Nichtparametrische Regression durch tiefe neuronale Netzwerke mit ReLU Aktivierungsfunktion

Minh Duc Bui

January 16, 2020

Inhaltsverzeichnis I

Inhaltsverzeichnis II

Einleitung

Ziel der Arbeit

Annahme: Multivariates nichtparametrisches Regressionsmodel und die Regressionsfunktion besteht aus einer Komposition von Funktionen. Betrachten ein *sparsly connected* tiefes neuronales Netzwerk mit einer ReLU Aktivierungsfunktion.

Ziele:

- Für eine bestimmte gewählte Netzwerkarchitektur, eine obere Schranke für den L_2-Fehler beweisen
- Untere Schranke für L_2-Fehler angeben
- Minimax-Konvergenzrate für Schätzer aus solchen Netzwerken

Nichtparametrische Regression I

- Zufallsvektor $(\mathbf{X}, Y)$ mit Werten in $\mathbb{R}^d \times \mathbb{R}$, wobei $\mathbb{E}Y^2 < \infty$
- $Y = f(\mathbf{X}) + \epsilon$, wobei Störgröße standardnormalverteilte und unabhängig von $\mathbf{X}$ Zufallsvariable
- Minimiere $\mathbb{E}[L(Y, f'(\mathbf{X}))]$, wobei $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ messbare Funktion
- Wähle $L(y, s) = (y - s)^2$ quadratische Verlustfunktion, es folgt $\mathbb{E}[|Y - f'(\mathbf{X})|^2]$

Nichtparametrische Regression II

- $m : \mathbb{R}^d \to \mathbb{R}, \;\; m = \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$ nennt man Regressionsfunktion
- $\mathbb{E}(|f'(\mathbf{X}) - Y|^2) = \mathbb{E}(|f'(\mathbf{X}) - m(\mathbf{X}))|^2) + \mathbb{E}(|m(\mathbf{X}) - Y)|^2)$
- $\mathbb{E}(|f'(\mathbf{X}) - m(\mathbf{X}))|^2)$ nennt man L_2-Fehler von f'
- Regressionsfunktion minimiert L_2-Fehler, aber nicht berechenbar

Nichtparametrische Regression III

- Gegebene Beobachtungen $D_n = (\mathbf{X}_1, Y_1, \dots, \mathbf{X}_n, Y_n)$, wobei $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ u.i.v. Zufallsvariablen
- Ziel: $f_n(\mathbf{x}) = f_n(\mathbf{x}, D_n)$ konstruieren, sodass die L_2-Fehler

$$\mathbb{E}(|f_n(\mathbf{X}) - f_0(\mathbf{X})|^2) = \int |f_n(\mathbf{x}) - f_0(\mathbf{x})|^2 P_X(dx)$$

minimal

Konvergenzrate I

- Definiere L_2 Fehler als $R(f_n, f_0) := \int |f_n(\mathbf{X}) - f_0(\mathbf{X})|^2 P_X(dx)$
- Analyse der optimalen Konvergenzrate des L_2-Fehler gegeben einer Verteilungsklasse
- Optimal in diesem Kontext heißt, falls die Rate der Minimax-Konvergenzrate des L_2 Fehlers entspricht

Konvergenzrate II

- Klassische Annahme der nichtparametrischen Statistik: Regressionsfunktion ist β-glatt
- Optimale Konvergenzrate liegt bei $n^{-\frac{2\beta}{2\beta+d}}$
- Hochdimensionale Probleme verursachen langsame Konvergenzrate, das nennt man Fluch der Dimension

