MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Aibat Kossumov

Stejnoměrný zákon velkých čísel, VC dimenze a strojové učení

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. Ing. Marek Omelka, Ph.D
Studijní program: Matematika
Studijní obor: Obecná matematika

Praha 2022

Prohlašuji, že jsem tuto bakalářskou 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

Chtěl bych poděkovat svému vedoucímu bakalářské práce doc. Ing. Markovi Omelkovi, Ph.D., za velké množství konzultací a cenných rad, které mně dal, a za čas, který mně věnoval. Dále bych chtěl poděkovat své matce za její obrovskou podporu a za možnost vystudovat v Praze. V neposlední řadě bych rád poděkoval svému kamarádovi Vojtovi Urxovi za pomoc, kterou mně poskytoval během studia.

I would like to thank my bachelor thesis supervisor doc. Ing. Marek Omelka, Ph.D., for many consultations and valuable advice he gave me and the time he gave me. I would also like to thank my mother for her huge support and for the opportunity to study in Prague. Last but not least, I would like to thank my friend Vojtěch Urx for the help he provided me during my studies.

Název práce: Stejnoměrný zákon velkých čísel, VC dimenze a strojové učení

Autor: Aibat Kossumov

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. Ing. Marek Omelka, Ph.D, Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V této práci se zabýváme zobecněnou Glivenkovou-Cantelliho větou a její aplikací v matematických základech strojového učení. Nejprve dokážeme zobecněnou Glivenkovu-Cantelliho větu pomocí pokrývacích čísel a lemmatu o symetrizaci. Dále vyslovíme stejnoměrný zákon velkých čísel. Následně budeme se zabývat Vapnikovými-Červonenkisovými třídami funkcí (VC třídami). Ukážeme, že pro VC třídy jsou pokrývací čísla stejnoměrně omezená. Nakonec popíšeme úlohu strojového učení a uvedeme příklad jedné konkretní úlohy, která se dá „naučit“. Hlavní aplikací bude dokázat základní větu statistického učení. Obvykle je tato věta dokazovaná pro třídy prediktorů, které jsou tzv. Probably Aproximately Correct learnable (PAC learnable). V této práci zesílíme vlastnost PAC learnable a dokážeme pro ni základní větu statistického učení.

Klíčová slova: Glivenkova-Cantelliho věta, pokrývací čísla, stejnoměrný zákon velkých čísel, VC třídy, PAC learnable, základní věta statistického učení

Title: Uniform law of large numbers, VC dimension and machine learning

Author: Aibat Kossumov

Department: Department of probability and mathematical statistics

Supervisor: doc. Ing. Marek Omelka, Ph.D, Department of probability and mathematical statistics

Abstract: In this thesis we study the generalized Glivenko-Cantelli theorem and its application in mathematical foundations of machine learning. Firstly, we prove the generalized Glivenko-Cantelli's theorem using covering numbers and lemma of symmetrization. Next we show the uniform law of large numbers. Then, we deal with Vapnik-Chervonenkis classes of functions (VC classes). We show that for VC classes covering numbers are uniformly bounded. Finally, we describe the task of machine learning and give an example of one specific task that can be "learned". The main application will be to prove the fundamental theorem of statistical learning. Usually this theorem is proved for classes of predictors that are Probably Approximately Correct learnable (PAC learnable). In this work we strengthen the property of PAC learnable and for it we prove the basic theorem of statistical learning.

Keywords: Glivenko-Cantelli theorem, covering numbers, uniform law of large numbers, VC classes, PAC learnable, fundamental theorem of statistical learning

Obsah

Úvod

Začneme motivací k problému, který budeme zkoumat. Mějme nezávislé stejně rozdělené náhodné veličiny $X_1, ..., X_n$ (kde $n \in \mathbb{N}$) s distribuční funkcí F a rozdělením P. Uvažujme empirickou distribuční funkci

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\}, \text{ kde } t \in \mathbb{R}.$$

Podle silného zákonu velkých čísel pro $\forall t \in \mathbb{R}$ platí

$$F_n(t) \xrightarrow[n\to\infty]{s.j.} F(t).$$

Klasická Glivenkova-Cantelliho věta říká, že F_n konverguje k F stejnoměrně na $\mathbb{R}$, tj.

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{s.j.} 0. \tag{1}$$

Všimněme si, že není jasné zda je vůbec levá strana v (1) měřitelná, neboť supremum bereme přes nespočetnou množinu $\mathbb{R}$. Později v práci ukážeme, že levá strana v (1) je měřitelná pro $\forall n \in \mathbb{N}$ (viz. poznámka 1.5). Proto konvergence skoro jistě ve výrazu (1) má smysl.

Uvažujme následující třídu funkcí $\mathcal{T} := \Big\{\mathbb{1}\{x \leq t\} : t \in \mathbb{R}\Big\}$. Všimněme si, že výraz (1) lze přepsat následovně

$$\sup_{f\in\mathcal{T}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_1) \right| \xrightarrow[n\to\infty]{s.j.} 0. \tag{2}$$

Různá zobecnění Glivenkovy-Cantelliho věty udávají postačující podmínky pro třídu reálných funkcí $\mathcal{F}$, aby platilo

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_1) \right| \xrightarrow[n\to\infty]{s.j.^*} 0,$$

kde $s.j.^*$ je zobecnění konvergence skoro jistě (viz. definice 1.3). Toto zobecnění zavádíme protože, pokud třída funkcí $\mathcal{F}$ je nespočetná, potom supremum přes třídu funkcí $\mathcal{F}$ nemusí být měřitelným zobrazením.

V první kapitole zformulujeme zobecněnou Glivenkovou-Cantelliho větu pomocí tzv. pokrývacích čísel (viz. definice 1.7). Dále vysvětlíme, co znamená, že třída funkcí je stejnoměrná Glivenkova-Cantelliho (viz. definice 1.12). Na konci první kapitoly zformulujeme tzv. stejnoměrný zákon velkých čísel (tj. zákon velkých čísel, který funguje stejnoměrně přes třídu všech pravděpodobnostních měr viz. věta 1.7).

V druhé kapitole zavedeme Vapnikovy-Červonenkisovy třídy množin neboli VC třídy (viz. definice 2.1). Dále rozšíříme pojem VC třídy pro funkce. Na konci druhé kapitoly uvedeme vztah mezi VC třídami funkcí a Glivenkovými-Cantelliho třídami funkcí (viz. věta 2.8, věta 2.7).

V třetí kapitole ukážeme aplikace prvních dvou kapitol v matematických základech strojového učení. Nejprve popíšeme úlohu strojového učení. Dále zavedeme formální definice „učení" (viz. definice 3.1). Jako hlavní aplikace dokážeme základní větu statistického učení (viz. věta 3.2).

1. Stejnoměrný zákon velkých čísel

1.1 Vnější pravděpodobnost a zobecnění konvergence skoro jistě

V této kapitole vysvětlíme, co přesně znamená symbol $\xrightarrow[n\to\infty]{s.j.^*}$. Potom ukážeme proč klasická Glivenkova-Cantelliho věta je speciálním případem zobecněné věty (viz. poznámka 1.1).

Během celé práce se budeme setkávat s množinami, které nemusí být měřitelné. Proto připomeňme definici vnější pravděpodobnosti.

Definice 1.1. *Nechť $(\Omega, \mathcal{A}, \mathbb{P})$ je pravděpodobnostní prostor. Pro libovolnou množinu $B \subset \Omega$ definujme její vnější pravděpodobnost (outer probability) jako*

$$\mathbb{P}^*(B) = \inf\{\mathbb{P}(A) : B \subset A, A \in \mathcal{A}\}.$$

Všimněme si, že vnější pravděpodobnost opravdu zobecňuje pravděpodobnost, protože pro každou měřitelnou množinu B platí $\mathbb{P}^*(B) = \mathbb{P}(B)$.

Často budeme pracovat se zobrazeními, které jsou definovány na pravděpodobnostních prostorech, ale nemusí být měřitelné. To je motivací pro následující definici.

Definice 1.2. *Nechť $(\Omega, \mathcal{A}, \mathbb{P})$ je pravděpodobnostní prostor, nechť S je libovolná množina. Pak libovolné zobrazení $Y : \Omega \to S$ nazveme náhodným zobrazením (random map).*

Nechť $X_1, ..., X_n$ je náhodný výběr. Nechť f je reálná měřitelná funkce. Potom pro libovolné reálné číslo $a \in \mathbb{R}$ zobrazení

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - a \right|$$

je náhodná veličina (protože absolutní hodnota je spojitá funkce odsud měřitelná). Předpokládejme, že $\mathcal{F}$ je nespočetná třída reálných měřitelných funkcí. Potom

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - a \right|$$

nemusí být náhodnou veličinou (protože supremum přes nespočetnou třídu nemusí být měřitelným), ale je náhodným zobrazením.

Také zobecníme konvergence *skoro jistě* a *v pravděpodobnosti.* V dalším textu $\mathbb{D}$ značí metrický prostor s metrikou d (v našem případě bude typicky $\mathbb{D} = \mathbb{R}^m$ s euklidovskou metrikou, kde $m \in \mathbb{N}$).

Definice 1.3. *Nechť $X_1, X_2, ..., X$ jsou náhodná zobrazení definovaná na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$ s hodnotami v $(\mathbb{D}, d)$.*

(i) Říkáme že X_n konverguje ve vnější pravděpodobnosti (in outer probability) k X, jestliže pro každé $\eta > 0$ platí $P^*\Big(d(X_n,X) > \eta\Big) \underset{n\to\infty}{\longrightarrow} 0$. Tato konvergence se značí jako $X_n \xrightarrow[n\to\infty]{P^*} X$.

(ii) Říkáme, že X_n konverguje zvnějšku skoro jistě (outer almost surely) k X, jestliže existuje posloupnost měřitelných náhodných veličin $\{\Delta_n\}$ takových, že $d(X_n,X) \leq \Delta_n$ a $\Delta_n \xrightarrow[n\to\infty]{s.j.} 0$. Tato konvergence se značí jako $X_n \xrightarrow[n\to\infty]{s.j.^*} X$.

