DEPARTAMENTO DE
MATEMÁTICA

O poder da Aprendizagem Profunda Parte 2: Os polinômios atacam novamente

Felipe Kaminsky Riffel

Universidade Federal de Santa Catarina

23 de abril de 2025

Sumário

Artigo: Power of Deep Learning on Expressing Natural Functions

Artigo:
TEGMARK, Max; ROLNICK, David. The Power of Deeper Networks for Expressing Natural Functions. arXiv. 2018. (ROLNICK; TEGMARK, 2018)

Artigo submetido para apresentação na modalidade pôster para o International Conference on Learning Representations (ICLR) 2018.

No artigo (LIN; TEGMARK; ROLNICK, 2017), foram demonstrados 2 Teoremas principais:

Teorema

dado $\varepsilon > 0$, *para qualquer* σ *não linear (aplicada ponto a ponto), existem* $A_1 : \mathbb{R}^2 \to \mathbb{R}^4, A_2 : \mathbb{R}^4 \to \mathbb{R}$ *tais que a rede* $f(x) = A_2 \sigma A_1 \mathbf{x}$ *é tal que, dado* $x = (u\ v)^T$ *qualquer*

$$|f(x) - uv| < \varepsilon$$

para u, v *em um compacto qualquer.*

Teorema

Dados $x_1, \ldots, x_n \in \mathbb{R}$ e $\sigma \in C^\infty$, o monômio $\Pi_{i=1}^n x_i$ pode ser aproximado por uma rede neural de 1 camada com 2^n neurônios, seguindo a fórmula

$$\Pi_{i=1}^n x_i \approx \frac{1}{2^n} \sum_{\{s\}} s_1 \cdots s_n \sigma(s_1 x_1 + \cdots + s_n x_n),$$

onde $s_i \in \{-1, 1\}$, para cada $i = 1, \ldots, k$, e a soma é tomada sobre todas as 2^n configurações possíveis de $s_1 \cdots s_n$.
Além disso, essa é a menor rede de 1 camada capaz de fazer tal aproximação.

Obs: A aproximação acima é no sentido que ambas as funções têm mesma expansão de Taylor em torno de 0 até termos de ordem n.

A segugir, nós

- Aprofundamos a discussão sobre número de camadas;

A segugir, nós

- Aprofundamos a discussão sobre número de camadas;
- Construímos aproximações mais enxutas para uma classe maior de polinômios;

Neste artigo:

A segugir, nós

- Aprofundamos a discussão sobre número de camadas;
- Construímos aproximações mais enxutas para uma classe maior de polinômios;
- Obtemos estimativas para números de neurônios em redes de diferentes profundidades.

Definição

Consideramos o modelo de redes neurais feedforward ou multilayer perceptron:

$$N(x) = A_k \circ \sigma \circ A_{k-1} \circ \cdots \sigma A_1 \sigma A_0 x \tag{1}$$

onde:

- $A_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_{i+1}}$ são matrizes/operadores afins;
- $\sigma : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ são funções não lineares aplicados ponto a ponto;

Definição

Consideramos o modelo de redes neurais feedforward ou multilayer perceptron:

$$N(x) = A_k \circ \sigma \circ A_{k-1} \circ \cdots \sigma A_1 \sigma A_0 x \tag{1}$$

onde:

- $A_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_{i+1}}$ são matrizes/operadores afins;
- $\sigma : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ são funções não lineares aplicados ponto a ponto;

Denominamos:

- k a profundidade da rede;

Definição

Consideramos o modelo de redes neurais feedforward ou multilayer perceptron:

$$N(x) = A_k \circ \sigma \circ A_{k-1} \circ \cdots \sigma A_1 \sigma A_0 x \tag{1}$$

onde:

- $A_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_{i+1}}$ são matrizes/operadores afins;
- $\sigma : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ são funções não lineares aplicados ponto a ponto;

Denominamos:

- k a profundidade da rede;
- camadas escondidas cada vetor $\sigma A_i \sigma A_{i-1} \cdots A_0 x$;

Definição

Consideramos o modelo de redes neurais feedforward ou multilayer perceptron:

$$N(x) = A_k \circ \sigma \circ A_{k-1} \circ \cdots \sigma A_1 \sigma A_0 x \tag{1}$$

onde:

- $A_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_{i+1}}$ são matrizes/operadores afins;
- $\sigma : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ são funções não lineares aplicados ponto a ponto;

Denominamos:

- k a profundidade da rede;
- camadas escondidas cada vetor $\sigma A_i \sigma A_{i-1} \cdots A_0 x$;
- neurônio as entradas de cada $\sigma A_i \sigma A_{i-1} \cdots A_0 x$.

Definição

Dado $\varepsilon > 0$ e um compacto K, dizemos que uma rede $N(\mathbf{x})$ $\varepsilon-$*aproxima* uma função $f : \mathbb{R}^n \to \mathbb{R}^m$ se

$$\sup_{x \in K} |N(\mathbf{x}) - f(\mathbf{x})| < \varepsilon$$

Definição

Dado $\varepsilon > 0$ e um compacto K, dizemos que uma rede $N(\mathbf{x})$ $\varepsilon-$*aproxima* uma função $f : \mathbb{R}^n \to \mathbb{R}^m$ se

$$\sup_{x \in K} |N(\mathbf{x}) - f(\mathbf{x})| < \varepsilon$$

Definição

Dizemos que uma rede $N(x)$ *Taylor-aproxima* um polinômio $p(x)$, com $p : \mathbb{R}^n \to \mathbb{R}$ de grau d, se $p(x)$ é o polinômio de Taylor de grau d de $N(x)$ em torno da origem.

Definição

Dado $\varepsilon > 0$ e um compacto K, dizemos que uma rede $N(\mathbf{x})$ $\varepsilon-$*aproxima* uma função $f : \mathbb{R}^n \to \mathbb{R}^m$ se

$$\sup_{x \in K} |N(\mathbf{x}) - f(\mathbf{x})| < \varepsilon$$

Definição

Dizemos que uma rede $N(x)$ *Taylor-aproxima* um polinômio $p(x)$, com $p : \mathbb{R}^n \to \mathbb{R}$ de grau d, se $p(x)$ é o polinômio de Taylor de grau d de $N(x)$ em torno da origem. I.e.,

$$\begin{aligned} N(x) &= \sum_{|\alpha| \leq d} \frac{D^\alpha N(0) x^\alpha}{\alpha!} + \mathcal{O}(x^d) \\ &= p(x) + \mathcal{O}(x^d) \end{aligned}$$

MTM

Proposição

Seja $p : \mathbb{R}^n \to \mathbb{R}$ polinômio homogêneo e suponha que a rede N Taylor-aproxima p em um compacto K. Então, para cada $\varepsilon > 0$, existe uma rede N_ε que $\varepsilon-$aproxima p, tais que N e N_ε tem o mesmo número de neurônios em cada camada.

Prop 3.3 - Demonstração

Seja $p(x)$ polinômio de grau d.

Prop 3.3 - Demonstração

Seja $p(x)$ polinômio de grau d.Como N Taylor-aproxima p, se σ for inf. diferenciável:

$$N(x) = p(x) + E(x) \tag{2}$$

com $\sum_{i=d+1}^{\infty} E_i(x)$ e cada E_i homogêneo de grau i.

