Modelová složitost neuronových sítí - zdánlivý paradox

Věra Kůrková

Ústav informatiky, Akademie věd České republiky
Pod Vodárenskou věží 2, 18207 Praha
Email: vera@cs.cas.cz

Abstrakt

V článku jsou studovány limity schopností umělých neuronových sítí s jednou skrytou vrstvou výpočetních jednotek při řešení vysoce dimenzionálních úloh. Na základě vlastnosti „koncentrace míry" eukleidovských prostorů vyšších dimenzí je ukázáno, že reprezentace náhodně vybrané funkce na dostatečně velké doméně s velkou pravděpodobností vyžaduje síť s počtem jednotek nebo velikostí vah závisející na vstupní dimenzi sítě exponencionálně. Je zdánlivým paradoxem, že najít konkrétní příklad takové funkce je obtížné. Možným vysvětlením je, že jak reálné úlohy modelované neuronovými sítěmi, tak funkce popsané matematickými formulemi patří do malé množiny funkcí, které mají strukturu, která se dá realizovat neuronovými sítěmi přijatelných velikostí. Situace připomíná známý paradox z teorie kódování „Každý kód, který nemůžeme vymyslet, je dobrý".

1 Úvod

Ačkoliv biologicky inspirované perceptronové sítě byly původně navrženy jako vícevrstvé architektury, nejrozšířenějším typem sítí se postupně staly sítě s vrstvou vstupů následovanou jednou skrytou vrstvou výpočetních jednotek a jednou výstupní jednotkou. Sítě s jednou skrytou vrstvou byly úspěšně využity v mnoha praktických aplikacích. Teprve v nedávné době došlo k obnovení zájmu o vícevrstvé architektury v podobě tzv. hlubokých sítí (Hinton a spol., 2006; Bengio, 2009). Efektivní učení sítí s několika skrytými vrstvami umožnil rozvoj hardwaru. Algoritmy učení těchto sítí využívají totiž GPU (graphic processing units), které byly vyvinuty pro potřeby počítačových her. Zcela ale chybí teoretická analýza výhod a nevýhod hlubokých sítí a porovnání jejich vlastností s vlastnostmi sítí s jednou skrytou vrstvou (kterým se začalo říkat mělké pro odlišení od hlubokých sítí).

Je známo, že mělké sítě s mnoha různými typy výpočetních jednotek dovedou dobře aproximovat spojité funkce na kompaktních množinách (viz např. (Pinkus, 1999)). V praktických aplikacích počítají neuronové sítě funkce na konečných množinách obsahujících pixely zpracovávaných obrázků nebo data, která mají být klasifikována. Všechny funkce na konečných doménách v $\mathbb{R}^d$ lze přesně reprezentovat mělkými sítěmi s populárními typy jednotek jako jsou perceptrony (Ito, 1992) nebo radiální jednotky (Micchelli, 1986).

Výsledky o univerzálních aproximačních a reprezentačních vlastnostech neuronových sítí předpokládají, že počet jednotek sítě je neomezený nebo, v případě konečných domén, je roven jejich velikosti. Modelová složitost sítě měřená počtem jednotek je ovšem limitujícím faktorem pro praktické aplikace. Řada horních odhadů této složitosti byla odvozena pomocí metod teorie nelineární aproximace funkcí (viz např. (Kainen a spol., 2009, 2012; Kůrková, 2012)) a umožnila popis tříd úloh, které lze řešit pomocí sítí s přijatelnou velikostí.

Na rozdíl od horních odhadů, které pouze vyžadují nalezení vhodného způsobu aproximace nebo reprezentace dané třídy funkcí mělkými sítěmi s určitým počtem výpočetních jednotek, získání dolních odhadů bývá mnohem obtížnější. Vyžaduje důkazy, že dané typy funkcí nelze žádným způsobem sítěmi omezené velikosti reprezentovat nebo aproximovat.