Rozdíl předchozí definice od klasické definice konvergence skoro jistě a v pravděpodobnosti je v tom, že ani množina $\{d(X_n,X) > \eta\}$, ani funkce $d(X_n,X)$ nemusí být měřitelné (protože X_n a X nejsou nutně náhodné veličiny). Pokud $X_1, X_2, ..., X$ jsou náhodné veličiny, potom se předchozí definice shoduje s klasickou definicí konvergence skoro jistě a v pravděpodobnosti.

Nyní již víme co znamená

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_1)\right| \xrightarrow[n\to\infty]{s.j.^*} 0.$$

Poznámka 1.1. Teď ukážeme, že klasická Glivenkova-Cantelliho věta je speciálním případem zobecněné Glivenkovy-Cantelliho věty.

Mějme reálný náhodný výběr $X_1, ..., X_n$. Nechť F je distribuční funkce tohoto náhodného výběru. Uvažujme následující třídu měřitelných funkcí

$$\mathcal{T} = \Big\{\mathbb{1}\{x \leq t\} : t \in \mathbb{R}\Big\}.$$

Potom platí

$$\sup_{f\in\mathcal{T}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_1)\right| = \sup_{t\in\mathbb{R}}\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\} - \mathbb{E}[\mathbb{1}\{X_1 \leq t\}]\right|$$
$$= \sup_{t\in\mathbb{R}}|F_n(t) - F(t)|.$$

Ukazuje se, že $\sup_{t\in\mathbb{R}}|F_n(t) - F(t)|$ jsou měřitelné pro každé $n \in \mathbb{N}$ (viz. poznámka 1.5). Proto platí následující ekvivalence:

$$\sup_{f\in\mathcal{T}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_1)\right| \xrightarrow[n\to\infty]{s.j.^*} 0 \iff \sup_{t\in\mathbb{R}}|F_n(t) - F(t)| \xrightarrow[n\to\infty]{s.j.} 0.$$

V předchozím textu jsme zavedli pojem náhodného zobrazení. Všimněme si, že pro náhodné zobrazení nemusí být definována střední hodnota. Proto zobecníme pojem střední hodnoty.

Dále budeme používat následující značení:

- $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$.
- $\mathcal{B}$ značí borelovskou $\sigma-$algebru na $\mathbb{R}$.
- $\mathcal{B}^n$ značí borelovskou $\sigma-$algebru na $\mathbb{R}^n$.
- $\overline{\mathcal{B}} = \sigma\Bigg(\mathcal{B} \cup \Big\{\{+\infty\},\{-\infty\}\Big\}\Bigg)$.

- Nechť $(S, \mathcal{S})$ je měřitelný prostor. Pak symbolem $\mathcal{S}^n$ značíme součinovou $\sigma-$algebru na S^n.

Definice 1.4. *Nechť $(\Omega, \mathcal{A}, \mathbb{P})$ je pravděpodobnostní prostor a $Y : \Omega \to \overline{\mathbb{R}}$ je náhodné zobrazení. Potom vnější střední hodnotou (outer expectation) Y nazýváme*

$$\mathbb{E}^*[Y] = \inf \Big\{ \mathbb{E}[U] \mid U \geq Y,\ U : (\Omega, \mathcal{A}) \to (\overline{\mathbb{R}}, \overline{\mathcal{B}}),\ \mathbb{E}[U] \ \textit{existuje} \Big\},$$

kde $U : (\Omega, \mathcal{A}) \to (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ znamená, že U je měřitelné zobrazení.

Poznámka 1.2. Mezi vnější pravděpodobností a vnější střední hodnotou platí následující vztah $\mathbb{E}^*[\mathbb{1}_B] = \mathbb{P}^*(B)$, kde $B \subset \Omega$. Důkaz je proveden v knize van der Vaart a Wellner (1996), lemma 1.2.3.(i).

Všimněme si, že pokud Y je náhodná veličina a $\mathbb{E}[Y]$ existuje, pak

$$\mathbb{E}[Y] = \mathbb{E}^*[Y].$$

Důležitým faktem je, že množina v definici 1.4 obsahuje své infimum, což je tvrzením následující věty.

Věta 1.1. *Nechť $S : \Omega \to \mathbb{R}$ je náhodné zobrazení, potom existuje měřitelná funkce S^* (tzv. minimální měřitelná majoranta) taková, že $S^*(\omega) \geq S(\omega)$ pro každé $\omega \in \Omega$ a $S^* \leq U$ s.j. pro každou měřitelnou funkce $U : \Omega \to \mathbb{R}$ splňující $U \geq S$ s.j. Pro libovolnou takovou funkce S^* platí $\mathbb{E}^*[S] = \mathbb{E}[S^*]$ za předpokladu že $\mathbb{E}[S^*]$ existuje.*

Důkaz je proveden v knize van der Vaart a Wellner (1996), lemma 1.2.1.

Lemma 1.2. *Nechť $S, T : \Omega \to \mathbb{R}$ jsou náhodné zobrazení, potom*

- $\forall c \in \mathbb{R} : \mathbb{P}^*(S > c) = \mathbb{P}(S^* > c).$
- $\forall \alpha, \beta \in \mathbb{R} : (\alpha S + \beta T)^* \leq \alpha S^* + \beta T^*.$

Důkaz lze najít v knize van der Vaart a Wellner (1996), lemma 1.2.2.

Podle poznámky 1.2 plyne, že množina v definici 1.1 také obsahuje své infimum.

Pro formulaci zobecněné Glivenkovy-Cantelliho věty také potřebujeme mít nějakou představu o „rychlosti“ konvergenci na pravděpodobnostním prostoru.

Definice 1.5. *Nechť $\{X_n\}_{n\in\mathbb{N}}$ je posloupnost náhodných zobrazení. Řekněme, že posloupnost $\{X_n\}_{n\in\mathbb{N}}$ je $o_{P^*}(1)$, píšeme $X_n = o_{P^*}(1)$, jestliže platí*

$$X_n \xrightarrow{P^*} 0.$$

Nechť $\{r_n\}_{n\in\mathbb{N}}$ je posloupnost kladných reálných čísel. Řekněme, že posloupnost $\{X_n\}_{n\in\mathbb{N}}$ je $o_{P^*}(r_n)$, píšeme $X_n = o_{P^*}(r_n)$, jestliže existuje posloupnost náhodných zobrazení $\{Y_n\}_{n\in\mathbb{N}}$ taková, že

$$X_n = r_n Y_n \ \textit{a} \ Y_n = o_{P^*}(1).$$

Pokud X_n je posloupnost reálných čísel, potom se předchozí definice shoduje s definicí malého o z matematické analýzy (viz. Ghorpade a Limaye, 2006, str. 53 poznámka 2.11).

1.2 Lemma o symetrizaci

V této kapitole dokážeme technický výsledek tzv. lemma o symetrizaci (viz. lemma 1.3), které využijeme v důkazu zobecněné Glivenkovy-Cantelliho věty.

Definice 1.6. *Nechť* $(\Omega, \mathcal{A}, P)$ *je pravděpodobnostní prostor,* $(S, \mathcal{S})$ *je měřitelný prostor,* $A \in \mathcal{S}$. *Nechť* $X_1, ..., X_n$ *jsou nezávislé stejně rozdělené náhodné veličiny na* $(\Omega, \mathcal{A}, P)$ *s hodnotami v* $(S, \mathcal{S})$. *Empirickou mírou (empirical measure) nazýváme náhodnou pravděpodobnostní míru na* $(S, \mathcal{S})$

$$P_n(A, \omega) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A\Big(X_i(\omega)\Big) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(\omega)}(A),$$

kde $\mathbb{1}_A$ *je charakteristická funkce množiny* $A \in \mathcal{S}$ *a* $\delta_{X_i(\omega)}$ *je Diracova míra v bodě* $X_i(\omega)$.

Nechť $X_1, ..., X_n$ je reálný náhodný výběr, pak jeho empirickou distribuční funkci budeme nazývat následující zobrazení

$$F_n(t, \omega) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i(\omega) \leq t\}, \text{ kde } t \in \mathbb{R},\ \omega \in \Omega.$$

Všimněme si, že empirická distribuční funkce F_n souvisí s empirickou mírou P_n následovně:

$$\forall t \in \mathbb{R},\ \forall \omega \in \Omega :\ P_n\Big((-\infty, t], \omega\Big) = F_n(t, \omega).$$

Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor, $(S, \mathcal{S})$ je měřitelný prostor. V následujícím textu (nebude-li řečeno jinak) jsou $X_1, ..., X_n$ nezávislé stejně rozdělené náhodné veličiny definované na $(\Omega, \mathcal{A}, P)$ s hodnotami v $(S, \mathcal{S})$ a P_n je jejich empirická míra. Dále symbolem P_{X_1} budeme označovat obraz míry P v zobrazení X_1. Tedy

$$\forall B \in \mathcal{S} : P_{X_1}(B) = P\{\omega \in \Omega : X_1(\omega) \in B\}.$$

Nechť $f : (S, \mathcal{S}) \longrightarrow (\mathbb{R}, \mathcal{B})$ je měřitelná funkce taková že $\mathbb{E}\Big[f(X_1)\Big]$ existuje. Pak budeme používat následující značení

$$P_n(f) := \int_S f(s) dP_n(s),$$

$$P(f) := \int_S f(s) dP_{X_1}(s).$$

Všimněme si, že

$$P_n(f) = \int_S f(s) dP_n(s) = \int_S f(s) d\left(\frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}\right)(s) = \frac{1}{n} \sum_{i=1}^{n} \int_S f(s) d\delta_{X_i}(s)$$
$$= \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$
$$P(f) = \int_S f(s) dP_{X_1}(s) = \mathbb{E}\Big[f(X_1)\Big].$$

Dále (nebude-li řečeno jinak) bude $\mathcal{F}$ daná třída měřitelných funkcí z $(S, \mathcal{S})$ do $(\mathbb{R}, \mathcal{B})$. Pro třídu funkcí $\mathcal{F}$ budeme používat následující značení

$$||P_n - P||_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |P_n(f) - P(f)|.$$

V dalším textu nezávislost náhodných veličin (náhodných vektorů) označovat symbolem $\perp$.

