which is a way to implement the polarization formula (2.37). With convenient neural network approximations to approximate f_1^{obj} and f_3^{obj}, the proof is ended for $n = 3$. The construction is $x \mapsto x^4$ is done by squaring a square. Then $x^5 = \frac{1}{4}(x+x^4)^2 - \frac{1}{4}(x-x^4)^2$ allows to construct x^5. All monomials for $n \geq 6$ can be constructed in the same way. □

2.5.2 Yarotsky Theorem

The next result is a prelude to the Yarotsky Theorem [104].

Definition 2.5.10. *Given a small number* $0 < \varepsilon < 1$*, we will say that a function*

$$\text{mul} : \mathbb{R}^2 \to \mathbb{R}$$

is an approximate multiplier operator if it satisfies the properties

- *for all* $(x, y) \in \mathbb{R}^2$*, one has* $\text{mul}(0, y) = \text{mul}(x, 0) = 0$,
- *if* $|x| \leq 1$ *and* $|y| \leq 1$*, then* $|\text{mul}(x, y) - xy| \leq \varepsilon$,

Proposition 2.5.11. *There exists a ReLU architecture which implements an approximate multiplier operator. The number of computational units of the network is bounded as* $\#(N) \leq C\,(\log 1/\varepsilon + 1)$ *where the constant* $C > 0$ *is independent of* ε.

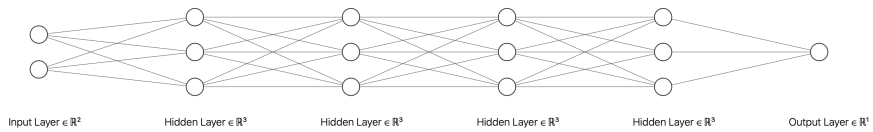

Fig. 2.9: Graph structure of an approximate multiplier $(x, y) \mapsto xy$ with $m = 2$, $p = 4$ and $n = 1$.

Proof. The proof is an elaborated version of the polarization formula (2.37). Consider the function f_p defined in 2.29) and set

$$\text{mul}(x, y) = f_p\left(\frac{|x+y|}{2}\right) - f_p\left(\frac{|x-y|}{2}\right). \tag{2.39}$$

• One has by definition $\text{mul}(0, y) = \text{mul}(x, 0) = 0$.
• Next one notes that

$$\text{mul}(x,y) - xy = \left(f_p\left(\frac{|x+y|}{2}\right) - \left(\frac{|x+y|}{2}\right)^2 \right) - \left(f_p\left(\frac{|x-y|}{2}\right) - \left(\frac{|x-y|}{2}\right)^2 \right).$$

For $|x| \le 1$ and $|y| \le 1$, one has $0 \le \frac{|x \pm y|}{2} \le 1$. Therefore Lemma 2.5.7 can be used two times for the two terms between parentheses. It yields the bound $|\text{mul}(x,y) - xy| \le \frac{2}{3\times 4^p}$. So one takes $\varepsilon = \frac{2}{3\times 4^p}$ which shows that (2.39) is indeed an approximate multiplier operator.
• It remains to bound the number of computational units which is the product of the number of layers of the network times the width of the network. This is evident since

$$p = \frac{1}{4}\log\left(\frac{2}{3\varepsilon}\right) = \frac{1}{4}\log 1/\varepsilon + \frac{1}{4}\log\left(\frac{2}{3}\right).$$

The number of computational units is the number of layers times the number of neurons (3 for f_p), that is $O(p) \le C(\log 1/\varepsilon + 1)$. □

To go further, one defines a convenient partition of unity as follows. The partition of unity will be used to construct with local pieces an approximation of an objective function, as visible on the formula (2.49).