V tomto článku se zabýváme limity schopností mělkých neuronových sítí. Motivací je hledání porozumnění situacím, kdy jsou dvě a více skrytých vrstev výhodnější než jedna. Zkoumáme proto případy, kdy je využití mělkých sítí nevýhodné, protože má přílišné nároky na počet jednotek sítě nebo na velikost jejích parametrů. Zaměřujeme se na dolní odhady počtu jednotek a velikosti parametrů sítí reprezentujících funkce na konečných množinách. Množiny funkcí na konečných doménách lze reprezentovat jako eukleidovské prostory dimenzí rovných velikostem těchto domén. Vzhledem k tomu, že domény funkcí bývají v typických aplikacích neuronových sítí velké, projevují se při zkoumání složitosti sítí reprezentujících funkce na těchto doménách geometrické vlastnosti vysoce dimenzionálních prostorů. Jednou z nich je tzv. vlastnost koncentrace míry, která spočívá v tom, že s rostoucí dimenzí d se dostává většina povrchu d-dimenzionální koule do malé vzdálenosti od rovníku.

Tyto geometrické vlastnosti eukleidovských prostorů kombinované s relativně malou velikostí množin funkcí, které lze počítat běžnými výpočetními jednotkami (jako jsou perceptrony nebo jádrové jednotky používané v algoritmu Support Vector Machine), využíváme v tomto článku pro odvození dolních odhadů modelové složitosti mělkých sítí. Dokazujeme, že s rostoucí dimenzí exponencionálně klesá pravděpodobnost, že uniformě odně vybranou funkci lze reprezentovat sítí s počtem jednotek a velikostí výstupních vah závisejících na dimenzi

polynomiálně.

Je zdánlivým paradoxem, že ačkoliv jen relativně malá část množiny všech funkcí na dané konečné doméně se dá reprezentovat sítěmi přijatelné velikosti, sestrojení funkcí, které do této malé množiny nepatří, je obtížné. Situace připomíná teorii kódování, kde název článku „Any code of which we cannot think is good" (Coffey and Goodman, 1990) vyjadřuje skutečnost, že kódy, které v sobě nemají nějakou pravidelnost, nedokážeme vymyslet. Obdobně je obtížné vymyslet funkce nebo popsat vztahy mezi reálnými daty, které v sobě nemají nějakou strukturu, která se dá dobře modelovat mělkými sítěmi s vhodnými typy jednotek.

2 Slovníky výpočetních jednotek

Mělké sítě s lineární výstupní jednotkou počítají funkce vstup-výstup, které patří do množin tvaru

$$\operatorname{span}_n G := \left\{ \sum_{i=1}^{n} w_i g_i \,|\, w_i \in \mathbb{R},\, g_i \in G \right\},$$

kde G je množina funkcí, které počítají výpočetní jednotky daného typu (často nazývaná *slovník*), w_i jsou výstupní váhy a n je počet skrytých jednotek, jímž bývá měřena *modelová složitost*. Typické slovníky výpočetních jednotek jsou parametrické množiny funkcí tvaru $G_\phi(X, U)$, kde $\phi : X \times U \to \mathbb{R}$ je funkce dvou proměnných, $X \subset \mathbb{R}^d$ je *množina vstupů* a $U \subset \mathbb{R}^r$ je *množina parametrů*, které jsou optimalizovány během učení.

Původní výpočetní jednotky využívané v neuronových sítích jsou *perceptrony*, které počítají funkce tvaru

$$\sigma(v \cdot . + b) : X \to \mathbb{R},$$

kde $\sigma : \mathbb{R} \to \mathbb{R}$ je *aktivační funkce*. Nejčastěji má tvar *sigmoidy*, tj. je neklesající a $\lim_{t\to-\infty} \sigma(t) = 0$ a $\lim_{t\to\infty} \sigma(t) = 1$. Důležité typy sigmoid jsou *Heavisidova funkce* $\vartheta : \mathbb{R} \to \{0, 1\}$

$$\vartheta(t) := 0 \text{ pro } t < 0 \quad \text{a} \quad \vartheta(t) := 1 \text{ pro } t \geq 0$$

a funkce *signum* $\operatorname{sgn} : \mathbb{R} \to \{-1, 1\}$

$$\operatorname{sgn}(t) := -1 \text{ pro } t < 0 \quad \text{a } \operatorname{sgn}(t) := 1 \quad \text{pro } t \geq 0.$$

$H_d(X)$ značí slovník funkcí na $X \subset \mathbb{R}^d$ počítatelných *Heavisidovými perceptrony*, tj.

