Basic Theory of Neural Networks

Xiaoke Zhang

University of Science and Technology of China

Table of Contents

Neural Networks

- Linear function: the dot product of the weights and the input that gives an output.
- Neurons: introduce non-linearity to increase expressivity.

Neural network (NN) is a network of neurons arranged in layers, it can be represented as

$$\mathbf{y} = f_{NN}(\mathbf{x}) = f_{L+1} \circ f_L \circ \ldots \circ f_1(\mathbf{x}),$$

where $\mathbf{x}$ is the input, $\mathbf{y}$ is the predicted output, and $f_0, \ldots, f_{L+1}$ are layers of the NN.

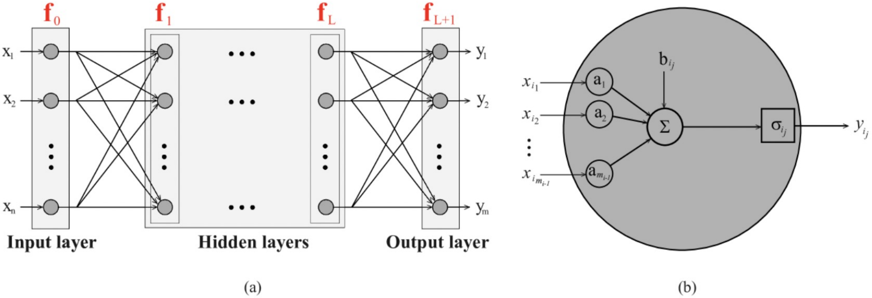

Figure: (a) Neural Network (b) Neuron.

Each layer of the NN can be represented as $y_i = f_i(\mathbf{x}_i) = \sigma_i(A_i \mathbf{x}_i + b_i)$, where $\sigma_i = (\sigma_{i1}, \ldots, \sigma_{im_i})^T$ contains the element-wise activation functions.

Activation functions

- Relu (Rectified Linear Unit) function: $\sigma(x) = \max\{0, x\}$.
- Step function: $\sigma(x) = \mathbf{1}(x > 0)$.
- Logistic function: $\sigma(x) = \frac{1}{1+e^{-x}}$.
- Tanh (Hyperbolic tangent) function: $\sigma(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

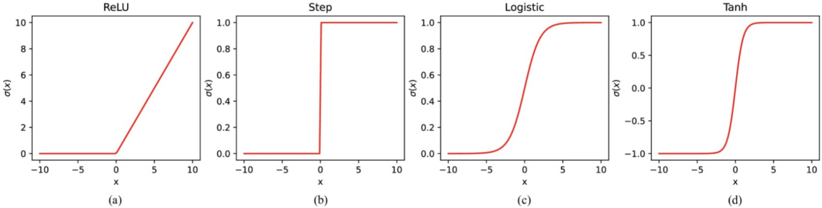

Figure: (a) ReLU (b) Step (c) Logistic (d) Tanh.

Theorem 1 (Taylor's theorem).

Any continuous function $f(x) : \mathbb{R} \to \mathbb{R}$ that is k-times differentiable at a can be represented as a sum of polynomials,

$$f(x) = \sum_{i=0}^{k} c_i (x-a)^i + R_k(x),$$

where $c_i = \frac{f^i(a)}{i!} = \frac{1}{i!} \frac{d^i}{dx^i} f(x)\Big|_{x=a}$ and $R_k(x) = o\left(|x-a|^k\right)$ is the residual term.

Theorem 2 (Weierstrass, 1885).

Any continuous real-valued function $f(x) : [a,b] \to R$ defined on the interval $[a,b]$ can be approximated with a polynomial function $p_N(x) = \sum_{i=0}^{N} c_i x^i$ with finite degree N such that:

$$|f(x) - p_N(x)| < \epsilon$$

- Arbitrary width case: an arbitrary number of neurons with a limited number of hidden layers.
- Arbitrary depth case: an arbitrary number of hidden layers with a limited number of neurons.

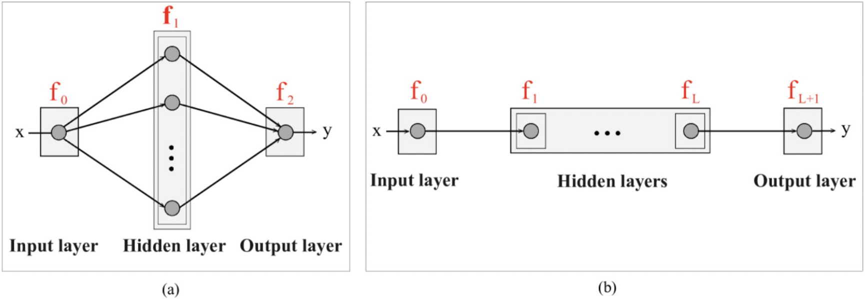

Figure: (a) NN with arbitrary width (b) NN with arbitrary depth.

UAT: Arbitrary width case

Theorem 3 (Funahashi, Hornick et al., and Cybenko, 1989).

Let X be any compact subset of R^n and σ be any sigmoid activation function, then the finite sum of the form:

$$f_{\mathrm{NN}}(\mathbf{x}) = \mathbf{A}_2 \sigma\left(\mathbf{A}_1 \mathbf{x} + \mathbf{b}_1\right) = \sum_{j=1}^{m_1} a_2 \sigma\left(\mathbf{A}_1 \mathbf{x} + b_{1_j}\right)$$

is dense in X. In other words, given any $f : X \to R$ and $\epsilon > 0$, there is a finite sum: f_{NN} for which $|f(\mathbf{x}) - f_{\mathrm{NN}}(\mathbf{x})| < \epsilon$ for all $\mathbf{x} \in X$.

NNs with one hidden layer and sigmoid activation function can approximate any continuous univariate function on a bounded domain with arbitrary accuracy.

Theorem 4 (Leshno et al., 1993).

Let X be any compact subset of R^n and σ be an activation function, then the finite sum f_{NN} is dense in X iff σ is not a polynomial function.

MLP with non-polynomial activation functions are universal approximators.

UAT: Arbitrary depth case

Theorem 5 (Lu et al., 2017).