Beschreibung des Modells

Definition eines neuronalen Netzwerkes

Definition

Sei $L \in \mathbb{N}_0$, $\mathbf{p} = (p_0, ..., p_{L+1})^T \in \mathbb{N}^{L+2}$. *Ein neuronales Netzwerk mit Netzwerkarchitektur* $(L, \mathbf{p})$ *und verschobener Aktivierungsfunktion* $\sigma_{\mathbf{v}_i} : \mathbb{R}^{p_i} \to \mathbb{R}^{p_i}$ *ist eine Funktion* $g : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}$ *mit*

$$g : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}, \ \mathbf{x} \mapsto g(\mathbf{x}) = W_{L+1} \cdot \sigma_{\mathbf{v}_L}(W_L \cdot \sigma_{\mathbf{v}_{L-1}}(\cdots W_2 \cdot \sigma_{\mathbf{v}_1}(W_1 \cdot \mathbf{x}) \cdots)), \tag{1}$$

wobei $W_l \in \mathbb{R}^{p_l \times p_{l-1}}$, $l = 1, ..., L+1$ *Gewichtsmatrizen und* $\mathbf{v}_l \in \mathbb{R}^{p_l}$ *Verschiebungsvektoren heißen.* L *nennen wir die Anzahl der hidden Layer/Tiefe des neuronalen Netzwerkes und* $\mathbf{p}$ *heißt width vector.*

Anschauliche Betrachtung

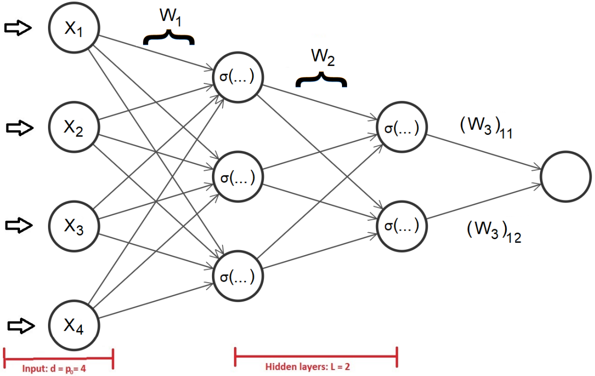

Figure: Graphische Darstellung eines neuronalen Netzwerkes mit *width vector* $\mathbf{p} = (4, 3, 2, 1)$

Rahmenbedingungen des Modells

Aktivierungsfunktion

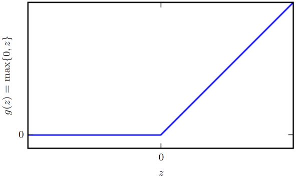

Figure: Die ReLU Funktion

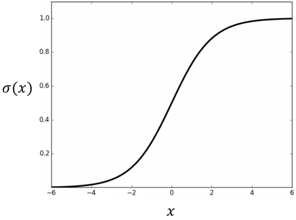

Figure: Die Sigmoid Funktion

- ReLU Aktivierungsfunktion:

$$\sigma : \mathbb{R} \to \mathbb{R}; \quad x \mapsto \sigma(x) = \max(0, x) = (x)_+$$

Aktivierungsfunktion: Vorteile ReLU I

- Produziert viele inaktive *hidden units*

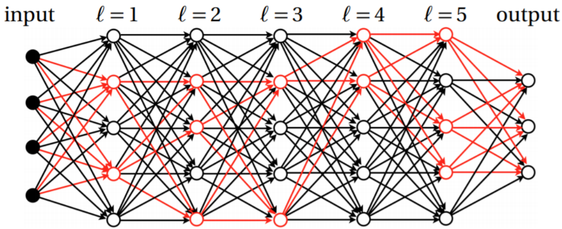

Figure: Ein *sparse* Netzwerk mit ReLU Aktivierungsfunktion. Die roten Linien illustrieren die Verbindungen zu den aktiven Knoten.