One starts with a continuous and piecewise affine cut-off function

$$\psi(x) = \begin{cases} 1 & |x| \le 1, \\ 2 - |x| & 1 \le |x| \le 2, \\ 0 & 2 \le |x|. \end{cases}$$

For $N \ge 1$ large enough, define

$$\phi_m(x) = \psi(3Nx - 3m) \text{ for } 0 \le m \le N. \tag{2.40}$$

The functions (ϕ_m) realize a partition of unity on the interval $[0,1]$

$$\sum_{m=0}^{N} \phi_m(x) = 1, \quad 0 \le x \le 1. \tag{2.41}$$

One defines the function f_{mn}

$$f_{mn}(x) = \text{mul}\left(\phi_m(x), \underbrace{\text{mul}\left(x - \frac{m}{N}, \text{mul}\left(x - \frac{m}{N}, \dots\right)\dots\right)}_{n \text{ times}} \right). \tag{2.42}$$

The function f_{mn} is a local approximation of $\phi_m(x)(x - m/N)^n$ so it will be possible to combine the partition of unity (2.41) with a local (around m/N) Taylor expansion based on shifted monomials $(x - m/N)^n$. One has first to characterize the properties of these functions.

Lemma 2.5.12. *The function f_{mn} satisfies the properties*

- *For $x \notin supp(\phi_m)$, then $f_{mn}(x) = 0$.*
- *The error bound holds $\left| f_{mn}(x) - \phi_m(x)\left(x - \frac{m}{N}\right)^n \right| \leq n\varepsilon$ for $0 \leq x \leq 1$.*
- *The function f_{mn} can be implemented on a ReLU architecture with a number $\#(N_{mn})$ of computational unit bounded as $\#(N_{mn}) \leq O(n)\left(\log 1/\varepsilon + 1\right)$.*

Proof. • The first property of a multiplier (see Definition 2.5.10) and the definition (2.42) yield $f_{mn}(x) = 0$.

• The second point comes from the telescopic decomposition with n terms

$$\begin{aligned}
& f_{mn}(x) - \phi_m(x)(x - \tfrac{m}{N})^n \\
&= \left[\mathrm{mul}\left(\phi_m(x), \underbrace{\mathrm{mul}\left(x - \frac{m}{N}, \dots\right)}_{n \text{ times}}\right) - \phi_m(x)\underbrace{\mathrm{mul}\left(x - \frac{m}{N}, \dots\right)}_{n \text{ times}}\right] \\
&+ \phi_m(x)\left[\underbrace{\mathrm{mul}\left(x - \frac{m}{N}, \dots\right)}_{n \text{ times}} - (x - \tfrac{m}{N})\underbrace{\mathrm{mul}\left(x - \frac{m}{N}, \dots\right)}_{n-1 \text{ times}}\right] \\
&+ \dots \\
&+ \phi_m(x)(x - \tfrac{m}{N})^{n-2}\left[\underbrace{\mathrm{mul}\left(x - \frac{m}{N}, \dots\right)}_{2 \text{ times}} - (x - \tfrac{m}{N})^2\right].
\end{aligned}$$

By construction $|\phi_m(x)| \leq 1$ and $\left|x - \frac{m}{N}\right| \leq 1$, so it is possible to bound the absolute value of each line by using the second estimate of Definition 2.5.10. All lines are bound by ε. Adding the contributions, one gets $|f_{mn}(x) - \phi_m(x)(x - m/N)^n| \leq n\varepsilon$ which is the second point.

• Finally let us evaluate the number of computational units $\#(N_{mn})$. A network that construct f_{mn} can be obtained by placing n successive multipliers one after the other and adding one more line of neurons (a collation channel) to propagate $x - m/N$ inside the neural network. Then one gets $\#(N_{mn}) \leq Cn\left(\log 1/\varepsilon + 1\right)$. □

Next one constructs by tensorisation a partition of unity in higher dimension for $\mathbf{x} = (x_1, \dots, x_d) \in [0,1]^d$

$$\phi_{\mathbf{m}}(\mathbf{x}) = \Pi_{k=1}^d \phi_{m_k}(x_k), \quad \mathbf{m} = (m_1, \dots, m_d) \in \{0, \dots, N\}^d \tag{2.43}$$

such that

$$\sum_{\mathbf{m}/N \in [0,1]^d} \phi_{\mathbf{m}}(\mathbf{x}) = 1, \quad \mathbf{x} \in [0,1]^d. \tag{2.44}$$

One also defines

$$f_{\mathbf{mn}}(\mathbf{x}) = \mathrm{mul}\left(f_{m_1 n_1}(x_1), \mathrm{mul}\left(f_{m_2 n_2}(x_2), \dots\right)\right). \tag{2.45}$$