Except for a negligible set, all functions $f : \mathbb{R}^n \to \mathbb{R}$ cannot be approximated by any ReLU network whose width $W \leq n$.

Width-1 NNs can approximate only a small class of univariate functions, i.e., the minimum width required for universal approximation should be greater than 1.

Theorem 6 (Lu et al., 2017).

For any Lebesgue-integrable function $f : \mathbb{R}^n \to \mathbb{R}$ and $\epsilon > 0$, there exists a neural network f_{NN} of width $W \leq n + 4$ with ReLU activation function which satisfies:

$$\int |f(\mathbf{x}) - f_{NN}(\mathbf{x})| d\mathbf{x} \leq \epsilon.$$

NNs with arbitrary hidden layers and at most $n + 4$ number of neurons per layer can approximate any functions in a Lebesgue integrable space with sufficient accuracy.

Theorem 7 (Park et al., 2021).

The minimum width required for universal approximation of Lebesgue integrable functions $f : \mathbb{R}^n \to \mathbb{R}$ is $\max\{n + 1, m\}$.

Table of Contents

Least-squares estimation

For any random function f, let $Z \equiv (X, Y)$ be a random vector independent of f. The L_2 risk is defined by $L(f) = \mathbb{E}_Z|Y - f(X)|^2$. At the population level, the least-squares estimation is to find a measurable function $f^* : \mathbb{R}^d \to \mathbb{R}$ satisfying

$$f^* := \arg\min_f L(f) = \arg\min_f \mathbb{E}_Z|Y - f(X)|^2.$$

The distribution of (X, Y) is typically unknown and only a random sample $S \equiv \{(X_i, Y_i)\}_{i=1}^n$ is available. Let

$$L_n(f) = \sum_{i=1}^n |Y_i - f(X_i)|^2 /n,$$

be the empirical risk of f on the sample S.

Preliminaries

Let $\mathcal{F}_n$ be a function class consisting of feedforward neural networks. For any estimator $\hat{f}_n$, the excess risk defined as the difference between the L_2 risks of $\hat{f}_n$ and f_0,

$$L(\hat{f}_n) - L\left(f_0\right) = \mathbb{E}_Z \left|Y - \hat{f}_n(X)\right|^2 - \mathbb{E}_Z \left|Y - f_0(X)\right|^2 .$$

Because of the simple form of the least squares loss, it can be simply expressed as

$$\left\|\hat{f}_n - f_0\right\|_{L^2(\nu)}^2 = \mathbb{E}_X \left|\hat{f}_n(X) - f_0(X)\right|^2 ,$$

where ν denotes the marginal distribution of X.

Excess risk

The excess risk can be decomposed as:

$$L(\hat{f}_n) - L(f_0) = \left\{ L(\hat{f}_n) - \inf_{f \in \mathcal{F}_n} L(f) \right\} + \left\{ \inf_{f \in \mathcal{F}_n} L(f) - L\left(f_0\right) \right\}.$$

- The first term is the stochastic error, which depends on the estimator $\hat{f}_n$. It measures the difference of the error of $\hat{f}_n$ and the best one in $\mathcal{F}_n$;
- The second term is the approximation error, which depends on the function class $\mathcal{F}_n$ and the target f_0. It measures how well the function f_0 can be approximated using $\mathcal{F}_n$ with respect to the loss L.

Lemma 8.

For any random sample $S=\{(X_i,Y_i)\}_{i=1}^n$, the excess risk of ERM satisfies

$$\begin{aligned}\mathbb{E}_S\left[\left\|\hat{f}_n-f_0\right\|_{L^2(\nu)}^2\right] &= \mathbb{E}_S\left[L(\hat{f}_n)-L\left(f_0\right)\right]\\ &\leq \mathbb{E}_S\left[L\left(f_0\right)-2L_n(\hat{f}_n)+L(\hat{f}_n)\right]+2\inf_{f\in\mathcal{F}_n}\|f-f_0\|_{L^2(\nu)}^2 .\end{aligned}$$

- Stochastic error bound: $\mathbb{E}_S\left[L\left(f_0\right)-2L_n(\hat{f}_n)+L(\hat{f}_n)\right]$ can be bounded by the complexity of $\mathcal{F}_n$ using the empirical process theory.
- Approximation error: $\inf_{f\in\mathcal{F}_n}\|f-f_0\|_{L^2(\nu)}^2$,the approximation of high-dimensional functions using neural networks has been studied by many works.

Excess risk

Proof.

Since f_0 is the minimizer of quadratic functional $L(\cdot)$, by direct calculation we have

$$\mathbb{E}_S\left[\|\hat{f}_n - f_0\|^2_{L^2(\nu)}\right] = \mathbb{E}_S\left[L_n(\hat{f}_n) - L(f_0)\right].$$

By the definition of the empirical risk minimizer, we have

$$L_n(\hat{f}_n) - L_n(f_0) \leq L_n(\bar{f}_n) - L_n(f_0),$$

where $\bar{f}_n \in \arg\min_{f \in \mathcal{F}_n} \|f_n - f_0\|^2_{L^2(\nu)}$. Taking expectation on both side we get

$$\mathbb{E}_S\left[L_n(\hat{f}_n) - L(f_0)\right] \leq L(\bar{f}) - L(f_0) = \|\bar{f} - f_0\|^2_{L^2(\nu)}$$

□

Table of Contents

Notation

- Pseudo dimension $\mathrm{Pdim}(\mathcal{F})$: the largest integer m for which there exists $(x_1, \ldots, x_m, y_1, \ldots, y_m) \in \mathcal{X}^m \times \mathbb{R}^m$ such that for any $(b_1, \ldots, b_m) \in \{0,1\}^m$ there exists $f \in \mathcal{F}$ such that $\forall i : f(x_i) > y_i \Longleftrightarrow b_i = 1$. Specially, if $\mathcal{F}$ is the class of functions generated by a neural network with a fixed architecture and fixed activation functions, we have $\mathrm{Pdim}(\mathcal{F}) = \mathrm{VCdim}(\mathcal{F})$.
- Define $\mathcal{F}_n|_x = \{(f(x_1), \ldots, f(x_n) : f \in \mathcal{F}_n\}$ as the subset of $\mathbb{R}^n$.
- For a positive number δ, let $\mathcal{N}(\delta, \|\cdot\|_\infty, \mathcal{F}_n|_x)$ be the covering number of $\mathcal{F}_n|_x$ under the norm $\|\cdot\|_\infty$ with radius δ. Define the uniform covering number $\mathcal{N}_n(\delta, \|\cdot\|_\infty, \mathcal{F}_n) = \max\left\{\mathcal{N}\left(\delta, \|\cdot\|_\infty, \mathcal{F}_n|_x\right) : x \in \mathcal{X}\right\}$.