Aktivierungsfunktion: Vorteile ReLU II

- Häufiges Probleme bei anderen Aktivierungsfunktionen, ist das Problem des Verschwinden des Gradienten
  - Stellt ein großes Hindernis im Lernalgorithmus dar und führt häufig zu Ungenauigkeiten des Modells
- Nützliche Besonderheit der Funktion ist, dass sie eine Projektion ist

$$\sigma \circ \sigma = \sigma.$$

Netzwerkparameter I

- Netzwerkparameter im Betrag kleiner 1 halten, indem wir im Lernalgorithmus die Netzwerkparameter in jeder Iteration auf $[-1, 1]$ projizieren:

$$F(L, \mathbf{p}) := \{f \text{ in der Form } (1) : \max_{j=0,\dots,L} \|W_j\|_\infty \vee |\mathbf{v}_j|_\infty \leq 1\}$$

Dünnbesetzte Parameter I

- Größere Netzwerke können komplexere Aufgaben lösen, haben jedoch Neigung zum Overfitting

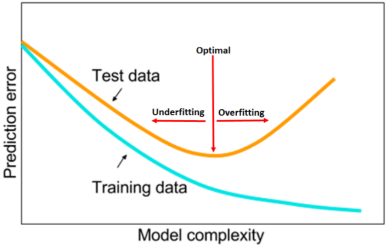

Figure: Problem des Overfittings durch tiefe neuronale Netzwerke ohne Regulierungen.

Dünnbesetzte Parameter II

Lösungsansatz:

- Einführen von dünnbesetzten Netzwerkparameter
- Falls $\|W_j\|_0$ die Anzahl der Nicht-Nullen Einträge von W_j bezeichnet, dann definieren wir ein *s-sparse* Netzwerk als

$$
\begin{aligned}
F(L, \mathbf{p}, s) &:= F(L, \mathbf{p}, s, F) \\
&:= \{f \in F(L, p) : \sum_{j=0}^{L} \|W_j\|_0 + |\mathbf{v}_j|_0 \leq s, \||f|_\infty\|_\infty \leq F\},
\end{aligned}
$$

Hierarchische Komposition der Regressionsfunktion I

- Für β-glatte Regressionsfunktion $f_0 : [0,1]^d \to \mathbb{R}$ gilt eine Minimax-Konvergenzrate von $n^{-2\beta/2\beta+d}$
- Müssen zusätzliche strukturelle Annahmen an die Regressionsfunktion treffen

Hierarchische Komposition der Regressionsfunktion II

- Heuristische Idee: Performen gut bei komplexen Objekten, die durch einfache Objekte in einer iterativen Weise aufgebaut werden können

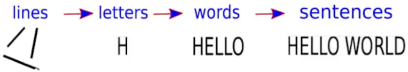

Figure: Beispiel einer hierarchischen Struktur: Aus Linien werden Buchstaben gebaut, aus Buchstaben werden Wörter geformt und zum Schluss werden die Wörter zu Sätzen zusammengesetzt.

Hierarchische Komposition der Regressionsfunktion III

- Regressionsfunktion f_0, die aus einer Komposition von Funktionen besteht, das heißt

$$f_0 = g_q \circ g_{q-1} \circ \ldots \circ g_1 \circ g_0$$

mit $g_i : [a_i, b_i]^{d_i} \to [a_{i+1}, b_{i+1}]^{d_{i+1}}$

- Einzelnen Komponenten von g_i bezeichnen wir mit $g_i = (g_{ij})^T_{j=1,\ldots,d_{i+1}}$
- $t_i \leq d_i$ als maximale Anzahl an Variablen, an denen die einzelnen g_{ij} abhängen
- Klassischer Ansatz: g_{ij} in der Hölderklasse mit Glattheitsindex β_i