Prop 3.3 - Demonstração

Seja $p(x)$ polinômio de grau d.Como N Taylor-aproxima p, se σ for inf. diferenciável:

$$N(x) = p(x) + E(x) \tag{2}$$

com $\sum_{i=d+1}^{\infty} E_i(x)$ e cada E_i homogêneo de grau i. Assim,

$$E_i(\delta x) = \delta^i E_i(x), \forall i \in \mathbb{N}.$$

Prop 3.3 - Demonstração

Seja $p(x)$ polinômio de grau d.Como N Taylor-aproxima p, se σ for inf. diferenciável:

$$N(x) = p(x) + E(x) \tag{2}$$

com $\sum_{i=d+1}^{\infty} E_i(x)$ e cada E_i homogêneo de grau i. Assim,

$$E_i(\delta x) = \delta^i E_i(x), \forall i \in \mathbb{N}.$$

Como $\sum_{i=d+1}^{\infty} E_i(x) < \infty, \forall x \in \mathbb{R}^n$, em particular, para $\delta < 1$

$$\sum_{i=d+1}^{\infty} E_i(\delta x) = \sum_{i=d+1}^{\infty} \delta^i E_i(x) < \infty$$

Seja $p(x)$ polinômio de grau d.Como N Taylor-aproxima p, se σ for inf. diferenciável:

$$N(x) = p(x) + E(x) \tag{2}$$

com $\sum_{i=d+1}^{\infty} E_i(x)$ e cada E_i homogêneo de grau i. Assim,

$$E_i(\delta x) = \delta^i E_i(x), \forall i \in \mathbb{N}.$$

Como $\sum_{i=d+1}^{\infty} E_i(x) < \infty, \forall x \in \mathbb{R}^n$, em particular, para $\delta < 1$

$$\sum_{i=d+1}^{\infty} E_i(\delta x) = \sum_{i=d+1}^{\infty} \delta^i E_i(x) < \infty$$

de modo que

$$\frac{1}{\delta^d} E(\delta x) = \sum_{i=d+1}^{\infty} \delta^{i-d} E_i(x) < \infty$$

Prop 3.3 - Demonstração

Como $i > d$, para δ suficientemente pequeno, cada δ^{i-d} se torna tão pequeno quanto queira, assim como $\frac{1}{\delta^d} E_i(x) = \delta^{i-d} E_i(\delta x)$.

Logo, $\frac{E(\delta x)}{\delta^d}$é arbitrariamente pequeno.

Como $i > d$, para δ suficientemente pequeno, cada δ^{i-d} se torna tão pequeno quanto queira, assim como $\frac{1}{\delta^d} E_i(x) = \delta^{i-d} E_i(\delta x)$.
Logo, $\frac{E(\delta x)}{\delta^d}$é arbitrariamente pequeno.
Seja $\varepsilon > 0$ e tome δ t.q.

$$|\frac{E(\delta x)}{\delta^d}| < \varepsilon \tag{3}$$

Defina $A'_0 = \delta A_0$, $A'_k = \frac{1}{\delta^d} A_k \cdots A_0(\delta x)$,

Como $i > d$, para δ suficientemente pequeno, cada δ^{i-d} se torna tão pequeno quanto queira, assim como $\frac{1}{\delta^d} E_i(x) = \delta^{i-d} E_i(\delta x)$.
Logo, $\frac{E(\delta x)}{\delta^d}$é arbitrariamente pequeno.
Seja $\varepsilon > 0$ e tome δ t.q.

$$|\frac{E(\delta x)}{\delta^d}| < \varepsilon \tag{3}$$

Defina $A'_0 = \delta A_0$, $A'_k = \frac{1}{\delta^d} A_k \cdots A_0(\delta x)$,

$$N_\varepsilon(x) = A'_k \sigma A_{k-1} \sigma \cdots \sigma A_1 \sigma A'_0 x \tag{4}$$

$$= \frac{1}{\delta^d} A_k \sigma A_{k-1} \sigma \cdots \sigma A_1 \sigma (\delta A_0) x \tag{5}$$

$$= \frac{N(\delta x)}{\delta^d} \tag{6}$$

MTM

Assim,

$$
\begin{aligned}
|N_\varepsilon(x) - p(x)| &= |\frac{1}{\delta^d} N(\delta x) - p(x)| \\
&= |\frac{1}{\delta^d}(p(\delta x) + E(\delta x)) - p(x)| \\
&= |\frac{1}{\delta^d} p(\delta x) - p(x) + \frac{1}{\delta^d} E(\delta x)|
\end{aligned}
$$

Assim,

$$
\begin{aligned}
|N_\varepsilon(x) - p(x)| &= |\frac{1}{\delta^d} N(\delta x) - p(x)| \\
&= |\frac{1}{\delta^d}(p(\delta x) + E(\delta x)) - p(x)| \\
&= |\frac{1}{\delta^d} p(\delta x) - p(x) + \frac{1}{\delta^d} E(\delta x)|
\end{aligned}
$$

$$
= |\frac{1}{\delta^d} E(\delta x)| < \varepsilon.
$$

Ou seja, N_ε é $\varepsilon-$aproximação de p, como queríamos. $\square$

Teorema

Suponha que $p(x)$ é um polinômio de grau d multivariado e que $\sigma_i := \sigma^{(i)}(0) \neq 0$, para cada $i \leq d$. Seja $m_k^{\varepsilon}(p)$ o número mínimo de neurônios em uma rede de profundidade k que $\varepsilon-$aproxima p em um compacto K.

Teorema

Suponha que $p(x)$ é um polinômio de grau d multivariado e que $\sigma_i := \sigma^{(i)}(0) \neq 0$, para cada $i \leq d$. Seja $m_k^\varepsilon(p)$ o número mínimo de neurônios em uma rede de profundidade k que $\varepsilon-$aproxima p em um compacto K. Então, o limite

$$\lim_{\varepsilon>0} m_k^\varepsilon(p)$$

existe e é finito.

Mostremos que $\lim_{\varepsilon \to 0} m_1^{\varepsilon}$ existe e é finito.

Mostremos que $\lim_{\varepsilon \to 0} m_1^{\varepsilon}$ existe e é finito.

Sejam $p_1, \ldots, p_s$ monômios t.q $p(x) = \sum_i^s p_i(x)$. Cada p_i pode ser Taylor-aproximado por uma rede N_i de 1 camada,conforme Teorema 2 de Lin, Tegmark e Rolnick (2017).

Mostremos que $\lim_{\varepsilon \to 0} m_1^{\varepsilon}$ existe e é finito.

Sejam $p_1, \ldots, p_s$ monômios t.q $p(x) = \sum_i^s p_i(x)$. Cada p_i pode ser Taylor-aproximado por uma rede N_i de 1 camada,conforme Teorema 2 de Lin, Tegmark e Rolnick (2017).

Suponha que N^i tem m_i neurônios, tome $\varepsilon > 0$ e seja $\delta = \frac{\varepsilon}{s}$. Pela Prop. 3.3, sendo p_i homogêneo, existe uma rede N_δ^i que $\delta-$aproxima N_i.