V důkazu zobecněné Glivenkovy-Cantelliho věty se využívá tzv. symetrizace. Symetrizace je speciální technika, která spočívá v tom, že místo $||P_n - P||_{\mathcal{F}}$ nám stačí kontrolovat

$$||P_n^0||_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right|,$$

kde $\varepsilon_1, ..., \varepsilon_n$ je náhodný výběr definovaný na nějakém pravděpodobnostním prostoru $(\Omega_\varepsilon, \mathcal{A}_\varepsilon, P_\varepsilon)$ takový že

$$P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = \frac{1}{2},$$

přičemž $(\varepsilon_1, ..., \varepsilon_n) \perp (X_1, ..., X_n)$ na $(\Omega \times \Omega_\varepsilon, \mathcal{A} \otimes \mathcal{A}_\varepsilon, P \otimes P_\varepsilon)$. Veličiny $\varepsilon_1, ..., \varepsilon_n$ nazýváme *Rademacherovými náhodnými veličinami (Rademacher random variables)*. Více informace o symetrizaci lze najít v knize van der Vaart a Wellner (1996), kapitola 2.3.

Lemma 1.3. ***(o symetrizaci)*** *Nechť P_n je empirická míra nezávislých stejně rozdělených náhodných veličin $X_1, ..., X_n$. Předpokládejme, že $\mathcal{F}$ je třída reálných měřitelných funkcí shora omezených konečnou konstantou M tj.*

$$\forall f \in \mathcal{F}\ \forall x \in S : |f(x)| \leq M.$$

Potom pro $n \geq \frac{8M^2}{\eta^2}$ platí

$$P^*\Big(||P_n - P||_{\mathcal{F}} > \eta\Big) \leq 4\ P^*\left(||P_n^0||_{\mathcal{F}} > \frac{\eta}{4}\right).$$

Poznámka 1.3. Všimneme si, že zatímco vnější pravděpodobnost na levé straně nerovnosti se počítá vzhledem k $X_1, ..., X_n$, vnější pravděpodobnost na pravé straně se počítá vzhledem k sdruženému rozdělení $X_1, ..., X_n$ a $\varepsilon_1, ..., \varepsilon_n$.

Důkaz. Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor a $\mathbb{X} := (X_1, ..., X_n)$ je náhodný vektor takový, že

$$\mathbb{X} : (\Omega, \mathcal{A}, P) \longrightarrow (S^n, \mathcal{S}^n).$$

Předpokládejme, že $\mathbb{X}' := (X_1', ..., X_n')$ je nezávislá kopie náhodného vektoru $\mathbb{X}$. To jest existuje pravděpodobnostní prostor $(\widetilde{\Omega}, \widetilde{\mathcal{A}}, \widetilde{P})$ takový, že

$$\mathbb{X}' : (\widetilde{\Omega}, \widetilde{\mathcal{A}}, \widetilde{P}) \longrightarrow (S^n, \mathcal{S}^n).$$

Tedy $\mathbb{X} \perp \mathbb{X}'$ na součinovém pravděpodobnostním prostoru

$$(\Omega \times \widetilde{\Omega}, \mathcal{A} \otimes \widetilde{\mathcal{A}}, P \otimes \widetilde{P}).$$

V důkazu použijeme následující značení:

- $\mathbb{P} = P \otimes \widetilde{P}$.
- $\mathbb{E}_P$ značí střední hodnotu počítanou vzhledem k $(\Omega, \mathcal{A}, P)$.
- $\mathbb{E}^*_P$ značí vnější střední hodnotu počítanou vzhledem k $(\Omega, \mathcal{A}, P)$.

Analogicky budeme používat značení pro $\mathbb{E}_{\widetilde{P}}$, $\mathbb{E}^*_{\widetilde{P}}$, $\mathbb{E}_{\mathbb{P}}$, $\mathbb{E}^*_{\mathbb{P}}$.

Nechť P_n' je empirická míra odpovídající náhodnému výběru $X_1', ..., X_n'$. Potom pro každou funkci $f \in \mathcal{F}$

$$\widetilde{P}\left(|P_n'(f) - P(f)| \geq \frac{\eta}{2}\right) = \widetilde{P}\left\{\widetilde{\omega} \in \widetilde{\Omega} : \left|\frac{1}{n}\sum_{i=1}^n f(X_i'(\widetilde{\omega})) - P(f)\right| \geq \frac{\eta}{2}\right\}.$$

Z Čebyševovy nerovnosti (viz. lemma A.1) platí, že

$$\widetilde{P}\left\{\widetilde{\omega} \in \widetilde{\Omega} : \left|\frac{1}{n}\sum_{i=1}^n f(X_i'(\widetilde{\omega})) - P(f)\right| \geq \frac{\eta}{2}\right\} \leq \frac{\operatorname{var}\left(f(X_1')\right)}{n(\eta/2)^2} \leq \frac{\mathbb{E}_{\widetilde{P}}\left[f(X_1')\right]^2}{n(\eta/2)^2}.$$

Z omezenosti funkcí v $\mathcal{F}$ dostáváme, že $\mathbb{E}_{\widetilde{P}}\left[f(X_1')\right]^2 \leq M^2$. To jest pro $n \geq \frac{2M^2}{\eta^2}$

$$\widetilde{P}\left(|P_n'(f) - P(f)| \geq \frac{\eta}{2}\right) \leq \frac{M^2}{n\eta^2} \leq \frac{1}{2}. \tag{1.1}$$

Uvažujme následující podmnožinu $\left\{||P_n - P||_{\mathcal{F}} > \eta\right\}$ (nemusí být měřitelná) množiny Ω. Zafixujme $\omega \in \left\{||P_n - P||_{\mathcal{F}} > \eta\right\}$, pak existuje (z vlastnosti suprema) $f^\omega \in \mathcal{F}$ (horní index ukazuje závislost této funkce na ω) takové, že

$$|P_n(f^\omega)(\omega) - P(f^\omega)| > \eta.$$

Dále na jev

$$\left\{|P_n'(f^\omega) - P_n(f^\omega)(\omega)| > \frac{\eta}{2}\right\}$$

se díváme pro pevné $\omega \in \left\{||P_n - P||_{\mathcal{F}} > \eta\right\}$ jako na podmnožinu $\widetilde{\Omega}$. Všimněme si, že s využitím (1.1) pro $n \geq \frac{2M^2}{\eta^2}$ máme

$$\begin{aligned}
&\widetilde{P}\left(|P_n'(f^\omega) - P_n(f^\omega)(\omega)| > \frac{\eta}{2}\right) \\
&\geq \widetilde{P}\left(|P_n(f^\omega)(\omega) - P(f^\omega)| - |P_n'(f^\omega) - P(f^\omega)| > \frac{\eta}{2}\right) \\
&\geq \widetilde{P}\left(\eta - |P_n'(f^\omega) - P(f^\omega)| > \frac{\eta}{2}\right) = \widetilde{P}\left(|P_n'(f^\omega) - P(f^\omega)| < \frac{\eta}{2}\right) \overset{(1.1)}{\geq} \frac{1}{2}. \qquad (1.2)
\end{aligned}$$

Pomocí (1.2) tedy pro dané $\omega \in \left\{||P_n - P||_{\mathcal{F}} > \eta\right\}$ máme

$$\begin{aligned}
&\frac{1}{2}\mathbb{1}\left\{||P_n - P||_{\mathcal{F}} > \eta\right\} \leq \widetilde{P}\left(|P_n'(f^\omega) - P_n(f^\omega)(\omega)| > \frac{\eta}{2}\right)\mathbb{1}\left\{||P_n - P||_{\mathcal{F}} > \eta\right\} \\
&\leq \widetilde{P}^*\left(\sup_{f \in \mathcal{F}}|P_n'(f) - P_n(f)(\omega)| > \frac{\eta}{2}\right). \qquad (1.3)
\end{aligned}$$

kde

$$\widetilde{P}^* \left(\sup_{f\in\mathcal{F}} |P_n^{'}(f) - P_n(f)(\omega)| > \frac{\eta}{2} \right)$$
$$= \inf \left\{ \ \widetilde{P}(B) : \left\{ \sup_{f\in\mathcal{F}} |P_n^{'}(f) - P_n(f)(\omega)| > \frac{\eta}{2} \right\} \subset B \ , B \in \widetilde{\mathcal{A}} \right\}.$$

Nerovnost (1.3) můžeme přepsat jako

$$\mathbb{1}\Big\{ ||P_n - P||_{\mathcal{F}} > \eta \Big\} \leq 2 \ \widetilde{P}^* \left(\sup_{f\in\mathcal{F}} |P_n^{'}(f) - P_n(f)(\omega)| > \frac{\eta}{2} \right).$$

Po aplikaci vnější střední hodnoty $\mathbb{E}_P^*$ a využitím Poznámky 1.2 dostáváme

$$P^*\Big(||P_n - P||_{\mathcal{F}} > \eta \Big) \leq 2 \ \mathbb{E}_P^* \left[\widetilde{P}^* \left(\sup_{f\in\mathcal{F}} |P_n^{'}(f) - P_n(f)(\omega)| > \frac{\eta}{2} \right) \right] \tag{1.4}$$
$$\leq 2 \ \mathbb{P}^* \left(|| \ P_n^{'} - P_n \ ||_{\mathcal{F}} > \frac{\eta}{2} \right),$$

přičemž důkaz poslední nerovnosti provedeme později.

Uvažujme Rademacherovy náhodné veličiny $\varepsilon_1, ..., \varepsilon_n$, které jsou definovány na pravděpodobnostním prostoru $(\Omega \times \widetilde{\Omega}, \mathcal{A} \otimes \widetilde{\mathcal{A}}, \mathbb{P})$ a jsou nezávislé na $X_1, ..., X_n$ i na $X_1^{'}, ..., X_n^{'}$. Všimneme si, že pro libovolnou měřitelnou funkci f je rozdělení náhodné veličiny $f(X_i^{'}) - f(X_i)$ symetrické a stejné jako rozdělení náhodné veličiny $\varepsilon_i(f(X_i^{'}) - f(X_i))$ pro každé $i \in \{1, 2, ..., n\}$. Odsud dostáváme, že

$$\begin{aligned}
\mathbb{P}^* \left(|| \ P_n^{'} - P_n \ ||_{\mathcal{F}} > \frac{\eta}{2} \right) &= \mathbb{P}^* \left(\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \left(f(X_i^{'}) - f(X_i) \right) \right| > \frac{\eta}{2} \right) \\
&\leq \mathbb{P}^* \left(\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i^{'}) \right| + \sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| > \frac{\eta}{2} \right) \\
&\leq \mathbb{P}^* \left(\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i^{'}) \right| > \frac{\eta}{4} \vee \sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| > \frac{\eta}{4} \right) \\
&\leq \mathbb{P}^* \left(\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i^{'}) \right| > \frac{\eta}{4} \right) + \mathbb{P}^* \left(\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| > \frac{\eta}{4} \right) \\
&= 2 \ \mathbb{P}^* \left(|| \ P_n^0 \ ||_{\mathcal{F}} > \frac{\eta}{4} \right).
\end{aligned}$$

Odsud spolu s (1.4) dostáváme tvrzení lemmatu.

