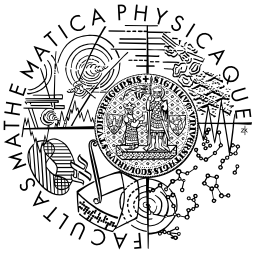


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Bc. Petr Raab

Věty o univerzalitě a konzistenci neuronových sítí

Katedra pravděpodobnosti a matematické statistiky MFF UK

Vedoucí diplomové práce: doc. RNDr. Michal Pešta, Ph.D.

Studijní program: Pravděpodobnost, matematická statistika a ekonometrie

Praha 2024

Prohlašuji, že jsem tuto diplomovou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Zde bych chtěl poděkovat doc. Peštovi za vedení této práce, jeho rady při řešení problémů a tipy na směrování celé práce. Dále bych ještě rád poděkoval mým rodičům, bez jejichž podpory bych se k dokončení této práce v životě nedostal.

Název práce: Věty o univerzalitě a konzistenci neuronových sítí

Autor: Bc. Petr Raab

Katedra: Katedra pravděpodobnosti a matematické statistiky MFF UK

Vedoucí diplomové práce: doc. RNDr. Michal Pešta, Ph.D., Katedra pravděpodobnosti a matematické statistiky MFF UK

Abstrakt: Práce se zabývá neuronovými sítěmi a modelem hlubokého učení, kdy se autor snaží na neuronové sítě dívat jako na statistický model podobný zobecněným lineárním modelům. Po představení tohoto modelu a zavedení značení se práce věnuje schopnosti neuronových sítí aproximovat spojité funkce, kdy je předveden důkaz věty o univerzalitě. Následně jsou zkoumány asymptotické vlastnosti neuronových sítí, pomocí sítového odhadu je také dokázána jejich konzistence a asymptotická normalita. Právě tyto dvě vlastnosti jsou objektem zkoumání v simulační studii na generovaných datech.

Klíčová slova: věty o univerzalitě|věty o konzistenci|neuronové sítě|hluboké učení|asymptotické chování|sítový odhad

Title: Universality and consistency theorems for neural networks

Author: Bc. Petr Raab

Department: Department of Probability and Mathematical Statistics, MFF UK

Supervisor: doc. RNDr. Michal Pešta, Ph.D., Department of Probability and Mathematical Statistics, MFF UK

Abstract: The work deals with neural networks and deep learning models, where the author attempts to view neural networks as a statistical model similar to generalized linear models. After introducing this model and introducing the notation, the work focuses on the ability of neural networks to approximate continuous functions, with a proof of the universality theorem presented. Subsequently, the asymptotic properties of neural networks are examined, and using network estimation, their consistency and asymptotic normality are also proven. These two properties are precisely the subject of investigation in a simulation study on generated data.

Keywords: universality theorems|consistency theorems|neural networks|deep learning|asymptotic behavior|sieve estimator

Obsah

Úvod

Při modelování závislosti odezvy na nějakých vysvětlujících veličinách pomocí zobecněných lineárních modelů se během hledání finálního modelu nevyhneme volbě vhodné parametrizace jednotlivých vysvětlujících veličin, protože v praxi pouze velmi vzácně je podmíněná střední hodnota odezvy lineární funkcí dostupných vysvětlujících veličin. Takový proces obvykle obnáší podrobnou vizuální analýzu grafů a popisných statistik a následné testování jednotlivých modelů, která parametrizace minimalizuje devianci. V případě velkého počtu proměnných je tento proces velmi časově náročný. Navíc pokud vliv jedné veličiny na odezvu závisí na hodnotách dalších vysvětlujících veličin ve složitých nelineárních vztazích, tak je v podstatě nemožné vypozorovat tyto komplexní interakce z popisných statistik nebo z nějakých grafů. Právě snaha o zautomatizování procesu hledání vhodné parametrizace nás vede k použití neuronových sítí a modelu hlubokého učení, kdy je proces tento proces parametrizován a tyto parametry jsou odhadnuty z dat.

Použití tohoto v současnosti populárního modelu nás navíc neomezuje pouze na modelování lineárních závislostí jako v případě zobecněných lineárních modelů. Teorie okolo hlubokého učení je momentálně budována spíše informatiky než statistiky, proto se v první kapitole této práce pokusíme zavést model hlubokého učení do formální podoby, kdy bude kladen důraz na to, aby se tato definice podobala obvyklé definici zobecněných lineárních modelů. K tomu bude třeba formální zadefinování dopředné neuronové sítě a jejích stavebních bloků, neuronu a jednotlivých neuronových vrstev. Také se pokusíme propojit některé pojmy používané v machine-learningové komunitě se statistickými pojmy používanými v teorii zobecněných lineárních modelů, stejně tak si propojíme obvyklé úlohy strojového učení se základními rozděleními exponenciálního typu. Uvedeme si několik příkladů aktivačních funkcí a jejich vztah s linkovou funkcí zobecněného lineárního modelu.

V druhé kapitole se podíváme, zda-li neuronové sítě umožňují modelování libovolně složité závislosti podmíněné střední podmíněné vysvětlujícími veličinami. Jinými slovy nás bude zajímat, jestli je možné aproximovat každou spojitou funkci s libovolnou přesností za použití dostatečně velké a složité neuronové sítě. Větám, které se touto otázkou zabývají, se říká věty o univerzalitě. V této práci se omezíme na pouze dva specifické typy neuronových sítí, pro které si věty o univerzalitě po zavedení potřebných definic formulujeme a následně dokážeme.

Třetí kapitola se zabývá asymptotickým chováním neuronových sítí. Zajímá nás totiž, jestli je možné nejen libovolnou závislost odezvy na vysvětlujících proměnných aproximovat, ale zda-li je možné i tuto závislost konzistentně odhadnout z dat. Po zadefinování parametrického sítového odhadu si pro tento odhad parametrů neuronové sítě dokážeme konzistenci a formulujeme větu o jeho asymptotické normalitě za vhodných relativně obecných předpokladů pro jeden specifický typ neuronových sítí.

V poslední, čtvrté kapitole se v krátké simulační studii pokusíme ověřit, jestli je možné větám z třetí kapitoly věřit i v praktických aplikacích, kdy nemáme k dispozici nekonečné množství dat, ale pouze datovou sadu v rozsahu stovek či tisíců pozorování. Po vygenerování takovýchto dat se podíváme, zda-li sítový

odhad již vykazuje známky konzistence a asymptotické normality.

1. Hluboké učení: teorie

V této kapitole si položíme potřebné definice, které následně využijeme při budování statistické teorie zabývající se moderní třídou statistických modelů spadajících do kategorie strojového učení, konkrétně neuronovými sítěmi a hlubokým učením.

1.1 Definice rozdělení exponenciálního typu

Nejprve, jak se obvykle postupuje při definici zobecněného lineárního modelu, se kterým, jak bude v této práci později ukázáno, mají neuronové sítě nemálo společného, si zadefinujeme rozdělení exponenciálního typu náhodných veličin. Právě s rozděleními spadající do této třídy budeme v této práci pracovat.

Definice 1 (rozdělení exponenciálního typu)**.** *Nechť $\boldsymbol{X}$ je reálná náhodná veličina, pak řekneme, že její rozdělení je exponenciálního typu, jestliže její hustota vzhledem k nějaké $\sigma-$konečné míře může být zapsána ve tvaru:*

$$f(x,\theta,\varphi) = \exp\Big\{\frac{x\theta - b(\theta)}{\varphi} + c(x,\varphi)\Big\},$$

kde θ se nazývá kanonický parametr, b a c jsou nějaké reálné funkce a φ je disperzní parametr, pro který platí, že $\varphi > 0$.

Mezi rozdělení exponenciálního typu patří i rozdělení uvažované při klasických úlohách strojového učení, ať již se jedná o binární klasifikační úlohu modelovanou pomocí binomického modelu, či vícetřídovou klasifikaci modelovanou pomocí log-lineárního modelu, nebo regresní úlohu na reálných číslech, kdy můžeme využít normální, gamma či například inverzní gaussovské podmíněné rozdělení vysvětlující proměnné. Všechny tyto rozdělení, jak je známo, totiž patří do třídy rozdělení exponenciálního typu.

Právě na rozdělení exponenciálního typu jsou definovány zobecněné lineární modely. Tyto modely se používají k odhadování podmíněné střední hodnoty, podobně jako v případě běžných lineárních modelů. Konkrétně nám umožňují modelovat podmíněnou střední hodnotu závislé náhodné veličiny Y za podmínky vysvětlujících náhodných veličin $\boldsymbol{X}$, kdy předpokládáme, že pokud na tuto podmíněnou střední hodnotu aplikujeme ryze monotónní, dvakrát diferencovatelnou linkovou funkci g, tak dostáváme rovnost s lineární kombinací vysvětlujících veličin. Tato závislost se dá zapsat jako:

$$\mathsf{E}\,[Y \mid \boldsymbol{X}] = g^{-1}(\boldsymbol{\beta}^\top \boldsymbol{X}), \tag{1.1}$$

kde $\boldsymbol{\beta}$ je vektor regresních parametrů. V praxi se ale velmi zřídka stane, že bychom měli k dispozici pro každé pozorování vektor vysvětlujících veličin $\boldsymbol{X}$, který by tuto rovnost splňoval pro zvolenou linkovou funkci g. Obvykle se snažíme najít vhodnou transformaci vektoru $\boldsymbol{X}$, abychom se platnosti uvažované rovnosti, co nejvíce přiblížili. To vyžadujeme nemalé úsilí, kdy člověk sestavující model musí analyzovat vlivy jednotlivých vysvětlujících proměnných na závislou proměnnou pomocí různých popisných statistik a grafických metod (hlavně grafů).

Kromě časové náročnosti a celkové pracnosti je dalším problémem hledání vhodné transformace i subjektivita analýzy grafických vizualizací. Co by se někomu mohlo zdát jako exponenciální závislost, by někomu mohla připadat jako závislost kvadratická, v případě, že by vyzkoušení obou parametrizací vedlo k velmi podobným výsledkům, tak by bylo pouze na člověku sestavujícím daný model, jako parametrizaci nakonec subjektivně zvolí. Navíc v případě velmi jemných závislostí není ani velmi zkušený člověk během datové explorace schopný tyto závislosti zachytit (a vůbec se například pokusit testovat jejich statistickou významnost). Podobný problém nastává i pokud je k dispozici velké množství vysvětlujících proměnných, mezi kterými jsou velmi složité, nelineární interakční závislosti.

Právě snaha o zautomatizování procesu hledání vhodných parametrizací jednotlivých vysvětlujících veličin při modelování podmíněné střední hodnoty vede k použití dopředných neuronových sítí, které hledání nejvhodnější parametrizace převádí na regresní problémy.

1.2 Dopředná neuronová síť

V této kapitole si tyto dopředné neuronové sítě představíme a formálně zadefinujeme. Začneme ale nejprve od základní stavební jednotky všech neuronových sítí, neuronů.

Definice 2 (neuron)**.** *Jako neuron $z(\boldsymbol{x})$ s aktivační funkcí ϕ a váhami $\boldsymbol{w} \in \mathbb{R}^{n+1}$ chápeme zobrazení z prostoru $1 \times \mathbb{R}^n$ do $\mathbb{R}$ zadané předpisem*

$$z(\boldsymbol{x}) = \phi(\boldsymbol{w}^\top \boldsymbol{x}).$$

Již na první pohled je zřejmá souvislost se zobecněnými lineárními modely. Aktivační funkce ϕ v uvedeném předpisu se velmi nápadně podobá inverzu linkové funkce g v rovnosti (1.1). Později se ukáže, že opravdu automatizované hledání vhodné transformace vysvětlujících proměnných při modelování podmíněné střední hodnoty pomocí zobecněných lineárních modelů je vlastně pouhé odhadování vah neuronu $\boldsymbol{w}$.

Když již máme zadefinovaný neuron, můžeme přistoupit k zavedení vyšší složky dopředných neuronových sítí.

Definice 3 (vrstva dopředné neuronové sítě)**.** *Vrstvu dopředné neuronové sítě (také FN vrstva[1]) $\boldsymbol{z}(\boldsymbol{x})$ s aktivační funkcí ϕ definujeme jako zobrazení z $1 \times \mathbb{R}^n$ do $1 \times \mathbb{R}^m$, dané předpisem:*

$$\boldsymbol{z}(\boldsymbol{x}) = (1,z_1(\boldsymbol{x}),z_2(\boldsymbol{x}),...,z_m(\boldsymbol{x}))^\top,$$

kde $z_i, i = 1,...,m$ jsou neurony s aktivační funkcí ϕ a váhami $\boldsymbol{w}_i$.

FN vrstva je opravdu složitější stavební jednotka neuronových sítí, jelikož se již z definice přímo skládá z neuronů. Mohli bychom ji explicitně definovat ve tvaru

$$\boldsymbol{z}(\boldsymbol{x}) = (1,\phi(\boldsymbol{w}_1^\top \boldsymbol{x}),\phi(\boldsymbol{w}_2^\top \boldsymbol{x}),...,\phi(\boldsymbol{w}_m^\top \boldsymbol{x}))^\top,$$

[1] z angl. Fast-forward Neural Network- FN Network

ale dali jsme přednost si ji zadefinovat pomocí neuronů, aby byla více vidět její spojitost s jednotlivými neurony.

Samozřejmě bychom mohli definovat vrstvy dopředných neuronových sítí ještě obecněji, kdy bychom dovolovali jednotlivým neuronům v jedné vrstvě mít rozdílné aktivační funkce, ale vzhledem k tomu že takové funkce se v praxi v podstatě nevyužívají, se omezíme pouze na tento případ. Vrstvy dopředné neuronové sítě tvořené těmito obecnějšími neurony by navíc vyžadovaly hledání vhodných jednotlivých aktivačních funkcí, což by byl ještě obtížnější úkol než manuální klasické hledání vhodné parametrizace vysvětlujících veličin, protože počet neuronů i v poměrně malých neuronových sítí je již dost vysoký. U menších neuronek by stále přicházelo v úvahu hledání těchto aktivačních funkcí pomocí tzv. grid searche (nebo nějaké jeho varianty), ale s komplikovanější strukturou neuronových sítí, jak bude popsáno později, velmi rychle roste počet neuronů, takže by se hledání optimálních aktivačních funkcí stávalo výpočetně nezvladatelné.

Nyní nám již nic nebrání v definici dopředné neuronové sítě.

Definice 4 (dopředná neuronová síť)**.** *Nechť* $\boldsymbol{z}^{(1)}$,$\boldsymbol{z}^{(2)}$,...,$\boldsymbol{z}^{(d)}$ *jsou vrstvy dopředné neuronové sítě s aktivačními funkcemi* ϕ_1, ϕ_2,..., ϕ_d, *splňující:*

$$\boldsymbol{z}^{(i+1)} : 1 \times \mathbb{R}^{q_i} \to 1 \times \mathbb{R}^{q_{i+1}},$$

kde $q_i \in \mathbb{N}$, *pro* $i = 0,1,..,d-1$.

Pak jako dopřednou neuronovou síť (FN network) $\boldsymbol{z}^{(d:1)}$ *o hloubce d chápeme zobrazení z* $1 \times \mathbb{R}^{q_0}$ *do* $1 \times \mathbb{R}^{q_d}$ *vzniklé složením jednotlivých vrstev* $\boldsymbol{z}^{(i)}$, *tedy toto zobrazení je zadáno předpisem:*

$$\boldsymbol{z}^{(d:1)}(\boldsymbol{x}) = \boldsymbol{z}^{(d)} \circ ...\boldsymbol{z}^{(2)} \circ \boldsymbol{z}^{(1)}(\boldsymbol{x}).$$

Dopředná neuronová síť je tedy složením jejích jednotlivých vrstev. Požadavek na dimenzi prostorů, jejichž podmnožinou jsou definiční obory i obory hodnot jednotlivých vrstev byl v definici právě kvůli tomu, aby následné postupné složení vrstev mělo smysl.

Předpoklad na to, aby obor hodnot aktivační funkce předcházející vrstvy byl v definičním oboru aktivační funkce vrstvy následující nebyl opomenut, protože vůbec není potřeba. Tuto vlastnost totiž zaručí vhodné váhy neuronů v napojené vrstvě, které například ze záporného vstupu snadno vytvoří kladnou hodnotu, která již bude v definičním oboru příslušné aktivační funkce, a to ať již zápornou hodnotou některých váhových parametrů či pomocí posunutí způsobeným kladnou hodnotou parametru u absolutního členu, který je vždy z definice neuronů přítomen. Na tuto vlastnost je ale třeba myslet v případě, kdy jsou váhové parametry odhadovány pomocí iteračních algoritmů, kdy by bylo vhodné volit iniciační hodnoty parametrů, tak aby k problémům s $\boldsymbol{w}^\top \boldsymbol{x}$ ležícím mimo definiční obor ϕ nedocházelo.

Právě struktura neuronových sítí stojí za jejím názvem. Podobně jako neurony v mozku si i neurony v rozdílných vrstvách v "našich"neuronových sítích předávají impulzy reagující na vstupní podnět. Dopředné neuronové sítě, pak dostaly své jméno díky tomu, že neurony si tyto impulzy předávají postupně od první do $d-$té vrstvy. Existuje i jiný typ neuronových sítí, kdy jsou v jejich strukturách cykly, které umožňují zobrazit přes neuron vstupy rekurentně. Neuron takové rekurentní sítě je pak definován jako:

$$z(\boldsymbol{x},\boldsymbol{z}) = \phi(\boldsymbol{w}^\top \boldsymbol{x} + \boldsymbol{u}^\top \boldsymbol{z}),$$

kde $\boldsymbol{z}$ značí nějakou vrstvu neuronů z dané neuronové sítě. Váhy $\boldsymbol{u}$ již neobsahují absolutní člen, ten je již zastoupen v první složce $\boldsymbol{w}$. Tyto rekurentní neuronové sítě nacházejí uplatnění například při analýze časových řad nebo v různých aplikacích zaměřených na rozeznávání textu, zvukových signálů či řeči. Neuronovými sítěmi tohoto typu se nebudeme dále v této práci zabývat a omezíme se pouze na klasické dopředné neuronové sítě.

Často se rozlišují dopředné neuronové sítě na dva případy, na tzv. *mělké* a *hluboké* dopředné neuronové sítě, podle počtu jejich vrstev, kdy jako mělké označujeme takové neuronové sítě, které mají jednu vrstvu, $d = 1$, a tedy pro hluboké platí, že $d > 1$. Někdy se v literatuře označují jako hluboké pouze sítě, které mají výrazně vyšší počet vrstev, $d \gg 1$, ale v této práci budeme jako hluboké považovat všechny dopředné neuronové sítě, které jsou tvořeny více než jednou vrstvou.