Mostremos que $\lim_{\varepsilon \to 0} m_1^{\varepsilon}$ existe e é finito.

Sejam $p_1, \ldots, p_s$ monômios t.q $p(x) = \sum_i^s p_i(x)$. Cada p_i pode ser Taylor-aproximado por uma rede N_i de 1 camada,conforme Teorema 2 de Lin, Tegmark e Rolnick (2017).

Suponha que N^i tem m_i neurônios, tome $\varepsilon > 0$ e seja $\delta = \frac{\varepsilon}{s}$. Pela Prop. 3.3, sendo p_i homogêneo, existe uma rede N_δ^i que $\delta-$aproxima N_i.

Definimos:

$$N_\varepsilon(x) = \sum_i N_\delta^i(x),$$

a qual tem $\sum_i m_i$ neurônios.

Definimos:

$$N_\varepsilon(x) = \sum_i N_\delta^i(x),$$

a qual tem $\sum_i m_i$ neurônios. Então,

$$\begin{aligned} |N_\varepsilon(x) - p(x)| &= |\sum_i^s N_\delta^i(x) - \sum_i^s p_i(x)| \\ &\leq \sum_i^s |N_\delta^i(x) - p_i(x)| \\ &\leq \sum_i^s \frac{\varepsilon}{s} = \varepsilon. \end{aligned}$$

Definimos:

$$N_\varepsilon(x) = \sum_i N_\delta^i(x),$$

a qual tem $\sum_i m_i$ neurônios. Então,

$$\begin{aligned} |N_\varepsilon(x) - p(x)| &= |\sum_i^s N_\delta^i(x) - \sum_i^s p_i(x)| \\ &\leq \sum_i^s |N_\delta^i(x) - p_i(x)| \\ &\leq \sum_i^s \frac{\varepsilon}{s} = \varepsilon. \end{aligned}$$

Logo, N_ε é uma $\varepsilon-$aproximação de p com $\sum_m^i$ neurônios.

Na construção de N_δ^i, o número de neurônios é o mesmo para cada δ escolhido. Isto é, N_δ tem sempre $\sum_i m_i$ neurônios, independente de δ.

Na construção de N_δ^i, o número de neurônios é o mesmo para cada δ escolhido. Isto é, N_δ tem sempre $\sum_i m_i$ neurônios, independente de δ. Portanto,

$$m_1^\varepsilon(p) \leq \sum_i m_i, \forall \varepsilon > 0.$$

Na construção de N_δ^i, o número de neurônios é o mesmo para cada δ escolhido. Isto é, N_δ tem sempre $\sum_i m_i$ neurônios, independente de δ. Portanto,

$$m_1^\varepsilon(p) \leq \sum_i m_i, \forall \varepsilon > 0.$$

Ou seja, $\lim_{\varepsilon \to 0} m_1^\varepsilon(p) \leq \infty$, como queríamos.

Na construção de N_δ^i, o número de neurônios é o mesmo para cada δ escolhido. Isto é, N_δ tem sempre $\sum_i m_i$ neurônios, independente de δ. Portanto,

$$m_1^\varepsilon(p) \leq \sum_i m_i, \forall \varepsilon > 0.$$

Ou seja, $\lim_{\varepsilon \to 0} m_1^\varepsilon(p) \leq \infty$, como queríamos.
Podemos sempre construir uma rede N_k de profundidade $k > 1$ que aproxima uma rede $N_1(x)$ de uma camada:

$$N_k(x) = A_k \sigma A_{k_1} \sigma \cdots A_2 \sigma N_1(x)$$

onde $A_i \sigma$ é uma camada de passagem, $\forall i > 1$.

Na construção de N_δ^i, o número de neurônios é o mesmo para cada δ escolhido. Isto é, N_δ tem sempre $\sum_i m_i$ neurônios, independente de δ. Portanto,

$$m_1^\varepsilon(p) \leq \sum_i m_i, \forall \varepsilon > 0.$$

Ou seja, $\lim_{\varepsilon \to 0} m_1^\varepsilon(p) \leq \infty$, como queríamos.
Podemos sempre construir uma rede N_k de profundidade $k > 1$ que aproxima uma rede $N_1(x)$ de uma camada:

$$N_k(x) = A_k \sigma A_{k_1} \sigma \cdots A_2 \sigma N_1(x)$$

onde $A_i\sigma$ é uma camada de passagem, $\forall i > 1$.
Logo, $m_k(p) \lesssim m_1(p) < \infty$. ■.

Definição

Seja σ função não linear. Dado p polinômio multivariado, seja $m_k^{uniform}(p)$ o número mínimo de neurônios em uma rede de profundidade k que $\varepsilon-$aproxima p para todo $\varepsilon > 0$.

Definição

Seja σ função não linear. Dado p polinômio multivariado, seja $m_k^{uniform}(p)$ o número mínimo de neurônios em uma rede de profundidade k que $\varepsilon-$aproxima p para todo $\varepsilon > 0$. Denotamos:

$$m^{uniform}(p) = \min_{k \in \mathbb{N}} m_k^{uniform}(p)$$

Definição

Seja σ função não linear. Dado p polinômio multivariado, seja $m_k^{uniform}(p)$ o número mínimo de neurônios em uma rede de profundidade k que $\varepsilon-$aproxima p para todo $\varepsilon > 0$. Denotamos:

$$m^{uniform}(p) = \min_{k \in \mathbb{N}} m_k^{uniform}(p)$$

I.e., é o mínimo de número de neurônios considerando todas as profundidades de redes.

Definição

Seja σ função não linear. Dado p polinômio multivariado, seja $m_k^{uniform}(p)$ o número mínimo de neurônios em uma rede de profundidade k que $\varepsilon-$aproxima p para todo $\varepsilon > 0$. Denotamos:

$$m^{uniform}(p) = \min_{k \in \mathbb{N}} m_k^{uniform}(p)$$

I.e., é o mínimo de número de neurônios considerando todas as profundidades de redes.

Definimos m_k^{Taylor} e m^{Taylor} analogamente.

Teorema

Seja $p(x)$ *o monômio* $p(x) = x_1^{r_1} x_2^{r_2} \cdots x_n^{r_n}$, *com* $d = \sum_{i=1}^{n} r_i$ *e* σ *função não linear qualquer. Suponha que* $\sigma_i \neq 0$, *para todo* $i \leq 2d$. *Então,*

(i) $m_1^{uniform}(p) = \prod_{i=1}^{n}(r_i + 1)$,

(ii) $m^{uniform}(p) \leq \sum_{i=1}^{n}(7d\lceil \log_2(r_i) \rceil + 4)$,

onde $\lceil x \rceil$ *denota o menor inteiro maior ou igual a* x.

Teorema

Seja $p(x)$ o monômio $p(x) = x_1^{r_1} x_2^{r_2} \cdots x_n^{r_n}$, com $d = \sum_{i=1}^{n} r_i$. e σ função não linear qualquer. Suponha que $\sigma_i \neq 0$, para todo $i \leq d$. Então,

(i) $m_1^{Taylor}(p) = \prod_{i=1}^{n}(r_i + 1)$,

(ii) $m^{Taylor}(p) \leq \sum_{i=1}^{n}(7d\lceil \log_2(r_i) \rceil + 4)$.