Assumption 1 (Sub-exponential).

The response variable Y is sub-exponentially distributed, i.e., there exists a constant $\sigma_Y > 0$ such that $\mathbb{E}\exp(\sigma_Y Y) \leq \infty$.

Assumption 2 (Hölder smoothness).

The target function f_0 belongs to the Hölder class $\mathcal{H}^\beta\left([0,1]^d, B_0\right)$ for a given $\beta > 0$ and a finite constant $B_0 > 0$, where $\mathcal{H}^\beta([0,1]^d, B_0)$ is

$$\left\{ f : [0,1]^d \to \mathbb{R}, \max_{\|\alpha\|_1 \leq s} \|\partial^\alpha f\|_\infty \leq B_0, \max_{\|\alpha\|_1 = s} \sup_{x \neq y} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{\|x-y\|_2^r} \leq B_0 \right\}.$$

Lemma 9.

Let $\mathcal{F}_n = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ be the class of feedforward neural networks with a continuous piecewise-linear activation function with finitely many inflection points and $\hat{f}_n \in \arg\min_{f\in\mathcal{F}_n} L_n(f)$ be the empirical risk minimizer over $\mathcal{F}_n$. Assume that Assumption 1 holds and $\|f_0\|_\infty \le \mathcal{B}$ for $\mathcal{B} \ge 1$. Then, for $n \ge \mathrm{Pdim}\left(\mathcal{F}_n\right)/2$,

$$\mathbb{E}_S\left[L\left(f_0\right) - 2L_n(\hat{f}_n) + L(\hat{f}_n)\right] \le c_0 \mathcal{B}^4 (\log n)^4 \frac{1}{n} \log \mathcal{N}_{2n}\left(n^{-1}, \|\cdot\|_\infty, \mathcal{F}_n\right)$$

where $c_0 > 0$ is a constant independent of $d, n, \mathcal{B}, \mathcal{D}, \mathcal{W}$ and $\mathcal{S}$, and

$$\mathbb{E}\left\|\hat{f}_n - f_0\right\|^2_{L^2(\nu)} \le C_0 \mathcal{B}^5 (\log n)^5 \frac{1}{n} \mathcal{S}\mathcal{D} \log(\mathcal{S}) + 2 \inf_{f\in\mathcal{F}_n} \|f - f_0\|^2_{L^2(\nu)}$$

where $C_0 > 0$ is a constant independent of $d, n, \mathcal{B}, \mathcal{D}, \mathcal{W}$ and $\mathcal{S}$.

Proof

- Let $S' = \{Z_i' = (X_i', Y_i')\}_{i=1}^n$ be another sample independent of S. Define $g(f, Z_i) = (f(X_i) - Y_i)^2 - (f_0(X_i) - Y_i)^2$ for any f and sample Z_i. Observing

$$\mathbb{E}_S\left[L(f_0) - 2L_n\left(\hat{f}_n\right) + L\left(\hat{f}_n\right)\right] = \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n\left\{-2g\left(\hat{f}_\phi, Z_i\right) + \mathbb{E}_{S'}g\left(\hat{f}_\phi, Z_i'\right)\right\}\right].$$

- We define $g_{\beta_n}(f, Z_i) = (f(X_i) - T_{\beta_n}Y_i)^2 - (f_{\beta_n}(X_i) - T_{\beta_n}Y_i)^2$ and $G_{\beta_n}(f, Z_i) = \mathbb{E}_{S'}\{g_{\beta_n}(f, Z_i')\} - 2g_{\beta_n}(f, Z_i)$.
- For any $f \in \mathcal{F}_n$ we have

$$\begin{aligned}|g(f, Z_i) - g_{\beta_n}(f, Z_i)| = |\, & 2\{f(X_i) - f_0(X_i)\}(T_{\beta_n}Y_i - Y_i) \\ & + (f_{\beta_n}(X_i) - T_{\beta_n}Y_i)^2 - (f_0(X_i) - T_{\beta_n}Y_i)^2\,| \\ \leq & 4\mathcal{B}|Y_i| I(|Y_i| > \beta_n) + 4\beta_n|Y_i| I(|Y_i| > \beta_n)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_S\{g(f, Z_i)\} &\leq \mathbb{E}_S\{g_{\beta_n}(f, Z_i)\} + 4\mathcal{B}\mathbb{E}_S\{|Y_i| I(|Y_i| > \beta_n)\} + 4\beta_n\mathbb{E}_S\{|Y_i| I(|Y_i| > \\ &\leq \mathbb{E}_S\{g_{\beta_n}(f, Z_i)\} + 16\frac{\beta_n}{\sigma_Y}\mathbb{E}_S\exp(\sigma_Y|Y_i|)\exp(-\sigma_Y\beta_n/2).\end{aligned}$$

Proof

- Note that $|T_{\beta_n} Y| \leq \beta_n, \|g_{\beta_n}\|_\infty \leq \beta_n$ and $\beta_n \geq \mathcal{B} \geq 1$. Then by Theorem 11.4 of Györfi et al. (2002), for each $n \geq 1$,

$$
\begin{aligned}
& P\left\{\frac{1}{n}\sum_{i=1}^{n} G_{\beta_n}\left(\hat{f}_n, Z_i\right) > t\right\} \\
\leq & P\left\{\exists f \in \mathcal{F}_n : \frac{1}{n}\sum_{i=1}^{n} G_{\beta_n}\left(f, Z_i\right) > t\right\} \\
= & P\left\{\exists f \in \mathcal{F}_n : \mathbb{E}_{S'}\left\{g_{\beta_n}\left(f, Z_i'\right)\right\} - \frac{2}{n}\sum_{i=1}^{n} g_{\beta_n}\left(f, Z_i\right) > t\right\} \\
\leq & 14\mathcal{N}_{2n}\left(\frac{t}{80\beta_n}, \|\cdot\|_\infty, \mathcal{F}_n\right)\exp\left(-\frac{tn}{5136\beta_n^4}\right).
\end{aligned}
$$