Corollary 2.5.13 (of Lemma 2.5.12)**.** *The function* $f_{\mathbf{mn}}$ *satisfies the properties*

- *For* $\mathbf{x} \notin supp(\phi_{\mathbf{m}})$*, then* $f_{\mathbf{mn}}(\mathbf{x}) = 0$.
- *The error bound holds* $\left| f_{\mathbf{mn}}(\mathbf{x}) - \phi_{\mathbf{m}}(\mathbf{x}) \Pi_{k=1}^{d} \left(x_k - \frac{m_k}{N}\right)^{n_k} \right| \leq (d + |\mathbf{n}|)\varepsilon$ *for* $\mathbf{x} \in [0,1]^d$.
- *The function* $f_{\mathbf{mn}}$ *can be implemented on a ReLU architecture with a number* $\#(N_{\mathbf{mn}})$ *of computational unit bounded as* $\#(N_{\mathbf{mn}}) \leq O(d + |\mathbf{n}|)(\log 1/\varepsilon + 1)$.

Proof. • Necessarily, one has that $x_k \notin \mathrm{supp}(\phi_{m_k})$, at least for one index $k \in \{1, \ldots, d\}$. The first property of an approximate multiplier yields the claim.
• All terms $f_{m_k n_k}(x_k)$ are less than one in absolute value for $|X_k| \leq 1$. Therefore a telescopic decomposition like in the proof of Lemma 2.5.12 yields a first estimate

$$\left| f_{\mathbf{mn}}(\mathbf{x}) - \Pi_{k=1}^{d} f_{m_k n_k}(x_k) \right| \leq d\varepsilon.$$

By using Lemma 2.5.12, the term $f_{m_k n_k}(x_k)$ approximates $\phi_{m_k}\left(x_k - \frac{m_k}{N}\right)^{n_k}$ with an error less than $n_k\varepsilon$. Since all these terms are less than one, a telescopic decomposition yields a second estimate

$$\left| \Pi_{k=1}^{d} f_{m_k n_k}(x_k) - \Pi_{k=1}^{d} \phi_{m_k} \left(x_k - \frac{m_k}{N}\right)^{n_k} \right| \leq \left(\sum_{k=1}^{d} n_k \right) \varepsilon.$$

The triangular inequality and the notation $|\mathbf{n}| = \sum_{k=1}^{d} n_k$ shows the result.
• The cost is obtained by using Lemma 2.5.12 to estimate the cost of calculating all individuals terms $f_{m_k n_k}(x_k)$ and the cost of multiplying these d terms (Proposition 2.5.11). □

Theorem 2.5.14 (Yarotsky)**.** *Let* $\delta > 0$ *and consider functions in the space* $W^{r,\infty}[0,1]^d$*. There exists a ReLU architecture with less than* $O(\log 1/\delta + 1)$ *layers and less than* $C_r \delta^{-d/r}(\log 1/\delta + 1)$ *computational units which is capable to reproduce any smooth function such that* $\left\| f^{\mathrm{obj}} \right\|_{W^{r,\infty}[0,1]^d} \leq 1$ *with accuracy* δ.

Remark 2.5.15. *This remarkable result shows that the best theoretical accuracy increases with respect to the regularity. Or more precisely the cost for obtaining a certain accuracy is asymptotically less for functions with increased regularity. However the inequality shows also a devastating possibility, which is that high dimensionality* $d \gg 1$ *dramatically increases the cost of obtaining this accuracy: this phenomenon is related to the* **curse of dimension***. It will be discussed by other means in the sequel.*

Proof. The proof [104] is based on explicit manipulations and is reminiscent of high order Bernstein interpolation.

