabla\zeta)\} \le \mathbb{E}\|\mu\mathcal{S}\boldsymbol{\varepsilon}\|^2.$$

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

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

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

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

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

Theorem

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

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

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

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

Theorem

Introduce

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

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

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

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

$$(1-\alpha_{\mathcal{Q}})^2\mathbb{p}_{\mathcal{Q}} \le \mathbb{E}\Big\{\|\mathcal{Q}\,(\tilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*)\|^2\, \mathbb{1}_{\Omega(\mathtt{x})}\Big\} \le (1+\alpha_{\mathcal{Q}})^2\mathbb{p}_{\mathcal{Q}}\,.$$

Denote $\mathcal{I} = \mu \mathscr{F}^{-1}$ and $\boldsymbol{\xi} = \mu \mathcal{S} \boldsymbol{\varepsilon}$. Then

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

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

Corollary

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

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

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

$$\mathscr{L}(\boldsymbol{v}) = \mathscr{L}(W, \mathbb{X}, \boldsymbol{\beta}, \boldsymbol{z}) = \frac{\mu^2}{2} \|\boldsymbol{Z} - \boldsymbol{z}\|^2$$

$$+ \frac{1}{2} \|\boldsymbol{z} - \mathcal{S} {\mathbb{X}^{(K)}}^{\top} \boldsymbol{\beta}\|^2 + \frac{1}{2} \sum_{k=1}^{K} \|\mathbb{X}^{(k)} - \sigma(W^{(k)} \mathbb{X}^{(k-1)})\|^2$$

$$+ \frac{\lambda_{\mathbb{X},K}}{2} \sum_{k=1}^{K} \|\mathbb{X}^{(k)} - \mathbb{X}_0^{(k)}\|^2 + \frac{\lambda_{\mathbb{W}}}{2} \sum_{k=1}^{K} \|W^{(k)} - W_0^{(k)}\|^2$$

$$+ \frac{\lambda_{\boldsymbol{\beta}}}{2} \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + \frac{\lambda_{\boldsymbol{z}}}{2} \|\boldsymbol{z} - \boldsymbol{z}_0\|^2 .$$

Lemma

Fix $\boldsymbol{\upsilon} = (W, \mathbb{X}, \boldsymbol{\beta}, \boldsymbol{z})$ *with* $\mathbb{X}^{(k)} = \sigma(W^{(k)} \boldsymbol{X}^{(k-1)})$, $\boldsymbol{z} = \mathcal{S} \mathbb{X}^{(K)^\top} \boldsymbol{\beta}$.
For any $\boldsymbol{u} = (\mathbbm{w}, \mathbbm{x}, \boldsymbol{b}, \boldsymbol{h})$ *with* $\mathbbm{w} = (\mathbbm{w}^{(k)}) \in \mathbb{R}^{n_{\mathbbm{w}}}$,
$\mathbbm{x} = (\mathbbm{x}^{(k)}) \in \mathbb{R}^{n \times d}$, $\boldsymbol{b} \in \mathbb{R}^{M_K}$, *and* $\boldsymbol{h} \in \mathbb{R}^q$, *it holds*

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

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

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

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

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

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

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

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

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

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

Lemma

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

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

Then it holds with and $\mathbb{A} = \mathcal{S}\mathcal{S}^{\top}$

$$\mathscr{F}(\boldsymbol{v}) \geq \operatorname{block}\Bigl\{ \frac{1}{2} \mathbb{F}^{(1)}(\boldsymbol{v}^{(1)}), \; \mathbb{I}_{\mathbb{X}^{(1)}} , \ldots , \frac{1}{2} \mathbb{F}^{(K)}(\boldsymbol{v}^{(K)}), \; \mathbb{I}_{\mathbb{X}^{(K)}}, \\ \frac{1}{2} \, \mathbb{X}^{(K)} \mathbb{A} \mathbb{X}^{(K)^{\top}} + \lambda_{\boldsymbol{\beta}} \mathbb{I}_{\boldsymbol{\beta}} \, , \; (\mu^2 - 3) \mathbb{I}_{\boldsymbol{z}} \Bigr\} \, .$$

- Only effective dimension $\mathbb{p}$ matters, full dimension p irrelevant;
- Blessing of dimension: extension of parameter space (calming) may improve geometric properties of the objective function without changing the effective dimension;
- Small noise condition: DNN training requires “small noise” condition (purified data), may be improved by presmoothing;
- (Local) strong convexity helps to study theoretical guarantees of the trained DNN and convergence of alternating optimization;
- Each layer $\mathbb{X}^{(k)}$ and each weighting matrix $W^{(k)}$ can be studied separately;
- No need of backward propagation;
- ...

- (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.

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.

Gratton, S., Lawless, A. S., and Nichols, N. K. (2007).
Approximate gauss-newton methods for nonlinear least squares problems.
SIAM J. Optim., 18:106–132.

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.

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.

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).
Finite sample expansions and risk bounds in high-dimensional SLS models.
arXiv:2404.14227.

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

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

Wilks, S. S. (1938).
The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses.
The Annals of Mathematical Statistics, 9(1):60 – 62.

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.