$$H_d(X) := \{\vartheta(v \cdot . + b) : X \to \{0, 1\} \,|\, v \in \mathbb{R}^d, b \in \mathbb{R}\}$$

a $P_d(X)$ slovník funkcí na X počítatelných *signum perceptrony*, tj.

$$P_d(X) := \{\operatorname{sgn}(v \cdot . + b) : X \to \{-1, 1\} \,|\, v \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

V tomto článku se z technických důvodů zaměřujeme na signum perceptrony, protože všechny funkce z $P_d(X)$ mají stejné normy rovné $(\operatorname{card} X)^{1/2}$. Z hlediska modelové složitosti je jen zanedbatelný rozdíl mezi sítěmi se signum a Heavisideovými perceptrony, protože platí $\operatorname{sgn}(t) = 2\vartheta(t) - 1$ a $\vartheta(t) = \frac{\operatorname{sgn}(t)+1}{2}$. Libovolnou síť s n signum perceptrony lze tedy nahradit sítí s $n + 1$ Heavisidovými perceptrony.

Dalším hojně využívaným typem jednotek jsou *jádrové jednotky*. Pro jádro $K_d : X \times U \to \mathbb{R}$ značíme $F_{K_d}(X, U)$ slovník *jádrových jednotek s parametry v* U, tj.

$$F_{K_d}(X, U) := \{K_d(., u) : X \to \mathbb{R} \,|\, u \in U\}.$$

Je-li $X = U$, píšeme $F_{K_d}(X)$. Slovníky tohoto typu využívá algoritmus *Support Vector Machine* (*SVM*), který hledá vhodné parametry jednotek v množině $U = \{u_1, \ldots, u_l\}$ vstupních dat. Nejrozšířenejším typem jádra je gaussovské jádro.

V tomto článku se zabýváme schopnostmi mělkých sítí reprezentovat funkce na konečných množinách v $\mathbb{R}^d$. Pro $X \subset \mathbb{R}^d$ značíme

$$\mathcal{F}(X) := \{f \,|\, f : X \to \mathbb{R}\}$$

lineární prostor všech funkcí na X. $\mathcal{B}(X)$ značí podmnožinu $\mathcal{F}(X)$ tvořenou funkcemi s hodnotami 1 a -1, tj.

$$\mathcal{B}(X) := \{f : X \to \{-1, 1\}\}.$$

Je-li $X \subset \mathbb{R}^d$ konečná množina, potom $\mathcal{F}(X)$ je izomorfní s konečně dimenzionálním prostorem $\mathbb{R}^{\operatorname{card} X}$. Tento izomorfismus indukuje na $\mathcal{F}(X)$ eukleidovský skalární součin

$$\langle f, g \rangle := \sum_{u \in X} f(u) g(u)$$

a eukleidovskou normu

$$\|f\| := \sqrt{\langle f, f \rangle}.$$

S_r^{m-1} značí sféru o poloměru r v $\mathbb{R}^m$ a

$$S_r(X) := \{f \in \mathcal{F}(X) \,|\, \|f\| = r\}$$

značí sféru o poloměru r v $\mathcal{F}(X)$. Abychom odlišili skalární součin na $\mathbb{R}^d$ od skalárního součinu $\langle ., . \rangle$ na $\mathcal{F}(X)$, značíme ho

$$u \cdot v := \sum_{i=1}^{d} u_i v_i.$$

3 Modelová složitost a variace funkcí

Užitečným nástrojem pro získávání odhadů závislosti přesnosti aproximace funkcí neuronovými sítěmi na počtu výpočetních jednotek je norma měřící korelaci aproximované funkce s typem výpočetních jednotek. Tuto normu lze také využít pro získání dolních odhadů počtu výpočetních jednotek nebo velikostí výstupních vah sítě.