Teď ověříme poslední nerovnost v (1.4). Označíme $C = \Big\{ || \ P_n - P_n^{'} \ ||_{\mathcal{F}} > \frac{\eta}{2} \Big\}$. Potom

$$\mathbb{P}^*(C) = \inf \Big\{ \mathbb{P}(B) \mid C \subset B, \ B \in \mathcal{A} \otimes \widetilde{\mathcal{A}} \Big\}.$$

Víme z Fubiniovy věty, že pro B měřitelnou $\mathbb{P}(B) = \mathbb{E}_P \Big[\mathbb{E}_{\widetilde{P}}[\mathbb{1}_B] \Big]$. Tedy dostáváme

$$\begin{aligned}
\mathbb{P}^*(C) &= \inf \Big\{ \mathbb{E}_P \Big[\mathbb{E}_{\widetilde{P}}[\mathbb{1}_B] \Big] : \mathbb{1}_C \leq \mathbb{1}_B, \ B \in \mathcal{A} \otimes \widetilde{\mathcal{A}} \Big\} \\
&\geq \inf \Big\{ \mathbb{E}_P \Big[\mathbb{E}_{\widetilde{P}}[V] \Big] : \mathbb{1}_C \leq V, \ V : (\Omega \times \widetilde{\Omega}, \mathcal{A} \otimes \widetilde{\mathcal{A}}) \longrightarrow (\mathbb{R}, \mathcal{B}) \Big\}.
\end{aligned}$$

Z Fubiniovy věty pro V měřitelnou je $\mathbb{E}_{\widetilde{P}}[V]$ měřitelná. Označme

$$W := \inf\left\{\mathbb{E}_{\widetilde{P}}[V] : \mathbb{1}_C \leq V,\ V : (\Omega \times \widetilde{\Omega}, \mathcal{A} \otimes \widetilde{\mathcal{A}}) \longrightarrow (\mathbb{R}, \mathcal{B})\right\}.$$

Přičemž W nemusí být měřitelná. Odsud plyne následující inkluze

$$\left\{\mathbb{E}_P\left[\mathbb{E}_{\widetilde{P}}[V]\right] : \mathbb{1}_C \leq V,\ V : (\Omega \times \widetilde{\Omega}, \mathcal{A} \otimes \widetilde{\mathcal{A}}) \longrightarrow (\mathbb{R}, \mathcal{B})\right\} \\ \subset \left\{\mathbb{E}_P[U] : U \geq W,\ U : (\Omega, \mathcal{A}) \longrightarrow (\mathbb{R}, \mathcal{B})\right\}.$$

Poslední inkluze platí protože pro každou V měřitelnou ($V \geq \mathbb{1}_C$): $\mathbb{E}_{\widetilde{P}}[V]$ je měřitelná a $\mathbb{E}_{\widetilde{P}}[V] \geq W$.

Odsud dostáváme

$$\begin{aligned}
&\inf\left\{\mathbb{E}_P\left[\mathbb{E}_{\widetilde{P}}[V]\right] : \mathbb{1}_C \leq V,\ V : (\Omega \times \widetilde{\Omega}, \mathcal{A} \otimes \widetilde{\mathcal{A}}) \longrightarrow (\mathbb{R}, \mathcal{B})\right\} \\
&\geq \inf\left\{\mathbb{E}_P[U] : U \geq W,\ U : (\Omega, \mathcal{A}) \longrightarrow (\mathbb{R}, \mathcal{B})\right\} = \mathbb{E}^*_P[W] \\
&= \mathbb{E}^*_P\left[\inf\left\{\mathbb{E}_{\widetilde{P}}[V] : \mathbb{1}_C \leq V,\ V :\ (\Omega \times \widetilde{\Omega}, \mathcal{A} \otimes \widetilde{\mathcal{A}}) \longrightarrow (\mathbb{R}, \mathcal{B})\right\}\right].
\end{aligned}$$

Tedy

$$\begin{aligned}
&\mathbb{P}^*(C) \geq \mathbb{E}^*_P\left[\inf\left\{\mathbb{E}_{\widetilde{P}}[V] : \mathbb{1}_C \leq V,\ V : (\Omega \times \widetilde{\Omega}, \mathcal{A} \otimes \widetilde{\mathcal{A}}) \longrightarrow (\mathbb{R}, \mathcal{B})\right\}\right] \\
&\geq \mathbb{E}^*_P\left[\mathbb{E}^*_{\widetilde{P}}[\mathbb{1}_C]\right].
\end{aligned}$$

□

Všimněme si, že poslední nerovnost $\mathbb{P}^*(C) \geq \mathbb{E}^*_P\left[\mathbb{E}^*_{\widetilde{P}}[\mathbb{1}_C]\right]$ v důkazu lemmatu 1.3 (o symetrizaci) lze ekvivalentně přepsat jako $\mathbb{E}^*_{\mathbb{P}}[\mathbb{1}_C] \geq \mathbb{E}^*_P\left[\mathbb{E}^*_{\widetilde{P}}[\mathbb{1}_C]\right]$ (podle poznámky 1.2). Tedy alternativně by šla použít Fubiniova věta pro vnější střední hodnoty (viz. van der Vaart a Wellner, 1996, lemma 1.2.6).

1.3 Zobecněná Glivenkova-Cantelliho věta

V této části nejprve zavedeme všechny potřebné pojmy a vyslovíme pomocné tvrzení, které budeme potřebovat pro důkaz zobecněné Glivenkovy-Cantelliho věty. Kapitolu ukončíme důkazem této věty.

Dále budeme předpokládat, že $\mathcal{F}$ tvoří vektorový prostor funkcí (kde sčítání a skalární násobení jsou definovány bodově). Tedy má smysl na prostoru funkcí $\mathcal{F}$ uvažovat normu $|| \cdot ||$. V dalším textu budeme používat následující značení

$$B_\epsilon(g) = \{f \in \mathcal{F} :|| f - g ||< \epsilon\},\ \text{kde } g \in \mathcal{F}.$$

To jest $B_\epsilon(g)$ značí abstraktní kouli v třídě funkcí $\mathcal{F}$ se středem g a poloměrem ϵ.

Zobecněná Glivenkova-Cantelliho věta platí za předpokladu, že třída funkcí $\mathcal{F}$ není příliš „velká“. Proto zavedeme pojem pokrývacích čísel, které v jistém smyslu měří velikost třídy $\mathcal{F}$.

Definice 1.7. *Pokrývací číslo $N(\epsilon, \mathcal{F}, || \cdot ||)$ (covering number) je minimální počet koulí $B_\epsilon(g)$ (kde $g \in \mathcal{F}$) poloměru ϵ potřebných k pokrytí třídy $\mathcal{F}$. Středy g těchto koulí musí mít konečné normy.*

Ukážeme si, proč normy středu koulí musí být konečné. Předpokládejme, že $||g|| = \infty$. Pro $f \in B_\epsilon(g)$ z trojúhelníkové nerovnosti platí

$$\infty = ||g|| \leq ||g - f|| + ||f|| < \epsilon + ||f||.$$

Odsud by plynulo, že $||f|| > \infty$. To jest $B_\epsilon(g)$ by byla prázdná množina

Poznámka 1.4. V Glivenkově-Cantelliho větě budeme za normu $|| \cdot ||$ uvažovat $L_1(P_n)$ normu. Tedy pro libovolné funkce $f, g \in \mathcal{F}$

$$||f - g||_{L_1(P_n)} = \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - g(X_i)|.$$

Všimněme si, že $||f - g||_{L_1(P_n)}$ je náhodná veličina nikoli číslo. Pokud budeme chtít zdůraznit hodnotu funkce $||f - g||_{L_1(P_n)}$ v nějakém konkrétním bodě ω, pak použijeme následující značení

$$||f - g||_{L_1(P_n)}(\omega) = \frac{1}{n} \sum_{i=1}^{n} |f(X_i(\omega)) - g(X_i(\omega))|.$$

Jelikož $||f - g||_{L_1(P_n)}$ je náhodná veličina, proto pro libovolné kladné ϵ pokrývací čísla $N(\epsilon, \mathcal{F}, L_1(P_n))$ jsou náhodnými zobrazeními v následujícím smyslu:

$$N(\epsilon, \mathcal{F}, L_1(P_n)) = \min \left\{ k : \exists f_1, ..., f_k,\ \mathcal{F} \subset \bigcup_{i=1}^{k} \left\{ g \in \mathcal{F} : ||f_i - g||_{L_1(P_n)} \leq \epsilon \right\} \right\}.$$

V důkazu lemmatu 1.3 (o symetrizaci) jsme nepotřebovali integrovat pouze vzhledem k rozdělení náhodných veličin $\varepsilon_1, ..., \varepsilon_n$. Proto jsme pro jednoduchost předpokládali, že $\varepsilon_1, ..., \varepsilon_n$ jsou definovány na stejném pravděpodobnostním prostoru jako $X_1, ..., X_n$. V důkazu zobecněné Glivenkovy-Cantelliho věty budeme chtít rozlišit pravděpodobnostní prostor $(\Omega_\varepsilon, \mathcal{A}_\varepsilon, P_\varepsilon)$ (kde jsou definovány Rademacherovy náhodné veličiny) od $(\Omega, \mathcal{A}, P)$.

