Towards a regularity theory for ReLU networks – chain rule and global error estimates

Julius Berner
joint work with Dennis Elbrächter, Philipp Grohs, and Arnulf Jentzen

Faculty of Mathematics,
University of Vienna, Austria

October 9, 2019

Artificial Feed-Forward Neural Network

- parametrization of neural network with architecture $(N_0, \ldots, N_L)$:

$$\Phi = ((A_\ell, b_\ell))_{\ell=1}^L$$

where $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and $b_\ell \in \mathbb{R}^{N_\ell}$

Artificial Feed-Forward Neural Network

- parametrization of neural network with architecture $(N_0, \ldots, N_L)$:

$$\Phi = ((A_\ell, b_\ell))_{\ell=1}^L$$

  where $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and $b_\ell \in \mathbb{R}^{N_\ell}$

- activation function ϱ: locally Lipschitz continuous with at most countably many points of non-differentiability

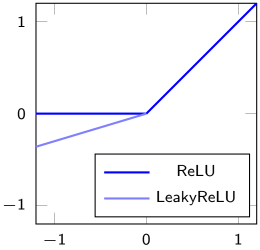
Artificial Feed-Forward Neural Network

- parametrization of neural network with architecture $(N_0, \dots, N_L)$:

$$\Phi = ((A_\ell, b_\ell))_{\ell=1}^{L}$$

  where $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and $b_\ell \in \mathbb{R}^{N_\ell}$

- activation function ϱ: locally Lipschitz continuous with at most countably many points of non-differentiability, e.g.
  - $\text{ReLU}(x) = max\{x, 0\}$
  - $\text{LeakyReLU}(x) = \max\{\alpha x, x\}, \quad \alpha \in (0, 1)$



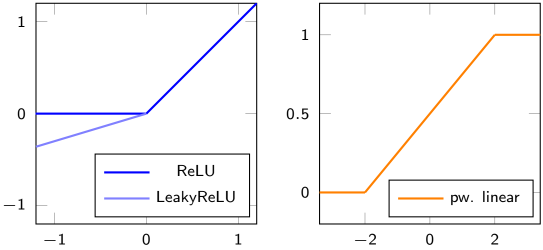
Artificial Feed-Forward Neural Network

- parametrization of neural network with architecture $(N_0, \dots, N_L)$:

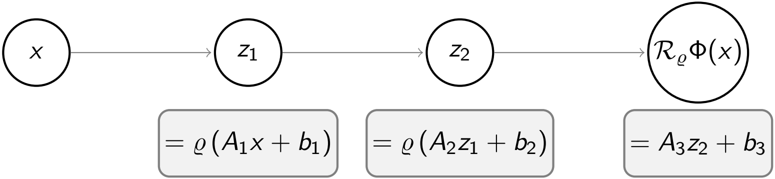
$$\Phi = ((A_\ell, b_\ell))_{\ell=1}^L$$

where $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and $b_\ell \in \mathbb{R}^{N_\ell}$

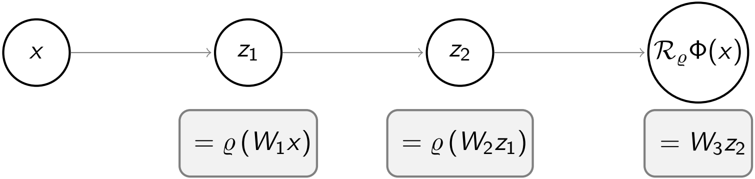
- activation function ϱ: locally Lipschitz continuous with at most countably many points of non-differentiability, e.g.
  - $\mathrm{ReLU}(x) = max\{x, 0\}$
  - $\mathrm{LeakyReLU}(x) = \max\{\alpha x, x\}, \quad \alpha \in (0, 1)$



Artifical Feed-Forward Neural Network



Artifical Feed-Forward Neural Network



Definition (realization of a parametrization)

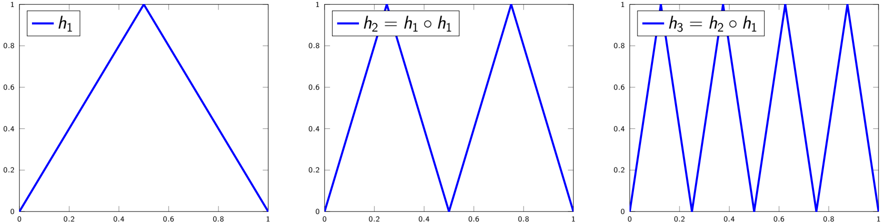
Realization $\mathcal{R}\Phi \in \mathcal{W}^{1,\infty}_{\text{loc}}(\mathbb{R}^{N_0}, \mathbb{R}^{N_L})$ of parametrization $\Phi = ((A_\ell, b_\ell))_{\ell=1}^L$:

$$\mathcal{R}\Phi := W_L \circ \varrho \circ W_{L-1} \circ \ldots \circ \varrho \circ W_1$$

where $W_k(x) := A_k x + b_k$ and ϱ is applied component-wise.