Pro omezenou podmnožinu G normovaného lineárního prostoru $(\mathcal{X}, \|.\|_{\mathcal{X}})$, *$G$-variace (variace vzhledem k množině G)*, označovaná $\|.\|_G$, je definovaná

$$\|f\|_G := \inf\{c \in \mathbb{R}_+ \mid f/c \in \mathrm{cl}_{\mathcal{X}}\,\mathrm{conv}\,(G \cup -G)\}\ ,$$

kde $-G := \{-g \mid g \in G\}$, $\mathrm{cl}_{\mathcal{X}}$ značí uzávěr vzhledem k topologii indukované normou $\|\cdot\|_{\mathcal{X}}$ a conv značí konvexní obal. Tento pojem zavedl Barron (1992) pro množiny funkcí počítatelných perceptrony a Kůrková (1997) definovala variaci vzhledem k libovolné omezené množině a aplikovala na různé typy výpočetních jednotek. Pomocí této normy byly odvozeny odhady rychlostí klesání aproximačních chyb s rostoucím počtem jednotek mělkých sítí (viz např. (Kůrková, 2003; Kainen a spol., 2012)).

Variační norma může být také využita pro studium modelové složitosti sítí reprezentujících funkce na konečných množinách. Z její definice lze snadno odvodit následující tvrzení.

Tvrzení 3.1 *Nechť G je konečná množina funkcí na $X \subset \mathbb{R}^d$, $\mathrm{card}\, G = k$, potom pro každou funkci $f : X \to \mathbb{R}$ platí $\|f\|_G = \min\left\{\sum_{i=1}^k |w_i| \,\middle|\, f = \sum_{i=1}^k w_i\, g_i\,,\ w_i \in \mathbb{R},\ g_i \in G\right\}$.*

Toto tvrzení ukazuje, že pokud má funkce velkou variaci vzhledem ke slovníku výpočetních jednotek G, potom každá reprezentace této funkce mělkou sítí s jednotkami z G musí mít velký počet jednotek nebo musí být velké některé z výstupních vah. Charakterizace funkcí s velkou variací vede tedy k popisu tříd funkcí, jejichž reprezentace vyžaduje velký počet jednotek nebo některou velkou výstupní váhu. Oboje může limitovat možnosti implementace. Je pozoruhodné, že rovněž v teorii složitosti obvodů hrají důležitou roli třídy funkcí definované pomocí podmínek omezujících polynomiálně současně počet jednotek a velikosti vah.

Pro popis tříd funkcí s velkou variací využijeme následující větu vycházející z geometrické charakterizace variace (Kůrková, 2012). $G^\perp$ značí ortogonální doplněk G.

Věta 3.1 *Nechť $X \subset \mathbb{R}^d$ a G je omezená podmnožina prostoru $\mathcal{F}(X)$, potom pro všechna $f \in \mathcal{F}(X) \setminus G^\perp$ platí*

$$\|f\|_G \geq \frac{\|f\|^2}{\sup_{g\in G} |\langle g, f\rangle|}.$$

Věta 3.1 ukazuje, že funkce, které jsou „téměř ortogonální" ke všem funkcím z G, mají velkou G-variaci. Složitost sítí s jednotkami ze slovníku G tedy záleží na tom, jak je reprezentovaná funkce korelována s funkcemi z G.

4 Variace funkcí na velkých doménách

V této části ukazujeme, že geometrické vlastnosti eukleidovských prostorů vyšších dimenzí vedou k tomu, že množiny funkcí korelovaných s jednotlivými výpočetními jednotkami jsou poměrně malé. To plyne z tzv. vlastnosti koncentrace míry, jejímž speciálním případem je exponencionální pokles velikosti „polárních čepiček" sfér s rostoucí dimenzí eukleidovských prostorů (viz např. (Ball, 1997)). Pro $g \in S_r^{m-1}$ a $\varepsilon \in (0,1)$ označme

$$C(g,\varepsilon) := \{h \in S_r^{m-1} \,|\, |\langle h^o, g^o\rangle| \geq \varepsilon\}.$$

Potom pro všechna $g \in S_r^{m-1}$, $\varepsilon \in (0,1)$ a μ uniformní pravděpodobnostní míru na S_r^{m-1} platí

$$\mu(C(g,\varepsilon)) \leq e^{-\frac{m\varepsilon^2}{2}}. \tag{1}$$

Následující věta využívající odhad (1) ukazuje, že pokud slovník výpočetních jednotek na velké doméně je „relativně malý" vzhledem k velikosti množiny všech funkcí na této doméně, potom s velkou pravděpodobností reprezentace náhodně vybrané funkce vyžaduje mělkou síť s „velkým" počtem jednotek nebo „velkou" velikostí vah.