Naše poznámka o rychlém růstu počtu neuronů v neuronových sítích s rostoucí složitostí struktury se po zadefinování kompletní dopředné neuronové sítě ukazuje jako pádná. Počet neuronů sítě o hloubce d a s výstupními dimenzemi (bez absolutních členů) jednotlivých dopředných vrstev $q_0, q_1, ..., q_d$ je roven $\sum_{i=1}^{d} q_i$.

Kromě jednotlivých vrstev $\boldsymbol{z}^{(i)}$ lze do FN sítí přidat i jiné vrstvy, které například umožňují přeskakovat nějakou zvolenou vrstvu, normalizují zobrazené hodnoty ze zvolených vrstev, transformují kategoriální veličiny nebo text do vektorové podoby (vnořování slov[2]) či pomáhají snižovat riziko přeučení. Používání těchto vrstev často nemá žádnou oporu v teorii a rozhodnutí, zda se tyto vrstvy přidají do finální podoby neuronové vrstvy, záleží případ od případu, kdy se většinou aplikuje metoda pokus-omyl.

1.3 Aktivační funkce

Při stavění modelu založeném na neuronových sítí se nevyhneme volbě vhodných aktivačních funkcí jednotlivých neuronů. Jak již bylo diskutováno, obvykle se volí jednotná aktivační funkce pro všechny neurony ze stejné dopředné vrstvy. Teoreticky bychom mohli aktivační funkci zvolit úplně libovolně, ale k numerickému hledání váhových parametrů neuronových sítí se nám bude hodit jejich diferencovatelnost a jednoduchost výpočtu jejich derivace. Nyní si představíme několik aktivačních funkcí, které jsou v machine learningové komunitě běžně používané.

Nejjednodušším typem aktivační funkce je tzv. kroková funkce definovaná jako $\phi(x) = 1\{x \geq 0\}$, jedná se tedy o indikátor nezáporných hodnot $\boldsymbol{w}^\top \boldsymbol{x}$. Tato aktivační funkce, pak dělá z neuronové sítě klasifikační rozhodovací strom. I když je tato funkce diferencovatelná skoro všude (v nule můžeme uvažovat subdiferenciál), tak hodnota její derivace je identicky rovná nule, což není ideální vzhledem k tomu, že parametry neuronových sítí jsou obvykle hledány pomocí optimalizačních iteračních metod, kdy se minimalizuje nějaká ztrátová funkce. I to je jeden z důvodů, proč se tato aktivační funkce v praxi neužívá.

Jako další možnost se nabízí volit lineární funkci $\phi(x) = ax$ pro nějaké $a \in \mathbb{R}$. Tato volba aktivační funkce nabízí snadno vypočítatelnou derivaci a v případě použití lineární aktivační funkce ve všech vrstvách sítě i interpretovatelnost jednotlivých váhových parametrů, což je pro pro neuronové sítě obecně nezvyklé.

[2] angl. *word embedding*

Na druhou stranu naší motivací k modelování podmíněné střední hodnoty pomocí hlubokého učení na místo zobecněných lineárních modelů byla kromě automatizovaného hledání optimální transformace vysvětlujících veličin i schopnost modelovat nelineární závislosti a interakce mezi jednotlivými vysvětlujícími veličinami. Neuronová síť tvořená neurony s lineárními aktivačními funkcemi tuto schopnost ale nemá. Nepřekvapí tedy, že ani tato aktivační funkce není příliš využívaná.

V praxi používanou aktivační funkcí je tzv. sigmoid funkce, definovaná jako

$$\sigma(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}},$$

což není nic jiného než inverzní funkce k logitové funkci $\log(\frac{x}{1-x})$, což je kanonická funkce alternativního rozdělení, a je tedy využívána v logistické regresi. Není tedy překvapením, že se jedná i o velmi často využívanou aktivační funkci výstupní vrstvy v modelu hlubokého učení, který je definován níže, protože binomická klasifikace je klasická úloha strojového učení. Tato funkce je zobrazuje množinu reálných čísel na interval (0,1). Její velmi dobrá vlastnost je diferencovatelnost na celém definičním oboru, kdy její derivace je ve tvaru:

$$\sigma'(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}} - \frac{1}{(1+e^{-x})^2} = \sigma(x)(1-\sigma(x)),$$

což je výpočetně velmi výhodné, není třeba $\sigma(x)$ numericky derivovat, stačí ji pouze evaluovat v x a s výsledkem provést jednoduchou transformaci.

Sigmoid funkce není antisymetrická okolo nuly, této vlastnosti ale lze dosáhnout posunutím a složením s transformací $2x$, kdy dostáváme funkci

$$\tanh(x) = 2\sigma(2x) - 1 = \frac{2}{1+e^{-2x}} = \frac{e^x - e^{-x}}{e^x + e^{-x}},$$

tedy hyperbolický tangent. Ten je již asymetrický okolo nuly, když zobrazuje reálná čísla na interval (-1,1). Další jeho výhodou oproti sigmoid funkci je i vyšší hodnota jeho derivace okolo nuly, což vede k rychlejšímu hledání optimálních parametrů.

Pravděpodobně nejpoužívanější aktivační funkcí je v posledních letech tzv. ReLU funkce[3], která není nic jiného než identická funkce na kladných číslech a identicky rovna nule na zbytku reálné osy, $ReLU(x) = \max\{0,x\}$. V literatuře se uvádí, že v aplikacích se ukazuje, že vrstvy s aktivační ReLU funkcí jsou nejvíce eficientní ve významu, že váhové parametry jsou odhadnuty rychleji, co se počtu iterací numerických algoritmů týče, než v případě vrstev se sigmoid aktivační funkcí. Zároveň tato funkce má triviální derivaci.

Tvar ReLU funkce umožňuje daný neuron v případě záporného vstupu $\boldsymbol{w}^\top \boldsymbol{x}$ vypnout, aby nepředal žádnou informaci do dalších vrstev. To může být dvojsečná zbraň, jelikož během odhadování váhových parametrů může při špatné volbě iniciačních vah neuronu či při smolném kroku algoritmu založeném na metodě největšího spádu, který je velmi často používán při tréninku neuronových sítí, k trvalému vypnutí neuronu. Právě toto vedlo k hledání modifikace ReLU funkce, která by nabízela nenulovou derivaci na záporných hodnotách a zároveň by si ponechala

[3] *Rectified Linear Unit*

asymetrii původní ReLU funkce. Jednoduchým a poměrně účinným řešením je *protékající ReLU funkce*[4] definovaná předpisem:

$$LReLU(x) = \max\{0,x\} + \min\{0,\frac{x}{100}\},$$

jejíž derivace je zřejmě nenulová pro záporné hodnoty. Jelikož volba škálující konstanty 0.01 není podložena nějakou analýzou nabízí se možnost nahradit ji trénovatelným parametrem (machine learningový termín pro parametr odhadovaný z dat) a. Taková to aktivační funkce $PReLU = \max\{0,x\} + a\min\{0,x\}$ se nazývá parametrická ReLU funkce.

Velmi podobný předpis má i ELU funkce, která je definovaná jako:

$$ELU(x) = \begin{cases} x & x > 0 \\ \alpha(e^x - 1) & x \leq 0 \end{cases}$$

Parametr α může být opět odhadovaný z dat, ale častěji se volí jako nějaká předem zvolená vhodná kladná konstanta. V některých aplikacích (Clevert a kol. (2015) s vhodně zvoleným $\alpha > 0$) dosahují neuronové sítě s touto aktivační funkcí lepších predikčních výsledků než sítě s neurony aktivovanými LReLU nebo klasickou ReLU funkcí. Zároveň se ukazuje, že tyto sítě jsou i více eficientní, potřebují méně dat a iterací optimalizačních algoritmů při odhadování váhových parametrů k dosažení podobně přesných predikcí. Tato funkce je skoro všude diferencovatelná, pro záporné x je hodnota její derivace rovna $ELU(x) + \alpha$, tudíž je dokonce při volbě $\alpha = 1$ dosáhnout a dodefinováním derivace v nule hodnotou 1 spojité derivace na $\mathbb{R}$.

Další modifikací ReLU funkce je tzv. RRelu funkce[5], která také vychází z LReLU funkce. V tomto případě není parametr a odhadovaný z dat, je ale pro každý neuron v dané vrstvě různý, generovaný náhodně pomocí rovnoměrného rozdělení $U[l,u]$, kde l,$u \in (0,1]$ jsou předem zvolené hodnoty. V praxi (např.Xu a kol. (2015)) se ukazuje, že neuronové sítě s aktivační funkcí RReLU dosahují často přesnějších mimovzorkových predikcí než sítě využívající jiné modifikace ReLU funkce. Obecně se ale nedá říct, že by nějaká její modifikace byla lepší než jiná a vždy záleží na konkrétní aplikaci.

Zatím všechny představené používané aktivační funkce mají jednu společnou vlastnost, jedná se o monotónní funkce. Uvedeme si tedy i jednu v posledních letech hojně používanou aktivační funkci vyvinutou v Googlu s názvem *swish funkce,* která je definovaná předpisem:

$$swish(x) = \frac{x}{1 + e^{-\beta x}} = x\sigma(\beta x)$$

pro nějaký parametr nezáporný parametr β, tento parametr může být opět i trénovatelný. Podrobněji se podíváme na limitní chování této funkce v závislosti na parametru β. V případě $\beta = 0$ dostáváme funkci

$$\frac{x}{1 + e^0} = x/2,$$

[4] *Leaky ReLU function*
[5] *Randomized ReLU function*

což je lineární aktivační funkce s parametrem $a = 1/2$, naopak posláním bety do nekonečna dostáváme

$$\lim_{\beta\to\infty} x\sigma(\beta x) = \lim_{\beta\to\infty} \frac{x}{1+e^{-\beta x}} = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases},$$

což je předpis klasické ReLU funkce. Tato funkce nám tedy dává do spojitosti lineární a ReLU aktivační funkci. Odhadování parametru β je vlastně nelineární interpolace mezi těmito dvěma funkcemi. Zároveň vidíme, že v těchto dvou krajních případech jsme vždy dostali opět monotónní funkci, pro kladné β reálné se ale jedná o nemonotónní funkci, kdy pro nějaké dostatečně záporné x dostaneme zápornou hodnotu derivace $\sigma(\beta x)(1 + x\beta(1 - \sigma(\beta x))$, což implikuje klesající a zároveň na $\mathbb{R}$ nemonotónní funkci.

1.4 Model hlubokého učení

V této sekci budeme uvažovat, že máme n nezávislých stejně rozdělených pozorování $(Y_i,\boldsymbol{X}_i)$, kde $\boldsymbol{X}_i = (X_{i1},X_{i2},...,X_{iq_o})^\top$. Definujeme si model, kterým se budeme snažit modelovat $\mu_i = \mathsf{E}[Y_i \mid \boldsymbol{X}_i]$, a který bude využívat dopřednou neuronovou síť k automatizovanému hledání vhodné transformace $\boldsymbol{X}_i$, a tím bude schopný modelovat i velmi složité nelineární závislosti.

Definice 5 (model hlubokého učení)**.** *Řekneme, že náhodné vektory* $(Y_i,\boldsymbol{X}_i), i = 1,..,n$, *splňují předpoklady modelu hlubokého učení*[6]*, jestliže:*

1. *Jednotlivé* $Y_1,...,Y_n$ *jsou nezávislé a jejich podmíněné rozdělení za podmínky* $\boldsymbol{X}_i$ *je exponenciálního typu ve tvaru*

$$f(y,\theta_i,\varphi) = \exp\left\{\frac{y\theta_i - b(\theta_i)}{\varphi} + c(y,\varphi)\right\},$$

   kde b *je známá dvakrát spojitě diferencovatelná funkce,* θ_i *závisí na* $\boldsymbol{X}_i$ *a parametrech* $\boldsymbol{\beta} = (\beta_0,\beta_1,...,\beta_{q_d})^\top$ *a váhových parametrech dopředné neuronové sítě* $\boldsymbol{z}^{(d:1)}$ *o hloubce* d *a* $\varphi > 0$ *je známá či neznámá konstanta.*

2. *Existuje známá, striktně monotónní, dvakrát spojitě diferencovatelná linková funkce* g *taková, že platí*

$$g(\mathsf{E}[Y_i \mid \boldsymbol{X}_i]) = \boldsymbol{\beta}^\top \boldsymbol{z}^{(d:1)}(\boldsymbol{X}_i). \tag{1.2}$$

Na první pohled je zřejmá podobnost s definicí zobecněného lineárního modelu, každý neuron v neuronové síti $\boldsymbol{z}^{(d:1)}$ s aktivační funkcí ϕ je vlastně zobecněný lineární model s linkovou funkcí ϕ^{-1} a regresními parametry $\boldsymbol{w}$ pouze bez předpokladu striktní monotónnosti a spojité druhé derivace funkce ϕ. Vysvětlující proměnné pro každý neuron jsou hodnoty výstupů neuronů z předešlé vrstvy. Výstup neuronové sítě $\boldsymbol{z}^{(d:1)}(\boldsymbol{X}_i) \in 1 \times \mathbb{R}^{q_d}$ je vysvětlující proměnná v klasickém zobecněném lineárním modelu s linkovou funkcí g a regresními parametry $\boldsymbol{\beta} = (\beta_0,\beta_1,...,\beta_{q_d})^\top$. Této poslední vrstvě modelu hlubokého

[6] angl. deep learning model

učení $g^{-1}(\boldsymbol{\beta}^\top \boldsymbol{z}^{(d:1)}(\boldsymbol{X}_i))$ se také často v literatuře říká výstupní vrstva. Pokud jde o volbu vhodné linkové funkce g, tak ta se u v případě tohoto modelu volí v závislosti na podmíněném rozdělení s hustotou f.

Aby byl měla formule (1.2) smysl musí platit $\boldsymbol{X}_i \in 1 \times \mathbb{R}^{q_0}$, tudíž musíme vždy uvažovat model s absolutním členem. V případě modelů hlubokého učení se vektoru $(1,\boldsymbol{X}_i)^\top$ často říká vstupní vrstva. V následujícím textu budeme vždy pro jednoduchost symbolem $\boldsymbol{X}_i$ myslet již právě tuto vstupní vrstvu, pokud nebude uvedeno jinak.

Model hlubokého učení je vlastně jakýmsi složením většího počtu zobecněných lineárních modelů. Jedná se o parametrický model, kdy všechny parametry modelu $(\boldsymbol{w}_1^{(1)},\boldsymbol{w}_2^{(1)},...,\boldsymbol{w}_{q_d}^{(d)},\boldsymbol{\beta})^\top$, kde $\boldsymbol{w}_i^{(j)}, 1 \leq i \leq q_j, j \in \{1,...,d\}$, značí váhy $i-$tého neuronu v $j-$té vrstvě, jsou odhadovány najednou.

Dokonce existují dva případy, kdy model hlubokého učení splývá s klasickým zobecněným lineárním modelem. Prvním, triviálním způsobem je volba hloubky sítě $d = 0$, kdy je podmíněná střední hodnota odezvy Y_i za podmínky $\boldsymbol{X}_i$ rovna

$$\mathsf{E}\left[Y_i \mid \boldsymbol{X}_i\right] = g^{-1}(\boldsymbol{\beta}^\top \boldsymbol{X}_i),$$

což je její vyjádření pomocí lineárního prediktoru z definice zobecněného lineárního modelu. Druhou možností je volba lineárních funkcí $\phi_i(x) = a_i x, i = 1,...,d$ jako aktivačních funkci všech vrstev dopředné neuronové sítě. V takovém případě se dá podmíněná střední hodnota odezvy v závislosti na vysvětlujících proměnných $\boldsymbol{X}_i$ zapsat jako

$$\mathsf{E}\left[Y_i \mid \boldsymbol{X}_i\right] = g^{-1}(\boldsymbol{\beta}^\top \boldsymbol{z}^{(d:1)}(\boldsymbol{X}_i)) = g^{-1}(a_d \boldsymbol{\beta}^\top \mathbb{W}^d \boldsymbol{z}^{(d-1:1)}(\boldsymbol{X}_i)),$$

kde matice $\mathbb{W}^d \in \mathbb{R}^{(q_d+1)\times(q_{d-1}+1)}$ je matice ve tvaru:

$$\mathbb{W}^d = \begin{pmatrix} 1/a_d & 0 & ... & 0 \\ w_{1,0}^d & w_{1,1}^d & ... & w_{1,q_{d-1}}^d \\ \vdots & & \ddots & \vdots \\ w_{q_d,0}^d & w_{q_d,1}^d & ... & w_{q_d,q_{d-1}}^d \end{pmatrix},$$

která tedy má v druhém až posledním řádku váhy jednotlivých neuronů z $d-$té vrstvy. Obdobně se dají rozepsat i $\boldsymbol{z}^{(i:1)}(\boldsymbol{X}_i)), i = 1,2,...,d$. Pak dostáváme vyjádření (1.2) ve tvaru

$$\mathsf{E}\left[Y_i \mid \boldsymbol{X}_i\right] = g^{-1}((\prod_{i=1}^{d} a_i)\boldsymbol{\beta}^\top \mathbb{W}^d \mathbb{W}^{d-1} ... \mathbb{W}^1 \boldsymbol{X}_i),$$

což již odpovídá definici zobecněného lineárního modelu s linkovou funkcí g a parametry $(\prod_{i=1}^{d} a_i)\boldsymbol{\beta}^\top \mathbb{W}^d \mathbb{W}^{d-1} ... \mathbb{W}^1$, což je $(q_0+1)-$složkový horizontální vektor. Nyní je již jasné, proč využití neuronové sítě pouze s lineárními aktivačními funkcemi nedává smysl. Model hlubokého učení s takovou dopřednou neuronovou sítí by nám neumožnil modelovat jiné než lineární závislosti podmíněné střední hodnoty, a to za daleko vyššího počtu parametrů a výrazně delší doby odhadování parametrů (vzhledem k tomu, jakým způsobem jsou parametry v modelu hlubokého učení odhadovány) než v případě běžného zobecněného lineárního modelu, který by modeloval podmíněnou střední hodnotu ze vztahu (1.1). Zároveň použití různých nelineárních aktivačních funkci nám zároveň umožní modelovat složitější

závislosti podmíněné střední hodnoty odezvy Y_i na vysvětlujících proměnných X_i.

Další motivací k zadefinování modelu hlubokého učení pro nás byla jeho schopnost automatického hledání vhodné transformace vysvětlujících proměnných. K tomu dochází právě odhadem váhových parametrů jednotlivých neuronů neuronové sítě. Protože zobrazením $\phi(\boldsymbol{w}^\top \boldsymbol{z})$ dojde k projekci vícerozměrného vstupu $\boldsymbol{z}$ na reálné číslo, dojde zároveň ke ztrátě informace. To je důvod, proč zpravidla v každé vrstvě neuronové sítě bývá neuronů několik. Každý neuron totiž přenáší do dalších vrstev neuronové sítě jinou část informace obsažené v $\boldsymbol{z}$.