Hierarchische Komposition der Regressionsfunktion IV

Der Ball der β-Hölder Funktionen mit Radius K ist definiert als

$$C_r^\beta(D,K) = \Bigg\{ f : D \subset \mathbb{R}^r \to \mathbb{R} :$$

$$\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|<\beta} \|\partial^{\boldsymbol{\alpha}} f\|_\infty + \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor\beta\rfloor} \sup_{\substack{\mathbf{x},\mathbf{y}\in D \\ \mathbf{x}\neq\mathbf{y}}} \frac{|\partial^{\boldsymbol{\alpha}} f(\mathbf{x}) - \partial^{\boldsymbol{\alpha}} f(\mathbf{y})|}{|\mathbf{x}-\mathbf{y}|_\infty^{\beta-\lfloor\beta\rfloor}} \leq K \Bigg\}$$

wobei $\partial^{\boldsymbol{\alpha}} = \partial^{\alpha_1} \ldots \partial^{\alpha_r}$ ein Multi-Index mit $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_r) \in \mathbb{N}^r$ und $|\boldsymbol{\alpha}| := |\boldsymbol{\alpha}|_1$.

Hierarchische Komposition der Regressionsfunktion V

Annahme: f_0 aus einer Komposition von Funktionen in der Klasse

$$
\begin{aligned}
G(q, \mathbf{d}, \mathbf{t}, \beta, K) := \{ f_0 = g_q \circ \dots \circ g_0 : g_i = (g_{ij})_j : \\
[a_i, b_i]^{d_i} \to [a_{i+1}, b_{i+1}]^{d_{i+1}}, \; g_{ij} \in C_{t_i}^{\beta_i}([a_i, b_i]^{t_i}, K), \\
\text{für beliebige } |a_i|, |b_i| \leq K \}
\end{aligned}
$$

mit $\mathbf{d} := (d_0, \dots, d_{q+1})$, $\mathbf{t} := (t_0, \dots, t_q)$, $\boldsymbol{\beta} := (\beta_0, \dots, \beta_q)$ besteht.

Glattheit einer kompositionalen Funktion

- $f_0 = g_q \circ g_{q-1} \circ \ldots \circ g_1 \circ g_0$
- Berechne Glattheit von f_0, die wiederum durch die Glattheit der Funktionen g_i induziert wird
- Sogenannte effektive Glattheitsindex

$$\beta_i^* := \beta_i \prod_{l=i+1}^{q} (\beta_l \wedge 1), \qquad (\beta_l \wedge 1) := \min(\beta_l, 1)$$

- $\phi_n := \max_{i=0,\ldots q} n^{-\frac{2\beta_i^*}{2\beta_i^* + t_i}}.$

Empirisches Risiko

- Gegeben $D_n = (\mathbf{X}_1, Y_1, \ldots, \mathbf{X}_n, Y_n)$, wobei $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ u.i.v. Zufallsvariablen
- Netzwerkfunktion f konstruieren, sodass das empirische Risiko $\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(\mathbf{X}_i))^2$ minimal ist
- Für einen beliebigen Schätzer $\widehat{f}_n \in F(L, \mathbf{p}, s, F)$ definieren wir

$$\Delta_n(\widehat{f}_n, f_0) := \mathbb{E}_{f_0}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{f}_n(\mathbf{X}_i))^2 - \inf_{f \in F(L,\mathbf{p},s,F)} \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(\mathbf{X}_i))^2\right].$$

Hauptresultat

Obere Schranke des L_2-Fehler I

Theorem

Betrachte ein d-variates nichtparametrisches Regressionsmodell, wobei die Regressionsfunktion eine Komposition aus Funktionen besteht und dabei in der Klasse $G(q, \mathbf{d}, \mathbf{t}, \beta, K)$ liegt. Sei nun $\widehat{f}_n$ ein Schätzer, der Funktionen in der Netzwerkklasse $F(L, (p_i)_{i=0,\dots,L+1}, s, F)$ schätzt, wobei die Netzwerkklasse die folgenden Bedingungen erfüllt:

Obere Schranke des L_2-Fehler II

Theorem

(i) *Die zur Schätzung verwendeten neuronalen Netzwerke erlauben Funktionswerte, die mindestens so groß sind wie die maximalen Funktionswerte der Regressionsfunktion f_0:* $F \geq \max(K, 1)$