Věta 4.1 *Nechť $X \subset \mathbb{R}^d$, $\mathrm{card}\, X = m$, $G(X) \subset \mathcal{F}(X)$ takový, že $\mathrm{card}\, G(X) = k$, $b, r > 0$, a pro všechna $g \in G(X)$, $\|f\| = r$. Potom*
(i) pro každou uniformní pravděpodobnostní míru μ na $S_r(X)$ platí,

$$\mu(\{f \in \mathcal{F}(X) \,|\, \|f\|_{G(X)} \geq b\}) \geq 1 - 2k\, e^{-\frac{m}{2b^2}};$$

(ii) je-li $G(X) \subset \mathcal{B}(X)$, a f je uniformě náhodně vybraná funkce z $\mathcal{B}(X)$, potom

$$\Pr(\|f\|_{G(X)} \geq b) \geq 1 - 2k\, e^{-\frac{m}{2b^2}}.$$

Obě části Věty 4.1 plynou z Věty 3.1 a z geometrických vlastností eukleidovských prostorů (důkaz viz (Kůrková, 2014; Kůrková and Sanguineti, 2014)). Prostor $\mathcal{F}(X)$ je isometrický s prostorem $\mathbb{R}^{\mathrm{card}\, X}$. S rostoucí velikostí domény $\mathrm{card}\, X = m$, pravděpodobnostní míra množin vektorů korelovaných s výpočetními jednotkami ze slovníku $G(X)$ klesá exponencionálně. Odhady velikostí těchto měr plynou z vlastnosti koncentrace míry (1) a z Chernoffova odhadu (Chernoff, 1952) z teorie pravděpodobnosti.

Z Věty 4.1 plyne, že uniformní pravděpodobnostní míra množiny funkcí s variací větší než b je alespoň

$$1 - 2\,\mathrm{card}\, G(X)\, e^{-\frac{m}{2b^2}}.$$

Například pro $b = m^{1/4}$ je tato míra alespoň

$$1 - 2\,\mathrm{card}\, G(X)\, e^{-\frac{m^{1/2}}{2}}. \tag{2}$$

Pro „relativně malé" slovníky a „velké" domény X je dolní (2) mez blízko 1.

Mezi „relativně malé" slovníky patří slovníky $F_{K_d}(X)$ tvořené jádrovými jednotkami se středy v trénovacích vstupních datech, které se používají v algoritmu Support Vector Machine (SVM), který vybírá výpočetní jednotky (tzv. support vectors) pouze mezi jednotkami s parametry danými vstupními trénovacími daty. V tomto případě platí $\operatorname{card} G(X) = k = \operatorname{card} X = m$. Takže pro slovníky používané v SVM plyne z Věty 4.1, že míra množiny funkcí s variací větší než b je alespoň

$$1 - 2m\, e^{-\frac{m}{2b^2}}.$$

Pro $b = m^{1/4}$ dostaneme dolní odhad této míry

$$1 - 2m\, e^{-\frac{m^{1/2}}{2}}.$$

Je-li doména X d-dimenzionální boolovská krychle $X = \{0,1\}^d$, je její velikost $m = 2^d$ a z (2) plyne dolní odhad

$$1 - 2^{d+1}\, e^{-2^{d/2-1}}$$

velikosti množiny funkcí, které mají variaci větší než $2^{d/4}$.

Také slovník signum perceptronů $P_d(X)$ je „relativně malý". Odhad jeho velikosti v závislosti na velikosti m domény X a dimenzi d prostoru $\mathbb{R}^d$, v němž jsou body z X umístěny, plyne z odhadů počtu lineárně separovatelných dichotomií, které odvodil již v 19. století švýcarský matematik Schläfli (viz Schläfli (1901)). Modernější výklad jeho výsledků lze nalézt v článku (Cover, 1965).