Proof

- This leads to a tail probability bound of $\sum_{i=1}^{n} G_{\beta_n}(f_{j^*}, Z_i)/n$. Then for $a_n > 0$,

$$
\begin{aligned}
&\mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^{n} G_{\beta_n}(f_{j^*}, Z_i)\right] \\
&\leq a_n + \int_{a_n}^{\infty} P\left\{\frac{1}{n}\sum_{i=1}^{n} G_{\beta_n}(f_{j^*}, Z_i) > t\right\} dt \\
&\leq a_n + \int_{a_n}^{\infty} 14\mathcal{N}_{2n}\left(\frac{t}{80\beta_n}, \|\cdot\|_\infty, \mathcal{F}_n\right)\exp\left(-\frac{tn}{5136\beta_n^4}\right) dt \\
&\leq a_n + 14\mathcal{N}_{2n}\left(\frac{a_n}{80\beta_n}, \|\cdot\|_\infty, \mathcal{F}_n\right)\exp\left(-\frac{a_n n}{5136\beta_n^4}\right)\frac{5136\beta_n^4}{n}.
\end{aligned}
$$

- Let $a_n = \log\left(14\mathcal{N}_{2n}\left(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{F}_n\right)\right) \cdot 5136\beta_n^4/n$, note that $a_n/(80\beta_n) \geq 1/n$. and $\mathcal{N}_{2n}\left(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{F}_n\right) \geq \mathcal{N}_{2n}\left(\frac{a_n}{80\beta_n}, \|\cdot\|_\infty, \mathcal{F}_n\right)$. Then we have

$$
\mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^{n} G_{\beta_n}(f_{j^*}, Z_i)\right] \leq \frac{5136\beta_n^4\left(\log\left(14\mathcal{N}_{2n}\left(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{F}_n\right)\right) + 1\right)}{n}.
$$

- Setting $\beta_n = c_2\mathcal{B}\log n$, we get

$$
\mathcal{R}\left(\hat{f}_n\right) \leq c_3\mathcal{B}^4\frac{\log\mathcal{N}_{2n}\left(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{F}_n\right)(\log n)^4}{n} + 2\left\|f_n^* - f_0\right\|_{L^2(\nu)}^2.
$$

- Lastly, we will give an upper bound on the covering number by the VC dimension of $\mathcal{F}_n$. By Theorem 12.2 in Anthony and Bartlett (1999), for $2n \geq \mathrm{Pdim}\left(\mathcal{F}_n\right)$,

$$\mathcal{N}_{2n}\left(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{F}_n\right) \leq \left(\frac{4e\mathcal{B}n^2}{\mathrm{Pdim}\left(\mathcal{F}_n\right)}\right)^{\mathrm{Pdim}\left(\mathcal{F}_n\right)}.$$

Moreover, based on Theorem 3 and 6 in Bartlett et al. (2019), there exist universal constants c, C such that

$$c \cdot \mathcal{S}\mathcal{D}\log(\mathcal{S}/\mathcal{D}) \leq \mathrm{Pdim}\left(\mathcal{F}_n\right) \leq C \cdot \mathcal{S}\mathcal{D}\log(\mathcal{S}).$$

Then, we have

$$\mathcal{R}\left(\hat{f}_n\right) \leq c_4\mathcal{B}^5 \frac{\mathcal{S}\mathcal{D}\log(\mathcal{S})(\log n)^5}{n} + 2\left\|f_n^* - f_0\right\|_{L^2(\nu)}^2,$$

for some constant $c_4 > 0$ not depending on $n, d, \mathcal{B}, \mathcal{S}$ or $\mathcal{D}$.

Theorem 10.

Assume that $f \in \mathcal{H}^{\beta}\left([0,1]^d, B_0\right)$ *with* $\beta = s + r, s \in \mathbb{N}_0$ *and* $r \in (0,1]$. *For any* $M, N \in \mathbb{N}^+$, *there exists a function* ϕ_0 *implemented by a ReLU network with width* $\mathcal{W} = 38(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor+1} N \lceil\log_2(8N)\rceil$ *and depth* $\mathcal{D} = 21(\lfloor\beta\rfloor + 1)^2 M \lceil\log_2(8M)\rceil$ *such that*

$$|f(x) - \phi_0(x)| \leq 18B_0(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor+(\beta\vee 1)/2}(NM)^{-2\beta/d}$$

for all $x \in [0,1]^d \backslash \Omega\left([0,1]^d, K, \delta\right)$, *where* $a \vee b := \max\{a, b\}$, $\lceil a \rceil$ *denotes the smallest integer no less than* a, *and*

$$\Omega\left([0,1]^d, K, \delta\right) = \bigcup_{i=1}^{d} \left\{x = [x_1, x_2, \ldots, x_d]^\top : x_i \in \bigcup_{k=1}^{K-1} (k/K - \delta, k/K)\right\},$$

with $K = \left\lceil (MN)^{2/d} \right\rceil$ *and* δ *an arbitrary number in* $(0, 1/(3K)]$.

The approximation error bound has the optimal approximation rate $(NM)^{-2\beta/d}$. This error bound is non-asymptotic in the sense that it is valid for arbitrary network width and depth specified by N and M.

Proof

The main idea of our proof is to approximate the Taylor expansion of Hölder smooth f. By Lemma A. 8 in Petersen and Voigtlaender (2018), for any $x, x_0 \in [0,1]^d$, we have

$$\left| f(x) - \sum_{\|\alpha\|_1 \le s} \frac{\partial^\alpha f(x_0)}{\alpha!}(x - x_0)^\alpha \right| \le d^s \| x - x_0 \|_2^\beta .$$

This reminder term could be well controlled when the approximation to Taylor expansion in implemented in a fairly small local region. Then we can focus on the approximation of the Taylor expansion locally.