(ii) *Für die Anzahl der Layer soll gelten:* $\sum_{i=0}^{q} \log_2(4t_i \vee 4\beta_i) \log_2 n \leq L \lesssim n\phi_n$

(iii) *Die Größe der Layer muss mindestens mit Rate $n\phi_n$ in n gegen unendlich gehen:* $n\phi_n \lesssim \min_{i=1,\dots,L} p_i$

(iv) *Anzahl der nicht verschwindende Einträge der Gewichtsmatrizen und Verschiebungsvektoren muss mit Rate $n\phi_n \log(n)$ in n gegen unendlich gehen:* $s \asymp n\phi_n \log n$

Obere Schranke des L_2-Fehler III

Theorem

Dann existieren Konstanten C und C', die nur abhängen von q, $\mathbf{d}$, $\mathbf{t}$, β, F, sodass, wenn $\Delta_n(\widehat{f}_n, f_0) \leq C\phi_n L \log^2(n)$ gilt, dann

$$R(\widehat{f}_n, f_0) \leq C'\phi_n L \log^2(n) \tag{2}$$

und falls $\Delta_n(\widehat{f}_n, f_0) \geq C\phi_n L \log^2(n)$, dann

$$\frac{1}{C'}\Delta_n(\widehat{f}_n, f_0) \leq R(\widehat{f}_n, f_0) \leq C'\Delta_n(\widehat{f}_n, f_0). \tag{3}$$

Folgerungen aus Theorem I

- Aus $\phi_n := \max_{i=0,\dots q} n^{-\frac{2\beta_i^*}{2\beta_i^*+t_i}}$ können wir sehen, dass Rate nicht mehr von ursprünglichen Inputdimension d abhängt, sondern von t_i
- Aus der Bedingung *(iv)* können wir folgern, dass wir ein *sparse* Netzwerk vorliegen haben müssen
- Flexible Möglichkeit eine gute Netzwerkarchitektur zu wählen, solange die Anzahl der aktiven Parameter s die Bedingung (iv) erfüllt

Untere Schranke des L_2-Fehler

Theorem

Betrachte ein d-variates nichtparametrisches Regressionsmodell mit Beobachtungen $\mathbf{X}_i$ *aus einer Wahrscheinlichkeitsverteilung mit einer Lebesque Dichte auf* $[0,1]^d$, *welche mit einer oberen und unteren positiven Konstante beschränkt ist. Für eine beliebige nicht-negative ganze Zahl q, beliebige Dimensionsvektoren* $\mathbf{d}$ *und* $\mathbf{t}$, *die* $t_i \leq \min(d_0, \ldots, d_{i-1})$ *für alle i erfüllen, ein beliebiger Glattheitsvektor* β *und alle hinreichend großen Konstanten* $K > 0$, *existiert eine positive Konstante c, sodass*

$$\inf_{\widehat{f}_n} \sup_{f_0 \in G(q,\mathbf{d},\mathbf{t},\beta,K)} R(\widehat{f}_n, f_0) \geq c\phi_n,$$

wobei das $\inf$ *über alle Schätzer* $\widehat{f}_n$ *genommen wird.*

Untere Schranke des L_2-Fehler

- Theorem 1 gibt uns obere Schranke für den L_2-Fehler für einen Schätzer aus der Netzwerkklasse $F(L, \mathbf{p}, s, F)$ über der Klasse $G(q, \mathbf{d}, \mathbf{t}, \beta, K)$:

$$R(\widehat{f}_n, f_0) \leq C' \phi_n L \log^2(n)$$

- Untere Schranke für L_2-Fehler über der Funktionsklasse $G(q, \mathbf{d}, \mathbf{t}, \beta, K)$:

$$\inf_{\widehat{f}_n} \sup_{f_0 \in G(q,\mathbf{d},\mathbf{t},\beta,K)} R(\widehat{f}_n, f_0) \geq c\phi_n$$