Jak bylo zmíněno dříve, v zobecněné Glivenkově-Cantelliho větě chceme využít lemma 1.3 (o symetrizaci). Takže budeme pracovat s výrazem $\mathbb{P}^* \left(||P_n^0||_{\mathcal{F}} > \frac{\eta}{4} \right)$ (kde $\mathbb{P} = P \otimes P_\varepsilon$). Klíčovým krokem důkazu bude rozepsat tento výraz pomocí Fubiniovy věty jako

$$\mathbb{E}_P \left[\mathbb{E}_{P_\varepsilon} \left[\mathbb{1} \left\{ ||P_n^0||_{\mathcal{F}} > \frac{\eta}{4} \right\} \right] \right].$$

To obecně nejde, protože množina $\left\{ ||P_n^0||_{\mathcal{F}} > \frac{\eta}{4} \right\}$ nemusí být měřitelná. Za tímto účelem zavedeme následující definici.

Definice 1.8. *Nechť $X_1, ..., X_n$ jsou nezávislé stejně rozdělené náhodné veličiny definované na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$ s hodnotami v $(S, \mathcal{S})$. Třídu měřitelných funkcí $\mathcal{F}$ nazveme $P-$měřitelnou třídou ($P-$measurable class), jestli pro každé $n \in \mathbb{N}$ a pro každé $(e_1, ..., e_n) \in \{-1, 1\}^n$ je funkce*

$$\left\| \sum_{i=1}^{n} e_i f(X_i) \right\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} e_i f(X_i) \right|$$

měřitelná na zúplnění pravděpodobnostního prostoru $(\Omega, \mathcal{A}, P)$.

Za předpokladu $P-$měřitelnosti třídy $\mathcal{F}$ platí, že množina $\left\{||P_n^0||_{\mathcal{F}} > \frac{\eta}{4}\right\}$ je měřitelná. Proto z Fubiniovy věty dostaneme, že

$$\mathbb{P}^*\left(||P_n^0||_{\mathcal{F}} > \frac{\eta}{4}\right) = \mathbb{P}\left(||P_n^0||_{\mathcal{F}} > \frac{\eta}{4}\right) = \mathbb{E}_P\left[\mathbb{E}_{P_\varepsilon}\left[\mathbb{1}\left\{||P_n^0||_{\mathcal{F}} > \frac{\eta}{4}\right\}\right]\right].$$

V praxi se často $P-$měřitelnost ověřuje podle tzv. bodové měřitelnosti.

Definice 1.9. *Třídu funkcí $\mathcal{F}$ nazveme bodově měřitelnou třídou (pointwise measurable class), pokud existuje spočetná třída $\mathcal{G} \subseteq \mathcal{F}$ taková, že*

$$\forall f \in \mathcal{F}\ \exists \{g_m\}_{m=1}^{\infty} \subset \mathcal{G}\ \forall x \in S : g_m(x) \underset{m \longrightarrow \infty}{\longrightarrow} f(x).$$

Lemma 1.4. *Pokud $\mathcal{F}$ je bodově měřitelná, potom $\mathcal{F}$ je $P-$měřitelná.*

Důkaz. Nejdříve si všimněme, že jelikož $\mathcal{G}$ je spočetná, tak stačí dokázat následující

$$\forall n \in \mathbb{N}\ \forall (e_1, ..., e_n) \in \{-1, 1\}^n : \ \left|\left|\sum_{i=1}^n e_i f(X_i)\right|\right|_{\mathcal{F}} = \left|\left|\sum_{i=1}^n e_i g(X_i)\right|\right|_{\mathcal{G}}.$$

Je zřejmé, že $\forall \omega \in \Omega\ \forall n \in \mathbb{N}\ \forall (e_1, ..., e_n) \in \{-1,1\}^n$ platí

$$\sup_{f \in \mathcal{F}}\left|\sum_{i=1}^n e_i f(X_i(\omega))\right| \geq \sup_{g \in \mathcal{G}}\left|\sum_{i=1}^n e_i g(X_i(\omega))\right|.$$

Zafixujme $f \in \mathcal{F}$, potom existuje posloupnost $\{g_m\}_{m=1}^{\infty} \subset \mathcal{G}$ taková, že $\forall x \in S$ platí

$$g_m(x) \underset{m \longrightarrow \infty}{\longrightarrow} f(x).$$

Tedy $\forall \omega \in \Omega\ \forall n \in \mathbb{N}\ \forall \varepsilon > 0\ \exists g \in \mathcal{G}\ \forall i \in \{1, ..., n\}$ platí

$$|f(X_i(\omega)) - g(X_i(\omega))| < \frac{\varepsilon}{n}.$$

Odsud dostaneme pro $\forall (e_1, ..., e_n) \in \{-1, 1\}^n$

$$\left|\sum_{i=1}^n e_i f(X_i(\omega))\right| = \left|\sum_{i=1}^n e_i(f(X_i(\omega)) - g(X_i(\omega))) + \sum_{i=1}^n e_i g(X_i(\omega))\right|$$
$$\leq \varepsilon + \sup_{g \in \mathcal{G}}\left|\sum_{i=1}^n e_i g(X_i(\omega))\right|.$$

Tedy máme

$$\sup_{f \in \mathcal{F}}\left|\sum_{i=1}^n e_i f(X_i(\omega))\right| \leq \varepsilon + \sup_{g \in \mathcal{G}}\left|\sum_{i=1}^n e_i g(X_i(\omega))\right|.$$

Limitním přechodem pro $\varepsilon \longrightarrow 0_+$ dostaneme

$$\sup_{f \in \mathcal{F}}\left|\sum_{i=1}^n e_i f(X_i(\omega))\right| \leq \sup_{g \in \mathcal{G}}\left|\sum_{i=1}^n e_i g(X_i(\omega))\right|.$$

$\square$

Poznámka 1.5. Pro ilustraci definice 1.9 uvažujme následující třídu funkcí

$$\mathcal{T} := \Big\{\mathbb{1}\{x \leq t\} : t \in \mathbb{R}\Big\}.$$

Položme

$$\mathcal{G} := \Big\{\mathbb{1}\{x \leq t\} : t \in \mathbb{Q}\Big\}.$$

Z hustoty racionálních čísel je zřejmé, že $\mathcal{G}$ splňuje definici 1.9 pro třídu funkcí $\mathcal{T}$. Potom podle lemmatu 1.4 platí

$$\sup_{t\in\mathbb{R}}|F_n(t) - F(t)| = \sup_{f\in\mathcal{T}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_1)\right| \text{ je } P\text{−měřitelná.}$$

Definice 1.10. *Řekněme, že funkce F je obálková funkce (envelope function) pro systém funkcí $\mathcal{F}$, jestliže platí* $\sup_{f\in\mathcal{F}}|f(x)| \leq F(x)\ \forall x \in S$.

Obálková funkce nám dovoluje kontrolovat střední hodnoty $\mathbb{E}\Big[f(X_1)\Big]$, kde $f \in \mathcal{F}$. To jest předpokladem, že obálková funkce má „konečnou střední hodnotu“ dostaneme, že všechny funkce z třídy $\mathcal{F}$ budou mít konečnou střední hodnotu (u obálkové funkce slovo *střední hodnota* používáme v uvozovkách protože obálková funkce nemusí být měřitelná).

Příklad 1.1. Nechť $S = [0{,}2\pi]$, $\mathcal{F} = \{\sin(ax) : a \in \mathbb{R}\}$. Potom za obálkovou funkci můžeme uvažovat $F(x) = 1$ pro $x \in [0, 2\pi]$.

Lemma 1.5. ***(Hoeffdingova nerovnost)*** *Nechť $Y_1,...,Y_n$ jsou nezávislé náhodné veličiny takové že $\mathbb{E}[Y_i] = 0$ a $a_i \leq Y_i \leq b_i$, pro každé $i = 1, 2, ..., n$. Potom pro $\eta > 0$ platí*

$$P\left(\Big|\sum_{i=1}^{n} Y_i\Big| > \eta\right) \leq 2\ \exp\left(\frac{-2\ \eta^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{1.5}$$

Tato věta plyne z článku Hoeffding (1963), věta 2.

Věta 1.6. ***(Glivenkova-Cantelliho zobecněná)*** *Nechť F je obálková funkce pro systém $\mathcal{F}$, přičemž $P^*(F) := \mathbb{E}^*\Big[F(X_1)\Big] < \infty$. Nechť $\mathcal{F}_M^*$ obsahuje funkce tvaru $x \longmapsto f(x)\mathbb{1}\{F^*(x) \leq M\}$, kde $f \in \mathcal{F}$. Předpokládejme že*

$$\forall M > 0 \text{ je } \mathcal{F}_M^*\ P - \text{měřitelná.}$$

Také nechť pro každé $M < \infty$ a pro každé $\epsilon > 0$

$$\log\ N\Big(\epsilon, \mathcal{F}_M^*, L_1(P_n)\Big) = o_{P^*}(n)$$

Potom platí

$$\sup_{f\in\mathcal{F}}|P_n(f) - P(f)| \xrightarrow[n\to\infty]{s.j.^*} 0.$$

Důkaz. Jak uvidíme na konci důkazu stačí ukázat, že

$$\sup_{f\in\mathcal{F}}|P_n(f) - P(f)| \xrightarrow[n\to\infty]{P^*} 0. \tag{1.6}$$

Důkaz (1.6) rozdělíme na následující kroky:

(i) Nechť $\eta > 0$ a pro dostatečně velké M platí

$$P^*\left(||P_n - P||_{\mathcal{F}} > \frac{8\eta}{3}\right) \leq P^*\Big(||P_n - P||_{\mathcal{F}^*_M} \geq \eta\Big) + P\Big(P_n\Big(F^*\mathbb{1}\{F^* > M\}\Big) \geq \eta\Big).$$

(ii) $\forall\eta > 0\ \exists M > 0\ \forall\varepsilon > 0\ \exists n_0 \in \mathbb{N}\ \forall n \geq n_0 :$

$$P\Big(P_n\Big(F^*\mathbb{1}\{F^* > M\}\Big) \geq \eta\Big) < \varepsilon.$$

(iii) $\forall M > 0 : ||P_n - P||_{\mathcal{F}^*_M} \xrightarrow[n\to\infty]{P^*} 0.$

Pokud jsou splněny všechny tři body, pak $\sup_{f\in\mathcal{F}}|P_n(f) - P(f)| \xrightarrow[n\to\infty]{P^*} 0.$

Důkaz (i) kroku.