Motivation - Approximation Results for ReLU networks

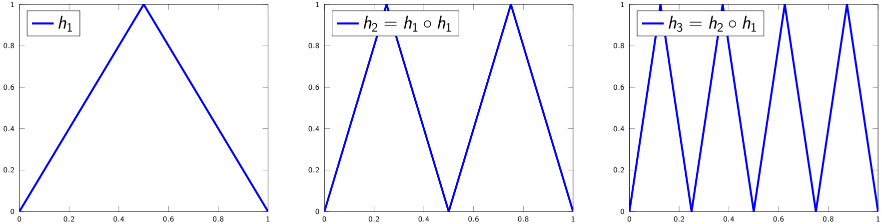
- sawtooth function $\#\text{parameters} \lesssim \log(\#\textit{teeth})$



Motivation - Approximation Results for ReLU networks

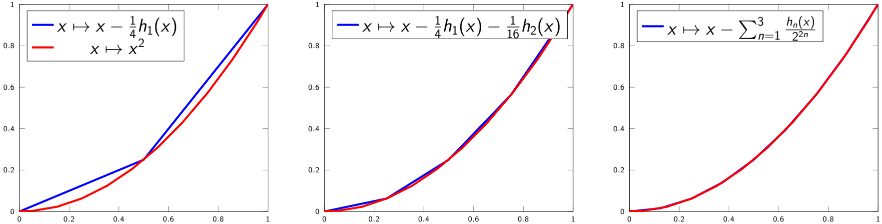
- sawtooth function

$\#parameters \lesssim \log(\#teeth)$



$\Rightarrow$ squaring function†

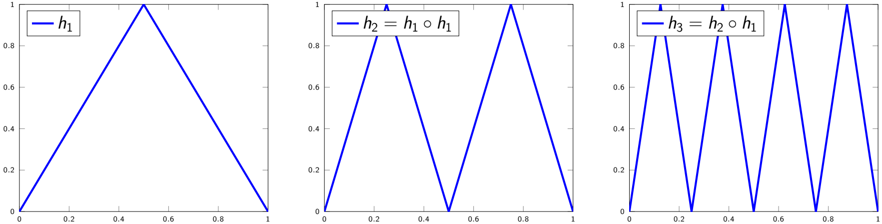
† $\#parameters \lesssim \log(precision^{-1})$



Motivation - Approximation Results for ReLU networks

- sawtooth function

#parameters $\lesssim \log(\#\mathit{teeth})$



$\Rightarrow$ squaring function†

$\Rightarrow$ multiplication†

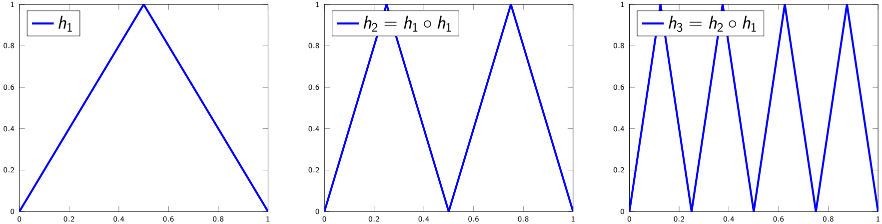
† #parameters $\lesssim \log(\mathit{precision}^{-1})$

$$xy = \left|\frac{x+y}{2}\right|^2 - \left|\frac{x-y}{2}\right|^2$$

Motivation - Approximation Results for ReLU networks

- sawtooth function

$\#\text{parameters} \lesssim \log(\#\textit{teeth})$



$\Rightarrow$ squaring function†

$\Rightarrow$ multiplication†

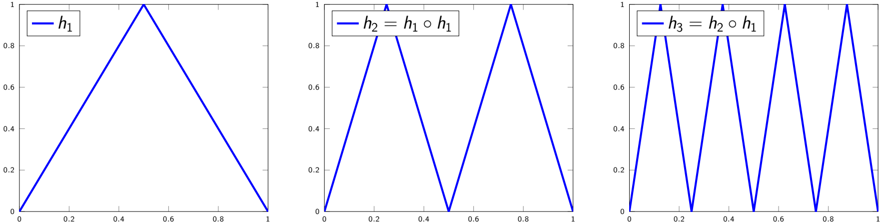
$\Rightarrow$ polynomials†

† $\#\text{parameters} \lesssim \log(\textit{precision}^{-1})$

Motivation - Approximation Results for ReLU networks

- sawtooth function

#parameters $\lesssim \log(\#\textit{teeth})$



$\Rightarrow$ squaring function†

$\Rightarrow$ multiplication†

$\Rightarrow$ polynomials†

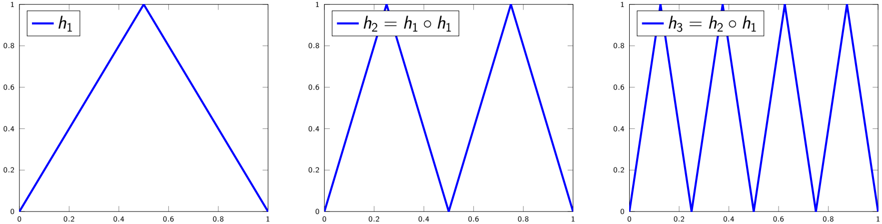
$\Rightarrow$ Sobolev-regular functions - Yarotsky '16