Einbettungseigenschaften einer Netzwerkfunktionsklasse

- *Größenvergleich*
- *Komposition*
- *Layers hinzufügen/Netzwerktiefe angleichen*
- *Parallelisierung*
- *Beseitigung der inaktiven Knoten*

Approximationsqualität neuronaler Netzwerke I

Theorem

Für jede beliebige Funktion $h \in C_r^\beta([0,1]^r, K)$ und jede beliebige ganze Zahl $m \geq 1$ und $N \geq (\beta+1)^r \vee (K+1)e^r$, existiert ein Netzwerk

$$\widetilde{f} \in F(L, (r, 6(r+\lceil\beta\rceil)N, ..., 6(r+\lceil\beta\rceil)N, 1), s, \infty)$$

mit der Tiefe

$$L = 8 + (m+5)(1 + \lceil \log_2(r \vee \beta)\rceil)$$

und die Anzahl an Parameter

$$s \leq 141(r+\beta+1)^{3+r} N(m+6),$$

sodass

$$\|\widetilde{f} - h\|_{L^\infty([0,1]^r)} \leq (2K+1)(1+r^2+\beta^2)6^r N 2^{-m} + K 3^\beta N^{-\frac{\beta}{r}}.$$

Approximationsqualität neuronaler Netzwerke II

- $f_0 = g_q \circ \ldots \circ g_0$ mit $g_{ij} \in C_{t_i}^{\beta_i}([a_i, b_i]^{t_i}, K_i)$ und $K_i \geq 1$
- $f_0 = g_q \circ \ldots \circ g_0 = h_q \circ \ldots \circ h_0$ mit
  $h_{0j} \in C_{t_0}^{\beta_0}([0,1]^{t_0}, 1)$ und
  $h_{ij} \in C_{t_i}^{\beta_i}([0,1]^{t_i}, (2K_{i-1})^{\beta_i})$ für $i = 1, \ldots, q-1$ und
  $h_{qj} \in C_{t_q}^{\beta_q}([0,1]^{t_q}, K_q(2K_{q-1})^{\beta_q})$

Lemma

Sei h_{ij} wie oben definiert mit $K_i \geq 1$. Dann gilt für jede beliebige Funktion $\widetilde{h}_i = (\widetilde{h}_{ij})_j^\top$ mit $\widetilde{h}_{ij} : [0,1]^{t_i} \to [0,1]$,

$$\|h_q \circ \ldots \circ h_0 - \widetilde{h}_q \circ \ldots \circ \widetilde{h}_0\|_{L^\infty[0,1]^d}$$
$$\leq K_q \prod_{l=0}^{q-1} (2K_l)^{\beta_{l+1}} \sum_{i=0}^{q} \||h_i - \widetilde{h}_i|_\infty\|_{L^\infty[0,1]^{d_i}}^{\prod_{l=i+1}^{q} \beta_l \wedge 1}.$$

Beweisidee zum Haupttheorem 1

- Für Stichprobenumfang $n \geq 3$ gilt:

$$\frac{1}{4}\Delta_n(\widehat{f}_n, f_0) - C'\phi_n L \log^2(n) \leq R(\widehat{f}_n, f_0)$$
$$\leq 4 \inf_{f^* \in F(L,\mathbf{p},s,F)} \|f^* - f_0\|_\infty^2 + 4\Delta_n(\widehat{f}_n, f_0) + C'\phi_n L \log^2(n). \tag{4}$$

Untere Grenze in (3)

- Falls $C\phi_n L\log^2(n) \leq \Delta_n(\widehat{f}_n, f_0)$ gilt und wir $C = 8C'$ wählen, dann folgt für

$$C'\phi_n L\log^2(n) = \frac{1}{8}C\phi_n L\log^2(n) \leq \frac{1}{8}\Delta_n(\widehat{f}, f_0).$$