Nechť $\eta > 0$. Pak pro dané M platí

$$\begin{aligned}
|| P_n - P ||_{\mathcal{F}} &= \sup_{f\in\mathcal{F}}|P_n(f) - P(f)| \\
&\leq \sup_{f\in\mathcal{F}^*_M}|P_n(f) - P(f)| + \sup_{f\in\mathcal{F}} P_n\Big(|f\mathbb{1}\{F^* > M\}|\Big) + \sup_{f\in\mathcal{F}} P\Big(|f\mathbb{1}\{F^* > M\}|\Big) \\
&\leq \sup_{f\in\mathcal{F}^*_M}|P_n(f) - P(f)| + P_n\Big(F^*\mathbb{1}\{F^* > M\}\Big) + P\Big(F^*\mathbb{1}\{F^* > M\}\Big).
\end{aligned}$$

Podle předpokladu věty $\mathbb{E}^*\Big[F(X_1)\Big] < \infty$. Podle lemmatu 1.2 platí

$$\mathbb{E}^*\Big[F(X_1)\Big] = \mathbb{E}\Big[F^*(X_1)\Big].$$

Všimněme si, že z $\mathbb{E}^*\Big[F(X_1)\Big] < \infty$ platí

$$\forall\eta > 0\ \exists M > 0 : P\Big(F^*\mathbb{1}\{F^* > M\}\Big) < \frac{2\eta}{3}. \tag{1.7}$$

Tedy pro M dostatečně velké platí řetězec nerovností

$$\begin{aligned}
&P^*\left(||P_n - P||_{\mathcal{F}} > \frac{8\eta}{3}\right) \\
&\leq P^*\left(||P_n - P||_{\mathcal{F}^*_M} + P_n\Big(F^*\mathbb{1}\{F^* > M\}\Big) + P\Big(F^*\mathbb{1}\{F^* > M\}\Big) > \frac{8\eta}{3}\right) \\
&\leq P^*\Big(||P_n - P||_{\mathcal{F}^*_M} + P_n\Big(F^*\mathbb{1}\{F^* > M\}\Big) > 2\eta\Big) \\
&\leq P^*\Big(||P_n - P||_{\mathcal{F}^*_M} \geq \eta\Big) + P\Big(P_n\Big(F^*\mathbb{1}\{F^* > M\}\Big) \geq \eta\Big).
\end{aligned}$$

Odsud dostáváme, že

$$P^*\left(||P_n - P||_{\mathcal{F}} > \frac{8\eta}{3}\right) \leq P^*\Big(||P_n - P||_{\mathcal{F}^*_M} \geq \eta\Big) + P\Big(P_n\Big(F^*\mathbb{1}\{F^* > M\}\Big) \geq \eta\Big). \tag{1.8}$$

Dokážeme (ii) krok.

$$P\Big(P_n\Big(F^*\mathbb{1}\{F^*>M\}\Big)\geq \eta\Big) = P\left(\frac{1}{n}\sum_{i=1}^n F^*(X_i)\mathbb{1}\{F^*(X_i)>M\}\geq \eta\right).$$

Ze zákona velkých čísel platí, že

$$\frac{1}{n}\sum_{i=1}^n F^*(X_i)\mathbb{1}\{F^*(X_i)>M\} \xrightarrow[n\to\infty]{s.j.} P\Big(F^*\mathbb{1}\{F^*>M\}\Big).$$

Ale víme, že pro dostatečně velká M platí

$$P\Big(F^*\mathbb{1}\{F^*>M\}\Big) < \frac{2\eta}{3}.$$

Proto

$$P\Big(P_n\Big(F^*\mathbb{1}\{F^*>M\}\Big)\geq \eta\Big) \xrightarrow[n\to\infty]{} 0. \tag{1.9}$$

To jest $\forall \eta>0\ \exists M>0\ \forall \varepsilon>0\ \exists n_0\in\mathbb{N}\ \forall n\geq n_0$:

$$P\Big(P_n\Big(F^*\mathbb{1}\{F^*>M\}\geq \eta\Big)\Big) < \varepsilon.$$

Důkaz (iii) kroku.

Nechť $(\Omega_\varepsilon, \mathcal{A}_\varepsilon, P_\varepsilon)$ je pravděpodobnostní prostor na kterém jsou definovány Rademacherovy náhodné veličiny $\varepsilon_1, ..., \varepsilon_n$. Předpokládejme, že

$$(\varepsilon_1, ..., \varepsilon_n) \perp (X_1, ..., X_n) \text{ na } (\Omega\times\Omega_\varepsilon, \mathcal{A}\otimes\mathcal{A}_\varepsilon, \mathbb{P}),$$

kde $\mathbb{P} = P\otimes P_\varepsilon$. Z lemmatu o symetrizaci víme, že platí

$$P^*\Big(||\ P_n - P\ ||_{\mathcal{F}^*_M} > \eta\Big) \leq 4\ \mathbb{P}^*\left(||\ P_n^0\ ||_{\mathcal{F}^*_M} > \frac{\eta}{4}\right). \tag{1.10}$$

Podle předpokladu věty $\mathcal{F}^*_M$ je $P-$měřitelná třída, proto

$$\left\{||\ P_n^0\ ||_{\mathcal{F}^*_M} > \frac{\eta}{4}\right\} = \left\{(\omega,\omega_\varepsilon)\in\Omega\times\Omega_\varepsilon : \sup_{f\in\mathcal{F}^*_M}|P_n^0(f)(\omega,\omega_\varepsilon)| > \frac{\eta}{4}\right\} \in \mathcal{A}\otimes\mathcal{A}_\varepsilon.$$

Tedy dle Fubiniho věty

$$\mathbb{P}^*\left(||P_n^0||_{\mathcal{F}^*_M} > \frac{\eta}{4}\right) = \mathbb{P}\left(\sup_{f\in\mathcal{F}^*_M}|P_n^0(f)| > \frac{\eta}{4}\right) = \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}\left\{\sup_{f\in\mathcal{F}^*_M}|P_n^0(f)| > \frac{\eta}{4}\right\}\right]$$
$$\overset{Fubini}{=} \mathbb{E}_P\left[\mathbb{E}_{P_\varepsilon}\left[\mathbb{1}\left\{\sup_{f\in\mathcal{F}^*_M}|P_n^0(f)| > \frac{\eta}{4}\right\}\right]\right]$$
$$= \mathbb{E}_P\left[P_\varepsilon\left\{\omega_\varepsilon\in\Omega_\varepsilon : \sup_{f\in\mathcal{F}^*_M}|P_n^0(f)(\omega,\omega_\varepsilon)| > \frac{\eta}{4}\right\}\right].$$

Tudíž z lemmatu o symetrizaci dostaneme

$$P^*\Big(||\ P_n - P\ ||_{\mathcal{F}^*_M} > \eta\Big) \leq 4\ \mathbb{E}_P\left[P_\varepsilon\left\{\sup_{f\in\mathcal{F}^*_M}|P_n^0(f)| > \frac{\eta}{4}\right\}\right]. \tag{1.11}$$

Připomeňme si, že podle poznámky 1.4 vzdálenost v $L_1(P_n)$ normě je náhodná veličina nikoli číslo (tj. vzdálenost v $L_1(P_n)$ normě je funkce od ω). Všimněme si, že $\forall \omega \in \Omega\ \exists g_1^\omega, ..., g_{N_\eta^\omega}^\omega \in \mathcal{F}_M^*$ takové že :

$$\forall f \in \mathcal{F}_M^*\ \exists j \in \{1, ..., N_\eta^\omega\} : ||f - g_j^\omega||_{L_1(P_n)}(\omega) \leq \frac{\eta}{8}.$$

Tedy $\forall j \in \{1, ..., N_\eta^\omega\} : |g_j^\omega| \leq M$. Zafixujme ω, f, j, pak pro $\forall \omega_\varepsilon \in \Omega_\varepsilon$:

$$\begin{aligned} &\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(\omega_\varepsilon) f(X_i(\omega))\right| \leq \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(\omega_\varepsilon)(f(X_i(\omega)) - g_j^\omega(X_i(\omega)))\right| \\ &+ \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(\omega_\varepsilon) g_j^\omega(X_i(\omega))\right| \leq \frac{\eta}{8} + |P_n^0(g_j^\omega)(\omega, \omega_\varepsilon)| \\ &\leq \frac{\eta}{8} + \max_{j=1,...,N_\eta^\omega} |P_n^0(g_j^\omega)(\omega, \omega_\varepsilon)|. \end{aligned} \tag{1.12}$$

Tedy s využitím (1.12) a Hoeffdingovy nerovnosti (viz.lemma 1.5)

$$\begin{aligned} &P_\varepsilon\left\{\omega_\varepsilon \in \Omega_\varepsilon : \sup_{f \in \mathcal{F}_M^*} |P_n^0(f)(\omega, \omega_\varepsilon)| > \frac{\eta}{4}\right\} \\ &\leq P_\varepsilon\left\{\omega_\varepsilon \in \Omega_\varepsilon : \max_{j=1,...,N_\eta^\omega} |P_n^0(g_j^\omega)(\omega, \omega_\varepsilon)| > \frac{\eta}{8}\right\} \\ &= P_\varepsilon\left(\bigcup_{j=1}^{N_\eta^\omega}\left\{\omega_\varepsilon \in \Omega_\varepsilon : |P_n^0(g_j^\omega)(\omega, \omega_\varepsilon)| > \frac{\eta}{8}\right\}\right) \\ &\leq \sum_{j=1}^{N_\eta^\omega} P_\varepsilon\left\{\omega_\varepsilon \in \Omega_\varepsilon : |P_n^0(g_j^\omega)(\omega, \omega_\varepsilon)| > \frac{\eta}{8}\right\} \\ &\leq N_\eta^\omega \max_{j=1,...,N_\eta^\omega} P_\varepsilon\left\{\omega_\varepsilon \in \Omega_\varepsilon : |P_n^0(g_j^\omega)(\omega, \omega_\varepsilon)| > \frac{\eta}{8}\right\} \\ &= N_\eta^\omega \max_{j \in \{1,...,N_\eta^\omega\}} P_\varepsilon\left\{\omega_\varepsilon \in \Omega_\varepsilon : \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(\omega_\varepsilon) g_j^\omega(X_i(\omega))\right| > \frac{\eta}{8}\right\} \\ &\overset{Lemma\ 1.5}{\leq} N_\eta^\omega\ 2\exp\left(-\frac{n\left(\frac{\eta}{8}\right)^2}{2P_n((g_j^\omega)^2)(\omega)}\right) \leq 2N_\eta^\omega\ \exp\left(-\frac{n\eta^2}{128M^2}\right). \end{aligned}$$