Již pouhý zápis všech parametrů modelu dává tušit, že počet parametrů je i pro menší počet vysvětlujících veličin, vrstev dopředné neuronové sítě a počty neuronů v nich velmi vysoký. Konkrétně je počet parametrů roven

$$\sum_{i=1}^{d} q_i(q_{i-1}+1) + (q_d+1).$$

Druhý sčítanec zřejmě vyjadřuje dimenzi parametru $\boldsymbol{\beta}$, hodnota sumy je pak počet váhových parametrů dopředné neuronové sítě. Vidíme, že rychlost růstu odhadovaných parametrů je velmi rychlá.

2. Věty o univerzalitě

2.1 Zavedení potřebných definic

Motivací k použití neuronových sítí byla jejich velká flexibilita a možnost modelovat i velmi komplikované nelineární závislosti. Vyvstává otázka, zda-li je opravdu možné modelovat libovolný typ závislosti podmíněné střední hodnoty na vysvětlujících proměnných. Jinými slovy, zda-li je možné pomocí dostatečně velké, vhodné neuronové sítě aproximovat libovolnou funkci. Takové vlastnosti se říká univerzalita, proto věty dokazující tuto vlastnost neuronových sítí jsou známé jako věty o univerzalitě. Těchto vět existuje větší množství s různými předpoklady na aproximované funkce či neuronové sítě používané k aproximaci, my v této práci představíme dvě věty o univerzalitě pocházející z práce Hornik a kol. (1989).

V této kapitole dokážeme tuto univerzální vlastnost pouze pro mělké neuronové sítě, což je ale dostačující, protože hluboké neuronové sítě mají ještě lepší aproximační schopnosti, jak ukázal například Elbrächter (2019) ve svém článku. V tomto článku porovnával aproximační schopnosti neuronových sítí, konkrétně s ReLU aktivačními funkcemi, s počtem vrstev $d = 1$ a vyšším. Dokázal, že hluboké neuronové sítě dosahují exponenciálním aproximačním schopnostem, ale mělké pouze polynomiálních aproximací. Zde exponenciální a polynomiální aproximační schopnost chápeme, tak jak se s rostoucím počtem neuronů či neuronů v jednotlivých vrstvách zvyšuje kvalita aproximace.

K dokázání zmíněných vět o univerzalitě budeme potřebovat hned několik definic, které si nyní zavedeme. Začneme nejprve zadefinováním několika množin funkcí.

Definice 6 (množina afiních funkcí)**.** *Nechť $q_0 \in \mathbb{N}$, pak jako množinu afinních funkcí o vstupní dimenzi q_0 chápeme množinu*

$$\mathcal{A}^{q_0} = \Big\{ A : 1 \times \mathbb{R}^{q_0} \to \mathbb{R}; A(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} \Big\}.$$

Jedná se tedy o množinu neuronů s lineární aktivační funkcí. Dále si obdobně zadefinujeme množinu všech mělkých dopředných neuronových sítí s měřitelnou aktivační funkcí první vrstvy ϕ a lineární aktivační funkcí výstupní vrstvy této neuronové sítě.

Definice 7 (množina mělkých lineárních neuronových sítí)**.** *Nechť opět $q_0 \in \mathbb{N}$ a ϕ je měřitelná funkce, pak jako množinu mělkých lineárních neuronových sítí s aktivační funkcí ϕ o vstupní dimenzi q_0 chápeme množinu*

$$\Sigma^{q_0}(\phi) = \Big\{ f : 1 \times \mathbb{R}^{q_0} \to \mathbb{R}; f(\boldsymbol{x}) = \sum_{j=0}^{q_1} \beta_j \phi(A_j(\boldsymbol{x})), A_j \in \mathcal{A}^{q_0}, \beta_j \in \mathbb{R}; \forall j; q_1 \in \mathbb{N} \Big\}.$$

Budeme potřebovat ještě jednu množinu neuronových sítí specifického tvaru, pro niž je důkaz věty o univerzalitě jednodušší.

Definice 8. *Nechť $q_0 \in \mathbb{N}$ a ϕ je měřitelná funkce, pak jako množinu mělkých neuronových sítí s aktivační funkcí ϕ o vstupní dimenzi q_0 a výstupní aktivační*

funkci v lineárně-multiplikativním tvaru chápeme množinu

$$\Sigma\Pi^{q_0}(\phi) = \left\{ f : 1 \times \mathbb{R}^{q_0} \to \mathbb{R}; f(\boldsymbol{x}) = \sum_{j=0}^{q_1} \beta_j \prod_{k=1}^{l_j} \phi(A_{j,k}(\boldsymbol{x})); A_{j,k} \in \mathcal{A}^{q_0}; q_1 \in \mathbb{N} \right\},$$

kde $\forall j : \beta_j \in \mathbb{R}, l_j \in \mathbb{N}$.

Položením $l_j = 1$,$j = 1$,2,...,q_0, dostáváme předpis množiny mělkých lineárních neuronových sítí, což implikuje $\Sigma^{q_0}(\phi) \subset \Sigma\Pi^{q_0}(\phi)$ pro každou měřitelnou funkci ϕ.

Potřebovat budeme ještě jednu definici, a to definici stejnoměrně husté podmnožiny množiny všech spojitých funkcí na prostoru $1 \times \mathbb{R}^{q_0}$, kterou budeme značit jako $\mathcal{C}(\mathbb{R}^{q_0})$. Dále pro nás bude důležitá ještě množina všech měřitelných funkcí na $1 \times \mathbb{R}^{q_0}$, značenou jako $\mathcal{M}(\mathbb{R}^{q_0})$.

Definice 9 (stejnoměrně hustá množina)**.** *Řekneme, že množina $S \subset \mathcal{M}(\mathbb{R}^{q_0})$ je stejnoměrně hustá na kompaktu v $\mathcal{C}(\mathbb{R}^{q_0})$, pokud pro každou kompaktní podmnožinu $K \subset 1 \times \mathbb{R}^{q_0}$ je množina S ρ_K-hustá v $\mathcal{C}(\mathbb{R}^{q_0})$, což znamená, že pro každé $\epsilon > 0$ a pro všechny $g \in \mathcal{C}(\mathbb{R}^{q_0})$ existuje $f \in S$, takové že*

$$\rho_K(g,f) = \sup_{x \in K} |g(\boldsymbol{x}) - f(\boldsymbol{x})| < \epsilon.$$

Věty o univerzalitě neuronových sítí nám vlastně tedy říkají, zda-li je množina příslušných sítí splňující dané předpoklady stejnoměrně hustá. V tomto textu se budeme zabývat univerzalitou množin $\Sigma^{q_0}(\phi)$ a $\Sigma\Pi^{q_0}(\phi)$.

2.2 Věty o univerzalitě

Nyní nám již nic nebrání formulovat větu o univerzalitě pro množinu $\Sigma\Pi^{q_0}(\phi)$.

Věta 1 (věta o univerzalitě pro $\Sigma\Pi^{q_0}$)**.** *Nechť ϕ je nekonstantní a spojitá aktivační funkce. Pak $\Sigma\Pi^{q_0}(\phi)$ je stejnoměrně hustá na kompaktu v $\mathcal{C}(\mathbb{R}^{q_0})$.*

Důkaz. Důkaz je uveden v Würthrich a Merz (2022).

□

Bylo tedy dokázáno, že každou spojitou funkci na $1 \times \mathbb{R}^{q_0}$ lze aproximovat mělkou lineárně-multiplikativní neuronovou sítí s aktivační funkcí, která je nekonstantní a spojitá. Naše požadavky na aktivační funkci jsou velmi mírné, ze všech běžně používaných aktivačních funkcích představených v první kapitole této práce tyto předpoklady nesplňuje pouze ta úplně nejjednodušší kroková funkce, která stejně není v praxi moc využívána.

Důkaz je založen na Stone-Weierstrassově větě, kdy se využilo, že $\Sigma\Pi^{q_0}(\phi)$ je pro libovolnou spojitou aktivační funkci ϕ algebra, tedy množina uzavřená na sčítání, násobení a násobení skalárem. Tato vlastnost ale nebyla v důkazu nikterak vysvětlena a nepovažujeme ji za triviální, proto platnost této vlastnosti nyní ukážeme.

Začneme s uzavřeností na sčítání. Mějme f,$g \in \Sigma\Pi^{q_0}(\phi)$, pak jejich součet $f + g$ lze zapsat ve tvaru:

$$(f+g)(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x}) = \sum_{j=0}^{q_1^f} \beta_j^f \prod_{k=1}^{l_j^f} \phi(A_{j,k}^f(\boldsymbol{x})) + \sum_{j=0}^{q_1^g} \beta_j^g \prod_{k=1}^{l_j^g} \phi(A_{j,k}^g(\boldsymbol{x})).$$

Protože součet dvou sum lze zapsat také jako suma, lze vyjádřit funkci $f + g$ ve tvaru z definice $\Sigma\Pi^{q_0}(\phi)$, takže $f + g$ leží také v $\Sigma\Pi^{q_0}(\phi)$. Tím je dokázána uzavřenost na sčítání.

Uzavřenost na násobení skalárem se ukáže obdobně, nechť $f \in \Sigma\Pi^{q_0}(\phi)$, pro nějakou spojitou aktivační funkci ϕ a nechť $\alpha \in \mathbb{R}$, pak platí

$$\alpha f(\boldsymbol{x}) = \alpha \sum_{j=0}^{q_1} \beta_j \prod_{k=1}^{l_j} \phi(A_{j,k}(\boldsymbol{x})),$$

což se dá přeznačením $\beta_j^* = \alpha\beta_j$ dostáváme:

$$\alpha f(\boldsymbol{x}) = \sum_{j=0}^{q_1} \beta_j^* \prod_{k=1}^{l_j} \phi(A_{j,k}(\boldsymbol{x})),$$

takže máme vyjádření αf ve tvaru z definice množiny $\Sigma\Pi^{q_0}(\phi)$, a máme tak dokázanou její uzavřenost na násobení skalárem.

Zbývá nám již dokázat jen uzavřenost na násobení. Nechť f,$g \in \Sigma\Pi^{q_0}(\phi)$, pro spojitou funkci ϕ. Jejich součin $h = fg$ pak lze zapsat ve tvaru:

$$h(\boldsymbol{x}) = fg(\boldsymbol{x}) = \left(\sum_{j=0}^{q_1^f} \beta_j^f \prod_{k=1}^{l_j^f} \phi(A_{j,k}^f(\boldsymbol{x}))\right)\left(\sum_{j=0}^{q_1^g} \beta_j^g \prod_{k=1}^{l_j^g} \phi(A_{j,k}^g(\boldsymbol{x}))\right),$$

pro vhodné $q_1^f, q_1^g \in \mathbb{N}$, to se dá dále snadno roznásobit na

$$\sum_{i=0}^{q_1^f}\sum_{j=0}^{q_1^g} \beta_i^f \beta_j^g \left(\prod_{k=1}^{l_j^f} \phi(A_{j,k}^f(\boldsymbol{x}))\right)\left(\prod_{k=1}^{l_j^g} \phi(A_{j,k}^g(\boldsymbol{x}))\right).$$

Když si nyní zadefinujeme $q_1^f q_1^g$ reálných konstant β_k^h, které vznikly jako součiny jednotlivých dvojic $\beta_i^f \beta_j^g$, pro každé i a j a označíme si jako $\prod_{m=1}^{l_k^h} \phi(A_{k,m}^h(\boldsymbol{x}))$ součiny dvou produktů z roznásobeného tvaru součinu $h = fg$, tak můžeme součin fg vyjádřit ve tvaru:

$$h(\boldsymbol{x}) = fg(\boldsymbol{x}) = \sum_{j=0}^{q_1^h} \beta_j^h \prod_{k=1}^{l_j^h} \phi(A_{j,k}^h(\boldsymbol{x})),$$

což již implikuje $h \in \Sigma\Pi^{q_0}(\phi)$. Dokázali jsme tedy i uzavřenost na násobení.

Množina $\Sigma\Pi^{q_0}(\phi)$ je opravdu algebrou a předpoklady Stone-Weierstrassovy věty jsou tedy splněny.

Nyní si formulujeme a dokážeme ještě jednu větu o univerzalitě, tentokrát pro množinu $\Sigma^{q_0}(\phi)$, která již nebude obsahovat předpoklad na spojitost aktivační funkce ϕ. I v tomto případě ale budeme mít na funkci ϕ jeden předpoklad (kromě měřitelnosti) a to, aby byla neklesající a splňovala $\lim_{x\to-\infty} \phi(x) = 0$ a $\lim_{x\to\infty} \phi(x) = 1$.

Věta 2 (věta o univerzalitě pro Σ^{q_0})**.** *Nechť ϕ je měřitelná neklesající aktivační funkce, splňující:*

$$\lim_{x\to-\infty} \phi(x) = 0 \wedge \lim_{x\to\infty} \phi(x) = 1.$$

Pak $\Sigma^{q_0}(\phi)$ je stejnoměrně hustá na kompaktu v $\mathcal{C}(\mathbb{R}^{q_0})$.

Důkaz. Důkaz bude veden v několika krocích, kdy si nejprve dokážeme 3 tvrzení, která následně využijeme k důkazu stejnoměrné hustoty množiny $\Sigma^{q_0}(\phi)$.

Nejprve si uvědomíme, že díky jednomu ze základních goniometrických vzorců platí $\cos(a)\cos(b) = \cos(a+b) - \cos(a-b)$. Výrazy $a+b$ a $a-b$ se dají považovat za afinní funkce parametru a. Z toho plyne, že libovolný součin $\prod_{k=1}^{n} \cos(A_k(\boldsymbol{x}))$, kde $\forall k : A_k(\boldsymbol{x}) \in \mathcal{A}^{q_0}$, lze pro vhodné $T \in \mathbb{N}$, $A_t \in \mathcal{A}^{q_0}$ a $\alpha_t \in \mathbb{R}$ přepsat do tvaru

$$\sum_{t=1}^{T} \alpha_t \cos(A_t(\boldsymbol{x})).$$

Mějme $f \in \Sigma\Pi^{q_0}(\cos)$ ve tvaru

$$f(\boldsymbol{x}) = \sum_{j=0}^{q_1} \beta_j \prod_{k=1}^{l_j} \cos(A_{j,k}(\boldsymbol{x})),$$

pak lze f díky výše popsané trigonometrické identitě přepsat jako:

$$f(\boldsymbol{x}) = \sum_{j=0}^{q_1} \beta_j \sum_{t=1}^{T_j} \alpha_t \cos(A_{j,t}(\boldsymbol{x})),$$

to se dá dále přepsat přeindexováním do jedné sumy, tak že celkem bude mít suma $N = (q_1+1)\sum_j^{q_1} T_j$ sčítanců. Tyto sčítance budou řazeny tak, že prvních T_0 sčítanců bude ve tvaru $\beta_0\alpha_i \cos(A_{0,i}(\boldsymbol{x}))$,$i = 1$,...,$T_0$. Další sčítance budou řazeny analogicky. Funkce f tedy bude tvaru

$$f(\boldsymbol{x}) = \sum_{j=1}^{N} \beta_j^* \cos(A_j^*(\boldsymbol{x})),$$

kde β_i^* jsou reálné konstanty (součiny původních bet a alf) a A_j leží v $\in \mathcal{A}^{q_0}$, což znamená, že $f \in \Sigma^{q_0}(\cos)$. Protože f byla brána libovolně dokázali jsme právě, že $\Sigma\Pi^{q_0}(\cos) \subset \Sigma^{q_0}(\cos)$. Protože opačná inkluze platí vždy, tak jsme právě dokázali i rovnost těchto dvou množin. To znamená, že dle věty 1 je $\Sigma^{q_0}(\cos)$ stejnoměrně hustá na kompaktu v $\mathcal{C}(\mathbb{R}^{q_0})$.

V dalším kroku dokážeme, že pro každou spojitou neklesající funkci ψ, která zároveň splňuje $\lim_{x\to-\infty} \psi(x) = 0$ a $\lim_{x\to\infty} \psi(x) = 1$, a pro každé $\epsilon > 0$ existuje $G_\epsilon(x) \in \Sigma^1(\phi)$, takové že

$$\sup_{x\in\mathbb{R}} |\psi(x) - G_\epsilon(x)| < \epsilon.$$

Toto se dokáže využitím neklesajícnosti funkcí ψ a ϕ a předpokladu na jejich chování v plus a mínus nekonečnu. Bez újmy na obecnosti nechť $\epsilon < 1$, zvolme $Q \in \mathbb{N}$ takové že $1/Q < \epsilon/2$. Dále nalezněme $R > 0$ takové, že $\phi(-R) < \epsilon/(2Q)$ a $\phi(R) > 1-\epsilon/(2Q)$, takové R lze najít právě díky předpokladu na limitní chování ϕ.

Nyní si pro $j = 1$,..,$Q-1$ zadefinujeme $r_j = \sup\{\lambda : \psi(\lambda) = j/Q\}$ a konečně $r_Q = \sup\{\lambda : \psi(\lambda) = 1 - 1/2Q\}$. Tyto suprema existují, protože ψ je spojitá neklesající funkce a zároveň splňuje požadavek na její limitní chování.

Nyní nám zbývá pouze dodefinovat afinní zobrazení $A_{a.b} \in \mathcal{A}^1$ pomocí předpisu, že $A_{a.b}(a) = -R$ a $A_{a.b}(b) = R$. Taková zobrazení určitě existují, jelikož každé afinní zobrazení se dá jednoznačně definovat pomocí hodnot ve dvou bodech.

Dále si zadefinujeme funkci $G_\epsilon(\lambda) = \sum_{j-1}^{Q-1} \beta_j \phi(A_{r_j,r_{j+1}}(\lambda))$, kde si pro každé j položíme $\beta_j = Q$. Tato funkce je zřejmě z $\Sigma^1(\phi)$ a zároveň splňuje požadovanou vlastnost $|\psi(\lambda) - G_\epsilon(\lambda)| < \epsilon$ pro každé λ reálné.

Že funkce G_ϵ tuto vlastnost opravdu splňuje, si nyní ukážeme. Protože pro $\lambda < r_1$ platí $G_\epsilon(\lambda) \in [0,\epsilon/2Q)$ a $G_\epsilon(\lambda) \in [0,1/Q]$, což implikuje, že jejich rozdíl v absolutní hodnotě leží mezi nulou a maximem z $\epsilon/2Q$ a $1/Q$ a to je vždy menší než ϵ díky volbě Q a předpokladu, že epsilon je menší než 1. Pro situace, kdy λ leží mezi r_i a r_{i+1} pro nějaké i, a kdy $\lambda > r_Q$ se tato vlastnost dokáže analogicky pouhým rozepsáním a odečtením krajních bodů intervalů od sebe.