† #parameters $\lesssim \log(\textit{precision}^{-1})$

Motivation - Approximation Results for ReLU networks

- sawtooth function

$\#\text{parameters} \lesssim \log(\#\textit{teeth})$



$\Rightarrow$ squaring function†

† $\#\text{parameters} \lesssim \log(\textit{precision}^{-1})$

$\Rightarrow$ multiplication†

$\Rightarrow$ polynomials†

$\Rightarrow$ Sobolev-regular functions - Yarotsky '16

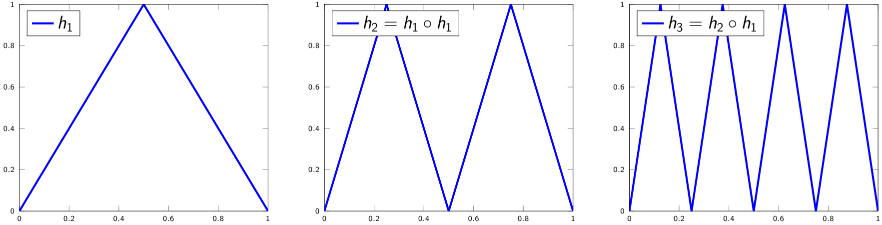
- in particular‡ $x \mapsto cos(2\pi x)$

‡ $\#\text{parameters} \lesssim \log^2(\textit{precision}^{-1})$

Motivation - Approximation Results for ReLU networks

- sawtooth function

$\#\text{parameters} \lesssim \log(\#\textit{teeth})$



$\Rightarrow$ squaring function†

† $\#\text{parameters} \lesssim \log(\textit{precision}^{-1})$

$\Rightarrow$ multiplication†

$\Rightarrow$ polynomials†

$\Rightarrow$ Sobolev-regular functions - Yarotsky '16

- in particular‡ $x \mapsto cos(2\pi x)$

‡ $\#\text{parameters} \lesssim \log^2(\textit{precision}^{-1})$

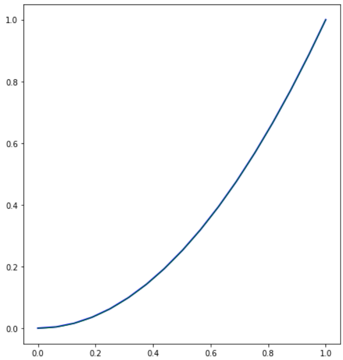
$\Rightarrow$ high-frequent cosine‡ - Perekrestenko, Grohs, Elbrächter, Bölcskei '18

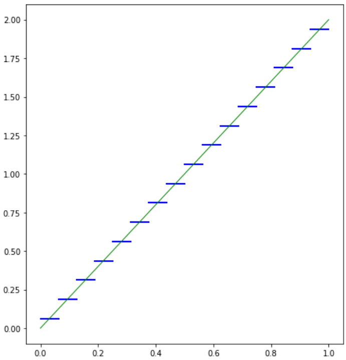
$$\cos(2\pi 2^n x) = cos(2\pi h_n(x))$$

Goal

- ReLU networks approximating previous functions in the $\mathcal{W}^{1,\infty}$ norm (applications in the numerical solution of PDEs)

$\Rightarrow$ same construction for squaring function

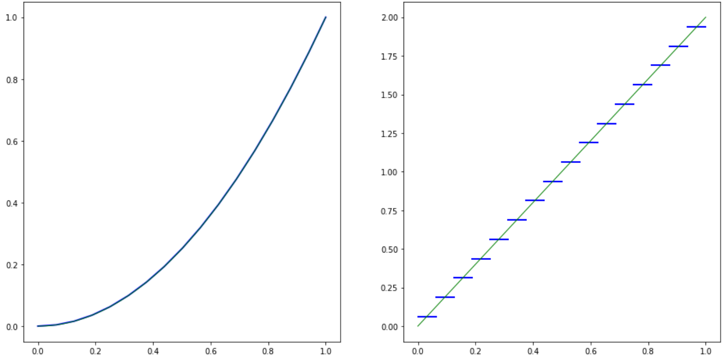




Goal

- ReLU networks approximating previous functions in the $\mathcal{W}^{1,\infty}$ norm (applications in the numerical solution of PDEs)

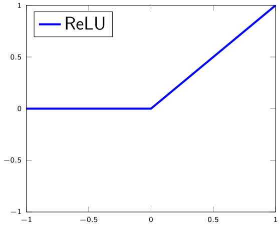
$\Rightarrow$ same construction for squaring function



Problem: classical chain rule fails!