Ideia: mesmo procedimento do Teorema 2 de Lin, Tegmark e Rolnick (2017) de fazer somas alternadas de $\sigma(\pm y_1 \pm y_2 \pm \cdots \pm y_m)$, sendo y_j as variávies x_i levando em conta as repetições.

Ideia: mesmo procedimento do Teorema 2 de Lin, Tegmark e Rolnick (2017) de fazer somas alternadas de $\sigma(\pm y_1 \pm y_2 \pm \cdots \pm y_m)$, sendo y_j as variávies x_i levando em conta as repetições.
Aqui, notamos que se temos r_i repetições de cada termo x_i, há apenas $r_i + 1$ formas de dispor os sinais nessa soma.

Ideia: mesmo procedimento do Teorema 2 de Lin, Tegmark e Rolnick (2017) de fazer somas alternadas de $\sigma(\pm y_1 \pm y_2 \pm \cdots \pm y_m)$, sendo y_j as variávies x_i levando em conta as repetições.
Aqui, notamos que se temos r_i repetições de cada termo x_i, há apenas $r_i + 1$ formas de dispor os sinais nessa soma. No total, são $\prod_{i=1}^{n}(r_i + 1)$ funções $\sigma(\pm x_1 \pm x_2 \pm \cdots \pm x_n)$.

Ideia: mesmo procedimento do Teorema 2 de Lin, Tegmark e Rolnick (2017) de fazer somas alternadas de $\sigma(\pm y_1 \pm y_2 \pm \cdots \pm y_m)$, sendo y_j as variávies x_i levando em conta as repetições.
Aqui, notamos que se temos r_i repetições de cada termo x_i, há apenas $r_i + 1$ formas de dispor os sinais nessa soma. No total, são $\prod_{i=1}^{n}(r_i + 1)$ funções $\sigma(\pm x_1 \pm x_2 \pm \cdots \pm x_n)$.
No que segue, o procedimento é o mesmo, considerando as particularidades.

Teorema

Seja $p(x)$ um polinômio multivariado de grau d e esparsidade c, com monômios $q_1(x), q_2(x), \ldots, q_c(x)$. Seja σ não linear e suponha que $\sigma_i \neq 0$, para todo $i \leq 2d$. Então, temos:

(i) $m_1^{uniform}(p) \geq \frac{1}{c} \max_j m_1^{uniform}(q_j)$,

(ii) $m^{uniform}(p) \leq \sum_j m^{uniform}(q_j)$.

Teorema

Seja $p(x)$ um polinômio multivariado de grau d e esparsidade c, com monômios $q_1(x), q_2(x), \ldots, q_c(x)$. Seja σ não linear e suponha que $\sigma_i \neq 0$, para todo $i \leq 2d$. Então, temos:

(i) $m_1^{uniform}(p) \geq \frac{1}{c} \max_j m_1^{uniform}(q_j)$,

(ii) $m^{uniform}(p) \leq \sum_j m^{uniform}(q_j)$.

Obs: Dizemos que p tem esparsidade c se pode ser representado como a soma de c monômios.

Teorema

Seja $p(x)$ um polinômio multivariado de grau d e esparsidade c, com monômios $q_1(x), q_2(x), \ldots, q_c(x)$. Seja σ não linear e suponha que $\sigma_i \neq 0$, para todo $i \leq 2d$. Então, temos:

(i) $m_1^{uniform}(p) \geq \frac{1}{c} \max_j m_1^{uniform}(q_j)$,

(ii) $m^{uniform}(p) \leq \sum_j m^{uniform}(q_j)$.

Obs: Dizemos que p tem esparsidade c se pode ser representado como a soma de c monômios.

Obs2: O resultado também é válido trocando $m^{uniform}$ para m^{Taylor}.

Teorema

Seja $p(x)$ o monômio $x_1^{r_1} x_2^{r_2} \cdots x_n^{r_n}$, com $d = \sum_{i=1}^{n} r_i$. Suponha que $\sigma_d \neq 0$ (os outros coeficientes de Taylor podem ser nulos). Então, $m_1^{uniform}(p)$ e $m_1^{Taylor}(p)$ são no mínimo $\frac{1}{d} \prod_{i=1}^{n} (r_i + 1)$. (Um limite inferior ainda melhor é o maior coeficiente no polinômio $\prod_i (1 + y + \cdots + y^{r_i})$.)

Teorema

Suponha que $\sigma_i \neq 0$ para cada $i \leq d$. Então, para cada polinômio $p(x)$ em uma variável de grau d,:

$$m_1^{Taylor}(p) \leq d + 1$$

Sejam $a_0, a_1, \ldots, a_n \in \mathbb{R}$ distintos.

Sejam $a_0, a_1, \ldots, a_n \in \mathbb{R}$ distintos. Considere $A \in \mathbb{R}^{(d+1)\times(d+1)}$ dada por

$$A_{ij} = a_i^j$$

Sejam $a_0, a_1, \ldots, a_n \in \mathbb{R}$ distintos. Considere $A \in \mathbb{R}^{(d+1)\times(d+1)}$ dada por

$$A_{ij} = a_i^j$$

Sabemos que $\det A \neq 0$ (matriz de Vandermond). Logo, A é inversível.

Sejam $a_0, a_1, \ldots, a_n \in \mathbb{R}$ distintos. Considere $A \in \mathbb{R}^{(d+1)\times(d+1)}$ dada por

$$A_{ij} = a_i^j$$

Sabemos que $\det A \neq 0$ (matriz de Vandermond). Logo, A é inversível. Portanto, multiplicando as linhas de A por um $b_i \neq 0$ produz outra inversível.

Sejam $a_0, a_1, \ldots, a_n \in \mathbb{R}$ distintos. Considere $A \in \mathbb{R}^{(d+1)\times(d+1)}$ dada por

$$A_{ij} = a_i^j$$

Sabemos que $\det A \neq 0$ (matriz de Vandermond). Logo, A é inversível. Portanto, multiplicando as linhas de A por um $b_i \neq 0$ produz outra inversível.

Sejam A_i as colunas de A e $\sigma(x) = \sum_j \sigma_j x_j$ expansão de Taylor de σ em 0. Defina

$$A' = \begin{bmatrix} | & | & & | \\ \sigma_0 A_0 & \sigma_1 A_1 & \cdots & \sigma_d A_d \\ | & | & & | \end{bmatrix} = \begin{bmatrix} 1 & \sigma_1 a_1 & \cdots & \sigma_d a_1^n \\ \vdots & \ddots & & \vdots \\ 1 & \sigma_1 a_d & \cdots & \sigma_d a_d^n \end{bmatrix}$$

Sejam $a_0, a_1, \ldots, a_n \in \mathbb{R}$ distintos. Considere $A \in \mathbb{R}^{(d+1)\times(d+1)}$ dada por

$$A_{ij} = a_i^j$$

Sabemos que $\det A \neq 0$ (matriz de Vandermond). Logo, A é inversível. Portanto, multiplicando as linhas de A por um $b_i \neq 0$ produz outra inversível.