• The number of bounded derivatives of the objective function f^{obj} is equal to r. Consider the polynomial $p_{\mathbf{m}}$ which is the truncation at order r of the Taylor-expansion at $\frac{\mathbf{m}}{N}$

$$p_{\mathbf{m}}(\mathbf{x}) = \sum_{|\mathbf{n}|\le r-1} \frac{1}{\mathbf{n}!} \frac{\partial^{|\mathbf{n}|} f^{\mathrm{obj}}}{\partial x_1^{n_1} \dots \partial x_d^{n_d}} \left(\frac{\mathbf{m}}{N}\right) \left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{|\mathbf{n}|}, \qquad \frac{\mathbf{m}}{N} \in [0,1]^d, \quad (2.46)$$

with the usual conventions $\mathbf{n}! = \Pi_{k=1}^d \frac{1}{n_k!}$ and $\left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{|\mathbf{n}|} = \Pi_{k=1}^d \left(x_k - \frac{m_k}{N}\right)^{n_1+\dots+n_k}$. Since $|\mathbf{n}| \le r-1$, then the total degree of these polynomials is bounded by $r-1$. One has the bound

$$\left\| f^{\mathrm{obj}} \right\|_{W^{r,\infty}[0,1]} = \max_{|\mathbf{n}|\le r} \left\| \frac{\partial^{|\mathbf{n}|} f^{\mathrm{obj}}}{\partial x_1^{n_1} \dots \partial x_d^{n_d}} \right\|_{L^\infty[0,1]} \le 1,$$

so all coefficients are bounded

$$\left| \frac{1}{\mathbf{n}!} \frac{\partial^{|\mathbf{n}|} f^{\mathrm{obj}}}{\partial x_1^{n_1} \dots \partial x_d^{n_d}} \left(\frac{\mathbf{m}}{N}\right) \right| \le 1, \qquad 0 \le |\mathbf{n}| \le r-1, \quad \frac{\mathbf{m}}{N} \in [0,1]^d. \qquad (2.47)$$

• Then one constructs the function $f(x) = \sum_{\mathbf{m}/N\in[0,1]^d} \phi_{\mathbf{m}}(x) p_{\mathbf{m}}(x)$ which can be expressed as a double sum

$$f(\mathbf{x}) = \sum_{\mathbf{m}/N\in[0,1]^d} \sum_{|\mathbf{n}|\le r-1} \frac{1}{\mathbf{n}!} \frac{\partial^{|\mathbf{n}|} f^{\mathrm{obj}}}{\partial x_1^{n_1} \dots \partial x_d^{n_d}} \left(\frac{\mathbf{m}}{N}\right) \phi_{\mathbf{m}}(\mathbf{x}) \left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{|\mathbf{n}|}.$$

• Using the partition of unity formula (2.43), the error between f^{obj} and f can be bounded as

$$\begin{aligned} \left| f^{\mathrm{obj}}(\mathbf{x}) - f(\mathbf{x}) \right| &= \left| \sum_{\mathbf{m}/N\in[0,1]^d} \phi_{\mathbf{m}}(\mathbf{x}) \left(f^{\mathrm{obj}}(\mathbf{x}) - p_{\mathbf{m}}(\mathbf{x}) \right) \right| \\ &\le \sum_{\mathbf{m}\in\bigcap\{|x_k - m_k/N|<1/N\}} \left| f^{\mathrm{obj}}(\mathbf{x}) - p_{\mathbf{m}}(\mathbf{x}) \right| \\ &\le 2^d \max_{\mathbf{m}\in\bigcap\{|x_k - m_k/N|<1/N\}} \left| f^{\mathrm{obj}}(\mathbf{x}) - p_{\mathbf{m}}(\mathbf{x}) \right|. \end{aligned}$$

A classical estimate of the remainder of the multivariate Tayler expansion between f^{obj} and $p_{\mathbf{m}}$ yields the bound

$$\left| f^{\mathrm{obj}}(\mathbf{x}) - f(\mathbf{x}) \right| \le \frac{2^d d^r}{r! N^r}$$