Motivation - Failure of the Classical Chain Rule

- $u(x) := \mathrm{ReLU}(x), \quad v(x) := 0$

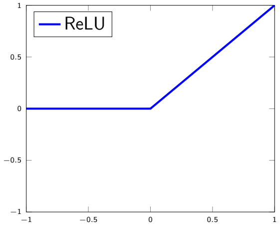


- chain rule (formally):

$$D(u \circ v)(x) = Du(v(x)) \cdot Dv(x)$$

Motivation - Failure of the Classical Chain Rule

- $u(x) := \mathrm{ReLU}(x), \quad v(x) := 0$

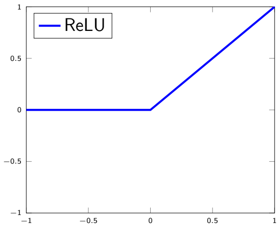


- chain rule (formally):

$$D(u \circ v)(x) = \underbrace{Du(v(x))}_{\text{not defined}} \cdot Dv(x)$$

Motivation - Failure of the Classical Chain Rule

- $u(x) := \mathrm{ReLU}(x), \quad v(x) := 0$



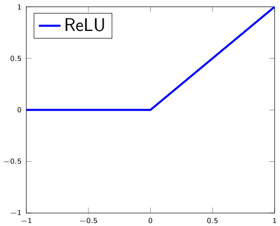
- chain rule (formally):

$$D(u \circ v)(x) = \underbrace{Du(v(x))}_{\text{not defined}} \cdot Dv(x)$$

$\Rightarrow$ define
$$\bar{D}[\mathrm{ReLU}](x) := \begin{cases} 1, & x > 0 \\ c, & x = 0 \\ 0, & x < 0 \end{cases} \qquad (c \in \mathbb{R})$$

Motivation - Failure of the Classical Chain Rule

- $u(x) := \text{ReLU}(x), \quad v(x) := 0$



- chain rule (formally):

$$D(u \circ v)(x) = \underbrace{Du(v(x))}_{\text{not defined}} \cdot Dv(x)$$

$\Rightarrow$ define $\bar{D}[\text{ReLU}](x) := \begin{cases} 1, & x > 0 \\ c, & x = 0 \\ 0, & x < 0 \end{cases} \qquad (c \in \mathbb{R})$

- deep learning libraries (TensorFlow, PyTorch): $c = 0 \quad \Rightarrow$ sparsity

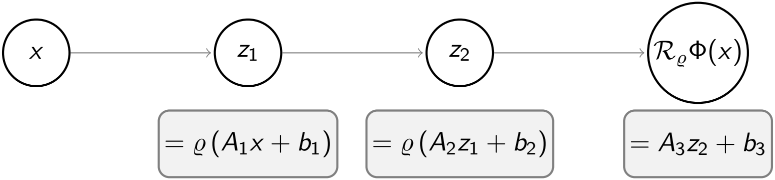
Definition of the Neural Network Derivative

Definition (ReLU network derivative)

Derivative $\mathcal{D}\Phi \in \mathcal{L}^\infty(\mathbb{R}^{N_0}, \mathbb{R}^{N_L \times N_0})$ of parametrization $\Phi = ((A_\ell, b_\ell))_{\ell=1}^L$:

$$\mathcal{D}\Phi := A_L \cdot \Delta_{L-1} \cdot A_{L-1} \cdot \ldots \cdot \Delta_1 \cdot A_1$$

where $\Delta_k := \operatorname{diag}(\bar{D}[\mathrm{ReLU}] \circ \mathcal{R}((A_\ell, b_\ell))_{\ell=1}^k)$ and $\bar{D}[\mathrm{ReLU}]$ is applied component-wise.



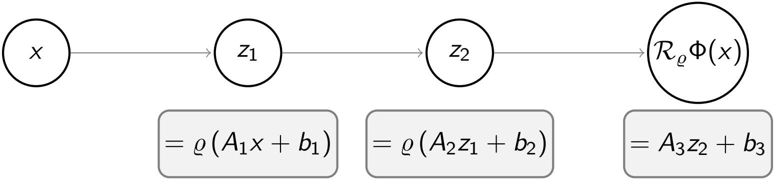
Definition of the Neural Network Derivative

Definition (network derivative with activation ϱ)

Derivative $\mathcal{D}\Phi \in \mathcal{L}^{\infty}_{\mathrm{loc}}(\mathbb{R}^{N_0}, \mathbb{R}^{N_L \times N_0})$ of parametrization $\Phi = ((A_\ell, b_\ell))_{\ell=1}^{L}$:

$$\mathcal{D}\Phi := A_L \cdot \Delta_{L-1} \cdot A_{L-1} \cdot \ldots \cdot \Delta_1 \cdot A_1$$

where $\Delta_k := \operatorname{diag}(\bar{D}\varrho \circ \mathcal{R}((A_\ell, b_\ell))_{\ell=1}^{k})$ and $\bar{D}\varrho$ is applied component-wise.