Poslední vlastnost, kterou si musíme dokázat, než dáme vše dohromady a dokážeme větu o univerzalitě je, že pro každé $\epsilon > 0, M > 0$ existuje funkce $\cos_{M,\epsilon} \in \Sigma^1(\phi)$, která splňuje:

$$\sup_{x\in[-M,M]} |\cos(x) - \cos_{M,\epsilon}(x)| < \epsilon.$$

Taková funkce skutečně existují. K ukázání platnosti tohoto tvrzení si budeme muset zadefinovat pomocnou funkci

$$\chi(x) = \frac{1}{2}\left(1 + \cos(x + \frac{3\pi}{2})\right) 1_{\{-\pi/2\leq x\leq \pi/2\}} + 1_{\{x>\pi/2\}}.$$

Tato funkce je zřejmě spojitá neklesající a splňuje limitní předpoklady minulého dokazovaného tvrzení. Musíme si uvědomit, že pomocí součtu konečného množství afinně posunutých funkcí χ lze přesně replikovat kosínus na intervalu $[-M,M]$. Zároveň každou z těchto funkcí χ lze aproximovat pomocí nějaké funkce z $\Sigma^1(\phi)$, díky předcházejícímu tvrzení.

Nyní jsme již schopni dokázat samotnou větu o univerzalitě. Mějme tedy nějakou funkci $g \in \mathcal{C}(\mathbb{R}^{q_0})$ a kompaktní podmnožinu $K \subset 1 \times \mathbb{R}^{q_0}$. Nalezneme teď funkci $\sum_{t=1}^{T} \alpha_t \cos(A_t(\boldsymbol{x}))$, s vhodnými α_t a A_t, takovými na množině K aproximovali funkci g ve významu ρ_k metriky s přesností $\epsilon/2$. Takové alfy a afinní zobrazení A_t lze najít, protože $\Sigma^{q_0}(\cos)$ je stejnoměrně hustá na kompaktu v $\mathcal{C}(\mathbb{R}^{q_0})$, jak bylo výše dokázáno. Nyní nechť $M > 0$ je takové reálné číslo, které splňuje, že $A_t(K) \subset [-M,M]$, pro všechna $t = 1,...,T$. Takové M existuje, protože K je kompaktní, A_t je pro každé t afinní funkce, tedy funkce spojitá.

Nyní si označíme jako $\check{\mathrm{T}} = T\sum_{t=1}^{T} |\alpha_t|$. Nakonec pro každé $t = 1,..,T$ nalezneme funkci $\cos^t_{M,\epsilon/(2T\check{\mathrm{T}})} \in \Sigma^1(\phi)$, takovou že

$$\sup_{\boldsymbol{x}\in[-M,M]} |\cos(A_t(\boldsymbol{x})) - \cos^t_{M,\epsilon/(2T\check{\mathrm{T}})}(A_t(\boldsymbol{x}))| < \epsilon/(2T\check{\mathrm{T}}).$$

Nyní dáme vše dohromady.

$$\sup_{\boldsymbol{x}\in K} |g(\boldsymbol{x}) - \sum_{t=1}^{t} \alpha_t \cos^t_{M,\epsilon/(2T\check{\mathrm{T}})}(A_t(\boldsymbol{x}))|.$$

To lze ekvivalentně upravit na

$$\sup_{\boldsymbol{x}\in K} |g(\boldsymbol{x}) - \sum_{t=1}^{t} \alpha_t \cos(A_t(\boldsymbol{x})) + \sum_{t=1}^{t} \alpha_t \cos(A_t(\boldsymbol{x})) - \sum_{t=1}^{t} \alpha_t \cos^t_{M,\epsilon/(2T\check{\mathrm{T}})}(A_t(\boldsymbol{x}))|,$$

což je díky trojúhelníkové nerovnosti menší nebo rovno než

$$\sup_{\boldsymbol{x}\in K}|g(\boldsymbol{x})-\sum_{t=1}^{t}\alpha_t\cos(A_t(\boldsymbol{x}))|+\sup_{\boldsymbol{x}\in K}|\sum_{t=1}^{t}\alpha_t\cos(A_t(\boldsymbol{x}))-\sum_{t=1}^{t}\alpha_t\cos^t_{M,\epsilon/(2T\check{\Upsilon})}(A_t(\boldsymbol{x}))|,$$

kde první sčítanec je menší než $\epsilon/2$. Druhý sčítanec lze opět pomocí trojúhelníkové nerovnosti shora odhadnou jako

$$\sup_{\boldsymbol{x}\in K}\sum_{t=1}^{t}|\alpha_t||\cos(A_t(\boldsymbol{x}))-\cos^t_{M,\epsilon/(2T\check{\Upsilon})}(A_t(\boldsymbol{x}))|,$$

a dále jako

$$\sup_{\boldsymbol{x}\in K}\sum_{t=1}^{t}|\alpha_t|\sum_{t=1}^{t}|\cos(A_t(\boldsymbol{x}))-\cos^t_{M,\epsilon/(2T\check{\Upsilon})}(A_t(\boldsymbol{x}))|,$$

což je díky definici $\cos^t_{M,\epsilon/(2T\check{\Upsilon})}(A_t(\boldsymbol{x}))$ a díky tomu, že $K\subset[-M,M]$, menší rovno než $\check{\Upsilon}T\epsilon/(2T\check{\Upsilon})=\epsilon/2$. Jelikož $\sum_{t=1}^{t}\alpha_t\cos^t_{M,\epsilon/(2T\check{\Upsilon})}\in\Sigma^{q_0}(\phi)$ tak jsme právě dokázali stejnoměrnou hustotu $\Sigma^{q_0}(\phi)$ na kompaktu v $\mathcal{C}(\mathbb{R}^{q_0})$.

□

Důkazem této věty končí tato kapitola. Jak si čtenář může všimnout, věty o univerzalitě nepojednávají o žádné náhodnosti, ta se nám objeví v další kapitole, kdy se dozvíme, jestli lze libovolnou funkci nejen aproximovat, ale i modelovat.

3. Asymptotické vlastnosti

3.1 Konzistence neuronových sítí

V minulé kapitole jsme si formulovali a dokázali dvě věty o univerzalitě. Tyto věty nám říkají, že pomocí dostatečně velké neuronové sítě lze aproximovat libovolnou spojitou funkci s předem danou požadovanou přesností. Předpoklady ani tvrzení těchto vět ale neobsahovaly žádnou náhodnost, jednalo se pouze o věty popisující algebraické vlastnosti neuronových sítí splňující určité předpoklady. Nás by ale samozřejmě zajímalo, zda-li jsme schopni tuto aproximaci nalézt v podobě odhadnutí parametrů modelu hlubokého učení z dostupných dat. Věty, které se touto otázkou zaobírají se nazývají věty o konzistenci.

Předpokládejme, že máme data v podobě náhodného výběru $(Y_i, \boldsymbol{X_i})$, kde $\boldsymbol{X_i}$ skoro jistě nabývají hodnot z nějaké konvexní podmnožiny $1 \times \mathbb{R}^{q_0}$, pro každé $i = 1,2,...n$, a zajímá nás, zda s rostoucím n neuronová síť s odhadnutými parametry skutečně bude aproximovat skutečnou závislost odezvy Y na vysvětlujících proměnných s předem zvolenou přesností. Tímto se od algebraicko-analýznických vět o univerzalitě přesouváme ke statistickým větám o konzistenci.

Stejně jako v případě vět o univerzalitě existuje i vět o konzistenci velké množství, jednotlivé věty se od sebe liší různými předpoklady na strukturu neuronové sítě, požadavky na aktivační funkce jednotlivých vrstev, či předpoklady na náhodnou chybu modelu. Ano, opravdu se budeme muset v této kapitole opřít o náhodnou chybu našeho modelu, notoricky značenou jako ϵ.

Nechť náš náhodný výběr $(Y_i, \boldsymbol{X}_i)$ splňuje

$$Y_i = \mu(\boldsymbol{X}_i) + \epsilon_i,$$

kde $\mu(\boldsymbol{x})$ značí spojitou regresní funkci definovanou na kompaktní podmnožině prostoru $1 \times \mathbb{R}^{q_0}$, která je definována předpisem

$$\mu(\boldsymbol{x}) = \mathsf{E}\left[Y \mid \boldsymbol{X} = \boldsymbol{x}\right] \tag{3.1}$$

a náhodné chyby ϵ_i, které pro každé i splňují:

$$\mathsf{E}\left[\epsilon_i \mid \boldsymbol{X}_i\right] = 0,$$

$$\mathsf{E}\left[|\epsilon_i|^{2+\delta} \mid \boldsymbol{X}_i\right] < \infty,$$

pro nějaké $\delta > 0$, zároveň ještě požadujeme, aby ϵ_i byly nezávislé na vysvětlujících proměnných $\boldsymbol{X}_i$.

Můžeme si všimnout, že předpoklad na druhý moment našich epsilonů není nic jiného než známá Ljapunovova podmínka, které se vyskytuje i v tzv. Ljapunovově centrální limitní větě.

Zároveň se v této sekci také omezíme pouze na případ, kdy budeme minimalizovat střední čtvercovou náhodných epsilonů, budeme tedy řešit úlohu hledání

$$\tilde{\mu}_n = \arg\min_{\mu \in \mathcal{C}(\mathcal{X})} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu(\boldsymbol{X}_i))^2,$$

kde $\mathcal{C}(\mathcal{X})$ značí množinu všech spojitých funkcí definovaných na kompaktní množině $\mathcal{X} \subseteq 1 \times \mathbb{R}^{q_0}$. Toto omezení se nemusí zdát na první pohled jako nikterak

omezující, protože střední čtvercová chyba je obvyklá ztrátová funkce v regresních úlohách či při konstruování predikcí náhodných procesů, ale vzhledem k tomu, že mezi typické úlohy strojového učení se řadí klasifikační úlohy, kde obvykle dáváme přednost devianci binomického či multinomického rozdělení (v ML komunitě často nazývané jako entropie) se přeci jen o výrazné omezení jedná. Teorie okolo hlubokého učení teprve vzniká v posledních dekádách, nebyly zatím tedy dokázány výrazně obecnější věty o konzistenci. V této oblasti je určitě velký prostor k dalšímu budoucímu výzkumu.

Na první pohled na definici $\tilde{\mu}_n$ vidíme, že hledání minima přes všechny spojité funkce na daném definičním oboru je netriviální, nelineární a hlavně neparametrická optimalizační úloha, jejíž řešení by bylo velmi obtížné. A to úplně pomíjíme fakt, že numerických hledáním řešení této úlohy bychom ve finále dostali takovou funkci, která by měla velmi špatné extrapolační vlastnosti, pro datovou sadu o libovolném počtu pozorování n, což je vlastnost, které se chceme vyvarovat, pokud máme v plánu konstruovat mimovzorkové předpovědi.

Právě toto jsou důvody, proč se omezíme pouze na určitou podmnožinu $\mathcal{C}(\mathcal{X})$. Protože v minulé kapitole jsme si dokázali větu o univerzalitě pro množinu mělkých lineárních neuronových sítí, tak se omezíme právě na podmnožinu množiny $\Sigma^{q_0}(\phi)$. Budeme muset ještě omezit šíři takovéto neuronové sítě, tedy omezit parametr q_1 z definice 7, protože s neomezenou šířkou neuronové sítě bychom měli stále neomezený počet parametrů této sítě, a taková síť by opět trpěla přeučením na (libovolně velké) konečné datové sadě. Zadefinujeme si tedy novou podmnožinu mělkých lineárních neuronových sítí $\mathcal{S}$ ve tvaru:

$$\mathcal{S}(d,\Delta,\tilde{\Delta},\phi) = \left\{ f \in \Sigma^{q_0}(\phi); q_1 = d, \sum_{j=0}^{q_1} |\beta_j| \leq \Delta, \max_{1\leq j\leq q_1} \sum_{l=0}^{q_0} |a_{j,l}| \leq \tilde{\Delta} \right\},$$

pro vhodné d přirozené a Δ i $\tilde{\Delta}$ jsou kladná reálná čísla a $a_{j,l}$ značí $l-$tou složku afinního zobrazení A_j z definice množiny mělkých lineárních neuronových sítí.

Vidíme, že jsme se kromě omezení na šířku neuronové vrstvy uchýlili i k omezení na hodnoty parametrů neuronových sítí. Zároveň by bylo dobré, aby omezení na šíři neuronové sítě či omezení hodnot parametrů neuronových sítí záviselo na počtu dostupných pozorování n. Nechť jsou tedy $\{d_n\}_{n=1}^{\infty}$ je rostoucí posloupnost přirozených čísel a nechť $\{\Delta_n\}_{n=1}^{\infty}, \{\tilde{\Delta}_n\}_{n=1}^{\infty}$ jsou rostoucí posloupnosti reálných čísel. To nám umožňuje zadefinování posloupnosti $\{\mathcal{S}_n(\phi)\}_{n=1}^{\infty}$, kde pro každé n je $\mathcal{S}_n(\phi)$ definováno jako $\mathcal{S}(d_n,\Delta_n,\tilde{\Delta}_n,\phi)$. Pro každé vhodné posloupnosti $\{d_n\}_{n=1}^{\infty}$,$\{\Delta_n\}_{n=1}^{\infty}$ a $\{\tilde{\Delta}_n\}_{n=1}^{\infty}$ platí:

$$... \subseteq \mathcal{S}_n(\phi) \subseteq \mathcal{S}_{n+1}(\phi) \subseteq \mathcal{S}_{n+2}(\phi) \subseteq ...$$

Tato vlastnost platí, což si nyní ukážeme.

Mějme neuronovou síť $\boldsymbol{z}(\boldsymbol{x}) \in \mathcal{S}_n(\phi)$. Tato neuronová síť se dá vyjádřit jako prvek množiny $\mathcal{S}_{n+1}(\phi)$ položením $\beta_j = 0$, pro každé $d_n \leq j \leq d_{n+1}$ a volbou nulového afinního rozdělení $A_j(\boldsymbol{x}) = 0$, pro každé $d_n \leq j \leq d_{n+1}$. Po tomto ekvivalentním dodefinováním je totiž skutečně q_1 takovéto sítě rovno d_{n+1}, dále

$$\sum_{j=0}^{d_{n+1}} |\beta_j| = \sum_{j=0}^{d_n} |\beta_j| \leq \Delta_n \leq \Delta_{n+1},$$

kdy jsme v první rovnosti využili nulovosti dodefinovaných bet a v poslední nerovnosti jsme pouze využili rostoucí vlastnosti posloupnosti $\{\Delta_n\}_{n=1}^{\infty}$. Obdobně způsobem se dá ukázat, že

$$\max_{1\leq j\leq d_{n+1}} \sum_{l=0}^{q_0} |a_{j,l}| = \max_{1\leq j\leq d_n} \sum_{l=0}^{q_0} |a_{j,l}| \leq \tilde{\Delta}_n \leq \tilde{\Delta}_{n+1},$$

kdy se opět v první rovnosti využilo nulovosti, v tomto případě afinních zobrazení, a v poslední nerovnosti se opět využívá monotónnosti posloupnosti $\{\tilde{\Delta}_n\}_{n=1}^{\infty}$. Na předchozích řádcích jsme tedy si vlastně i ukázali k čemu opravdu byly předpoklady, aby všechny tři posloupnosti parametrů byly rostoucí.

Zavedení definice $\mathcal{S}_n$ nám umožňuje řešit hledání $\tilde{\mu}_n$ jako optimalizační úlohu s daným počtem parametrů, který je roven počtu parametrů příslušné neuronové sítě. Nyní si formulujeme větu, která nám říká, že množiny $\mathcal{S}_n(\phi)$ jsou stejnoměrně husté.

Věta 3. *Nechť ϕ je měřitelná neklesající aktivační funkce, splňující:*

$$\lim_{x\to-\infty} \phi(x) = 0 \wedge \lim_{x\to\infty} \phi(x) = 1.$$

Nechť dále rostoucí posloupnosti $\{d_n\}_{n=1}^{\infty}$,$\{\Delta_n\}_{n=1}^{\infty}$ a $\{\tilde{\Delta}_n\}_{n=1}^{\infty}$ splňují to, že limity těchto posloupností, pro n jdoucí do nekonečna, jdou do nekonečna. Pak platí, že $\bigcup_{n\geq 1} \mathcal{S}_n(\phi)$ je stejnoměrně hustá na $\mathcal{C}(\mathcal{X})$.

Důkaz. Věta se dokáže pouze rozepsáním sjednocení množin $\mathcal{S}_n(\phi)$.

$$\bigcup_{n\geq 1} \mathcal{S}_n(\phi) = \bigcup_{n\geq 1} \left\{ f \in \Sigma^{q_0}(\phi); q_1 = d_n, \sum_{j=0}^{q_1} |\beta_j| \leq \Delta_n, \max_{1\leq j\leq q_1} \sum_{l=0}^{q_0} |a_{j,l}| \leq \tilde{\Delta}_n \right\}$$

Pro konečné $K \in \mathbb{N}$ platí, že sjednocení množin $\mathcal{S}_n(\phi)$ se dá zapsat ve tvaru

$$\bigcup_{n\geq 1}^{K} \mathcal{S}_n(\phi) = \left\{ f \in \Sigma^{q_0}(\phi); q_1 \leq d_K, \sum_{j=0}^{q_1} |\beta_j| \leq \Delta_K, \max_{1\leq j\leq q_1} \sum_{l=0}^{q_0} |a_{j,l}| \leq \tilde{\Delta}_K \right\}$$

limitním přechodem, když K pošleme do nekonečna, se pak sjednocení $\mathcal{S}_n(\phi)$ rovná

$$\lim_{K\to\infty} \bigcup_{n\geq 1}^{K} \mathcal{S}_n(\phi) = \lim_{K\to\infty} \left\{ f \in \Sigma^{q_0}(\phi); q_1 \leq d_K, \sum_{j=0}^{q_1} |\beta_j| \leq \Delta_K, \max_{1\leq j\leq q_1} \sum_{l=0}^{q_0} |a_{j,l}| \leq \tilde{\Delta}_K \right\}$$

což se vzhledem k předpokladům věty, kdy předpokládáme, že všechny tři parametrové posloupnosti konvergují do nekonečna, rovná

$$\left\{ f \in \Sigma^{q_0}(\phi); q_1 \leq \infty, \sum_{j=0}^{q_1} |\beta_j| \leq \infty, \max_{1\leq j\leq q_1} \sum_{l=0}^{q_0} |a_{j,l}| \leq \infty \right\}$$

Po tomto rozepsání již je zřejmé, že platí

$$\bigcup_{n\geq 1} \mathcal{S}_n(\phi) = \Sigma^{q_0}(\phi).$$

V minulé kapitole, konkrétně v Větě 2 jsme dokázali, že množina $\Sigma^{q_0}(\phi)$ je stejnoměrně hustá na kompaktu v $\mathcal{C}(\mathbb{R}^{q_0})$, tudíž je stejnoměrně hustá i na kompaktní podmnožině $\mathcal{X}$.