Sejam A_i as colunas de A e $\sigma(x) = \sum_j \sigma_j x_j$ expansão de Taylor de σ em 0. Defina

$$A' = \begin{bmatrix} | & | & & | \\ \sigma_0 A_0 & \sigma_1 A_1 & \cdots & \sigma_d A_d \\ | & | & & | \end{bmatrix} = \begin{bmatrix} 1 & \sigma_1 a_1 & \cdots & \sigma_d a_1^n \\ \vdots & \ddots & & \vdots \\ 1 & \sigma_1 a_d & \cdots & \sigma_d a_d^n \end{bmatrix}$$

Por hipótese, $\sigma_i \neq 0$, portanto A' é inversível.

MTM

$$A' = \begin{bmatrix} | & | & & | \\ \sigma_0 A_0 & \sigma_1 A_1 & \cdots & \sigma_d A_d \\ | & | & & | \end{bmatrix} = \begin{bmatrix} 1 & \sigma_1 a_1 & \cdots & \sigma_d a_1^n \\ \vdots & \ddots & & \vdots \\ 1 & \sigma_1 a_d & \cdots & \sigma_d a_d^n \end{bmatrix} \tag{7}$$

Note que cada linha i de A corresponde os coeficientes da expansão de Taylor de $\sigma(a_i x)$:

$$\sigma(a_i x) = \sum_j^d \sigma_j (a_i x)^j = \sum_j^d \sigma_j a_i^j x^j .$$

$$A' = \begin{bmatrix} | & | & & | \\ \sigma_0 A_0 & \sigma_1 A_1 & \cdots & \sigma_d A_d \\ | & | & & | \end{bmatrix} = \begin{bmatrix} 1 & \sigma_1 a_1 & \cdots & \sigma_d a_1^n \\ \vdots & \ddots & & \vdots \\ 1 & \sigma_1 a_d & \cdots & \sigma_d a_d^n \end{bmatrix} \quad (7)$$

Note que cada linha i de A corresponde os coeficientes da expansão de Taylor de $\sigma(a_i x)$:

$$\sigma(a_i x) = \sum_j^d \sigma_j (a_i x)^j = \sum_j^d \sigma_j a_i^j x^j .$$

Sendo as linhas L.I., os polinômios:

$$p_i(x) = \sum_j^d \sigma_i a_i^j x^j$$

são L.I.

$$A' = \begin{bmatrix} | & | & & | \\ \sigma_0 A_0 & \sigma_1 A_1 & \cdots & \sigma_d A_d \\ | & | & & | \end{bmatrix} = \begin{bmatrix} 1 & \sigma_1 a_1 & \cdots & \sigma_d a_1^n \\ \vdots & \ddots & & \vdots \\ 1 & \sigma_1 a_d & \cdots & \sigma_d a_d^n \end{bmatrix} \tag{7}$$

Note que cada linha i de A corresponde os coeficientes da expansão de Taylor de $\sigma(a_i x)$:

$$\sigma(a_i x) = \sum_j^d \sigma_j (a_i x)^j = \sum_j^d \sigma_j a_i^j x^j .$$

Sendo as linhas L.I., os polinômios:

$$p_i(x) = \sum_j^d \sigma_i a_i^j x^j$$

são L.I. Tendo $\dim P_d(\mathbb{R}) = d + 1$ e $\{p_i\}$ conjunto L.I., segue que $\{p_i\}$ forma uma base para esse espaço.

A rede de uma camada com uma variável tem representação

$$N(x) = \sum_{i}^{m} w_i \sigma(a_i x)$$

A rede de uma camada com uma variável tem representação

$$N(x) = \sum_{i}^{m} w_i \sigma(a_i x)$$

Com isso,

$$N^{(j)}(0) = \sum_{i=1}^{m} w_i a_i^j \sigma_j$$

A rede de uma camada com uma variável tem representação

$$N(x) = \sum_{i}^{m} w_i \sigma(a_i x)$$

Com isso,

$$N^{(j)}(0) = \sum_{i=1}^{m} w_i a_i^j \sigma_j$$

Logo,

$$\begin{aligned} N(x) &= \sum_{j}^{d} \frac{N^{(j)}}{j!} x^j + \mathcal{O}(x^{d+1}) \\ &= \sum_{j}^{d} \frac{(\sum_{i=1}^{m} w_i a_i^j \sigma_j)}{j!} x^j + \mathcal{O}(x^{d+1}) \end{aligned}$$

Se $p(x) = \sum_{j=0}^{d} b_j x^j$ é polinômio de grau d qualquer, queremos que:

$$N(x) = p(x) + \mathcal{O}(x^{d+1}) = \sum_{j}^{d} \frac{(\sum_{i=1}^{m} w_i a_i^j \sigma_j)}{j!} x^j + \mathcal{O}(x^{d+1}),$$

Se $p(x) = \sum_{j=0}^{d} b_j x^j$ é polinômio de grau d qualquer, queremos que:

$$N(x) = p(x) + \mathcal{O}(x^{d+1}) = \sum_{j}^{d} \frac{(\sum_{i=1}^{m} w_i a_i^j \sigma_j)}{j!} x^j + \mathcal{O}(x^{d+1}),$$

de modo que,

$$p(x) = \sum_{j}^{d} \frac{(\sum_{i=1}^{m} w_i a_i^j \sigma_j)}{j!} x^j$$

$$p(x) = \sum_{j}^{d} \frac{(\sum_{i=1}^{m} w_i a_i^j \sigma_j)}{j!} x^j \qquad \Longleftrightarrow \quad j! b_j = \sum_{i=0}^{d} w_i a_i^j \sigma_j, \forall j.$$

$$p(x) = \sum_{j}^{d} \frac{(\sum_{i=1}^{m} w_i a_i^j \sigma_j)}{j!} x^j \iff j! b_j = \sum_{i=0}^{d} w_i a_i^j \sigma_j, \forall j.$$

$$\iff \begin{bmatrix} \sigma_0 & \sigma_0 & \cdots & \sigma_0 \\ \sigma_1 a_0 & \sigma_1 a_1 & \cdots & \sigma_1 a_d \\ \vdots & \vdots & & \vdots \\ \sigma_d a_0^d & \sigma_d a_1^d & \cdots & \sigma_d a_d^d \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} b_0 0! \\ b_1 1! \\ \vdots \\ b_d d! \end{bmatrix}$$

$$p(x) = \sum_{j}^{d} \frac{(\sum_{i=1}^{m} w_i a_i^j \sigma_j)}{j!} x^j \quad \iff \quad j! b_j = \sum_{i=0}^{d} w_i a_i^j \sigma_j, \forall j.$$

$$\iff \begin{bmatrix} \sigma_0 & \sigma_0 & \cdots & \sigma_0 \\ \sigma_1 a_0 & \sigma_1 a_1 & \cdots & \sigma_1 a_d \\ \vdots & \vdots & & \vdots \\ \sigma_d a_0^d & \sigma_d a_1^d & \cdots & \sigma_d a_d^d \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} b_0 0! \\ b_1 1! \\ \vdots \\ b_d d! \end{bmatrix}$$

Isto é, $(A')^T w = b$. Como A' é inversível, $(A')^T$ também, de modo que o sistema tem solução.

$$p(x) = \sum_{j}^{d} \frac{(\sum_{i=1}^{m} w_i a_i^j \sigma_j)}{j!} x^j \quad \iff \quad j! b_j = \sum_{i=0}^{d} w_i a_i^j \sigma_j, \forall j.$$

$$\iff \begin{bmatrix} \sigma_0 & \sigma_0 & \cdots & \sigma_0 \\ \sigma_1 a_0 & \sigma_1 a_1 & \cdots & \sigma_1 a_d \\ \vdots & \vdots & & \vdots \\ \sigma_d a_0^d & \sigma_d a_1^d & \cdots & \sigma_d a_d^d \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} b_0 0! \\ b_1 1! \\ \vdots \\ b_d d! \end{bmatrix}$$

Isto é, $(A')^T w = b$. Como A' é inversível, $(A')^T$ também, de modo que o sistema tem solução.