Definition of the Neural Network Derivative

Definition (network derivative with activation ϱ)

Derivative $\mathcal{D}\Phi \in \mathcal{L}^{\infty}_{\text{loc}}(\mathbb{R}^{N_0}, \mathbb{R}^{N_L \times N_0})$ of parametrization $\Phi = ((A_\ell, b_\ell))_{\ell=1}^L$:

$$\mathcal{D}\Phi := A_L \cdot \Delta_{L-1} \cdot A_{L-1} \cdot \ldots \cdot \Delta_1 \cdot A_1$$

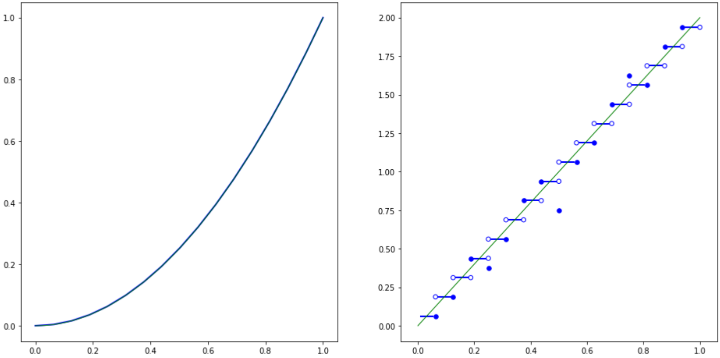
where $\Delta_k := \operatorname{diag}(\bar{D}\varrho \circ \mathcal{R}((A_\ell, b_\ell))_{\ell=1}^k)$ and $\bar{D}\varrho$ is applied component-wise.

Lemma (properties - B., Elbrächter, Grohs, Jentzen '19)

- well-defined: $\mathcal{D}\Phi = D[\mathcal{R}\Phi]$ a.e.
- chain-rule: $\mathcal{D}(\Psi \circ \Phi) = \mathcal{D}\Psi(\mathcal{R}\Phi) \cdot \mathcal{D}\Phi$
- stability: $\mathcal{D}(\Psi \circ \Phi)(x) = \lim_{y \to \mathcal{R}\Phi(x)} \mathcal{D}\Psi(y) \cdot \mathcal{D}\Phi(x)$ a.e. $x \in \mathbb{R}^{N_0}$

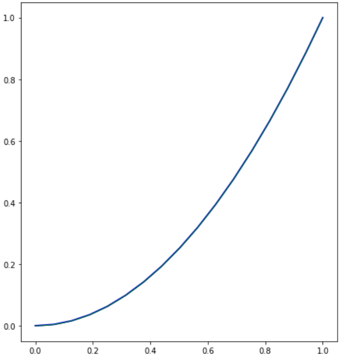
Behaviour of the Derivative on Nullsets

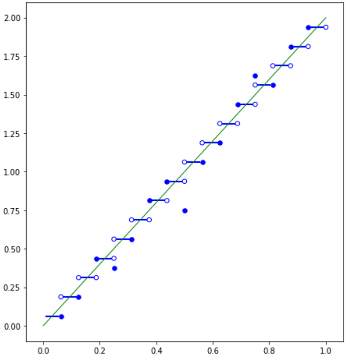
- $\mathcal{D}\Phi$ does not necessarily lie in the subdifferential regardless of choice of c (derivative at points of non-differentiability)



Behaviour of the Derivative on Nullsets

- $\mathcal{D}\Phi$ does not necessarily lie in the subdifferential regardless of choice of c (derivative at points of non-differentiability)

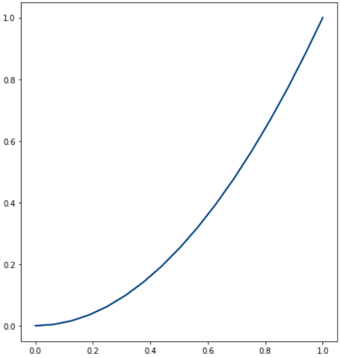


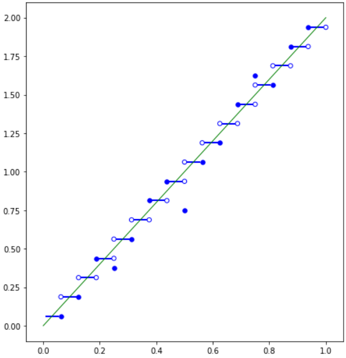


- similar behavior during back-propagation for derivative of $\Phi \mapsto \mathcal{R}\Phi(x^*)$ (fixed sample x^*)