□

Předcházející věta nám vlastně říká, že pro libovolnou spojitou regresní funkci $\mu(\boldsymbol{x})$ z rovnosti (3.1) jsme schopni nalézt $n \in \mathbb{N}$ a μ_n takové, že $\mu_n \in \mathcal{S}_n(\phi)$ aproximuje μ s předem zvolenou přesností.

Vyvstává přirozená otázka, co by se stalo, kdyby některý z předpokladů na rostoucí monotónnost nebyl splněn. Dá se ukázat, že v případě, že $\Delta_n = \Delta > 0$, pro každé přirozené n, se dá libovolně přesně aproximovat pouze velmi malá třída spojitých funkcí na $\mathcal{X}$, a to uzávěr $\overline{\bigcup_{n\geq 1} \mathcal{S}_n(\phi)} \subset \mathcal{C}(\mathcal{X})$.

Po zavedení potřebné posloupnosti $\mathcal{S}_n(\phi)$ si nyní zavedeme tzv. sítový odhad, jehož konzistenci si následně v této kapitole dokážeme. Nutno podotknout, že název sítový není odvozen od slova síť, jak by se mohlo na první pohled zdát, když se zabýváme konzistencí neuronových sítí, ale od slova síto. Jedná se totiž o neparametrický odhad minima na nekonečnědimenzionálním prostoru, tím způsobem, že hledáme minimum na menší podmnožině, se kterou se lépe pracuje. Tato menší množina by měla být hustá, hledáme tedy optimální řešení na jakési mřížce či sítu.

Definice 10 (sítový odhad)**.** *Mějme posloupnost množin vhodných neuronových sítí* $\{\mathcal{S}_n(\phi)\}_{n=1}^{\infty}$*, kdy pro každé* $n \in \mathbb{N}$ *je* $\mathcal{S}_n(\phi)$ *definováno předpisem*

$$\mathcal{S}_n(\phi) = \left\{ f \in \Sigma^{q_0}(\phi); q_1 = d_n, \sum_{j=0}^{q_1} |\beta_j| \leq \Delta_n, \max_{1\leq j\leq q_1} \sum_{l=0}^{q_0} |a_{j,l}| \leq \tilde{\Delta}_n \right\},$$

pro vhodné $q_0 \in \mathbb{N}$*, a kde* $\{d_n\}_{n=1}^{\infty}$ *je rostoucí posloupnost přirozených čísel a* $\{\Delta_n\}_{n=1}^{\infty}$,$\{\tilde{\Delta}_n\}_{n=1}^{\infty}$ *jsou rostoucí posloupnosti reálných čísel.*

Za těchto předpokladů definujeme sítový odhad[1] *funkce* $\mu(\boldsymbol{x}) = \mathsf{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$ *jako posloupnost odhadů* $\{\hat{\mu}_n\}_{n\geq 1}$*, kdy pro každé* $n \in \mathbb{N}$ *je* $\hat{\mu}_n$ *definován předpisem*

$$\hat{\mu}_n = \arg \min_{\mu \in \mathcal{S}_n(\phi)} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu(\boldsymbol{X}_i))^2.$$

Čistě po přečtení definice sítového odhadu může mít čtenář relevantní dotaz, zda-li vůbec pro libovolná data existuje tento sítový odhad pro každé $n \in \mathbb{N}$. Opravdu existence minima z této definice není pro každé n zaručena, součástí důkazu věty o konzistenci bude dokázána i existence tohoto sítového odhadu. Tato věta, ale bude mít několik poměrně striktních předpokladů, ať už na rozdělení chybových epsilonů v předpisu modelu či na strukturu jednotlivých $\mathcal{S}_n(\phi)$.

Ještě nám zbývá si vysvětlit v jakém smyslu budeme konzistenci našeho sítového odhadu uvažovat. Jedná se totiž o odhad funkce, nikoli pouze nějakého parametru. K tomuto účelu si musíme představit následující pseudo-normu na prostoru spojitých funkcí na $\mathcal{X}$:

$$\|\mu\|_n = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\mu(\boldsymbol{X}_i))^2}$$

[1] z angl. sieve estimator

Opravdu se jedná o pseudonormu. Její hodnota je zřejmě kladná pro libovolnou spojitou funkci a libovolný náhodný výběr, je absolutně homogenní při násobení skalárem, protože zřejmě platí

$$\|\alpha\mu\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\alpha\mu(\boldsymbol{X}_i))^2} = |\alpha| \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mu(\boldsymbol{X}_i))^2} = |\alpha| \, \|\mu\|_n,$$

ale není normou, protože když $\|\mu\|_n$ je rovno nule, tak to ještě neimplikuje, že μ je nulová funkce, protože je stále možnost, že μ je pouze nulová na datech $\boldsymbol{X}_i$, pro $i = 1,...,n$. Například v případě $\mu(x) : \mathbb{R} \to \mathbb{R}, \mu(x) = ReLU(x)$ a záporných pozorování v našem náhodném výběru by skutečně $\|\mu\| = 0$, ale $ReLU \neq 0$.

Důležitou vlastností této pseudo-normy je to, že závisí na pozorovaných datech a jejich rozsahu n, pravě díky tomu ji lze použit k důkazu konzistence sítového odhadu $\hat{\mu}_n$, v tuto chvíli již pro konkrétní n, v klasické podobě, t.j. důkazu konvergence v pravděpodobnosti ke skutečné regresní funkci μ.

Ještě než se dostaneme k formulování věty o konzistenci právě zavedeného sítového odhadu, formulujeme si pár vět, které budeme potřebovat během jejího důkazu. Začneme White-Wooldridgeovou větou. Před její samotnou formulací ještě zadefinujeme, co budeme myslet pojmem úplný pravděpodobnostní prostor.

Definice 11 (úplný pravděpodobnostní prostor)**.** *Řekneme, že pravděpodobnostní prostor* $(\Omega,\mathcal{A},\mathbb{P})$ *je úplný, jestliže pro každou* $B \in \mathcal{A},\mathbb{P}(B) = 0$, *platí* $A \subset B \implies A \in \mathcal{A}$.

Věta 4 (White-Woolridgeova věta)**.** *Nechť* $(\Omega,\mathcal{A},\mathbb{P})$ *je úplný pravděpodobnostní prostor a nechť* (Θ, ρ) *je pseudo-metrický prostor. Nechť* $\{\Theta_n\}_{n=1}^{\infty}$ *je posloupnost kompaktních podmnožin* Θ. *Dále nechť*

$$\mathcal{Q}_n : \Omega \times \Theta_n \to \bar{\mathbb{R}}$$

je $\mathcal{A} \otimes \mathcal{B}(\Theta_n)/\mathcal{B}(\hat{\mathbb{R}})-$*měřitelné zobrazení a dále předpokládejme, že pro každé* $n \in \mathbb{N},\omega \in \Omega$ *je* $\mathcal{Q}_n(\omega,.)$ *zdola polospojitá funkce na* Θ_n. *Pak pro každé* $n = 1,2,...$ *existuje* $\hat{\theta}_n : \Omega \to \Theta_n,\mathcal{A}/\mathcal{B}(\Theta_n)-$*měřitelné zobrazení takové, že pro každé* $\omega \in \Omega$ *platí*

$$\mathcal{Q}_n(\omega,\hat{\theta}_n(\omega)) = \inf_{\theta\in\Theta_n} \mathcal{Q}_n(\omega,\theta).$$

Důkaz. Důkaz je uveden v White a Wooldridge (1991).

□

Pouze pro připomenutí čtenáři uvedeme, co je nazýváno jako zdola polospojitá funkce. Funkce $f(x)$ se nazývá jako zdola polospojitá v bodě x jestliže

$$\liminf_{y\to x} f(y) \leq f(x).$$

Naopak funkce se nazývá shora polospojitá, pokud

$$\limsup_{y\to x} f(y) \geq f(x).$$

Tyto dva pojmy nejsou navzájem disjunktní, dokonce platí, že funkce je zároveň zdola i shora polospojitá právě tehdy, když je spojitá.

Budeme potřebovat ještě jedno tvrzení, v jehož formulaci se objevuje stochastická varianta symbolu malé $o(a_n)$, kde a_n je nějaká posloupnost reálných čísel. Sice v samotném tvrzení bychom si vystačili nahrazením $o_P(1)$ nějakým zbytkovým členem, který pro n jdoucí do nekonečna konverguje v pravděpodobnosti k 0, ale v další části této práce se nám zavedení o_P a O_P bude ještě hodit, takže si je zavedeme pomocí exaktní matematické definice.

Definice 12. *Mějme posloupnost náhodných veličin $\{X_n\}_{n\geq 1}$ a posloupnost reálných čísel $\{a_n\}_{n\geq 1}$, pak značíme $X_n = o_P(a_n)$, pokud pro každé $\epsilon > 0$, platí*

$$\lim_{n\to\infty} \mathbb{P}\Big(\Big|\frac{X_n}{a_n}\Big| \geq \epsilon\Big) = 0.$$

Obdobně značíme $X_n = O_P(a_n)$ v případě, že pro každé $\epsilon > 0$, existují konečné $M > 0$ a konečné $N > 0$ takové, že pro každé $n \geq N$ platí

$$\mathbb{P}\Big(\Big|\frac{X_n}{a_n}\Big| \geq M\Big) < \epsilon.$$

S výše zadefinovaným $o_P(a_n)$ si můžeme formulovat následující větu, kterou následně využijeme k důkazu konzistence sítového odhadu.

Věta 5. *Nechť $M_n(\theta)$ a $M(\theta)$ jsou funkce a nechť ρ je pseudo-metrika. Dále ať pro každé $\epsilon > 0$*

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathbb{P}} 0$$

konverguje v pravděpodobnosti k nule a zároveň

$$\inf_{\theta:\rho(\theta,\theta_0)\geq\epsilon} M(\theta) > M(\theta_0).$$

Pak každá posloupnost odhadů $\hat{\theta}_n$ splňující $M_n(\hat{\theta}_n) \leq M_n(\theta_0) + o_P(1)$ konverguje v pravděpodobnosti k θ_0.

Důkaz. Díky předpokladům věty platí, že

$$M_n(\hat{\theta}_n) \leq M_n(\theta_0) + o_P(1),$$

díky stejnoměrné konvergenci posloupnosti funkcí M_n a vztahu nerovností a limit pravá strana nerovnosti konverguje v pravděpodobnosti k $M(\theta_0) + o_P(1)$ a platí i následující nerovnost

$$M_n(\hat{\theta}_n) \leq M(\theta_0) + o_P(1).$$

To ale implikuje $M(\theta_0) \geq M_n(\hat{\theta}_n) - o_P(1)$. Tento spodní odhad $M(\theta_0)$ nyní využijeme, protože platí

$$M(\hat{\theta}_n) - M(\theta_0) \leq M(\hat{\theta}_n) - M_n(\hat{\theta}_n) + o_P(1)$$

to je určitě menší než

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| + o_P(1),$$

ale toto již díky předpokladu na stejnoměrnou konvergenci konverguje v pravděpodobnosti k 0 pro n jdoucí do nekonečna.

Nyní mějme ϵ reálné, z předpokladů věty víme, že existuje ν takové, že pro každé $\theta, \rho(\theta,\theta_0) \geq \epsilon$ platí $M(\theta) > M(\theta_0) + \nu$. To ale znamená, že v případě, kdy $\rho(\hat{\theta}_n,\theta_0) \geq \epsilon$, by muselo platit i $M(\hat{\theta}_n) < M(\theta_0) + \nu$, pravděpodobnost jevu $\rho(\hat{\theta}_n,\theta_0) \geq \epsilon$ tedy musí být nižší než pravděpodobnost jevu $M(\hat{\theta}_n) < M(\theta_0) + \nu$. Výše jsme ale ukázali, že $M(\hat{\theta}_n) - M(\theta_0) \xrightarrow{\mathbb{P}} 0$, takže $\mathbb{P}(M(\hat{\theta}_n) < M(\theta_0) + \nu)$ musí konvergovat k nule. Právě jsme tedy dokázali, že $\mathbb{P}(\rho(\hat{\theta}_n,\theta_0) \geq \epsilon)$ se pro n jdoucí do nekonečna limitně blíží nule. Jinými slovy jsme právě dokázali konzistenci odhadu $\hat{\theta}_n$.

□

Konečně se dostáváme k formulaci věty o existenci a konzistenci sítového odhadu, tuto větu si následně i dokážeme.

Věta 6 (věta o konzistenci sítového odhadu)**.** *Nechť* $(\Omega,\mathcal{A},\mathbb{P})$ *je úplný pravděpodobnostní prostor a* $\mathcal{X} = 1 \times [0,1]^{q_0}$. *Dále nechť platí následující:*

1. $(Y_i,\boldsymbol{X}_i), i = 1,...,n$ *jsou nezávislé stejně rozdělené pozorování splňující pro každé* i

$$Y_i = \mu_0(\boldsymbol{X}_i) + \epsilon_i,$$

kde $\mu(\boldsymbol{x}) \in \mathcal{C}(\mathcal{X})$ *a pro* ϵ_i *platí*

$$\mathsf{E}[\epsilon_i \mid \boldsymbol{X}_i] = 0,$$

$$\mathrm{var}[\epsilon_i \mid \boldsymbol{X}_i] = \sigma^2 \in \mathbb{R}^+,$$

$$\mathsf{E}[|\epsilon_i|^{2+\delta} \mid \boldsymbol{X}_i] < \infty,$$

pro nějaké $\delta > 0$, *zároveň ještě požadujeme, aby* ϵ_i *byly nezávislé na vysvětlujících proměnných* $\boldsymbol{X_i}$.

2. *aktivační funkce* ϕ *je sigmoid funkce*

3. *posloupnost* $\{d_n\}_{n\geq 1}$, *resp. posloupnosti* $\{\Delta_n\}_{n\geq 1}$ *a* $\{\tilde{\Delta}_n\}_{n\geq 1}$ *jsou rostoucí posloupnosti přirozených resp. reálných čísel konvergujících do nekonečna, které zároveň splňují*

$$\frac{d_n\Delta_n^2 \log(d_n\Delta_n)}{n} \to 0, n \to \infty.$$

Pak sítový odhad $\{\hat{\mu}_n\}_{n\geq 1}$ *existuje. Zároveň* $\|\hat{\mu}_n - \mu_0\|_n$ *konverguje v pravděpodobnosti k nule, tedy pro každé* $\tilde{\epsilon} > 0$

$$\lim_{n\to\infty} \mathbb{P}(\|\hat{\mu}_n - \mu_0\|_n > \tilde{\epsilon}) = 0.$$

Důkaz. Důkaz první části věty zabývající se existencí sítového odhadu je založen na White-Woolrdridgeově větě. Musíme si obecné znění této věty převézt na případ sítového odhadu se značením zavedeném v naší práci.

Pseudo-metrický prostor (Θ, ρ) bude v tomto případě pseudo-metrický prostor spojitých funkcí na $\mathcal{X}$, $(\mathcal{C}(\mathcal{X}),\|.\|_n)$.

Posloupnost kompaktní podprostorů $\{\Theta_n\}_{n\geq 1}$ prostoru $\mathcal{C}(\mathcal{X})$ bude v našem případě posloupnost $\{\mathcal{S}_n(\sigma)\}$, kde σ značí aktivační funkci sigmoid. Musíme ale dokázat, že jsou tyto prostory kompaktní.

Nechť n je libovolné přirozené číslo. Označme si parametry každé sítě splňující předpoklady věty $\boldsymbol{z} \in \mathcal{S}_n(\sigma)$ jako vektor $\boldsymbol{\theta}_n = [\beta_0,...,\beta_{d_n}, a_{1,0},...,a_{1,q_0},a_{2,0},...,a_{d_n,q_0}]$. Protože pro každé n jsou $d_n,\Delta_n,\tilde{\Delta}_n$ konečná čísla a díky předpisu množiny $\mathcal{S}_n(\sigma)$ jsme schopni nalézt konstanty A_n a B_n takové, že

$$\boldsymbol{\theta}_n \in [-A_n,A_n]^{d_n+1} \times [-B_n,B_n]^{d_n(q_0+1)}.$$

Tento systém intervalů si označíme jako I_n. Systém uzavřených intervalů I_n je uzavřená, omezená podmnožina $\mathbb{R}^{d_n(q_0+2)+1}$. To znamená, že I_n je kompaktní množina. Nyní nalezením vhodného spojitého zobrazení, které zobrazuje I_n právě na množinu $\mathcal{S}_n(\sigma)$ prokážeme i jejich kompaktnost.

Uvažujme tedy zobrazení $H : (\Theta_n,\|.\|_2) \to (\mathcal{S}_n(\sigma),\|.\|_n)$ definované předpisem

$$H(\boldsymbol{\theta_n}) = \beta_0 + \sum_{j=1}^{d_n} \beta_j \sigma(a_{j,0} + \boldsymbol{\alpha}_j^\top \boldsymbol{x}),$$

kde jako $\boldsymbol{\alpha}_j$ značíme vektor $(a_{j,1}, a_{j,2},..., a_{j,q_0})^\top$.