Podle předpokladu věty platí

$$\log N\left(\frac{\eta}{8}, \mathcal{F}_M^*, L_1(P_n)\right) = o_{P^*}(n).$$

Tedy existuje posloupnost náhodných zobrazení $\{Y_n\}_{n \in \mathbb{N}}$ taková, že $Y_n \overset{P^*}{\to} 0$ a $\log N_\eta^\omega = nY_n(\omega)$. Tedy $N_\eta^\omega = \exp(nY_n(\omega))$. Jelikož

$$\begin{aligned} &\left\{\omega \in \Omega : \left|2N_\eta^\omega \exp\left(-\frac{n\eta^2}{128M^2}\right)\right| > \eta\right\} \\ &= \left\{\omega \in \Omega : 2\exp\left(nY_n(\omega) - \frac{n\eta^2}{128M^2}\right) > \eta\right\} \\ &= \left\{\omega \in \Omega : Y_n(\omega) > \frac{1}{n}\log(\frac{\eta}{2}) + \frac{\eta^2}{128M^2}\right\}. \end{aligned}$$

Odsud dostáváme, že $2N_\eta\ \exp\left(-\frac{n\eta^2}{128M^2}\right) \xrightarrow{P^*} 0$. Tedy máme

$$P_\varepsilon\left\{\omega_\varepsilon \in \Omega_\varepsilon : \sup_{f\in\mathcal{F}_M^*} |P_n^0(f)(\omega,\omega_\varepsilon)| > \frac{\eta}{4}\right\} \xrightarrow{P^*} 0.$$

Protože $\left\{||P_n^0||_{\mathcal{F}_M^*} > \frac{\eta}{4}\right\}$ je měřitelná, tak dle Fubiniho věty je zobrazení

$$\omega \longmapsto P_\varepsilon\left\{\omega_\varepsilon \in \Omega_\varepsilon : \sup_{f\in\mathcal{F}_M^*} |P_n^0(f)(\omega,\omega_\varepsilon)| > \frac{\eta}{4}\right\}$$

také měřitelné. Tudíž

$$P_\varepsilon\left\{\omega_\varepsilon \in \Omega_\varepsilon : \sup_{f\in\mathcal{F}_M^*} |P_n^0(f)(\omega,\omega_\varepsilon)| > \frac{\eta}{4}\right\} \xrightarrow{P} 0. \tag{1.13}$$

Je zřejmě, že $\forall n \in \mathbb{N}$:

$$\left|P_\varepsilon\left\{\omega_\varepsilon \in \Omega_\varepsilon : \sup_{f\in\mathcal{F}_M^*} |P_n^0(f)(\omega,\omega_\varepsilon)| > \frac{\eta}{4}\right\}\right| \leq 1.$$

Podle Věty A.2 je systém

$$\left\{P_\varepsilon\left\{\omega_\varepsilon \in \Omega_\varepsilon : \sup_{f\in\mathcal{F}_M^*} |P_n^0(f)(\omega,\omega_\varepsilon)| > \frac{\eta}{4}\right\}\right\}_{n\in\mathbb{N}}$$

stejnoměrně integrovatelný (viz. definice A.1). Odsud podle Věty A.3 můžeme konvergenci (1.13) zesílit na

$$P_\varepsilon\left\{\omega_\varepsilon \in \Omega_\varepsilon : \sup_{f\in\mathcal{F}_M^*} |P_n^0(f)(\omega,\omega_\varepsilon)| > \frac{\eta}{4}\right\} \xrightarrow{L_1(P)} 0.$$

Tudíž z nerovnosti (1.11) dostáváme

$$P^*\left(||P_n - P||_{\mathcal{F}_M^*} > \eta\right) \underset{n\to\infty}{\longrightarrow} 0. \tag{1.14}$$

Kombinací kroků 1. 2. a 3. dostaneme, že $||P_n - P||_{\mathcal{F}} \xrightarrow[n\to\infty]{P^*} 0$. Teď si dokážeme, že platí dokonce

$$\sup_{f\in\mathcal{F}}|P_n(f) - P(f)| \xrightarrow[n\to\infty]{s.j.^*} 0.$$

Podle lemmatu 1.2

$$\forall\eta > 0 : P^*\Big(||P_n - P||_{\mathcal{F}} > \eta\Big) = P\Big(||P_n - P||_{\mathcal{F}}^* > \eta\Big) \underset{n\to\infty}{\longrightarrow} 0.$$

Odsud $||P_n - P||_{\mathcal{F}}^* \xrightarrow[n\to\infty]{P} 0$. Podle Věty A.4 víme, že existuje filtrace (viz. definice A.2) podle které $\{||P_n - P||_{\mathcal{F}}^*\}_{n\in\mathbb{N}}$ tvoří tzv. inverzní sub-martingal (viz. definice A.3) a proto musí konvergovat $s.j.$ ke konstantě.

Jelikož však $||P_n - P||_{\mathcal{F}}^* \xrightarrow[n\to\infty]{P} 0$, tak nutně $||P_n - P||_{\mathcal{F}}^* \xrightarrow[n\to\infty]{s.j.} 0$.

□

Poznámka 1.6. V předchozí větě jsme měli následující předpoklad

$$\forall M > 0 \text{ třída } \mathcal{F}_M^* \text{ je } P - \text{měřitelná.}$$

Ve skutečnosti tento předpoklad nepotřebujeme pro každé $M > 0$. Stačilo by mít nějakou rostoucí posloupnost kladných reálných čísel $\{M_j\}_{j=1}^{\infty}$ takovou, že

$$M_j \nearrow \infty \text{ a } \forall j \in \mathbb{N} \text{ třída } \mathcal{F}_{M_j}^* \text{ je } P - \text{měřitelná.}$$

Pro jednoduchost značení jsme dokázali zobecněnou Glivenkovu-Cantelliho větu pro každé $M > 0$.

Také si všimněme, že pokud $\mathcal{F}$ je bodově měřitelná, potom

$$\forall M > 0 \text{ je } \mathcal{F}_M^* \text{ bodově měřitelná (tudíž } P- \text{ měřitelná).}$$

Poznámka 1.7. Zobecněnou Glivenkovou-Cantelliho větu lze dokázat obecněji za předpokladu $P-$měřitelnosti třídy funkce $\mathcal{F}$ (viz. van der Vaart a Wellner, 1996, věta 2.4.3). V takovém důkazu se používá obecnější lemma o symetrizaci (viz. van der Vaart a Wellner, 1996, lemma 2.3.1) a vhodnou maximální nerovnost (viz. van der Vaart a Wellner, 1996, lemma 2.2.2).

Definice 1.11. *Třídu funkcí $\mathcal{F}$ nazveme $P-$Glivenkova-Cantelliho, jestliže platí*

$$\sup_{f \in \mathcal{F}} |P_n(f) - P(f)| \xrightarrow[n \to \infty]{s.j.^*} 0.$$

Všimněme si, že v předchozí definici je pravděpodobnostní míra P fixovaná. To jest, pokud $\mathcal{F}$ je $P-$Glivenkova-Cantelliho, pak nemusí být $\tilde{P}-$Glivenkova-Cantelliho (kde $\tilde{P}$ je jiná pravděpodobnostní míra).

Poznámka 1.8. Poznamenejme, že podmínky

$$\log\ N\Big(\epsilon, \mathcal{F}_M, L_1(P_n)\Big) = o_{P^*}(n) \quad \text{a} \quad P^*(F) < \infty$$

jsou jen postačující. Nutné podmínky pro konvergenci

$$\sup_{f \in \mathcal{F}} |P_n(f) - P(f)| \xrightarrow[n \to \infty]{s.j.^*} 0$$

lze najít v článku Giné a Zinn (1984), věta 8.3 str. 981.

1.4 Glivenkovy-Cantelliho třídy funkcí

Doposud jsme zabývali zobecněnou Glivenkovou-Cantelliho větou pro pevnou pravděpodobnostní míru P. Ukazuje se, že existuje ještě silnější výsledek, který je stejnoměrný v pravděpodobnostních mírách na měřitelném prostoru $(\Omega, \mathcal{A})$. V této kapitole zavedeme další Glivenkovy-Cantelliho třídy funkcí, pro které zformulujeme stejnoměrný zákon velkých čísel (viz. Věta 1.7).

V dalším textu bude $\mathcal{P}(\Omega)$ značit množinu všech pravděpodobnostních měr na měřitelném prostoru $(\Omega, \mathcal{A})$.

Definice 1.12. *Třídu funkcí $\mathcal{F}$ nazveme silně stejnoměrná Glivenkova-Cantelliho (strong uniform Glivenko-Cantelli), pokud*

$$\forall \varepsilon > 0 : \lim_{n \longrightarrow \infty} \sup_{P \in \mathcal{P}(\Omega)} P^*\Big(\sup_{m \geq n} ||P_m - P||_{\mathcal{F}} > \varepsilon\Big) = 0.$$

Poznámka 1.9. V literatuře se často silná stejnoměrnost formuluje pro nějakou podmnožinu pravděpodobnostních měr $\mathcal{P}_0 \subset \mathcal{P}(\Omega)$ (viz. van der Vaart a Wellner, 1996, str. 167).

Pro naši účely (hlavně pro aplikace v strojovém učení viz. definice 3.1) jsme zadefinovali silnou stejnoměrnost vzhledem k celé množině pravděpodobnostních měr $\mathcal{P}(\Omega)$.

Poznámka 1.10. Pokud $\mathcal{F}$ je silně stejnoměrná Glivenkova-Cantelliho třída funkcí, pak podle lemmatu A.5 platí, že

$$\forall P \in \mathcal{P}(\Omega) : \ ||P_n - P||_{\mathcal{F}} \xrightarrow[n \to \infty]{s.j.^*} 0.$$

Důkaz následující věty lze najít v knize van der Vaart a Wellner (1996), věta 2.8.1.