Assim, para $W = \begin{bmatrix} w_0 & w_1 & \cdots w_d \end{bmatrix}$ que resolve $(A')^T w = b$, temos $N(x) = p(x) + \mathcal{O}(x^{d+1})$, como queríamos. ■

MTM

Teorema

Seja $p(x) = x^d$, *e suponha que* $\sigma_i \neq 0$ *para* $i \leq 2d$. *Então:*

(i) $m_1^{uniform}(p) = d + 1$,

(ii) $m^{uniform}(p) \leq 7d\lceil \log_2(d) \rceil$.

Essas afirmações também são válidas se $m^{uniform}$ *for substituído por* m^{Taylor}.

A parte (i) segue dos Teo. 4.1 e 4.2, com $n = 1$ e $r_1 = 1$.

A parte (i) segue dos Teo. 4.1 e 4.2, com $n = 1$ e $r_1 = 1$. Para a parte (ii):

$$\sigma(x) = \sigma_0 + \sigma_1 x + \sigma_2 x^2 + \mathcal{O}(x^3 + x^4 + x^5 \cdots)$$
$$\sigma(-x) = \sigma_0 - \sigma_1 x + \sigma_2 x^2 + \mathcal{O}(-x^3 + x^4 - x^5 \cdots)$$

A parte (i) segue dos Teo. 4.1 e 4.2, com $n = 1$ e $r_1 = 1$. Para a parte (ii):

$$\sigma(x) = \sigma_0 + \sigma_1 x + \sigma_2 x^2 + \mathcal{O}(x^3 + x^4 + x^5 \cdots)$$
$$\sigma(-x) = \sigma_0 - \sigma_1 x + \sigma_2 x^2 + \mathcal{O}(-x^3 + x^4 - x^5 \cdots)$$

Logo,

$$\frac{\sigma(x) + \sigma(-x) - 2\sigma_0}{2\sigma_2} = x^2 + \mathcal{O}(x^4 + x^6 + \cdots)$$

A parte (i) segue dos Teo. 4.1 e 4.2, com $n = 1$ e $r_1 = 1$. Para a parte (ii):

$$\sigma(x) = \sigma_0 + \sigma_1 x + \sigma_2 x^2 + \mathcal{O}(x^3 + x^4 + x^5 \cdots)$$
$$\sigma(-x) = \sigma_0 - \sigma_1 x + \sigma_2 x^2 + \mathcal{O}(-x^3 + x^4 - x^5 \cdots)$$

Logo,

$$\frac{\sigma(x) + \sigma(-x) - 2\sigma_0}{2\sigma_2} = x^2 + \mathcal{O}(x^4 + x^6 + \cdots)$$

I.e., podemos aproximar um quadrado usando 3 neurônios.

Escreva $d = d_0 2^0 + d_1 2^1 + \cdots + d_k 2^k$, com $d_i \in \{0, 1\}$.

Escreva $d = d_0 2^0 + d_1 2^1 + \cdots + d_k 2^k$, com $d_i \in \{0, 1\}$.
Em cada camada $l \in \{1, \cdots k\}$, faça:

- Uma porta de produto (4 neurônios) para produzir x^{2^l} a partir de $x^{2^{l-1}}$;
- Uma porta de:

Escreva $d = d_0 2^0 + d_1 2^1 + \cdots + d_k 2^k$, com $d_i \in \{0, 1\}$.
Em cada camada $l \in \{1, \cdots k\}$, faça:

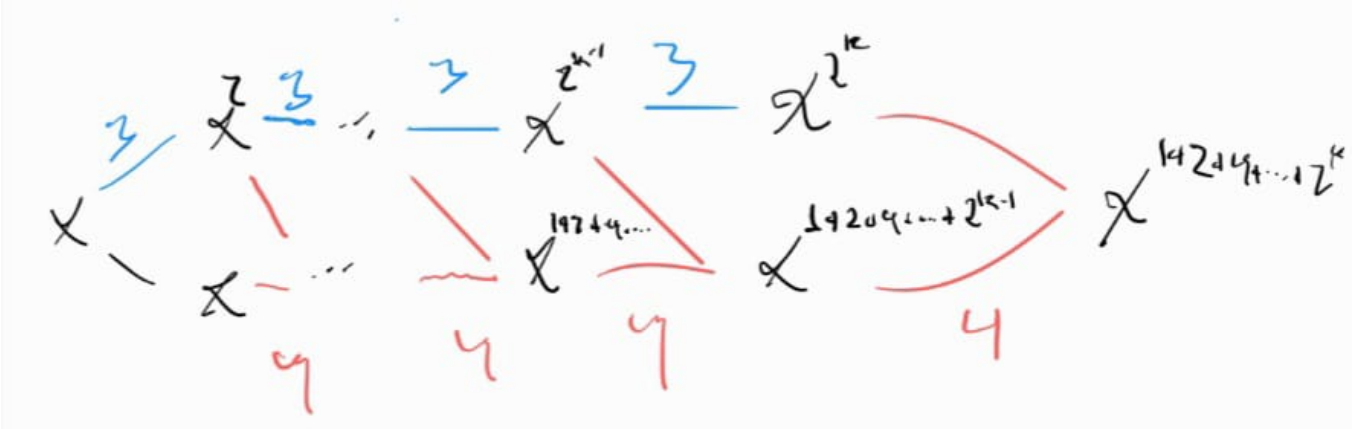
- Uma porta de produto (4 neurônios) para produzir x^{2^l} a partir de $x^{2^{l-1}}$;
- Uma porta de:
  - produto, se $d_l = 1$ (3 neurônios);
  - passagem, se $d_l = 0$ (1 neurônio);

Cada camada produz uma Taylor-aproximação de $x^{d_0 2^0 + d_1 2^1 + \cdots + d_l 2^l}$.

Escreva $d = d_0 2^0 + d_1 2^1 + \cdots + d_k 2^k$, com $d_i \in \{0, 1\}$.
Em cada camada $l \in \{1, \cdots k\}$, faça:

- Uma porta de produto (4 neurônios) para produzir x^{2^l} a partir de $x^{2^{l-1}}$;
- Uma porta de:
  - produto, se $d_l = 1$ (3 neurônios);
  - passagem, se $d_l = 0$ (1 neurônio);

Cada camada produz uma Taylor-aproximação de $x^{d_0 2^0 + d_1 2^1 + \cdots + d_l 2^l}$.