- Partition $[0,1]^d$ into small cubes $\bigcup_\theta Q_\theta$, and construct a network ψ that approximately maps each $x \in Q_\theta$ to a fixed point $x_\theta \in Q_\theta$. Hence, ψ approximately discretize $[0,1]^d$.
- For any multi-index α, construct a network ϕ_α that approximates the Taylor coefficient $x \in Q_\theta \mapsto \partial^\alpha f(\psi(x_\theta))$. Once $[0,1]^d$ is discretized, the approximation is reduced to a data fitting problem.
- Construct a network $P_\alpha(x)$ to approximate the polynomial $x^\alpha := x_1^{\alpha_1} \dots x_d^{\alpha_d}$ where $x = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$ and $\alpha = (\alpha_1, \dots, \alpha_d)^\top \in \mathbb{N}_0^d$. In particular, we can construct a network $\phi_\times(\cdot,\cdot)$ approximating the product function of two scalar inputs.

Then the construction of neural network can be written in the form,

$$\phi(x) = \sum_{\|\alpha\|_1 \le s} \phi_\times \left(\frac{\phi_\alpha(x)}{\alpha!}, P_\alpha(x - \psi(x)) \right).$$

Proof

Assume the Hölder norm of f is 1 , i.e. $f \in \mathcal{H}^{\beta}\left([0,1]^{d}, 1\right)$. The reason is that we can always approximate f / B_0 firstly by a network ϕ with approximation error ϵ, then the scaled network $B_0 \phi$ will approximate f with error no more than ϵB_0. Besides, it is a trivial case when the Hölder norm of f is 0 . Firstly, when $\beta > 1$, we divide the proof into three steps as follows.

- Given $K \in \mathbb{N}^{+}$and $\delta \in (0, 1/(3K)]$, for each $\theta = (\theta_1, \ldots, \theta_d) \in \{0, 1, \ldots, K-1\}^d$, we define

$$Q_\theta := \left\{ x = (x_1, \ldots, x_d) : x_i \in \left[\frac{\theta_i}{K}, \frac{\theta_i + 1}{K} - \delta \cdot 1_{\theta_i < K-1} \right], i = 1, \ldots, d \right\}.$$

- Note that $[0,1]^d \backslash \Omega\left([0,1]^d, K, \delta\right) = \bigcup_\theta Q_\theta$. By the definition of Q_θ, the region $[0,1]^d$ is approximately divided into hypercubes. By Lemma B.1, there exists a ReLU network ψ_1 with width $4\left\lfloor N^{1/d} \right\rfloor + 3$ and depth $4M + 5$ such that

$$\psi_1(x) = \frac{k}{K}, \quad \text{if } x \in \left[\frac{k}{K}, \frac{k+1}{K} - \delta \cdot 1_{\{k < K-1\}} \right], k = 0, 1, \ldots, K-1.$$

- We define

$$\psi(x) := \left(\psi_1\left(x_1\right), \ldots, \psi_1\left(x_d\right)\right), \quad x = (x_1, \ldots, x_d) \in \mathbb{R}^d.$$

Then we have $\psi(x) = \theta/K := (\theta_1/K, \ldots, \theta_d/K)^\top$ for $x \in Q_\theta$ and ψ is a ReLU network with width $d\left(4\left\lfloor N^{1/d} \right\rfloor + 3\right)$ and depth $4M + 5$.

Proof: Approximation of Taylor coefficients

- Since $\theta \in \{0, 1, \ldots, K-1\}^d$ is one-to-one correspondence to $i_\theta := \sum_{j=1}^d \theta_j K^{j-1} \in \left\{0, 1 \ldots, K^d - 1\right\}$, we define

$$\psi_0(x) := \left(K, K^2, \ldots, K^d\right) \cdot \psi(x) = \sum_{j=1}^{d} \psi_1\left(x_j\right) K^j, \quad x \in \mathbb{R}^d,$$

then

$$\psi_0(x) = \sum_{j=1}^{d} \theta_j K^{j-1} = i_\theta, \qquad \text{if } x \in Q_\theta, \theta \in \{0, 1, \ldots, K-1\}^d,$$

where $\psi_0(x)$ has width $d\left(4\left\lfloor N^{1/d}\right\rfloor + 3\right)$ and depth $4M + 5$.

- For any $\alpha \in \mathbb{N}_0^d$ satisfying $\|\alpha\|_1 \leq s$ and each $i = i_\theta \in \left\{0, 1, \ldots, K^d - 1\right\}$, we denote $\xi_{\alpha,i} := \left(\partial^\alpha f(\theta/K) + 1\right)/2 \in [0, 1]$.
- Since $K^d \leq N^2 M^2$, there exists a ReLU network φ_α with width $16(s+1)(N+1)$ $\lceil \log_2(8N) \rceil$ and depth $5(M+2)\lceil \log_2(4M) \rceil$ such that

$$\left|\varphi_\alpha(i) - \xi_{\alpha,i}\right| \leq (NM)^{-2(s+1)}$$

for all $i \in \left\{0, 1, \ldots, K^d - 1\right\}$.

- We define

$$\phi_\alpha(x) := 2\varphi_\alpha\left(\psi_0(x)\right) - 1 \in [-1,1], \quad x \in \mathbb{R}^d.$$

Then ϕ_α can be implemented by a network with width $16d(s+1)(N+1)\left\lceil \log_2(8N) \right\rceil \leq 32d(s+1)N\left\lceil \log_2(8N) \right\rceil$ and depth $5(M+2)\left\lceil \log_2(4M) \right\rceil + 4M + 5 \leq 15M\left\lceil \log_2(8M) \right\rceil$. And we have for any $\theta\{0,1,\ldots,K-1\}^d$, if $x \in Q_\theta$,

$$\left|\phi_\alpha(x) - \partial^\alpha f(\theta/K)\right| = 2\left|\varphi_\alpha\left(i_\theta\right) - \xi_{\alpha,i_\theta}\right| \leq 2(NM)^{-2(s+1)}.$$