Let us take N equal to the integer part of $\left(\frac{r!\delta}{2^{d+1}d^r}\right)^{-1/r}$ plus one (the ceiling function) so that $N^r \ge \frac{r!\delta}{2^{d+1}d^r}$. Under this condition, one has $\left| f^{\mathrm{obj}}(x) - f(x) \right| \le \frac{\delta}{2}$. It is a standard exercise to show that there exists a constant $C > 0$ such that

$$N \le C \frac{d}{\delta^{1/r}}. \qquad (2.48)$$

• In view of the technical Proposition 2.5.11, it is natural to define

$$g(\mathbf{x}) = \sum_{\mathbf{m}/N \in [0,1]^d} \sum_{|\mathbf{n}| \le r-1} \frac{1}{\mathbf{n}!} \frac{\partial^{|\mathbf{n}|} f^{\text{obj}}}{\partial x_1^{n_1} \dots \partial x_d^{n_d}} \left(\frac{\mathbf{m}}{N}\right) f_{\mathbf{mn}}(\mathbf{x}). \tag{2.49}$$

The function g is implementable in a neural network and is a priori close to f. The difference is

$$\begin{aligned}
& f(\mathbf{x}) - g(\mathbf{x}) \\
= & \sum_{\mathbf{m}/N \in [0,1]^d} \sum_{|\mathbf{n}| \le r-1} \frac{1}{\mathbf{n}!} \frac{\partial^{|\mathbf{n}|} f^{\text{obj}}}{\partial x_1^{n_1} \dots} \left(\frac{\mathbf{m}}{N}\right) \left[\phi_{\mathbf{m}}(\mathbf{x}) \left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{|\mathbf{n}|} - f_{\mathbf{mn}}(\mathbf{x}) \right] \\
= & \sum_{\mathbf{m} | \mathbf{x} \in \text{supp}(\phi_{\mathbf{m}})} \sum_{|\mathbf{n}| \le r-1} \frac{1}{\mathbf{n}!} \frac{\partial^{|\mathbf{n}|} f^{\text{obj}}}{\partial x_1^{n_1} \dots} \left(\frac{\mathbf{m}}{N}\right) \left[\phi_{\mathbf{m}}(\mathbf{x}) \left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{|\mathbf{n}|} - f_{\mathbf{mn}}(\mathbf{x}) \right] \\
= & 2^d \max_{\mathbf{m} | \mathbf{x} \in \text{supp}(\phi_{\mathbf{m}})} \sum_{|\mathbf{n}| \le r-1} \frac{1}{\mathbf{n}!} \left| \phi_{\mathbf{m}}(\mathbf{x}) \left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{|\mathbf{n}|} - f_{\mathbf{mn}}(\mathbf{x}) \right| \\
\le & 2^d e^d (d + r) \varepsilon,
\end{aligned}$$

where the last inequality comes Corollary 2.5.13 and $\sum_{|\mathbf{n}| \le r-1} \frac{1}{\mathbf{n}!} \le \sum_{\mathbf{n} \in \mathbb{N}^d} \frac{1}{\mathbf{n}!} = \left(\sum_{n \in \mathbb{N}} \frac{1}{n!}\right) = e^d$.

• Take the multiplier accuracy ε such that $2^d e^d (d + r)\varepsilon = \frac{\delta}{2}$. Then the triangular inequality yields

$$\left| f^{\text{obj}}(\mathbf{x}) - g(\mathbf{x}) \right| \le \left| f^{\text{obj}}(\mathbf{x}) - f(\mathbf{x}) \right| + |f(\mathbf{x}) - g(\mathbf{x})| \le \frac{\delta}{2} + \frac{\delta}{2} = \delta$$

which the accuracy claimed in the Theorem.

• Finally let us evaluate the complexity of the network needed to evaluate the function g. It is possible to consider a network which evaluates **in parallel** all the functions $f_{\mathbf{mn}}$, and then which makes a linear combination with a collation channel. So the number of layers scales as for the approximate multiplier.

The number of computational units is

$$\#(N_{\text{total}}) = O\left(\sum_{|\mathbf{n}| \le r-1} \sum_{|\mathbf{m}|/N \in [0,1]^d} \# \left(N_{\mathbf{mn}}\right) \right)$$

$$\le r(N+1)^d \left[C(d+r) \left(\log 1/\varepsilon + 1\right) \right] \le C(r,d) \delta^{-d/r} \left(\log 1/\delta + 1\right)$$

for some constant $C(r, d)$. □

2.5.3 Takagi function

The Takagi function [94] illustrates a fascinating property of neural networks, which is their ability to approximate very irregular functions. This feature is totally