Behaviour of the Derivative on Nullsets

- $\mathcal{D}\Phi$ does not necessarily lie in the subdifferential regardless of choice of c (derivative at points of non-differentiability)





- similar behavior during back-propagation for derivative of $\Phi \mapsto \mathcal{R}\Phi(x^*)$ (fixed sample x^*)
- Problems for neural network training? $\Rightarrow$ Correct Automatic Subdifferentiation - Kakade, Lee '18

Applications - $\mathcal{W}^{1,\infty}$ Approximation for ReLU networks

Theorem (multivariate polynomials - B., Elbrächter, Grohs, Jentzen '19)

For every $\varepsilon \in (0,1)$ and polynomial

$$p(x) = \sum_{\alpha \in I} c_\alpha x^\alpha \quad \Big(I = \big\{ \alpha \in \mathbb{N}_0^d \colon |\alpha| \leq n \big\}, \quad c \in \mathbb{R}^I \Big)$$

there is a parametrization Φ with

$$\|p - \mathcal{R}\Phi\|_{\mathcal{W}^{1,\infty}((0,1)^d)} \leq \varepsilon$$

Applications - $\mathcal{W}^{1,\infty}$ Approximation for ReLU networks

Theorem (multivariate polynomials - B., Elbrächter, Grohs, Jentzen '19)

For every $\varepsilon \in (0,1)$ and polynomial

$$p(x) = \sum_{\alpha \in I} c_\alpha x^\alpha \quad \Big(I = \big\{\alpha \in \mathbb{N}_0^d \colon |\alpha| \le n\big\}, \quad c \in \mathbb{R}^I \Big)$$

there is a parametrization Φ with

- $\|\Phi\|_\infty \le 4$

and

$$\|p - \mathcal{R}\Phi\|_{\mathcal{W}^{1,\infty}((0,1)^d)} \le \varepsilon$$

Applications - $\mathcal{W}^{1,\infty}$ Approximation for ReLU networks

Theorem (multivariate polynomials - B., Elbrächter, Grohs, Jentzen '19)

For every $\varepsilon \in (0,1)$ and polynomial

$$p(x) = \sum_{\alpha \in I} c_\alpha x^\alpha \quad \Big(I = \{\alpha \in \mathbb{N}_0^d \colon |\alpha| \le n\}, \quad c \in \mathbb{R}^I \Big)$$

there is a parametrization Φ with

- $\|\Phi\|_\infty \le 4$
- $\operatorname{depth}(\Phi) \lesssim \log^3(dn)\big[\log(\|c\|_1) + \log(\varepsilon^{-1})\big]$

and

$$\|p - \mathcal{R}\Phi\|_{\mathcal{W}^{1,\infty}((0,1)^d)} \le \varepsilon$$

Applications - $\mathcal{W}^{1,\infty}$ Approximation for ReLU networks

Theorem (multivariate polynomials - B., Elbrächter, Grohs, Jentzen '19)

For every $\varepsilon \in (0,1)$ and polynomial

$$p(x) = \sum_{\alpha \in I} c_\alpha x^\alpha \quad \Big(I = \{\alpha \in \mathbb{N}_0^d : |\alpha| \leq n\}, \quad c \in \mathbb{R}^I \Big)$$

there is a parametrization Φ with

- $\|\Phi\|_\infty \leq 4$
- $\mathrm{depth}(\Phi) \lesssim \log^3(dn)\big[\log(\|c\|_1) + \log(\varepsilon^{-1})\big]$
- $\mathrm{width}(\Phi) \lesssim d(\|c\|_0 + n)$

and

$$\|p - \mathcal{R}\Phi\|_{\mathcal{W}^{1,\infty}((0,1)^d)} \leq \varepsilon$$

Applications - $\mathcal{W}^{1,\infty}$ Approximation for ReLU networks

Theorem (multivariate polynomials - B., Elbrächter, Grohs, Jentzen '19)

For every $\varepsilon \in (0,1)$ and polynomial

$$p(x) = \sum_{\alpha \in I} c_\alpha x^\alpha \quad \left(I = \{\alpha \in \mathbb{N}_0^d : |\alpha| \leq n\}, \quad c \in \mathbb{R}^I \right)$$

there is a parametrization Φ with

- $\|\Phi\|_\infty \leq 4$
- $\operatorname{depth}(\Phi) \lesssim \log^3(dn)\big[\log(\|c\|_1) + \log(\varepsilon^{-1})\big]$
- $\operatorname{width}(\Phi) \lesssim d(\|c\|_0 + n)$
- $\|\Phi\|_0 \lesssim \operatorname{depth}(\Phi) \cdot \operatorname{width}(\Phi)$

and

$$\|p - \mathcal{R}\Phi\|_{\mathcal{W}^{1,\infty}((0,1)^d)} \leq \varepsilon$$

Applications - $\mathcal{W}^{1,\infty}$ Approximation for ReLU networks

Theorem (multivariate polynomials - B., Elbrächter, Grohs, Jentzen '19)