- $R(\widehat{f}_n, f_0) \geq \frac{1}{8}\Delta_n(\widehat{f}_n, f_0)$

Obere Grenze in (2) und (3)

- Schranke für den Approximationsfehler $\inf_{f^* \in F(L, \mathbf{p}, s, F)} \|f^* - f_0\|_\infty$
- Regressionsfunktion f_0 in der Form $f_0 = h_q \circ \ldots \circ h_0$ mit $h_i = (h_{ij})_j^\top$, h_{ij} definiert auf $[0,1]^{t_i}$
- Nach Theorem existiert Netzwerk $\widetilde{h}_{ij}$, sodass:

$$\|\widetilde{h}_{ij} - h_{ij}\|_{L^\infty([0,1]^{t_i})} \leq (2Q_i + 1)(1 + t_i^2 + \beta_i^2) 6^{t_i} N n^{-1} + Q_i 3^{\beta_i} N^{-\frac{\beta_i}{t_i}}. \tag{5}$$

- Transformation des Output x vom Netzwerk $\widetilde{h}_{ij}$ zu $(1 - (1 - x)_+)_+$ mit neuem Netzwerk $\sigma(h_{ij}^*)$

Obere Grenze in (2) und (3)

- Es gilt $\sigma(h^*_{ij}) = (h^*_{ij}(\mathbf{x}) \vee 0) \wedge 1$ und

$$\|\sigma(h^*_{ij}) - h_{ij}\|_{L^\infty([0,1]^r)} \leq \|\widetilde{h}_{ij} - h_{ij}\|_{L^\infty([0,1]^r)}.$$

- Durch Parallelisierung und Komposition erhält man $f^* = \widetilde{h}_{q1} \circ \sigma(h^*_{q-1}) \circ ... \circ \sigma(h^*_0)$, wobei durch Hinzufügen von Layern, diese in Netzwerkklasse $F(L, \mathbf{p}, s)$ liegt und die Bedingungen aus Theorem erfüllen

Obere Grenze in (2) und (3)

- $$\inf_{f^* \in F(L,\mathbf{p},s)} \|f^* - f_0\|_\infty^2 \leq C' \max_{i=0,\dots,q} c^{-\frac{2\beta_i^*}{t_i}} n^{-\frac{2\beta_i^*}{2\beta_i^*+t_i}}$$
- Man kann zeigen, dass

$$\begin{aligned}\inf_{f^* \in F(L,\mathbf{p},s,F)} \|f^* - f_0\|_\infty^2 &\leq \inf_{\tilde{f} \in F(L,\mathbf{p},s)} 4\|\tilde{f} - f_0\|_\infty^2 \\ &\leq 8C \max_{i=0,\dots,q} c^{-\frac{2\beta_i^*}{t_i}} n^{-\frac{2\beta_i^*}{2\beta_i^*+t_i}}.\end{aligned}$$

Obere Grenze in (2) und (3)

- Man kann folgern, dass

$$R(\widehat{f}_n, f_0) \leq 4 \inf_{f^* \in F(L,\mathbf{p},s,F)} \|f^* - f_0\|_\infty^2 + 4\Delta_n(\widehat{f}_n, f_0) + C'\phi_n L \log^2(n)$$

$$\leq 32C \max_{i=0,\dots,q} c^{-\frac{2\beta_i^*}{t_i}} n^{-\frac{2\beta_i^*}{2\beta_i^*+t_i}} + 4\Delta_n(\widehat{f}_n, f_0) + C'\phi_n L \log^2(n)$$

- Bedingung aus (2): $\Delta_n(\widehat{f}_n, f_0) \leq \widetilde{C}\phi_n L \log^2(n)$
- Bedingung aus (3): $\Delta_n(\widehat{f}_n, f_0) \geq \widetilde{C}\phi_n L \log^2(n)$