Proof: Approximation of f on $\bigcup_{\theta\in\{0,1,\dots,K-1\}^d} Q_\theta$

- Let $\varphi(t) = \min\{\max\{t, 0\}, 1\} = \sigma(t) - \sigma(t-1)$ for $t \in \mathbb{R}$ where $\sigma(\cdot)$ is the ReLU activation function. With a slightly abuse of the notation, we extend its definition to $\mathbb{R}^d$ coordinatewisely, i.e., $\varphi : \mathbb{R}^d \to [0,1]^d$ and $\varphi(x) = x$ for any $x \in [0,1]^d$.
- There exists a ReLU network with width $9N+1$ and depth $2(s+1)M$ such that for any $t_1, t_2 \in [-1, 1]$,

$$|t_1 t_2 - \phi_\times(t_1, t_2)| \le 24N^{-2(s+1)M}.$$

- For any $\alpha \in \mathbb{N}_0^d$ with $\alpha\|_2 \le s$, there exists a ReLU network P_α with width $9N + s + 8$ and depth $7(s+1)^2 M$ such that $P_\alpha(x) \in [-1, 1]$ and

$$|P_\alpha(x) - x^\alpha| \le 9(s+1)(N+1)^{-7(s+1)M}.$$

- For any $x \in Q_\theta, \theta \in \{0, 1, \dots, K-1\}^d$, we can now approximate the Taylor expansion of $f(x)$ by combined sub-networks. Thanks to Lemma A. 8 in Petersen and Voigtlaender (2018), we have the following error control for $x \in Q_\theta$,

$$\left| f(x) - f\left(\frac{\theta}{K}\right) - \sum_{1\le\|\alpha\|_1\le s} \frac{\partial^\alpha f\left(\frac{\theta}{K}\right)}{\alpha!}\left(x - \frac{\theta}{K}\right)^\alpha \right| \le d^s \left\| x - \frac{\theta}{K} \right\|_2^\beta \le d^{s+\beta/2} K^{-\beta}.$$

- Motivated by this, we define

$$\tilde{\phi}_0(x) := \phi_{\mathbf{0}_d}(x) + \sum_{1\le\|\alpha\|_1\le s} \phi_\times\left(\frac{\phi_\alpha(x)}{\alpha!}, P_\alpha(\varphi(x)-\phi(x))\right),$$

$$\phi_0(x) := \sigma\left(\tilde{\phi}_0(x)+1\right) - \sigma\left(\tilde{\phi}_0(x)-1\right) - 1 \in [-1,1],$$

  where $\mathbf{0}_d = (0,\ldots,0)\in\mathbb{N}_0^d$.

- Observe that the number of terms in the summation can be bounded by

$$\sum_{\alpha\in\mathbb{N}_o^d,\|\alpha\|_1\le s} 1 = \sum_{j=0}^{s}\sum_{\alpha\in\mathbb{N}_0^d,\|\alpha\|_1=j} 1 \le \sum_{j=0}^{s} d^s \le (s+1)d^s.$$

- Recall that width and depth of φ is $(2d,1)$, width and depth of ψ is $\left(d\left(4\left\lfloor N^{1/d}\right\rfloor+3\right), 4M+\right.$ 5), width and depth of P_α is $\left(9N+s+8, 7(s+1)^2M\right)$, width and depth of ϕ_α is width $(16d(s+1)(N+1)\lceil\log_2(8N)\rceil, 5(M+2)\lceil\log_2(4M)\rceil+4M+5)$ and width and depth of $\phi_\times$is $(9N+1, 2(s+1)M)$. Hence, by our construction, ϕ_0 can be implemented by a neural network with width $38(s+1)^2 d^{s+1} N\left[\log_2(8N)\right]$ and depth $21(s+1)^2 M\left[\log_2(8M)\right]$.

- For any $x \in Q_\theta, \varphi(x) = x$ and $\psi(x) = \theta/K$,

$$
\begin{aligned}
&|f(x) - \phi_0(x)| \le \left| f(x) - \tilde{\phi}_0(x) \right| \\
\le &|f(\theta/K) - \phi_{\mathbf{0}_d}(x)| + d^{s+\beta/2} K^{-\beta} \\
&+ \sum_{1\le\|\alpha\|_1\le s} \left| \frac{\partial^\alpha f(\theta/K)}{\alpha!}(x-\theta/K)^\alpha - \phi_\times\left(\frac{\phi_\alpha(x)}{\alpha!}, P_\alpha(x-\theta/K)\right) \right| \\
= &d^{s+\beta/2} \left\lfloor (MN)^{2/d} \right\rfloor^{-\beta} + \sum_{\|\alpha\|_1\le s} \mathcal{E}_\alpha,
\end{aligned}
$$

  where we denote $\mathcal{E}_\alpha = \left| \frac{\partial^\alpha f(\theta/K)}{\alpha!}(x-\theta/K)^\alpha - \phi_\times\left(\frac{\phi_\alpha(x)}{\alpha!}, P_\alpha(x-\theta/K)\right) \right|$ for each $\alpha \in \mathbb{N}_0^d$ with $\|\alpha\|_1 \le s$.

- Using the inequality $|t_1 t_2 - \phi_\times(t_3,t_4)| \le |t_1 t_2 - t_3 t_2| + |t_3 t_2 - t_3 t_4| + |\, t_3 t_4 - \phi_\times(t_3,t_4)\,| \le |t_1 - t_3| + |t_2 - t_4| + |t_3 t_4 - \phi_\times(t_3,t_4)|$ for any $t_1,t_2,t_3,t_4 \in [-1,1]$, then for $1 \le \|\alpha\|_1 \le s$ we have

$$
\begin{aligned}
\mathcal{E}_\alpha \le &2(NM)^{-2(s+1)} + 9(s+1)(N+1)^{-7(s+1)M} + 6N^{-2(s+1)M} \\
\le &(9s+17)(NM)^{-2(s+1)}.
\end{aligned}
$$