Essa construção leva $\lceil \log_2 d \rceil$ camadas e, no pior caso, $3 + 4 = 7$ neurônios por camada.

Essa construção leva $\lceil \log_2 d \rceil$ camadas e, no pior caso, $3 + 4 = 7$ neurônios por camada.

Logo, $m^{\text{Taylor}}(x^d) \leq 7\lceil \log_2 d \rceil$

Essa construção leva $\lceil \log_2 d \rceil$ camadas e, no pior caso, $3 + 4 = 7$ neurônios por camada.

Logo, $m^{\text{Taylor}}(x^d) \leq 7\lceil \log_2 d \rceil$

Pela Prop. 3.3, sendo x^d homogêneo, segue que $m^{\text{uniform}}(x^d) \leq m^{\text{Taylor}}(x^d)$.■

Teorema

Seja $p(x) = x_1 x_2 \dots x_n$ *e suponha que* $\sigma_i \neq 0$, *para* $i \leq n$. *Então,*

$$m_k^{uniform}(p) = \mathcal{O}(n^{\frac{k-1}{k}} 2^{n^{\frac{1}{k}}})$$

Ideia: Construção de uma rede separando as variáveis em grupos e multiplicando eles recursivamente.

Ideia: Construção de uma rede separando as variáveis em grupos e multiplicando eles recursivamente.

Separamos as n variáveis em grupos de b_1 elementos.

Ideia: Construção de uma rede separando as variáveis em grupos e multiplicando eles recursivamente.

Separamos as n variáveis em grupos de b_1 elementos.Cada grupo desse precisa de 2^{b_1} neurônios para ser multiplicado (LIN; TEGMARK; ROLNICK, 2017, Teorema 2).

Ideia: Construção de uma rede separando as variáveis em grupos e multiplicando eles recursivamente.

Separamos as n variáveis em grupos de b_1 elementos.Cada grupo desse precisa de 2^{b_1} neurônios para ser multiplicado (LIN; TEGMARK; ROLNICK, 2017, Teorema 2). No total, temos $\frac{n}{b_1}2^{b_1}$ neurônios na $1^{\underline{\text{a}}}$ camada.

Ideia: Construção de uma rede separando as variáveis em grupos e multiplicando eles recursivamente.

Separamos as n variáveis em grupos de b_1 elementos.Cada grupo desse precisa de 2^{b_1} neurônios para ser multiplicado (LIN; TEGMARK; ROLNICK, 2017, Teorema 2). No total, temos $\frac{n}{b_1}2^{b_1}$ neurônios na $1^{\underline{a}}$ camada.

Na camada seguinte, separamos os $\frac{n}{b_1}$ produtos em grupos de b_2 elementos, fazendo o produto entre eles. Cada grupo desse leva 2^{b_2} neurônios para multiplicar, totalizando $\frac{n}{b_1 b_2}2^{b_2}$ produtos.

Ideia: Construção de uma rede separando as variáveis em grupos e multiplicando eles recursivamente.

Separamos as n variáveis em grupos de b_1 elementos.Cada grupo desse precisa de 2^{b_1} neurônios para ser multiplicado (LIN; TEGMARK; ROLNICK, 2017, Teorema 2). No total, temos $\frac{n}{b_1}2^{b_1}$ neurônios na $1^{\underline{a}}$ camada.

Na camada seguinte, separamos os $\frac{n}{b_1}$ produtos em grupos de b_2 elementos, fazendo o produto entre eles. Cada grupo desse leva 2^{b_2} neurônios para multiplicar, totalizando $\frac{n}{b_1 b_2}2^{b_2}$ produtos.

Recursivamente, fazemos o processo em grupos de $b_1, b_2, \ldots, b_k$ de modo que $n = b_1 b_2 \cdots b_k$, tendo a cada $i-$ésima etapa:

Recursivamente, fazemos o processo em grupos de $b_1, b_2, \ldots, b_k$ de modo que $n = b_1 b_2 \cdots b_k$, tendo a cada $i-$ésima etapa:

- $\frac{n}{b_1 b_2 \cdots b_i}$ grupos de elementos;
- 2^{b_i} neurônios por grupo.

Recursivamente, fazemos o processo em grupos de $b_1, b_2, \ldots, b_k$ de modo que $n = b_1 b_2 \cdots b_k$, tendo a cada $i-$ésima etapa:

- $\frac{n}{b_1 b_2 \cdots b_i}$ grupos de elementos;
- 2^{b_i} neurônios por grupo.

sendo $\frac{n}{\prod_{j=1}^{i} b_j} 2^{b_i}$ neurônios por camada, totalizando nas k camadas:

Recursivamente, fazemos o processo em grupos de $b_1, b_2, \ldots, b_k$ de modo que $n = b_1 b_2 \cdots b_k$, tendo a cada $i-$ésima etapa:

- $\frac{n}{b_1 b_2 \cdots b_i}$ grupos de elementos;
- 2^{b_i} neurônios por grupo.

sendo $\frac{n}{\prod_{j=1}^{i} b_j} 2^{b_i}$ neurônios por camada, totalizando nas k camadas:

$$N = \sum_{i=1}^{k} \frac{n}{\prod_{j=1^i} b_j} 2^{b_i} = \sum_{i=1}^{n} \left(\prod_{j=i+1}^{k} b_j \right) 2^{b_i}$$

neurônios.

Essa construção faz, portanto, uma Taylor-aproximação de p com k camadas com o número apresentado de neurônios, de modo que:

$$m_k^{\text{Taylor}}(p) \leq \sum_{i=1}^{n} \left(\prod_{j=i+1}^{k} b_j \right) 2^{b_i}. \tag{8}$$

Essa construção faz, portanto, uma Taylor-aproximação de p com k camadas com o número apresentado de neurônios, de modo que:

$$m_k^{\text{Taylor}}(p) \leq \sum_{i=1}^{n} \left(\prod_{j=i+1}^{k} b_j \right) 2^{b_i}. \tag{8}$$

Como $p(x) = x_1 x_2 \cdots x_n$ é homogêneo, segue pela Proposição 3.3 que $m_k^{\text{uniform}}(p) \leq m_k^{\text{Taylor}}(p)$.

Essa construção faz, portanto, uma Taylor-aproximação de p com k camadas com o número apresentado de neurônios, de modo que:

$$m_k^{\text{Taylor}}(p) \leq \sum_{i=1}^{n} \left(\prod_{j=i+1}^{k} b_j \right) 2^{b_i}. \tag{8}$$

Como $p(x) = x_1 x_2 \cdots x_n$ é homogêneo, segue pela Proposição 3.3 que $m_k^{\text{uniform}}(p) \leq m_k^{\text{Taylor}}(p)$. Estabelecendo $b_i = n^{\frac{1}{k}}$:

$$m_k^{\text{Taylor}}(p) \leq \sum_{i=1}^{n} \left(\prod_{j=i+1}^{k} n^{\frac{1}{k}} \right) 2^{n^{\frac{1}{k}}} = 2^{n^{\frac{1}{k}}} \left(n^{\frac{k-1}{k}} + n^{\frac{k-2}{k}} + \cdots \right)$$

$$= \mathcal{O}(2^{n^{\frac{1}{k}}} n^{\frac{k-1}{k}}). \quad \square$$

Algumas questões:

- $n = b_1 b_2 \cdots b_k$?