Věta 4.2 *Pro každé d a každou podmnožinu $X \subset \mathbb{R}^d$ velikosti $\operatorname{card} X = m$ platí*

$$\operatorname{card} P_d(X) \leq 2\sum_{i=0}^{d}\binom{m-1}{i} \leq 2\frac{m^d}{d!}.$$

Z Věty 4.2 například plyne, že slovník funkcí počítatelných signum perceptrony na d-dimenzionální boolovské krychli $\{0,1\}^d$ má velikost menší než 2^{d^2}. To znamená, že jen malou část množiny $\mathcal{B}(\{0,1\}^d)$ velikosti 2^{2^d} tvoří funkce počítatelné signum perceptrony.

Na základě Vět 4.1 a 4.2 dostaneme následující odhad pravděpodobnostního rozložení funkcí s velkými variacemi vzhledem k signum perceptronům.

Důsledek 4.1 *Nechť $X \subset \mathbb{R}^d$, $\operatorname{card} X = m$, $G(X) \subset \mathcal{B}(X)$ takový, že $\operatorname{card} G(X) = k$, $b > 0$ a f je uniformě náhodně vybraná funkce z $\mathcal{B}(X)$, potom*

$$\Pr\left(\|f\|_{P_d(X)} \geq b\right) \geq 1 - 4\frac{m^d}{d!}e^{\frac{-m}{2b^2}}.$$

Např. pro doménu velikosti $m = 2^d$ dostaneme na základě Důsledku 4.1 dolní odhad

$$1 - 4\frac{2^{d^2}}{d!}e^{-2^{d/2-1}}$$

velikosti množiny funkcí, které mají variaci vzhledem k percepronům větší než $b = 2^{d/4}$.

5 Konstrukce funkcí s velkou variací vzhledem k perceptronům

Výsledky odvozené v předchozí části jsou existencionální. Plyne z nich, že s rostoucí velikostí domény X se zvyšuje pravděpodobnost, že perceptronová síť reprezentující náhodně vybranou binární klasifikační úlohu na X má „velký" počet jednotek nebo některé z výstupních vah této sítě jsou „velké". Přestože má většina funkcí velkou variaci vzhledem k perceptronům, sestrojit konkrétní příklad takové funkce není snadné. V této sekci popíšeme jediný typ konstrukce takových funkcí, který je nám znám.

Důležitý příklad z teorie boolovských funkcí je funkce "inner product mod 2", která nepatří do třídy $\widehat{LT_2}$ okruhů hloubky 2 polynomiální velikosti s prahovými jednotkami s polynomiálně omezenými vahami (Roychowdhury a spol., 1994). Tato funkce $\bar{\beta}_d : \{0,1\}^d \to \{0,1\}$ je definovaná na boolovských krychlích sudé dimenze $\{0,1\}^d$ takto:

$$\bar{\beta}_d := (l(x) \cdot r(x)) \mod 2,$$

kde $l(x), r(x) \in \{0,1\}^{d/2}$ jsou definovány $l(x)_i := x_i$ pro $i = 1, \ldots, \frac{d}{2}$ a $r(x)_i := x_{\frac{d}{2}+i}$ pro $i = 1, \ldots, \frac{d}{2}$. Nahrazením oboru hodnot $\{0,1\}$ oborem hodnot $\{-1,1\}$ dostaneme funkci

$$\beta_d := (-1)^{l(x)\cdot r(x)}.$$

Každou funkci f na boolovské krychli $\{0,1\}^d$ sudé dimenze lze reprezentovat jako $2^{d/2} \times 2^{d/2}$ matici M definovanou $M_{u,v} = f(u * v)$, kde $u, v \in \{0,1\}^{d/2}$ a $u * v$ značí konkatenaci vectorů u a v. Dá se ukázat, že matice reprezentující funkci β_d je tzv. Hadamardova matice, tj. její řádky (ekvivalentně sloupce) jsou navzájem kolmé. Na základě důkazu, že funkce $\bar{\beta}_d$ není ve třídě $\widehat{LT_2}$ (Hajnal a spol., 1993), dokázali Kůrková a spol. (1998), že variace vzhledem k perceptronům každé boolovské funkce, kterou lze reprezentovat jako Hadamardovu matici, závisí na d exponencionálně. Tento výsledek lze rozšířit na funkce na doménách $X \subset \mathbb{R}^d$ tvaru $X = Y \times Z$, kde $\operatorname{card} Y = \operatorname{card} Z = n$, definované pomocí $n \times n$ Hadamardových matic M jako $f(x,y) = M_{x,y}$. Variace vzhledem k perceptronům těchto funkcí je omezená zdola $\frac{\sqrt{m}}{\log_2 m}$, kde $m = n \times n$ (Kůrková, 2015).