- It is easy to check that the bound is also true when $\|\alpha\|_1 = 0$ and $s = 0$. Therefore,

$$\begin{aligned}
|f(x) - \phi_0(x)| &\leq \sum_{1\leq\|\alpha\|_1\leq s} (9s+17)(NM)^{-2(s+1)} + d^{s+\beta/2}(NM)^{-2\beta/d} \\
&\leq (s+1)d^s(9s+17)(NM)^{-2(s+1)} + d^{s+\beta/2}(NM)^{-2\beta/d} \\
&\leq 18(s+1)^2 d^{s+\beta/2}(NM)^{-2\beta/d}
\end{aligned}$$

for any $x \in \bigcup_{\theta\in\{0,1,\ldots,K-1\}^d} Q_\theta$. And for $f \in \mathcal{H}^\beta\left([0,1]^d, B_0\right)$, by approximate f/B_0 firstly, we know there exists a function implemented by a neural network with the same width and depth as ϕ_0, such that

$$|f(x) - \phi_0(x)| \leq 18B_0(s+1)^2 d^{s+\beta/2}(NM)^{-2\beta/d}$$

for any $x \in \bigcup_{\theta\in\{0,1,\ldots,K-1\}^d} Q_\theta$.

Corollary 11.

Assume that $f \in \mathcal{H}^{\beta}\left([0,1]^{d}, B_0\right)$ *with* $\beta = s + r, s \in \mathbb{N}_0$ *and* $r \in (0,1]$. *For any* $M, N \in \mathbb{N}^{+}$, *there exists a function* ϕ_0 *implemented by a ReLU network with width* $\mathcal{W} = 38(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor+1} N \lceil \log_2(8N) \rceil$ *and depth* $\mathcal{D} = 21(\lfloor\beta\rfloor + 1)^2 M \lceil \log_2(8M) \rceil + 2d$ *such that*

$$|f(x) - \phi_0(x)| \leq 19B_0(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor + (\beta\vee 1)/2} (NM)^{-2\beta/d}$$

for all $x \in [0,1]^d$.

Theorem 12 (Consistency).

Suppose that Y is sub-exponentially distributed, the target function f_0 is continuous on $[0,1]^d$, and $\|f_0\|_\infty \leq \mathcal{B}$ for some $\mathcal{B} \geq 1$, and the function class of feedforward neural networks $\mathcal{F}_n = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ with continuous piecewise-linear activation function with finitely many inflection points satisfies

$$\mathcal{S} \to \infty \quad \text{and} \quad \mathcal{B}^5 (\log n)^5 \frac{1}{n} \mathcal{S}\mathcal{D} \log(\mathcal{S}) \to 0, \text{ as } n \to \infty,$$

Then, the prediction error of the empirical risk minimizer $\hat{f}_n$ is consistent in the sense that

$$\mathbb{E}\left\|\hat{f}_n - f_0\right\|_{L^2(\nu)}^2 \to 0 \text{ as } n \to \infty.$$

The conditions are sufficient for the consistency of the deep neural regression, and they are relatively mild in terms of the assumptions on the underlying target f_0 and the distribution of Y. Van de Geer and Wegkamp (1996) gave the sufficient and necessary conditions for the consistency of the least squares estimation in nonparametric regression under the assumptions that $f_0 \in \mathcal{F}_n$, the error η is symmetric about 0 and it has zero point mass at 0 . Their results are for the convergence of the empirical error $\left\|\hat{f}_n - f_0\right\|_n^2 := \sum_{i=1}^n \left|\hat{f}_n(X_i) - f_0(X_i)\right|^2 / n$.

Theorem 13 (Non-asymptotic error bound).

suppose that Assumptions 1-2 hold, the probability measure of the covariate ν is absolutely continuous with respect to the Lebesgue measure and $\mathcal{B} \geq \max\{B_0, 1\}$. Then, for any $N, M \in \mathbb{N}^+$, the function class of ReLU multi-layer perceptrons $\mathcal{F}_n = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ with width $\mathcal{W} = 38(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor+1} N \lceil\log_2(8N)\rceil$ and depth $\mathcal{D} = 21(\lfloor\beta\rfloor + 1)^2 M \lceil\log_2(8M)\rceil$, for $n \geq \mathrm{Pdim}(\mathcal{F}_n)/2$, the prediction error of the $ERM\,\hat{f}_n$ satisfies

$$\mathbb{E}\left\|\hat{f}_n - f_0\right\|^2_{L^2(\nu)} \leq C\mathcal{B}^5(\log n)^5 \frac{1}{n}\mathcal{S}\mathcal{D}\log(\mathcal{S}) + 324B_0^2(\lfloor\beta\rfloor + 1)^4 d^{2\lfloor\beta\rfloor+\beta\vee 1}(NM)^{-4\beta/d}.$$

where $C > 0$ is a constant not depending on $n, d, \mathcal{B}, \mathcal{S}, \mathcal{D}, B_0, \beta, N$ or M.

Table of Contents

Assumption 3.

The predictor X is supported on $\mathcal{M}_\rho$, a ρ-neighborhood of $\mathcal{M} \subset [0,1]^d$, where $\mathcal{M}$ is a compact $d_{\mathcal{M}}$ -dimensional Riemannian submanifold and

$$\mathcal{M}_\rho = \left\{ x \in [0,1]^d : \inf\{\|x-y\|_2 : y \in \mathcal{M}\} \leq \rho \right\}, \rho \in (0,1).$$

Assumption 4.

The predictor X is supported on $\mathcal{M} \subset [0,1]^d$, where a $\mathcal{M}$ is a compact $d_{\mathcal{M}}$-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^d$ with condition number $(1/\tau)$ and area of surface $S_{\mathcal{M}}$.

Theorem 14 (Non-asymptotic error bound).