Takovéto zobrazení má požadovanou vlastnost $\mathcal{S}_n(\sigma) \in H(I_n)$. K důkazu kompaktnosti nám nyní stačí již jen ověřit spojitost zobrazení H. To uděláme přímo z definice, kdy pro nějaké $\boldsymbol{\theta_n} \in I_n$ a ϵ reálné budeme hledat δ takové, že pro každé $\boldsymbol{\theta_{n,\delta}} \in I_n$ platí

$$\|\boldsymbol{\theta_n} - \boldsymbol{\theta_{n,\delta}}\|_2 < \delta \implies \|H(\boldsymbol{\theta_n}) - H(\boldsymbol{\theta_{n,\delta}})\|_n < \epsilon.$$

Mějme $\boldsymbol{\theta}_{\boldsymbol{n}}^{(1)},\boldsymbol{\theta}_{\boldsymbol{n}}^{(2)} \in \mathcal{S}_n(\sigma)$, rozepíšeme si kvadrát normy $\|H(\boldsymbol{\theta}_{\boldsymbol{n}}^{(1)}) - H(\boldsymbol{\theta}_{\boldsymbol{n}}^{(2)})\|_n^2$. Dostáváme

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\beta_0^{(1)} + \sum_{j=1}^{d_n}\beta_j^{(1)}\sigma(a_{j,0}^{(1)} + \boldsymbol{\alpha}_j^{(1)\top}\boldsymbol{x}_i) - \beta_0^{(2)} + \sum_{j=1}^{d_n}\beta_j^{(2)}\sigma(a_{j,0}^{(2)} + \boldsymbol{\alpha}_j^{(2)\top}\boldsymbol{x}_i)\Big)^2.$$

Tento součet čtverců se dá přidáním absolutní hodnoty do sumy, která díky kvadrátu nezmění výslednou hodnotu celé sumy, a aplikací trojúhelníkové nerovnosti shora odhadnout jako

$$\leq \frac{1}{n}\sum_{i=1}^{n}\Big(\Big|\beta_0^{(1)} - \beta_0^{(2)}\Big| + \sum_{j=1}^{d_n}\Big|\beta_j^{(1)}\sigma(a_{j,0}^{(1)} + \boldsymbol{\alpha}_j^{(1)\top}\boldsymbol{x}_i) - \beta_j^{(2)}\sigma(a_{j,0}^{(2)} + \boldsymbol{\alpha}_j^{(2)\top}\boldsymbol{x}_i)\Big|\Big)^2.$$

Nyní si sumu ekvivalentně rozšíříme, dostáváme následně

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\Big|\beta_0^{(1)} - \beta_0^{(2)}\Big| + \sum_{j=1}^{d_n}\Big|\beta_j^{(1)}\Big|\Big|\sigma(a_{j,0}^{(1)} + \boldsymbol{\alpha}_j^{(1)\top}\boldsymbol{x}_i) - \sigma(a_{j,0}^{(2)} + \boldsymbol{\alpha}_j^{(2)\top}\boldsymbol{x}_i)\Big| +$$

$$+ \Big|\beta_j^{(1)} - \beta_j^{(2)}\Big|\sigma(a_{j,0}^{(2)} + \boldsymbol{\alpha}_j^{(2)\top}\boldsymbol{x}_i)\Big)^2.$$

V následujícím odhadu využijeme vlastnosti sigmoid funkce, tato funkce je totiž $1/4-$Lipschitzovská. K ověření tohoto tvrzení musíme funkci σ zderivovat a najít maximální hodnotu derivace. V první kapitole, během představování této aktivační funkce jsme si ukázali, že derivace této funkce má podobu $\sigma(1-\sigma)$. Zderivováním získáme tvar druhé derivace

$$\sigma'' = \sigma(1-\sigma)(1-\sigma) - \sigma\sigma(1-\sigma) = \sigma(1-\sigma)(1-2\sigma) = \sigma'(1-2\sigma).$$

Druhá derivace má kořen pouze v bodě, kde je funkční hodnota sigmoid funkce rovna 1/2, což je zřejmě v nule. Sigmoid funkce má tedy omezenou derivaci, protože limity její první derivace v plus i mínus nekonečnu jsou rovny nule. Diferencovatelná funkce s omezenou derivací je skutečně lipschitzovská, jelikož maximální hodnota derivace je $\sigma'(x) = 1/4$, tak jsme skutečně dokázali $1/4-$lipschitzovskost sigmoid funkce.

Této vlastnosti nyní využijeme v dalším kroku odhadování shoda naší sumy výše. Využitím $1/4-$Lipschitzovskosti sigmoid funkce se jí zbavíme v naší sumě

$$\leq \frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{j=0}^{d_n}\left|\beta_j^{(1)}-\beta_j^{(2)}\right|+\frac{\Delta_n}{4}\sum_{j=1}^{d_n}\left|(\boldsymbol{\alpha}_j^{(1)}-\boldsymbol{\alpha}_j^{(2)})^\top \boldsymbol{x}_i\right|+\left|a_{j,0}^{(1)}-a_{j,0}^{(2)}\right|\Big)^2.$$

Ve znění věty jsme předpokládali, že $\boldsymbol{X}_i \in [0,1]^{q_0}$, toho nyní využijeme k odstranění proměnných $\boldsymbol{x}_i$. Jelikož právě tyto $\boldsymbol{x}_i$ jako jediné závisely na i, tak je nyní první suma zbytečná. Po úpravě dostáváme další odhad v podobě

$$\leq \Big(\sum_{j=0}^{d_n}\left|\beta_j^{(1)}-\beta_j^{(2)}\right|+\frac{\Delta_n}{4}\sum_{j=1}^{d_n}\sum_{k=0}^{q_0}\left|a_{j,k}^{(1)}-a_{j,k}^{(2)}\right|\Big)^2.$$

Nyní můžeme předpokládat, že $\Delta_n > 0$, což není příliš omezující, jelikož předpokládáme, že posloupnost těch delt konverguje do nekonečna. Využitím tohoto předpokladu, umocněním celého výrazu na druhou a poměrně hrubým odhadem vypořádávajícím se se zbývajícími sdruženými členy z binomické věty dostáváme finální horní odhad v podobě:

$$\leq \left(\frac{\Delta_n}{4}\right)^2 (d_n(q_0+1))\|\boldsymbol{\theta_n^{(1)}}-\boldsymbol{\theta_n^{(2)}}\|_2^2.$$

Tento odhad nám již dává nápad, jak volit hledané δ, pro zadané epsilon z definice spojitosti, jednoduše ho zvolíme jako $\delta = \epsilon/\left(\frac{\Delta_n}{4}\sqrt{d_n(q_0+1)}\right)$. Takovouto volbou delty opravdu platí implikace

$$\|\boldsymbol{\theta_n}-\boldsymbol{\theta_{n,\delta}}\|_2 < \delta \implies \|H(\boldsymbol{\theta_n})-H(\boldsymbol{\theta_{n,\delta}})\|_n < \epsilon.$$

Dokázali jsme tedy, že pro každé n je opravdu prostor $\mathcal{S}_n(\sigma)$ kompaktní. Můžeme tyto množiny považovat za Θ_n ze znění White-Wooldridgeovy věty.

V této větě se objevují i zobrazení $\mathcal{Q}_n$, v případě naší věty o konzistenci se jedná o residuální součet čtverců spočítaný z našich dat. Jelikož pro fixní ω, tedy pro fixní pozorovaná data, je $\mathcal{Q}_n$, rovno

$$\mathcal{Q}_n(\omega,\mu) = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu(\boldsymbol{X}_i))^2,$$

pro $\mu(\boldsymbol{x}) \in \mathcal{S}_n(\sigma)$. Takovéto zobrazení je zřejmě spojité v μ na $\mathcal{S}_n(\sigma)$, tudíž je i na $\mathcal{S}_n(\sigma)$ zdola polospojité. Ověřením i tohoto předpokladu jsme již nyní schopni využit White-Wooldrigeovu větu, která nám v našem případě říká, že pro každé $n \in \mathbb{N}$ existuje $\hat{\mu}_n$, které pro každé $\omega \in \Omega$ splňuje

$$\hat{\mu}_n = \arg\min_{\mu\in\mathcal{S}_n(\phi)}\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu(\boldsymbol{X}_i))^2,$$

to je ale přesně definice sítového odhadu. Dokázali jsme, že za našich předpokladů pro libovolná data (splňující opět předpoklady věty) sítový odhad existuje.

Pokud jde o důkaz konzistence sítového odhadu $\hat{\mu}_n$, tak důkaz provedeme aplikací věty 5. V našem případě budou funkce M_n ze znění věty 5 právě funkce $\mathcal{Q}_n(\mu)$. Jako funkci M si zadefinujeme novou funkci danou předpisem

$$\mathcal{Q}(\mu) = \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu(\boldsymbol{X}_i))^2 \mid \boldsymbol{X_i}, i = 1,..,n\Big].$$

Nyní si budeme muset ověřit předpoklady věty 5. Nejprve si rozepíšeme funkci $\mathcal{Q}_n$ do vhodnější podoby.

$$\mathcal{Q}_n(\mu) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu(\boldsymbol{X}_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(\mu_0(\boldsymbol{X}_i) + \epsilon_i - \mu(\boldsymbol{X}_i))^2,$$

z čehož roznásobením získáme reprezentaci, kterou později využijeme

$$\mathcal{Q}_n(\mu) = \frac{1}{n}\sum_{i=1}^{n}(\mu_0(\boldsymbol{X}_i) - \mu(\boldsymbol{X}_i))^2 - \frac{2}{n}\sum_{i=1}^{n}\epsilon_i(\mu_0(\boldsymbol{X}_i) - \mu(\boldsymbol{X}_i)) + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2. \quad (3.2)$$

Skutečnost, že $\sup_{\mu\in\mathcal{S}_n(\sigma)}|\mathcal{Q}_n(\mu) - \mathcal{Q}(\mu)|$ konverguje k nule v pravděpodobnosti(v případě neměřitelnosti uvažujeme tuto konvergenci jako konvergenci vůči vnější pravděpodobnostní míře) dokázal ve svém článku Shen a kol. (2019). Důkaz je poměrně dlouhý a hodně technický, proto ho zde nebudeme uvádět, využívá se v něm třetí předpoklad ze znění dokazované věty na limitní chování posloupností $\{d_n\}_{n\geq 1}, \{\Delta_n\}_{n\geq 1}$.

K ověření dalšího z předpokladů si vyjádříme i funkci $\mathcal{Q}$ v podobné podobě jako je rovnost (3.2). Aplikací podmíněné střední hodnoty na tuto rovnost dostáváme

$$\mathcal{Q}(\mu) = \frac{1}{n}\sum_{i=1}^{n}(\mu_0(\boldsymbol{X}_i) - \mu(\boldsymbol{X}_i))^2 - \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}[\epsilon_i \mid \boldsymbol{X}_i](\mu_0(\boldsymbol{X}_i) - \mu(\boldsymbol{X}_i)) + \mathbb{E}[\epsilon_i^2 \mid \boldsymbol{X}_i] \quad (3.3)$$

jelikož $\mu(\boldsymbol{x})$,$\mu_0(\boldsymbol{x})$ jsou měřitelné funkce veličin $\boldsymbol{X}_i$, a protože díky předpokladům věty jsou chybové epsilony na těchto veličinách nezávislé. Díky předpokladům věty na první dva podmíněné momenty ϵ_i dostáváme vyjádření $\mathcal{Q}$ v podobě

$$\mathcal{Q}(\mu) = \frac{1}{n}\sum_{i=1}^{n}(\mu(\boldsymbol{X}_i) - \mu_0(\boldsymbol{X}_i))^2 + \sigma^2, \quad (3.4)$$

kde σ^2 značí hodnotu rozptylu ϵ_i nikoli kvadrát sigmoid funkce. Dále si můžeme všimnout, že položením $\mu = \mu_0$ se výraz zredukuje na

$$\mathcal{Q}(\mu_0) = \frac{1}{n}\sum_{i=1}^{n}(\mu_0(\boldsymbol{X}_i) - \mu_0(\boldsymbol{X}_i))^2 + \sigma^2 = \sigma^2. \quad (3.5)$$

Nyní mějme $\delta > 0$, pak

$$\inf_{\mu:\|\mu-\mu_0\|_n\geq\delta}\mathcal{Q}(\mu) - \mathcal{Q}(\mu_0) = \inf_{\mu:\|\mu-\mu_0\|_n\geq\delta}\frac{1}{n}\sum_{i=1}^{n}(\mu(\boldsymbol{X}_i) - \mu_0(\boldsymbol{X}_i))^2$$

využitím rovností (3.4) a (3.5). Z definice psedu-normy $\|.\|_n$ víme, že pro všechny funkce $\mu, \|\mu - \mu_0\|_n \geq \delta$:

$$\frac{1}{n}\sum_{i=1}^{n}(\mu(\boldsymbol{X}_i) - \mu_0(\boldsymbol{X}_i))^2 = \|\mu - \mu_0\|_n^2 \geq \delta^2,$$

z uzavřenosti množiny, přes kterou hledáme infimum, musí být i samotné infimum větší než δ^2, tedy

$$\inf_{\mu:\|\mu-\mu_0\|_n\geq\delta} \mathcal{Q}(\mu) - \mathcal{Q}(\mu_0) \geq \delta^2 > 0,$$

což implikuje

$$\inf_{\mu:\|\mu-\mu_0\|_n\geq\delta} \mathcal{Q}(\mu) > \mathcal{Q}(\mu_0),$$

tedy je splněn i druhý předpoklad věty 5. K dokázání konzistence sítového odhadu nám již zbývá dokázat jen platnost nerovnosti ze znění věty 5, tedy za našeho značení ukázat, že platí $\mathcal{Q}_n(\hat{\mu}_n) \leq \mathcal{Q}_n(\mu_0) + o_P(1)$. K tomu nám bude stačit, že $\mathcal{Q}_n(\hat{\mu}_n) - \mathcal{Q}_n(\mu_0)$ konverguje v pravděpodobnosti k nule. Dosazením μ_0 do rovnosti (3.2) zjistíme, že i v tomto případě se nám první dva členy vynulují. Pak rozdíl $\mathcal{Q}_n(\hat{\mu}_n) - \mathcal{Q}_n(\mu_0)$ je roven

$$\frac{1}{n}\sum_{i=1}^{n}(\mu_0(\boldsymbol{x}_i) - \hat{\mu}_n(\boldsymbol{x}_i))^2 - \frac{2}{n}\sum_{i=1}^{n}\epsilon_i(\mu_0(\boldsymbol{x}_i) - \hat{\mu}_n(\boldsymbol{x}_i)).$$

Druhý sčítanec konverguje v pravděpodobnosti k nule podle silného zákona velkých čísel, protože

$$\mathbb{E}\Big[\epsilon_i(\mu_0(\boldsymbol{X}_i) - \mu(\boldsymbol{X}_i))\Big] = \mathbb{E}[\mathbb{E}[\epsilon_i \mid \boldsymbol{X}_i]]\mathbb{E}[\mu_0(\boldsymbol{X}_i) - \mu(\boldsymbol{X}_i)] = 0,$$

kde jsme v první rovnosti využili nezávislosti epsilonů na vysvětlujících proměnných a vztahu střední a podmíněné střední hodnoty.

Pro první sčítanec si ukážeme rovnou klasickou nestochastickou konvergenci k nule. Využijeme totiž v této kapitole dříve dokázaných vlastností množiny $\mathcal{S}_n$. Podle věty 3 totiž $\bigcup_{n\geq 1}\mathcal{S}_n(\sigma)$ je stejnoměrně hustá množina na $\mathcal{C}(\mathcal{X})$, protože sigmoid funkce σ zřejmě splňuje předpoklady této věty. Zároveň jsme si ukázali, že pro každé $i \in \mathbb{N} : \mathcal{S}_n(\sigma) \subset \mathcal{S}_{n+1}(\sigma)$, protože všechny tři parametrové posloupnosti jsou dle znění věty rostoucí a rostou nade všechny meze.

Nyní mějme $\delta > 0$, ze stejnoměrné hustoty sjednocení $\bigcup_{n\geq 1}\mathcal{S}_n(\sigma)$ víme, že existuje $\bar{\mu}(\boldsymbol{x}) \in \bigcup_{n\geq 1}\mathcal{S}_n(\sigma)$ takové, že $\sup_{\boldsymbol{x}\in\mathcal{X}}|\bar{\mu}(\boldsymbol{x}) - \mu_0(\boldsymbol{x})| < \sqrt{\delta}$, to znamená, že musí existovat nějaké $n_0 \in \mathbb{N} : \bar{\mu}(\boldsymbol{x}) \in \mathcal{S}_{n_0}(\sigma)$. Pak pro každé $n \geq n_0$ platí

$$\frac{1}{n}\sum_{i=1}^{n}(\mu_0(\boldsymbol{X}_i) - \hat{\mu}_n(\boldsymbol{X}_i))^2 \leq \frac{1}{n}\sum_{i=1}^{n}(\mu_0(\boldsymbol{X}_i) - \bar{\mu}_n(\boldsymbol{X}_i))^2 < \frac{1}{n}\sum_{i=1}^{n}\delta = \delta,$$

kde první nerovnost plyne z definice sítového odhadu. Protože δ jsme brali jako libovolně malé, tak skutečně $\mathcal{Q}_n(\hat{\mu}_n) - \mathcal{Q}_n(\mu_0)$ konverguje v pravděpodobnosti k nule, čímž jsme dokázali i konzistenci sítového odhadu a důkaz věty je tedy (konečně) u konce.

□

Věta 6 nám tedy říká, že sítový odhad nám jako posloupnost odhadů za daných předpokladů existuje, a že tento sítový odhad je konzistentní. Na druhou stranu nám nedává žádný návod, jak daný sítový odhad hledat, museli bychom pro každé n řešit optimalizační úlohu v podobě hledání neuronové sítě z $\mathcal{S}_n$, která by minimalizovala čtvercovou chybu na pozorovaných datech. Nalezení globálního minima této nekonvexní úlohy, ale nejsme v praxi schopni zaručit, čímž ale nebudou splněny předpoklady věty a konzistence již nebude zaručena. Tato věta má tedy především teoretickou váhu, že jsme skutečně z dostatečného množství dat schopni odhadnout parametry neuronové sítě, která bude aproximovat libovolnou spojitou závislost podmíněné střední hodnoty na vysvětlujících proměnných s předem určenou přesností.

V celém důkazu věty 6 jsme využili předpokladu, že obecná funkce ϕ z definice množin $\mathcal{S}_n(\phi)$, tedy aktivační funkce skryté vrstvy uvažovaných neuronových sítí, je sigmoid funkce pouze jednou, a to když jsme využili její lipschitzovskovsti. Tento předpoklad nebyl využit ani v žádném z využitých lemmat, jejichž důkaz není v této práci uveden, a na který jsem pouze dali čtenáři referenci na vhodnou literaturu. To znamená, že znění věty by šlo ještě zobecnit a omezit se pouze na lipschitzovské funkce, ale vzhledem k tomu, že ze standartně používaných aktivačních funkcí je lipschitzovská pouze sigmoid funkce (hyperbolický tangent je pouze lineárně transformovaná sigmoid funkce), a vzhledem k tomu, že právě sigmoid funkci zvolíme jako aktivační síť neuronových sítí v simulační části v poslední kapitole, tak větu 6 necháme v tomto méně obecném tvaru.

3.2 Asymptotická normalita sítového odhadu

Po dokázání konzistence sítového odhadu parametrů (speciálního typu) neuronových sítí si ještě v této kapitole uvedeme větu o asymptotické normalitě sítového odhadu. Opět se dá v literatuře nalézt několik variant těchto vět, které se navzájem liší různými předpoklady na tvaru neuronových sítí, my si vybereme variantu věty o asymptotické normalitě, jejíž předpoklady jsou velmi podobné předpokladům Věty 6.

Stejně jako v případě věty o konzistenci sítového odhadu i vět o univerzalitě si musíme nejprve před formulací samotné věty zavést nové definice a značení. K první definici nás motivuje kompaktnost množin $\mathcal{S}_n(\phi)$ v prostoru $\mathcal{C}(\mathcal{X})$, kterou jsme dokázali v průběhu důkazu věty 6. Symbolem $\pi_n\mu_0$ budeme značit nejbližší aproximaci funkce μ_0 v prostoru $\mathcal{S}_n(\phi)$. Věta 3 nám říká, že sjednocení $\bigcup_{n\geq 1}\mathcal{S}_n(\phi)$ je stejnoměrně hustá, to implikuje, že $\pi_n\mu_0$ konverguje ve významu naší pseudonormy $\|.\|_n$ ke skutečné funkci μ_0. Následující věta nám řekne, jakým tempem konverguje k této funkci μ_0 náš sítový odhad v případě, že náš odhad $\{\hat{\mu}_n\}_{n\geq 1}$ není úplně přesně sítový z definice, protože jsme nebyli schopni nalézt přesné minimum střední čtvercové chyby, ale zároveň jsme se mu dostatečně přiblížili. Zřejmě tato rychlost závisí na konvergenci nejlepší aproximace $\pi_n\mu_0$.