Poděkování

Tato práce vznikla za podpory grantu MŠMT COST LD1302 a institucionální podpory Ústavu informatiky AVČR RVO 67985807.

Literatura

Ball, K. (1997). An elementary introduction to modern convex geometry. Levy, S. (ed.), In *Flavors of Geometry*, pp. 1–58. Cambridge University Press.

Barron, A. R. (1992). Neural net approximation. Narendra, K. S. (ed.), In *Proc. 7th Yale Workshop on Adaptive and Learning Systems*, pp. 69–72. Yale University Press.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2:1–127.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23:493–507.

Coffey, J. T. a Goodman, R. M. (1990). Any code of which we cannot think is good. *IEEE Transactions on Information Theory*, 36:326–334.

Cover, T. (1965). Geometrical and statistical properties of systems of linear inequalities with applictions in pattern recognition. *IEEE Transactions on Electronic Computers*, 14:326–334.

Hajnal, A., Maass, W., Pudlák, P., Szegedy, M. and Turán, G. (1993). Threshold circuits of bounded depth. *Journal of Computer and System Sciences*, 46:129–154.

Hinton, G. E., Osindero, S. and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554.

Ito, Y. (1992). Finite mapping by neural networks and truth functions. *Mathematical Scientist*, 17:69–77.

Kainen, P. C., Kůrková, V. and Sanguineti, M. (2009). Complexity of Gaussian radial-basis networks approximating smooth functions. *Journal of Complexity*, 25:63–74.

Kainen, P. C., Kůrková, V. and Sanguineti, M. (2012). Dependence of computational models on input dimension: Tractability of approximation and optimization tasks. *IEEE Transactions on Information Theory*, 58:1203–1214.

Kůrková, V. (1997). Dimension-independent rates of approximation by neural networks. Warwick, K. and Kárný, M. (ed.), In *Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality*, pp. 261–270. Birkhäuser, Boston, MA.

Kůrková, V. (2003). High-dimensional approximation and optimization by neural networks. Suykens, J. (ed.), In *Advances in Learning Theory: Methods, Models, and Application*, vol. 190 *NATO Science Series III: Computer & Systems Sciences*, pp. 69–88. IOS Press, Amsterdam.

Kůrková, V. (2012). Complexity estimates based on integral transforms induced by computational units. *Neural Networks*, 33:160–167.

Kůrková, V. (2014). Representations of highly-varying functions by one-hidden-layer networks. Rutkowski, L. (ed.), In *Proceedings of ICAISC 2014*, vol. 8467 *LNAI*, pp. I. 67–76. Springer.

Kůrková, V. (2015). Complexity of shallow networks representing finite mappings. Rutkowski, L. (ed.), In *Proceedings of ICAISC 2015*, *LNAI*, Springer (to appear).

Kůrková, V. a Sanguineti, M. (2014). Complexity of shallow networks representing functions with large variations. Wermter, S., Weber, C., Duch, W., Honkela, T., Koprinkova-Hristova, P., Magg, S., Palm, G. and Villa, A. (ed.), In *Artificial Neural Networks and Machine Learning, Proceedings of ICANN 2014*, vol. 8681 *LNCS*, pp. 331–338. Springer.

Kůrková, V., Savický, P. and Hlaváčková, K. (1998). Representations and rates of approximation of real-valued Boolean functions by neural networks. *Neural Networks*, 11:651–659.

Micchelli, C. A. (1986). Interpolation of scattered data: Distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2:11–22.

Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195.

Roychowdhury, V., Siu, K.-Y. and Orlitsky, A. (1994). Neural models and spectral methods. Roychowdhury, V., Siu, K. and Orlitsky, A. (ed.), In *Theoretical Advances in Neural Computation and Learning*, pp. 3–36. Springer, New York.

Schläfli, L. (1901). *Theorie der vielfachen Kontinuität*. Zürcher & Furrer, Zürich.