For every $\varepsilon \in (0,1)$ and polynomial

$$p(x) = \sum_{\alpha \in I} c_\alpha x^\alpha \quad \Big(I = \{\alpha \in \mathbb{N}_0^d : |\alpha| \le n\}, \quad c \in \mathbb{R}^I \Big)$$

there is a parametrization Φ with

- $\|\Phi\|_\infty \le 4$
- $\mathrm{depth}(\Phi) \lesssim \log^3(dn)\big[\log(\|c\|_1) + \log(\varepsilon^{-1})\big]$
- $\mathrm{width}(\Phi) \lesssim d(\|c\|_0 + n)$
- $\|\Phi\|_0 \lesssim \mathrm{depth}(\Phi) \cdot \mathrm{width}(\Phi)$

$\|c\|_0 \le \binom{n+d}{d}$

and

$$\|p - \mathcal{R}\Phi\|_{\mathcal{W}^{1,\infty}((0,1)^d)} \le \varepsilon$$

Applications - $\mathcal{W}^{1,\infty}$ Approximation for ReLU networks

Lemma (chain-rule in $\mathcal{W}^{1,\infty}$- Gühring, Kutyniok, Petersen '19)

$$|g \circ f|_{\mathcal{W}^{1,\infty}} \leq C|g|_{\mathcal{W}^{1,\infty}}|f|_{\mathcal{W}^{1,\infty}} \qquad (f, g \in \mathcal{W}^{1,\infty})$$

Applications - $\mathcal{W}^{1,\infty}$ Approximation for ReLU networks

Lemma (chain-rule in $\mathcal{W}^{1,\infty}$- Gühring, Kutyniok, Petersen '19)

$$|g \circ f|_{\mathcal{W}^{1,\infty}} \leq C|g|_{\mathcal{W}^{1,\infty}}|f|_{\mathcal{W}^{1,\infty}} \qquad (f, g \in \mathcal{W}^{1,\infty})$$

drawback: C depends on input dimensions of f and g and can be avoided

Applications - $\mathcal{W}^{1,\infty}$ Approximation for ReLU networks

Lemma (chain-rule in $\mathcal{W}^{1,\infty}$- Gühring, Kutyniok, Petersen '19)

$$|g \circ f|_{\mathcal{W}^{1,\infty}} \leq C|g|_{\mathcal{W}^{1,\infty}}|f|_{\mathcal{W}^{1,\infty}} \qquad (f, g \in \mathcal{W}^{1,\infty})$$

drawback: C depends on input dimensions of f and g and can be avoided

Theorem (Sobolev-regular functions - Gühring, Kutyniok, Petersen '19)

For every $\varepsilon \in (0,1)$ and

$$f \in \mathcal{W}^{n,\infty}((0,1)^d) \quad \text{with} \quad \|f\|_{\mathcal{W}^{n,\infty}((0,1)^d)} \leq B$$

there exists a parametrization Φ with

$$\|f - \mathcal{R}\Phi\|_{\mathcal{W}^{1,\infty}((0,1)^d)} \leq \varepsilon$$

Applications - $\mathcal{W}^{1,\infty}$ Approximation for ReLU networks

Lemma (chain-rule in $\mathcal{W}^{1,\infty}$- Gühring, Kutyniok, Petersen '19)

$$|g \circ f|_{\mathcal{W}^{1,\infty}} \leq C|g|_{\mathcal{W}^{1,\infty}}|f|_{\mathcal{W}^{1,\infty}} \qquad (f, g \in \mathcal{W}^{1,\infty})$$

drawback: C depends on input dimensions of f and g and can be avoided

Theorem (Sobolev-regular functions - Gühring, Kutyniok, Petersen '19)

For every $\varepsilon \in (0,1)$ and

$$f \in \mathcal{W}^{n,\infty}((0,1)^d) \quad \text{with} \quad \|f\|_{\mathcal{W}^{n,\infty}((0,1)^d)} \leq B$$

there exists a parametrization Φ with

- $\operatorname{depth}(\Phi) \lesssim \log(\varepsilon^{-n/(n-1)})$
- $\|\Phi\|_0 \lesssim \varepsilon^{-d/(n-1)} \cdot \log^2(\varepsilon^{-n/(n-1)})$

and

$$\|f - \mathcal{R}\Phi\|_{\mathcal{W}^{1,\infty}((0,1)^d)} \leq \varepsilon$$

Applications - $\mathcal{W}^{1,\infty}$ Approximation for ReLU networks

Lemma (chain-rule in $\mathcal{W}^{1,\infty}$- Gühring, Kutyniok, Petersen '19)

$$|g \circ f|_{\mathcal{W}^{1,\infty}} \leq C|g|_{\mathcal{W}^{1,\infty}}|f|_{\mathcal{W}^{1,\infty}} \qquad (f, g \in \mathcal{W}^{1,\infty})$$

drawback: C depends on input dimensions of f and g and can be avoided

Theorem (Sobolev-regular functions - Gühring, Kutyniok, Petersen '19)