Věta 7. *Nechť platí předpoklady věty 6 dále nechť posloupnost, odhadů $\{\hat{\mu}_n\}_{n\geq 1}$ splňuje nerovnost*

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{\mu}_n(\boldsymbol{X}_i))^2 \leq \inf_{\mu\in\mathcal{S}_n(\phi)}\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu(\boldsymbol{X}_i))^2+O_P(r_n),$$

kde r_n je definované jako

$$r_n = O\Big(\min\Big\{\|\pi_n\mu_0 - \mu_0\|_n^2, \frac{d_n \log(d_n\Delta_n)}{n}, \frac{d_n \log n}{n}\Big\}\Big).$$

Pak pro každé $n \geq 1$ platí

$$\|\hat{\mu}_n - \mu_0\|_n = O_P\Big(\max\Big\{\|\pi_n\mu_0 - \mu_0\|_n, \sqrt{\frac{d_n \log n}{n}}\Big\}\Big).$$

Důkaz. Důkaz je uveden v Shen a kol. (2019)

□

Vidíme, že chybová posloupnost r_n z definice konverguje k nule, protože ve větě 6 jsme předpokládali, že $\frac{d_n \log(d_n\Delta_n)}{n} \to 0$. To znamená, že $\{\hat{\mu}_n\}_{n\geq 1}$ se stejně musí skutečnému sítovému odhadu v nekonečnu blížit, aby platila rychlost konvergence z věty 7.

Vidíme také, že rychlost této konvergence je ovlivněna rychlostí růstu posloupnosti d_n, která nám udává výstupní dimenzi první vrstvy q_1, tedy zbytečně velké, široké neuronové mají asymptoticky pomalejší konvergenci.

Posledním pojmem, který si musíme zadefinovat před formulací věty o normalitě sítového odhadu, je perturbovaný operátor $\tilde{\mu}$ definovaný jako:

$$\tilde{\mu}_n(\mu,\lambda_n) = (1 - \lambda_n^{1/2})\mu + \lambda_n^{1/2}(\mu_0 + 1),$$

kde $\mu \in \mathcal{S}_n(\phi)$ a pro každé $n \in \mathbb{N}, \lambda_n \in (0,1)$. Nyní nám již nic nebrání formulovat následující větu. Důkaz asymptotické normality je velmi technický, kdy se využívají všechny netriviální předpoklady této věty, proto si tuto větu nebudeme v této práci dokazovat.

Věta 8. *Nechť $(\Omega,\mathcal{A},\mathbb{P})$ je úplný pravděpodobnostní prostor a $\mathcal{X} = 1 \times [0,1]^{q_0}$. Dále nechť platí následující:*

1. *$(Y_i,\boldsymbol{X}_i), i = 1,...,n$ jsou nezávislé stejně rozdělené pozorování splňující pro každé i*

   $$Y_i = \mu_0(\boldsymbol{X}_i) + \epsilon_i,$$

   kde $\mu(\boldsymbol{x}) \in \mathcal{C}(\mathcal{X})$ a pro ϵ_i platí

   $$\mathsf{E}[\epsilon_i \mid \boldsymbol{X}_i] = 0,$$

   $$\mathrm{var}[\epsilon_i \mid \boldsymbol{X}_i] = \sigma^2 \in \mathbb{R}^+,$$

   $$\mathsf{E}[|\epsilon_i|^{2+\delta} \mid \boldsymbol{X}_i] < \infty,$$

   pro nějaké $\delta > 0$, zároveň ještě požadujeme, aby ϵ_i byly nezávislé na vysvětlujících proměnných $\boldsymbol{X_i}$.

2. *aktivační funkce ϕ je sigmoid funkce*

3. *posloupnost $\{d_n\}_{n\geq 1}$, resp. posloupnosti $\{\Delta_n\}_{n\geq 1}$ a $\{\tilde{\Delta}_n\}_{n\geq 1}$ je rostoucí posloupnost přirozených resp. reálných čísel konvergujících do nekonečna, které $d_n\Delta_n \log(d_n\Delta_n) = o(n^{1/4})$ pro n jdoucí do nekonečna*

4. *nechť existuje posloupnost* $\{\rho\}_{n\geq 1}$ *taková, že* $\rho_n\|\hat{\mu}_n - \mu_0\|_n = O_P(1)$, *a pro kterou existuje* $\delta > 0$ *takové, že* $\frac{n}{\rho_n^2 \Delta_n^\delta} = o(1)$

5. *pro posloupnost splňující* $\lambda_n = o(n^{-1})$ *platí tyto dvě rovnosti:*

$$\sup_{\mu\in\mathcal{S}_n(\phi):\|\mu-\mu_0\|_n\leq\rho_n^{-1}} \|\pi_n\tilde{\mu}_n(\mu,\lambda_n) - \tilde{\mu}_n(\mu,\lambda_n)\|_n = O_P(\lambda_n\rho_n),$$

$$\sup_{\mu\in\mathcal{S}_n(\phi):\|\mu-\mu_0\|_n\leq\rho_n^{-1}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i(\pi_n\tilde{\mu}_n(\mu.\lambda_n)(\boldsymbol{X}_i) - \tilde{\mu}_n(\mu.\lambda_n)(\boldsymbol{X}_i)) = O_P(\lambda_n).$$

Pak pro sítový odhad $\{\hat{\mu}_n\}_{n\geq 1}$ *platí pro n jdoucí do nekonečna:*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{\mu}_n(\boldsymbol{X}_i) - \mu_0(\boldsymbol{X}_i)) \xrightarrow{d} \mathcal{N}(0,\sigma^2).$$

Důkaz. Důkaz je uveden v Shen a kol. (2019)

□

Všimněme si rozdílu ve třetím předpokladu této věty a věty o konzistenci. Předpoklady na limitní chování parametrových posloupností se liší, v případě věty o konzistenci je tempo růstu posloupnosti $\{d_n\}$ limitně rychlejší. Takto rychlé tempo růstu by mohlo způsobit, že by člen $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{\mu}_n(\boldsymbol{X}_i) - \mu_0(\boldsymbol{X}_i))$ mohl konvergovat rovnou k nule v distribuci (tedy i v pravděpodobnosti).

Znalost asymptotického rozdělení nám umožňuje sestrojit statistické testy, problém je, že obvykle v aplikacích neznáme skutečnou hodnotu σ^2. Shen a kol. (2019) dokázali, že odhad

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}_n(\boldsymbol{X}_i))^2$$

za předpokladů věty 8 konzistentně odhaduje neznámý parametr σ^2. Pak ale platí, že

$$\frac{1}{\sqrt{\hat{\sigma}_n^2 n}}\sum_{i=1}^{n}(\hat{\mu}_n(\boldsymbol{X}_i) - \mu_0(\boldsymbol{X}_i)) = \frac{\sigma}{\hat{\sigma}}\frac{1}{\sqrt{\sigma^2 n}}\sum_{i=1}^{n}(\hat{\mu}_n(\boldsymbol{X}_i) - \mu_0(\boldsymbol{X}_i)),$$

kde první zlomek právě díky konzistenci odhadu $\hat{\sigma}_n^2$ a větě o spojité transformaci konverguje k jedné v pravděpodobnosti, a $\frac{\sigma}{\hat{\sigma}}\frac{1}{\sqrt{\sigma^2 n}}\sum_{i=1}^{n}(\hat{\mu}_n(\boldsymbol{X}_i) - \mu_0(\boldsymbol{X}_i))$ konverguje v distribuci díky větě 8 k normovanému normálnímu rozdělení. Cramér-Sluckého věta nám říká, že v tomto případě máme konvergenci

$$\frac{1}{\sqrt{\hat{\sigma}_n^2 n}}\sum_{i=1}^{n}(\hat{\mu}_n(\boldsymbol{X}_i) - \mu_0(\boldsymbol{X}_i)) = \frac{\sigma}{\hat{\sigma}}\frac{1}{\sqrt{\sigma^2 n}}\sum_{i=1}^{n}(\hat{\mu}_n(\boldsymbol{X}_i) - \mu_0(\boldsymbol{X}_i)) \xrightarrow{d} \mathcal{N}(0,1).$$

4. Simulační studie

V minulé kapitole jsme se věnovali asymptotickým vlastnostem sítového odhadu a ukázali jsme si, že sítový odhad je konzistentní a asymptoticky normální. Zároveň jsme ale zmiňovali, že tyto věty mají spíše teoretický přínos, protože v praxi často nejsme schopni vyřešit složitou optimalizační úlohu minimalizace čtvercové chyby, tudíž vypočtený odhad není ve skutečnosti tím sítovým z definice. Dále také nevíme, jak rychlá daná konvergence je, zda-li se na ni můžeme spolehnout v případě omezeného množství pozorovaných dat, protože v reálných aplikacích většinou nemáme k dispozici miliardy pozorování. V této kapitole se v krátké simulační studii podíváme právě na to, jestli námi v konečném kroku iterací numericky odhadnutý sítový odhad skutečně vykazuje známky konzistence či asymptotické normality.

4.1 Konzistence

V první části naší simulační studie jsme se zaměřili na konzistenci našeho sítového odhadu. V každé iteraci naší simulace jsme si vygenerovali odezvu dle modelu:

$$Y_i = \mu_0(\boldsymbol{X}_i) + \epsilon_i,$$

kde všechny složky $\boldsymbol{X}_i$ leží v [0,1], tak aby byly splněny předpoklady věty 6, a kde ϵ_i jsou nezávislé a stejně rozdělené z $\mathcal{N}(0{,}0.7^2)$ rozdělení. Pro různé délky výběrů jsme spočetli pomocí balíčku scipy v jazyce Python sítový odhad $\hat{\mu}_n$ a vypočetli hodnotu $\|\hat{\mu}_n - \mu_0\|_n$.

Pokud jde o volbu μ_0, tak jsme celkem zvolili čtyři funkce. Ve dvou případech se jednalo o funkci s interceptem a jednorozměrnou vysvětlující proměnnou, kdy jsme nejprve jako μ_0 zvolili funkci, která vznikla jako afinní kombinace dvou sigmoid funkcí, a v druhém případě se jednalo o afinní kombinaci goniometrických funkcí sinus. Další dvě funkce byly také afinními kombinacemi sinusoid, ale v tomto případě se již jednalo o funkce dvou proměnných a v případě funkce f_4 dokonce nediferencovatelné, jelikož jeden ze sinů měl v sobě maximum těchto dvou proměnných. Touto volbou jsme se chtěli přesvědčit, jestli si mělká lineární neuronová síť opravdu poradí s modelováním různých spojitých, klidně i nediferencovatelných funkcí, a zda-li takové závislosti je opravdu schopná v datech zaznamenat. Konkrétně jsme tedy jako funkci μ_0 využili tyto funkce:

$$f_1(x) = 1 + \frac{1}{1 + e^{x-\frac{1}{3}}} - \frac{1}{2(1 + e^{2x-1})}$$

$$f_2(x) = 1 + 2\sin\left(\frac{\pi}{3}x - \frac{\pi}{2}\right) + \sin\left(\frac{5\pi}{6}x + 1\right)$$

$$f_3(x,y) = 1 + 2\sin\left(\frac{\pi}{3}x - \frac{\pi}{2}\right) - 3\sin\left(\frac{\pi}{2}y - \frac{\pi}{2}\right) + \sin\left(1 + \frac{5\pi}{6}x - \frac{3\pi}{4}y\right)$$

$$f_4(x,y) = 1 + 2\sin\left(\frac{\pi}{3}x - \frac{\pi}{2}\right) - 3\sin\left(\frac{\pi}{2}y - \frac{\pi}{2}\right) + \sin\left(1 + \frac{\pi}{2}\max\{x,y\}\right).$$

Museli jsme také nastavit vhodné hodnoty parametrových posloupností, tak aby byly splněny předpoklady věty 6. Zde jsme zvolili stejné posloupnosti, jaké

zvolil Shen a kol. (2019), pro každé n jsme položili $d_n = n^{1/4}$,$\Delta_n = 10n^{1/4}$, $\tilde{\Delta}_n = n$. V tomto článku nebyla tedy uvedena hodnota členů posloupnosti $\tilde{\Delta}_n$, protože jsme ale potřebovali posloupnost, která roste a konverguje do nekonečna, aby byly splněny předpoklady naší věty o konvergenci, tak jsme si ji takto sami dodefinovali.

Takovýchto iterací jsme provedli 50 a následně jsme spočetli průměrné hodnoty pseudo-normy rozdílu sítového odhadu a skutečné funkce μ_0. Výsledky této simulace jsou k nalezení v tabulce (4.1). Můžeme vidět, že v případě všech čtyř funkcí vidíme stejný trend: s přibývajícím počtem pozorování klesá hodnota chybové pseudo-normy $\|.\|_n$, která by podle věty o konzistenci měla v pravděpodobnosti klesat k nule s rostoucím n. Zdá se tedy, že i na tomto relativně malém množství pozorování skutečně můžeme konvergenci očekávat.

V tabulce výsledků si také můžeme všimnout, že průměry pseudonorem rozdílů sítových odhadů funkcí dvou proměnných jsou znatelně vyšší než funkcí jednorozměrných pro všechny sledované délky náhodných výběrů. To se dá vysvětlit vyšším počtem odhadovaných parametrů v těchto modelech. Zajímavé je, že průměry pseudonorem $\|\hat{\mu}_n - f_3\|_n$ a $\|\hat{\mu}_n - f_4\|_n$ vyšly pro všechny n téměř identicky i přes to, že funkce f_4 nebyla diferencovatelná.

Délka n. výběru	$\|\hat{\mu}_n - f_1\|_n$	$\|\hat{\mu}_n - f_2\|_n$	$\|\hat{\mu}_n - f_3\|_n$	$\|\hat{\mu}_n - f_4\|_n$
500	0.057	0.069	0.106	0.100
1000	0.042	0.060	0.081	0.074
2000	0.028	0.041	0.061	0.059
5000	0.016	0.024	0.042	0.049

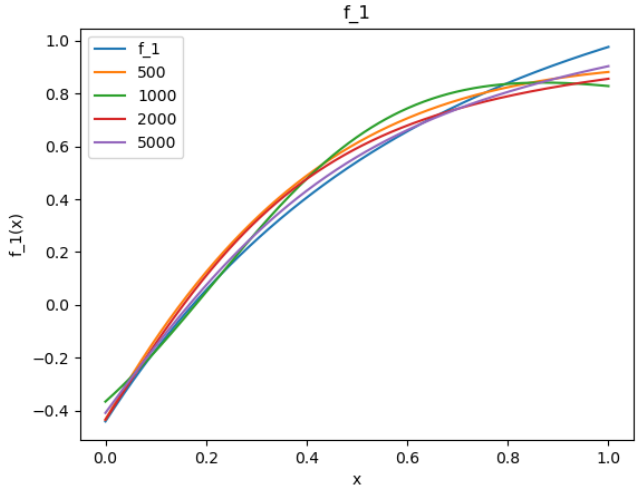
Tabulka 4.1: Průměrné pseudonormy pro 50 datových sad o různých velikostech

Pro jednu z iterací naší simulace jsme si vykreslili průběh sítového odhadu pro různá n a porovnali se skutečnými funkcemi f_1, f_2, f_3, f_4. Odhady funkcí f_1 a f_2 jsou vykresleny na obrázcích (4.1) a (4.2). V obou těchto případech vidíme, že nejlépe aproximují skutečné funkce skutečně sítové odhady $\hat{\mu}_{5000}$.

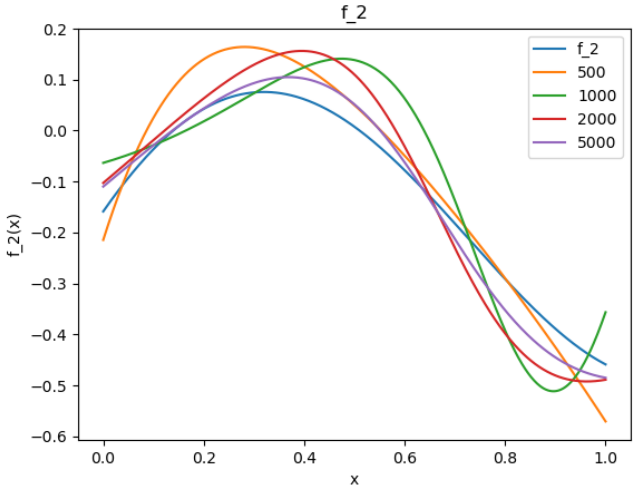
V případě funkce f_2, která má poměrně divoký nemonotónní průběh také vidíme, že všechny neuronové sítě si s touto nelineární, nemonotónní závislostí poradili. Opravdu se tedy ukazuje, že pomocí modelu hlubokého učení lze modelovat i složité nelineární závislosti odezvy na vysvětlujících proměnných i v případě menšího počtu pozorování.

V případě, že bychom se tuto závislost snažili modelovat zobecněným lineárním modelem pomocí vhodné transformace x_i vypozorované z popisných grafů, tak bychom se pravděpodobně uchýlili buď ke kvadratické transformaci nebo bychom se rozhodli použít lineární (nebo jiné) spliny s vnitřním uzlem někde v okolí maxima, zhruba 0.3, tedy volba tohoto uzlu by byla subjektivní. Díky využití neuronových sítí a modelu hlubokého učení jsme se hledání vhodné transformace vyhnuli a skutečnou funkci f_2 jsme aproximovali velmi dobře.