Algumas questões:

- $n = b_1 b_2 \cdots b_k$?
- $b_i = n^{\frac{1}{k}}$?

Algumas questões:

- $n = b_1 b_2 \cdots b_k$?
- $b_i = n^{\frac{1}{k}}$?
- O que seria um b_i fracionário?

Vimos que o número de neurônios é dado por

$$\sum_{i=1}^{n} \left(\prod_{j=i+1}^{k} b_j \right) 2^{b_i}.$$

para b_i tais que $n = \prod_{i=1}^{k} b_i$.

Vimos que o número de neurônios é dado por

$$\sum_{i=1}^{n} \left(\prod_{j=i+1}^{k} b_j \right) 2^{b_i}.$$

para b_i tais que $n = \prod_{i=1}^{k} b_i$.
Podemos obter uma escolha ótima dos b_i buscando pontos crítico do Lagrangeano:

$$\mathcal{L}(b_i, \lambda) = \left(n - \prod_{i=1}^{k} b_j \right) \lambda + \sum_{i=1}^{n} \left(\prod_{j=i+1}^{k} b_j \right) 2^{b_i}$$

Diferenciando com relação a b_i, λ e igualando a 0, obtemos:

$$0 = -\lambda \prod_{j \neq i} b_j + \sum_{h=1}^{i-1} \left(\frac{\prod_{j=h+1}^{k} b_j}{b_i} \right) 2^{b_h} + (\log 2) \left(\prod_{j=i+1}^{k} b_j \right) 2^{b_i}, 1 \leq i \leq k$$

$$0 = n - \prod_{i=1}^{k} b_j$$

Diferenciando com relação a b_i, λ e igualando a 0, obtemos:

$$0 = -\lambda \prod_{j \neq i} b_j + \sum_{h=1}^{i-1} \left(\frac{\prod_{j=h+1}^{k} b_j}{b_i} \right) 2^{b_h} + (\log 2) \left(\prod_{j=i+1}^{k} b_j \right) 2^{b_i}, 1 \leq i \leq k$$

$$0 = n - \prod_{i=1}^{k} b_j$$

Fazendo um malabarismo algébrico, podemos verificar que:

$$b_i = b_{i-1} + log_2 \left(b_{i-1} - \frac{1}{\log 2} \right)$$

Diferenciando com relação a b_i, λ e igualando a 0, obtemos:

$$0 = -\lambda \prod_{j \neq i} b_j + \sum_{h=1}^{i-1} \left(\frac{\prod_{j=h+1}^{k} b_j}{b_i} \right) 2^{b_h} + (\log 2) \left(\prod_{j=i+1}^{k} b_j \right) 2^{b_i}, 1 \leq i \leq k$$

$$0 = n - \prod_{i=1}^{k} b_j$$

Fazendo um malabarismo algébrico, podemos verificar que:

$$b_i = b_{i-1} + log_2 \left(b_{i-1} - \frac{1}{\log 2} \right)$$

Assim, b_i cresce lentamente junto a i.

Conjectura 5.2

Seja $p(x) = x_1 x_2 \dots x_n$ e suponha que $\sigma_i \neq 0$, para $i \leq n$. Então,

$$m_k^{uniform}(p) = 2^{\Theta(n^{\frac{1}{k}})},$$

i.e., o expoente cresce na ordem de $n^{\frac{1}{k}}$.

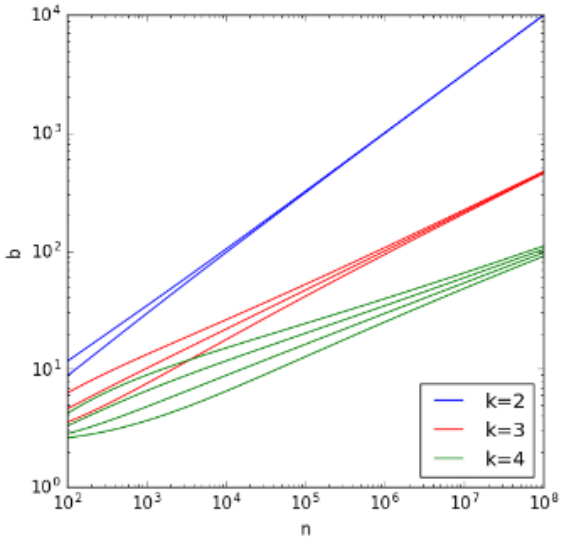


Figure 1: The optimal settings for $\{b_i\}_{i=1}^{k}$ as n varies are shown for $k = 1, 2, 3$. Observe that the b_i converge to $n^{1/k}$ for large n, as witnessed by a linear fit in the log-log plot. The exact values are given by equations (4) and (5).

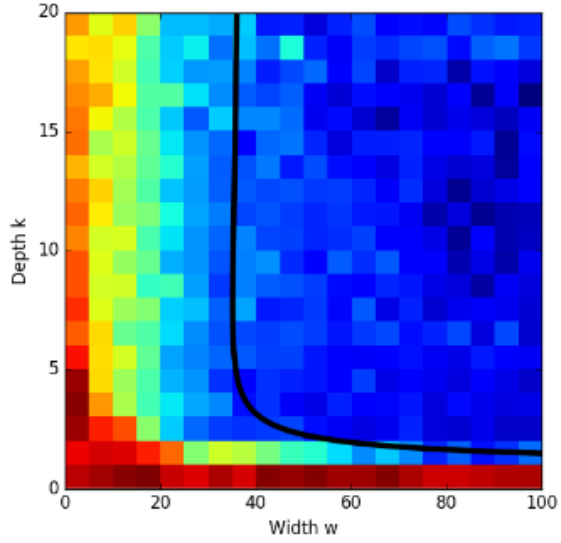


Figure 2: Performance of trained networks in approximating the product of 20 input variables, ranging from red (high error) to blue (low error). The error shown here is the expected absolute difference between the predicted and actual product. The curve $w = n^{(k-1)/k} \cdot 2^{n^{1/k}}$ for $n = 20$ is shown in black. In the region above and to the right of the curve, it is possible to effectively approximate the product function (Theorem 5.1).

Fonte: (ROLNICK; TEGMARK, 2018)

 LIN, H. W.; TEGMARK, M.; ROLNICK, D. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, Springer Science and Business Media LLC, v. 168, n. 6, p. 1223–1247, jul. 2017. ISSN 1572-9613. Disponível em: ⟨http://dx.doi.org/10.1007/s10955-017-1836-5⟩.

 ROLNICK, D.; TEGMARK, M. *The power of deeper networks for expressing natural functions*. 2018. Disponível em: ⟨https://arxiv.org/abs/1705.05502⟩.

Obrigado!

Contato: riffel.felipe@posgrad.ufsc.br

Repositório com os slides:
https://github.com/felipekriffel/Mestrado/