For every $\varepsilon \in (0,1)$ and

$$f \in \mathcal{W}^{n,\infty}((0,1)^d) \quad \text{with} \quad \|f\|_{\mathcal{W}^{n,\infty}((0,1)^d)} \leq B$$

there exists a parametrization Φ with

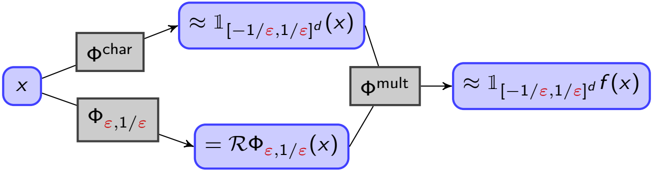
- $\text{depth}(\Phi) \lesssim \log(\varepsilon^{-n/(n-s)})$
- $\|\Phi\|_0 \lesssim \varepsilon^{-d/(n-s)} \cdot \log^2(\varepsilon^{-n/(n-s)})$

and

$$\|f - \mathcal{R}\Phi\|_{\mathcal{W}^{s,\infty}((0,1)^d)} \leq \varepsilon \qquad s \in [0,1]$$

Applications - Global Estimates for ReLU networks

- given local approximations $\|f - \mathcal{R}\Phi_{\varepsilon,B}\|_{\mathcal{W}^{1,\infty}((-B,B)^d)} \leq \varepsilon$ for f with at most polynomially (with degree κ) growing derivative



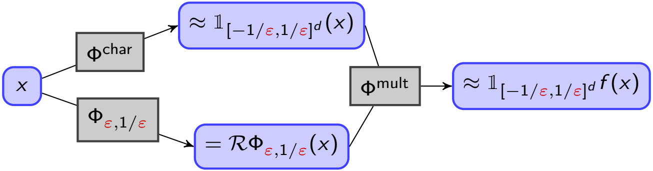
Applications - Global Estimates for ReLU networks

- given local approximations $\|f - \mathcal{R}\Phi_{\varepsilon,B}\|_{\mathcal{W}^{1,\infty}((-B,B)^d)} \leq \varepsilon$ for f with at most polynomially (with degree κ) growing derivative

Theorem (global estimates - B., Elbrächter, Grohs, Jentzen '19)

For every $\varepsilon \in (0,1)$ there exists a parametrization Ψ with

- $|f(x) - \mathcal{R}\Psi(x)| \leq \varepsilon(1 + \|x\|^{\kappa+2}) \quad \forall x \in \mathbb{R}^d$
- $\|Df(x) - \mathcal{D}\Psi(x)\| \leq \varepsilon(1 + \|x\|^{\kappa+2}) \quad$ a.e. $x \in \mathbb{R}^d$



Applications - Global Estimates for ReLU networks

- given local approximations $\|f - \mathcal{R}\Phi_{\varepsilon,B}\|_{\mathcal{W}^{1,\infty}((-B,B)^d)} \leq \varepsilon$ for f with at most polynomially (with degree κ) growing derivative

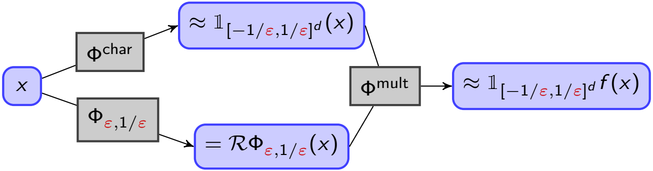
Theorem (global estimates - B., Elbrächter, Grohs, Jentzen '19)

For every $\varepsilon \in (0,1)$ there exists a parametrization Ψ with

- $\text{size}(\Psi) \lesssim \text{size}(\Phi_{\varepsilon,1/\varepsilon}) + \log(d + \varepsilon^{-1})$

and

- $|f(x) - \mathcal{R}\Psi(x)| \leq \varepsilon(1 + \|x\|^{\kappa+2}) \quad \forall x \in \mathbb{R}^d$
- $\|Df(x) - \mathcal{D}\Psi(x)\| \leq \varepsilon(1 + \|x\|^{\kappa+2}) \quad \text{a.e. } x \in \mathbb{R}^d$



Thank you for your Attention!

Julius Berner, Dennis Elbrächter, Philipp Grohs, and Arnulf Jentzen. "Towards a regularity theory for ReLU networks–chain rule and global error estimates". In: *arXiv:1905.04992* (2019).