Vykreslit graf porovnávající kvalitu aproximace funkce dvou proměnných je poměrně složité, chtěli jsme se vyhnout porovnávání dvourozměrných ploch vykreslených na 3D grafech, proto jsme raději vykreslili funkce $f_3(x,\bar{y})$ a $f_4(x,\bar{y})$ jako funkce pouze jedné proměnné x s několika zafixovanými hodnotami $\bar{y}$. Čtyři takové grafy pro případ funkce f_3 si lze prohlédnout na obrázku (4.3). Hned si můžeme všimnout odlišného grafu funkce pro různé hodnoty y. Vypozorovat ta-

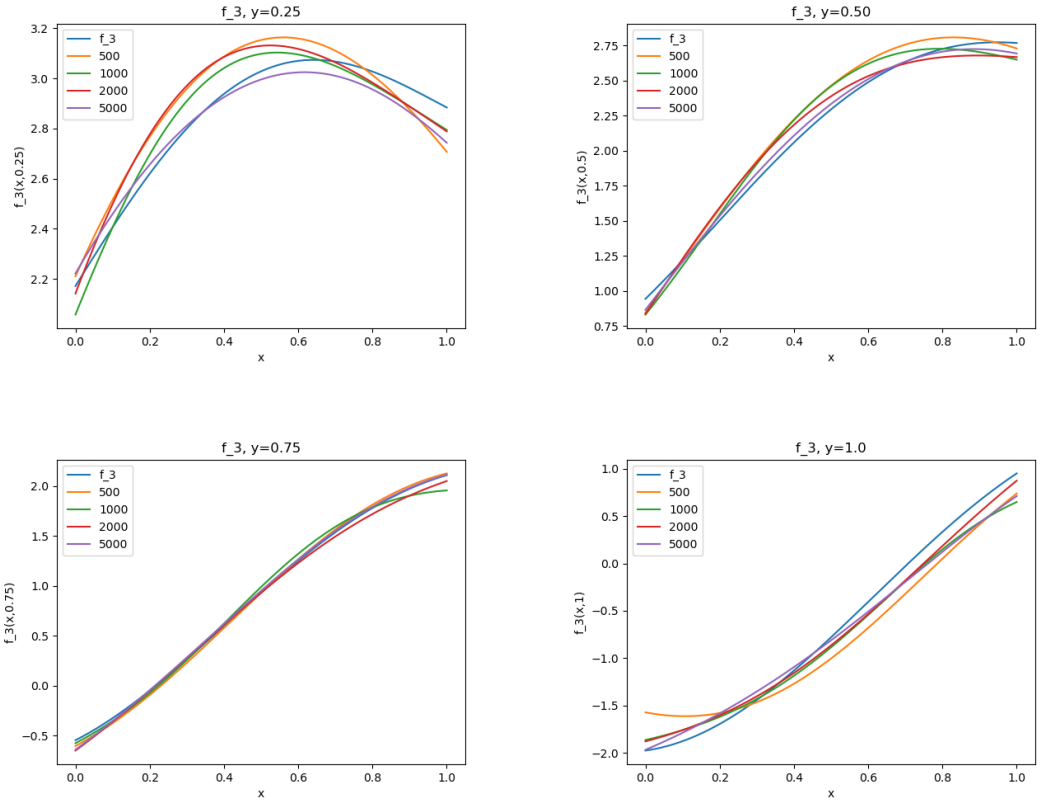


Obrázek 4.1: Sítové odhady f_1 pro různá n



Obrázek 4.2: Sítové odhady f_2 pro různá n

kové chování funkce $f_(x,y)$ by bylo velmi obtížné, tudíž manuální hledání vhodné transformace při použití zobecněného lineárního modelu by pravděpodobně vedlo k výrazně horšímu modelu. I v tomto případě vidíme, že nejlepší aproximace jsme opět dosáhli v případě nejvyššího počtu pozorování.



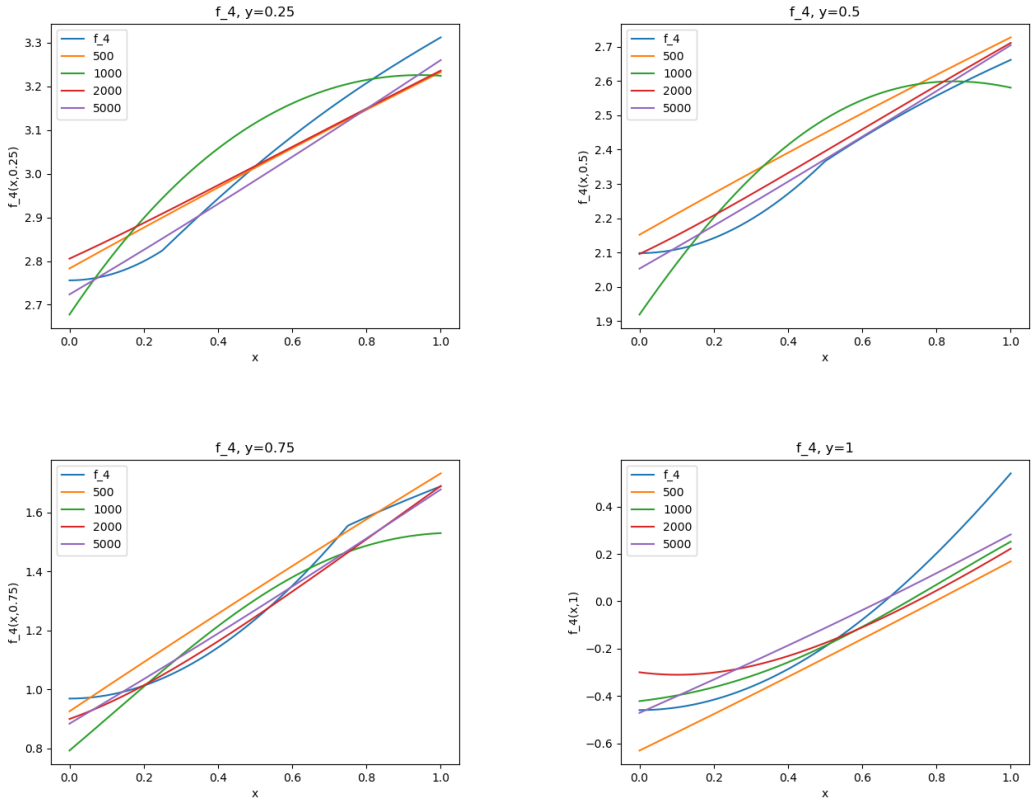
Obrázek 4.3: Sítové odhady f_3 pro různá n a různé hodnoty y

Na obrázku (4.4) vidíme vykreslené průběhy sítových odhadů funkce f_4, opět pro čtyři zafixované hodnoty y. Můžeme si opět všimnout rozdílných průběhů funkce pro různé hodnoty y a také nehladkých bodů způsobených nediferencovatelným maximem v předpisu funkce f_4. I v tomto případě vyšla jako nejlepší aproximace ta odhadnutá z 5000 pozorování, i zde se tedy zdá, že se můžeme spolehnout na větu o konzistenci dokázanou v kapitole 3. Je ale nutné poznamenat, že v tomto případě je kvalita všech aproximací výrazně horší než u ostatních příkladů funkcí, když porovnáme odhady ze stejné délky datové sady. Pravděpodobně vyšší počet parametrů a divočejší, navíc nehladký, průběh funkce f_4 je důvod, proč pro podobně dobré aproximace bychom vyžadovali vyšší počet pozorování.

4.2 Asymptotická normalita

V druhé části simulační studie jsme se zaměřili na asymptotickou normalitu sítového odhadu dokázanou ve větě 8. Pro každou z výše definovaných funkcí jsme si vygenerovali 5000 odezev odpovídajících následujícímu modelu:

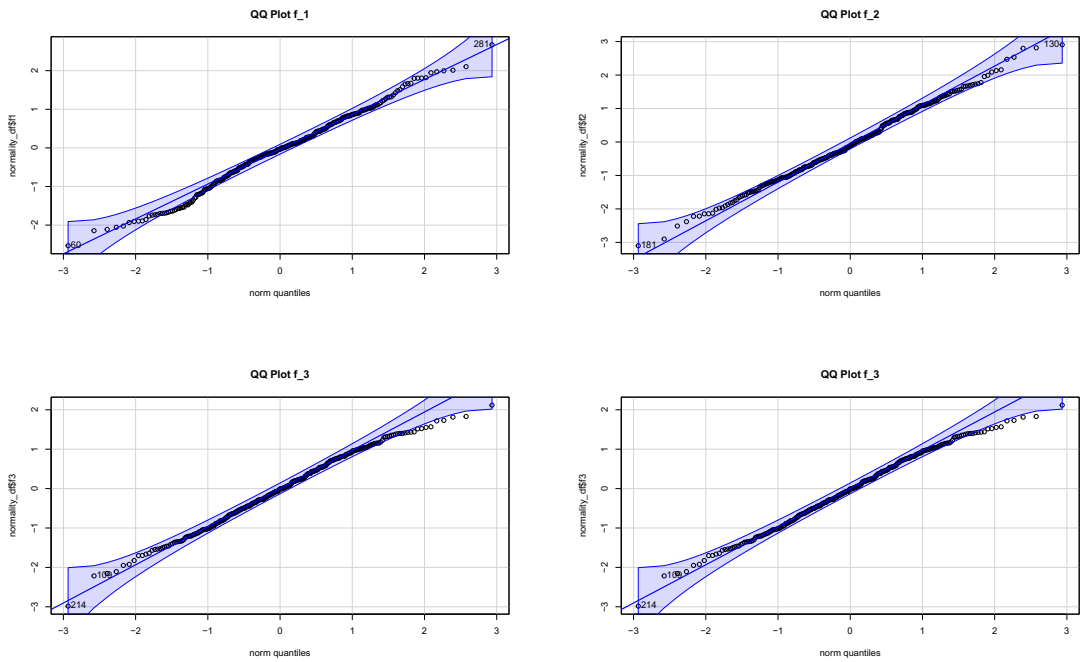
$$Y_i = \mu_0(\boldsymbol{X}_i) + \epsilon_i,$$



Obrázek 4.4: Sítové odhady f_4 pro různá n a různé hodnoty y

kde opět složky $\boldsymbol{X}_i$ leží pro každé i v intervalu [0,1]. Rozhodli jsme se ale změnit rozdělení epsilonů na $\mathcal{N}(0,1)$, které je nezávislé na $\boldsymbol{X}_i$ a to z důvodu absence nutnosti korigovat asymptotický rozptyl σ^2 pomocí odhadu $\hat{\sigma}_n^2$ představeném na konci třetí kapitoly. Následně jsme minimalizovali střední čtvercovou chybu, získali tak sítový odhad $\hat{\mu}_{5000}$, spočetli hodnotu statistiky $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{\mu}_n(\boldsymbol{X}_i) - \mu_0(\boldsymbol{X}_i))$. Tento proces jsme pro každou funkci opakovali 300krát. Následně jsme se si chtěli ověřit, zda statistika $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{\mu}_n(\boldsymbol{X}_i) - \mu_0(\boldsymbol{X}_i))$ opravdu vykazuje známky normálního rozdělení.

Na obrázku (4.5) nalezneme QQ-ploty těchto hodnot pro všechny čtyři funkce f_j. Můžeme vidět, že všechny čtyři grafy opravdu vykazují známky normality, jelikož pozorované hodnoty velmi věrně kopírují přímku, kde by měly ležet teoretické kvantily v případě normality.



Obrázek 4.5: Sítové odhady f_4 pro různá n a různé hodnoty y

Rozhodli jsme se i normalitu těchto vypočtených hodnot ověřit i pomocí formálního statistického testu normality. Konkrétně jsme se rozhodli využít Shapiro-Wilkův test normality. Nulová hypotéza tohoto testu je, že dané pozorovaní jsou normálně rozdělené, alternativa je její doplněk, tedy že pozorování nepochází z normálního rozdělení. Výsledné $p-$hodnoty jsou vypsány v tabulce (4.2). Testy normality pro všechny čtyři funkce nám dali stejný závěr, vysoké $p-$hodnoty nám nedovolují na 5% hladině zamítnout nulovou hypotézu, takže naše pozorování, že QQ-ploty na obrázku (4.5) potvrzují normalitu, se ukazuje jako správné. I v tomto případě se tedy ukazuje, že teoretickým větám o asymptotickém normálním rozdělení z kapitoly 3 můžeme věřit i při realistickém množství pozorování.

regresní funkce	$p-$hodnota
f_1	0.175
f_2	0.764
f_3	0.161
f_4	0.647

Tabulka 4.2: Výsledné $p-$hodnoty Shapiro-Wilkových testů normality

Závěr

V této práci jsme se zabývali neuronovými sítěmi a modelem hlubokého učení. V první kapitole jsme si nejprve tyto pojmy zadefinovali, kdy jsme si matematicky zavedli jednotlivé stavební bloky dopředných neuronových sítí, tedy neurony a vrstvy dopředných neuronových sítí. Při definování modelu hlubokého učení jsme se snažili co možná nejvíce samotnou definici přiblížit definici zobecněného lineárního modelu, protože se model hlubokého učení s dopřednou neuronovou sítí dá považovat za jakési zobecnění lineárně zobecněného modelu. Dále jsme si v první kapitole matematicky zavedli pojem aktivační funkce používaný v machine learningové literatuře a uvedli jsme si několik hojně používaných příkladů.

Ve druhé kapitole jsme se zaměřili na věty o univerzalitě neuronových sítí. Tyto věty pojednávají o flexibilitě neuronových sítí, co se týče schopnosti aproximovat libovolnou spojitou funkci. Po zavedení potřebných definic jsme se zaměřili na neuronové sítě z množiny mělkých lineárních sítí $\Sigma^{q_0}(\phi)$ s vhodnou aktivační funkcí ϕ. Důkaz byl proveden tak, že jsme se odkázali na větu o univerzalitě pro větší množinu neuronových sítí $\Sigma\Pi^{q_0}(\phi)$, kde se předpoklady na funkci ϕ sice lišily, ale následně se využilo gonionometrických identit, které umožňují vyjádřit součty součinů funkcí kosinus jako součty lineárně transformovaných kosinů, po této úpravě již předpoklady byly splněny i pro funkce z $\Sigma^{q_0}(\phi)$. Dokázali jsme tedy, že libovolnou spojitou funkci lze aproximovat s dostatečně širokou mělkou lineární sítí s aktivační funkcí jako je například funkce sigmoid či hyperbolický tangent, které jsou často při konstrukci neuronových sítí používány.

Věty o univerzalitě nám ale neřekli nic o tom, zda-li jsme schopni pomocí modelu hlubokého učení z dat tuto spojitou závislost podmíněné střední hodnoty na vysvětlujících proměnných konzistentně modelovat. Po každém smysluplném odhadu ve statistice požadujeme, aby byl konzistentní, tedy abychom se s rostoucím počtem pozorování blížili skutečné hodnotě parametru či funkce. Rádi bychom vlastnost konvergence měli i v případě neuronových sítí. Zadefinovali jsme si tzv. sítový odhad neuronové sítě, což je vlastně posloupnost odhadů, která pro každé n minimalizuje čtvercovou chybu přes všechny pozorování od jedné právě do n. Protože hledání funkce minimalizující čtvercovou chybu přes všechny spojité funkce by vedlo k funkci, která by trpěla na tzv. přeučení, měla by velmi špatné predikční vlastnosti pro pozorování, která nebyla přítomna v dostupném náhodném výběru při hledání sítového odhadu pro dané n. My jsme tedy hledali jednotlivé složky sítového odhadu v podmnožinách mělkých lineárních sítí $\mathcal{S}_n(\phi) \in \Sigma(\phi)$, které byly definovány pomocí posloupností parametrů. Za doplňujících předpokladů jsme dokázali, že tento sítový odhad existuje a je konzistentní. V druhé části třetí kapitoly jsme si ještě formulovali větu o asymptotické normalitě sítového odhadu, která nám dává asymptotické rozdělení (tedy i asymptotický rozptyl) sítového odhadu neuronových sítí. Znalost asypmtotického rozdělení umožňuje testování nulové hypotézy, že $\mathsf{E}[Y \mid \boldsymbol{X}] = \mu_0$ pro nějakou spojitou funkci μ_0.

V poslední části této práce byla provedena krátká simulační studie, ve které jsme si ověřili, jestli je dokázaná konzistence spolehlivá i při relativně malém počtu pozorování v řádu stovek a nižších tisíců. Vygenerovali jsme si celkem čtyři náhodné výběry, kdy pokaždé byla podmíněná střední hodnota odezvy jinou funkcí vysvětlujících veličin. Ve všech čtyřech případech byla konvergence jasně vidi-

telná, což vzhledem k poměrně divokému chování poslední funkce, která byla nediferencovatelná a nemonotónní funkce dvou proměnných, považujeme za úspěch. Na posledních několika stranách této práce jsme si ještě na vygenerovaných datech ověřili asymptotickou normalitu formulovanou ve třetí kapitole. Normalita byla z QQ-grafů velmi viditelná a dokonce ji nezamítnul ani statistický Shapiro-Wilkův test.

Jelikož neuronové sítě získávají popularitu až v posledních letech, tak teorie okolo nich není zatím vybudována na pevných matematických základech. Za přínos této práce považujeme snahu podívat se na hluboké učení jako na statistický model velmi podobný zobecněnému lineárnímu modelu. Porovnali jsme také tzv. aktivační funkce, které byly porovnány s linkovými funkcemi známými právě z teorie zobecněných lineárních modelů. U moderní funkce *swish* jsme dokázali, že její hodnota při volbě parametru $\beta = 0$ splývá s definicí lineární funkce a v případě limitního chování pro β jdoucí do nekonečna zase splývá se známou ReLU funkcí. Pokud jde o věty o univerzalitě, tak důkazy vět s našimi předpoklady nebyly v dostupné literatuře nikde pořádně sepsané, v článcích se často objevovaly jen různé náčrty důkazů, které se v kruzích odkazovaly na další články. My jsme jeden takový důkaz plně sepsali. Obdobný případ byl důkaz věty o konzistenci, kde jsme navíc dokázali, že všechny předpoklady White-Woolridgeovy věty jsou skutečně za předpokladů věty o konzistenci splněny. Během důkazu věty o konzistenci bylo třeba dokázat několik kratších tvrzení nebo dát jejich důkazy dostupné v literatuře řádně dohromady. Články ostatních autorů často obsahovaly na závěr simulační studie na vygenerovaných datech, ale jednalo se vždy o jednorozměrné případy, my jsme nejen vyzkoušeli konvergenci sítového odhadu na vícerozměrné funkci, ale dokonce jsme se rozhodli použít k ukázce i nehladkou funkci více proměnných.

Všechny dokázané věty byly dokázány pouze pro speciální případ neuronových sítí a jejich jednoduchá struktura byla v důkazech využívána. Do budoucna se samozřejmě nabízí se zaměřit buď na další třídu neuronových sítí či nějakou větu zobecnit například na všechny mělké neuronové sítě. Využívali jsme zároveň i předpokladů na homoskedasticitu a nezávislost náhodných chyb ϵ na vysvětlujících proměnných, odstranění těchto předpokladů by také mohlo být obsahem pozdějšího výzkumu.

Seznam použité literatury

Clevert, D.-A., Unterthiner, T. a Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*.

Elbrächter, Dennis, e. a. (2019). Deep neural network approximation theory. *arXiv preprint arXiv:1901.02220*.

Hornik, K., Stinchcombe, M. a White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, **5**(5), 359–366.

Shen, X., Jiang, C., Sakhanenko, L. a Lu, Q. (2019). Asymtotic properties of neural network sieve estimators. *Journal of Nonparametric Statistics*.

White, H. a Wooldridge, J. M. (1991). *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge University Press,. ISBN 9780521424318.

Würthrich, M. a Merz, M. (2022). *Statistical Foundations of Acturial Learning and Its Applications*. Springer International Publishing. ISBN 2523-3270.

Xu, B., Wang, N., Chen, T. a Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.