Suppose that Assumptions 1-3 hold, the probability measure ν of X is absolutely continuous with respect to the Lebesgue measure and $\mathcal{B} \geq \max\{1, B_0\}$. Then for any $N, M \in \mathbb{N}^+$, the function class of ReLU multi-layer perceptrons $\mathcal{F}_n = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ with width $\mathcal{W} = 38(\lfloor\beta\rfloor + 1)^2 d_\delta^{\lfloor\beta\rfloor+1} N \lceil \log_2(8N) \rceil$ and depth $\mathcal{D} = 21(\lfloor\beta\rfloor + 1)^2 M \lceil \log_2(8M) \rceil$, the prediction error of the empirical risk minimizer $\hat{f}_n$ satisfies

$$\mathbb{E}\left\|\hat{f}_n - f_0\right\|_{L^2(\nu)}^2 \leq C_1 \mathcal{B}^5 \frac{\mathcal{S}\mathcal{D}\log(\mathcal{S})(\log n)^5}{n} + \frac{(36 + C_2)^2 B_0^2}{(1-\delta)^{2\beta}} (\lfloor\beta\rfloor + 1)^4 d d_\delta^{3\lfloor\beta\rfloor} (NM)^{-4\beta/d_\delta}$$

for $n \geq P \dim(\mathcal{F}_n)/2$ and
$\rho \leq C_2 (NM)^{-2\beta/d_\delta} (s+1)^2 d^{1/2} d_\delta^{3s/2} \left(\sqrt{d/d_\delta} + 1 - \delta\right)^{-1} (1 - \delta)^{1-\beta}$*, where $d_\delta = O\left(d_{\mathcal{M}} \log(d/\delta)/\delta^2\right)$ is an integer such that $d_{\mathcal{M}} \leq d_\delta < d$ for any $\delta \in (0,1)$, and $C_1, C_2 > 0$ are constants that do not depend on $n, \mathcal{B}, \mathcal{S}, \mathcal{D}, B_0, \beta, \rho, \delta, N$ or M.*

- To achieve the optimal convergence rate with a minimal network size, we can set $\mathcal{F}_n = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ to consist of fixed-width networks with $\mathcal{W} = 114(\lfloor\beta\rfloor + 1)^2 d_\delta^{\lfloor\beta\rfloor+1}, \mathcal{D} = 21(\lfloor\beta\rfloor + 1)^2 \left\lceil n^{d_\delta/2(d_\delta+2\beta)} \log_2\left(8n^{d_\delta/2(d_\delta+2\beta)}\right)\right\rceil, \mathcal{S} = O\left(\mathcal{W}^2\mathcal{D}\right) = O((\lfloor\beta\rfloor + 1)^6 d_\delta^{2\lfloor\beta\rfloor+2} \left\lceil n^{d_\delta/2(d_\delta+2\beta)} (\log_2 n)\right\rceil)$
- Then the prediction error of $\hat{f}_n$ becomes

$$\mathbb{E}\left\|\hat{f}_n - f_0\right\|^2_{L^2(\nu)} \leq C_3(1-\delta)^{-2\beta}\mathcal{B}^5 d d_\delta^{3\lfloor\beta\rfloor+3}(\lfloor\beta\rfloor + 1)^9 n^{-2\beta/(d_\delta+2\beta)}(\log n)^8.$$

  where $C_3 > 0$ is a constant not depending on $n, d, d_\delta, \mathcal{B}, \mathcal{S}, \mathcal{D}, B_0, \delta$ or β.
- It shows that nonparametric regression using deep neural networks can alleviate the curse of dimensionality under an approximate manifold assumption.

6.2. Exact low-dimensional manifold assumption. Under the exact manifold support assumption, we show that the $\log(d)$ factor in (14) can be removed. We establish error bounds

Theorem 15 (Non-asymptotic error bound).

Suppose that Assumptions 1, 2 and 4 hold, and $\mathcal{B} \geq \max\{1, B_0\}$. *Then for any* $N, M \in \mathbb{N}^+$, *the function class of ReLU multi-layer perceptrons* $\mathcal{F}_n$ *with* $\mathcal{W} = 266(\lfloor\beta\rfloor + 1)^2 \left\lceil S_{\mathcal{M}}(6/\tau)^{d_{\mathcal{M}}} \right\rceil (d_{\mathcal{M}})^{\lfloor\beta\rfloor+2} N \left\lceil \log_2(8N) \right\rceil$ *and depth* $\mathcal{D} = 21(\lfloor\beta\rfloor + 1)^2 M \left\lceil \log_2(8M) \right\rceil + 2d_{\mathcal{M}} + 2$, *the prediction error satisfies*

$$\mathbb{E}\left\|\hat{f}_n - f_0\right\|^2_{L^2(\nu)} \leq C_1 \mathcal{B}^5 \frac{\mathcal{S}\mathcal{D}\log(\mathcal{S})(\log n)^5}{n} + C_2 B_0^2 (\lfloor\beta\rfloor+1)^4 d\,(d_{\mathcal{M}})^{3\lfloor\beta\rfloor+1} (NM)^{-4\beta/d_{\mathcal{M}}},$$

for $n \geq \mathrm{Pdim}(\mathcal{F}_n)/2$, *where* $C_2 > 0$ *is a constant. If we set*

$$\mathcal{W} = 798(\lfloor\beta\rfloor + 1)^2 \left\lceil S_{\mathcal{M}}(6/\tau)^{d_{\mathcal{M}}} \right\rceil (d_{\mathcal{M}})^{\lfloor\beta\rfloor+2},$$

$$\mathcal{D} = 21(\lfloor\beta\rfloor + 1)^2 \left\lceil n^{d_{\mathcal{M}}/2(d_{\mathcal{M}}+2\beta)} \log_2\left(8n^{d_{\mathcal{M}}/2(d_{\mathcal{M}}+2\beta)}\right) \right\rceil + 2d_{\mathcal{M}} + 2,$$

$$\mathcal{S} = O\left((\lfloor\beta\rfloor + 1)^6 d (6/\tau)^{2d_{\mathcal{M}}} (d_{\mathcal{M}})^{2\lfloor\beta\rfloor+5} n^{d_{\mathcal{M}}/2(d_{\mathcal{M}}+2\beta)} \log_2(n)\right),$$

the prediction error of $\hat{f}_n$ *satisfies*

$$\mathbb{E}\|\hat{f}_n - f_0\|^2_{L^2(\nu)} \leq C_3 \mathcal{B}^5 (\lfloor\beta\rfloor + 1)^9 (6/\tau)^{2d_{\mathcal{M}}} (d_{\mathcal{M}})^{3\lfloor\beta\rfloor+6} d (\log n)^8 n^{-2\beta/(d_{\mathcal{M}}+2\beta)},$$

where $C_3 > 0$ *is a constant.*