Věta 1.7. ***(stejnoměrný zákon velkých čísel)*** *Nechť $\forall P \in \mathcal{P}(\Omega) : \mathcal{F}$ je $P-$ měřitelná třída funkcí. Nechť F je měřitelná obálková funkce třídy $\mathcal{F}$. Předpokládejme, že platí:*

$$\lim_{M \longrightarrow \infty} \sup_{P \in \mathcal{P}(\Omega)} P\Big(F \mathbb{1}\{F > M\}\Big) = 0,$$

$$\forall \varepsilon > 0 : \sup_{Q \in \mathcal{Q}_n(S)} \log\ N\Big(\varepsilon ||F||_{L_1(Q)}, \mathcal{F}, L_1(Q)\Big) = o(n),$$

kde $\mathcal{Q}_n(S)$ je množina všech pravděpodobnostních měr na $(S, \mathcal{S})$ s atomy velikosti celočíselných násobků $1/n$. Potom $\mathcal{F}$ je silně stejnoměrná Glivenkova-Cantelliho třída funkcí.

V aplikacích (zejména v teorii strojového učení) nám často stačí, aby výraz $||P_n - P||_{\mathcal{F}}$ konvergoval ve vnější pravděpodobnosti k nule přes třídu všech pravděpodobnostních měr. Což je motivací pro následující definici.

Definice 1.13. *Třídu funkce $\mathcal{F}$ budeme nazývat slabě stejnoměrná Glivenkova-Cantelliho (weak uniform Glivenko-Cantelli), pokud*

$$\forall \varepsilon > 0 : \lim_{n \longrightarrow \infty} \sup_{P \in \mathcal{P}(\Omega)} P^*\Big(||P_n - P||_{\mathcal{F}} > \varepsilon\Big) = 0.$$

Je zřejmé, že platí následující nerovnost:

$$P^*\Big(||P_n - P||_{\mathcal{F}} > \varepsilon\Big) \leq P^*\Big(\sup_{m \geq n} ||P_m - P||_{\mathcal{F}} > \varepsilon\Big).$$

To jest pokud $\mathcal{F}$ je silně stejnoměrná Glivenkova-Cantelliho, pak $\mathcal{F}$ je slabě stejnoměrná Glivenkova-Cantelliho.

Mohlo by nás také zajímat, jestli ze slabé stejnoměrnosti plyne silná stejnoměrnost. Pro tuto implikaci je třeba dodat omezenost a další předpoklady na měřitelnost (využívá se tzv. *universální měřitelnost* detaily lze najít v článku Dudley a kol., 1991, věta 6 a tvrzení 10).

Věta 1.8. *Pokud $\mathcal{F}$ je slabě stejnoměrná Glivenkova-Cantelliho třída funkcí, potom $\mathcal{F}$ je $P-$Glivenkova-Cantelliho pro každou $P \in \mathcal{P}(\Omega)$.*

Důkaz. Ze slabé stejnoměrnosti $\mathcal{F}$ platí

$$\forall \varepsilon > 0 \; \forall P \in \mathcal{P}(\Omega) : \lim_{n \longrightarrow \infty} P^* \Big(||P_n - P||_{\mathcal{F}} > \varepsilon \Big) = 0.$$

S využitím lemmatu 1.2 tedy

$$\forall P \in \mathcal{P}(\Omega) : ||P_n - P||^*_{\mathcal{F}} \xrightarrow[n \longrightarrow \infty]{P} 0.$$

Podle věty A.4 existuje filtrace (viz. definice A.2) podle které tvoří

$$\Big\{ ||P_n - P||^*_{\mathcal{F}} \Big\}_{n \in \mathbb{N}}$$

inverzní sub-martingal (viz. definice A.3) a konverguje s.j. ke konstantě. Proto

$$\forall P \in \mathcal{P}(\Omega) : ||P_n - P||^*_{\mathcal{F}} \xrightarrow[n \longrightarrow \infty]{s.j.} 0,$$

což bylo dokázat.

□

Poznámka 1.11. Tedy bez dodání dalších předpokladů dostáváme následující řetězec implikací

$$\begin{aligned} &\mathcal{F} \text{ je silně stejnoměrná Gl.-Cant.} \implies \mathcal{F} \text{ je slabě stejnoměrná Gl.-Cant.} \\ &\implies \mathcal{F} \text{ je } P- \text{ Gl.-Cant. } \forall P \in \mathcal{P}(\Omega). \end{aligned}$$

2. VC třídy množin a funkcí

V předchozí kapitole jsme dokázali zobecněnou Glivenkovu-Cantelliho větu. Jeden ze základních předpokladů této věty je

$$\log\ N\Big(\epsilon, \mathcal{F}_M^*, L_1(P_n)\Big) = o_{P^*}(n),$$

kde $\mathcal{F}_M^*$ obsahuje funkce tvaru $x \longmapsto f(x)\mathbb{1}\{F^*(x) \leq M\}$.

Vidíme, že pokrývací číslo třídy $\mathcal{F}_M^*$ závisí na pravděpodobnostní míře P skrze empirickou normu $L_1(P_n)$. Přirozeně vznikají následující otázky:

- Existují třídy funkcí pro které pokrývací čísla jsou stejnoměrně omezené v pravděpodobnostních mírách?
- Pokud ano, jaký mají vztah k Glivenkovým-Cantelliho třídám?

Ukazuje se, že takové třídy existují (tzv. VC třídy). Základní vlastností těchto tříd je to, že jsou „malé“ z kombinatorického hlediska.

V první podkapitole zavedeme Vapnikovy-Červonenkisovy třídy množin (VC třídy) a ukážeme některé základní příklady VC tříd množin. Pak dokážeme důležitou kombinatorickou vlastnost VC tříd množin známou jako Sauerovo lemma.

V druhé podkapitole zobecníme pojem VC tříd pro funkce. Pak ukážeme vztah mezi VC třídami množin a třídami jejich indikátorových funkcí. Dokážeme, že pokrývací čísla VC tříd funkcí jsou stejnoměrně omezené. Kapitolu ukončíme tím, že uvedeme základní vztahy mezi VC třídami funkcí a Glivenkovými-Cantelliho třídami.

2.1 VC třídy množin

Nechť S je množina a $\mathcal{C}$ je systém jejich podmnožin. Označme $\{x_1,...,x_n\}$ nějakou konečnou podmnožinu S. Dále řekneme, že $\mathcal{C}$ *vybere* určitou podmnožinu $A \subseteq \{x_1,...,x_n\}$, pokud ji lze zapsat jako $A = C \cap \{x_1,...,x_n\}$ pro nějaké $C \in \mathcal{C}$. Řekneme, že třída $\mathcal{C}$ *tříští (shatters)* množinu $\{x_1,...,x_n\}$, pokud $\mathcal{C}$ vybere každou z jejich 2^n podmnožin.

VC-dimenzí $\mathcal{V}(\mathcal{C})$ třídy množin $\mathcal{C}$ nazveme nejmenší n pro které neexistuje žádná množina velikosti n, která je tříštěna třídou $\mathcal{C}$. Formálněji, nechť $\Delta_n(\mathcal{C},x_1,x_2,...,x_n)$ je počet podmnožin $\{x_1,...,x_n\}$, které jsou vybrané třídou $\mathcal{C}$, tzn.

$$\Delta_n(\mathcal{C},x_1,x_2,...,x_n) = \#\Big\{C \cap \{x_1,...,x_n\} : C \in \mathcal{C}\Big\}.$$

Potom VC-dimenze je dána jako

$$\mathcal{V}(\mathcal{C}) = \inf\Big\{n \in \mathbb{N} : \max_{\mathrm{x_1,...,x_n}} \Delta_n(\mathcal{C},x_1,x_2,...,x_n) < 2^n\Big\}.$$

Definice 2.1. *Nechť $\mathcal{C}$ je systém měřitelných množin, pak ho nazveme **VC-třídou**, jestli $\mathcal{V}(\mathcal{C}) < \infty$.*

Poznámka 2.1. Všimněme si, že v předchozí definici implicitně se předpokládá existence nějakého měřitelného prostoru $(S, \mathcal{S})$ takového, že $\mathcal{C} \subset \mathcal{S}$.

Z definice VC-třída vybere striktně méně než 2^n podmnožin z každé množiny velikosti $n \geq \mathcal{V}(\mathcal{C})$.

Poznámka 2.2. Nechť $\mathcal{C}$ je VC třída a $\mathcal{C}' \subset \mathcal{C}$. Zvolme $n \in \mathbb{N}$ a $x_1, ..., x_n \in S$, potom platí

$$\Delta_n(\mathcal{C}', x_1, ..., x_n) \leq \Delta_n(\mathcal{C}, x_1, ..., x_n).$$

Tedy

$$\max_{x_1,...,x_n} \Delta_n(\mathcal{C}', x_1, ..., x_n) \leq \max_{x_1,...,x_n} \Delta_n(\mathcal{C}, x_1, ..., x_n).$$

$$\Big\{n : \max_{\mathrm{x_1,...,x_n}} \Delta_n(\mathcal{C},x_1,x_2,...,x_n) < 2^n\Big\} \subset \Big\{n : \max_{\mathrm{x_1,...,x_n}} \Delta_n(\mathcal{C}',x_1,x_2,...,x_n) < 2^n\Big\}.$$

Odsud

$$\mathcal{V}(\mathcal{C}') \leq \mathcal{V}(\mathcal{C}).$$

To jest $\mathcal{C}'$ je VC třída.

Příklad 2.1. (i) Nechť $S = \mathbb{R}$. Předpokládejme, že máme následující třídu množin

$$\mathcal{C} = \Big\{(-\infty, t] \mid t \in \mathbb{R}\Big\}.$$

Potom třída $\mathcal{C}$ tříští každou jednoprvkovou množinu $\{x_1\}$, ale už netříští dvouprvkové množiny $\{x_1,x_2\}$, protože třída $\mathcal{C}$ nemůže vybrat množinu $\Big\{\max\{x_1, x_2\}\Big\}$. Odsud $\mathcal{V}(\mathcal{C}) = 2$.

(ii) Nechť $S = \mathbb{R}^d$ (kde $d \in \mathbb{N}$). Položme následující třídu množin

$$\mathcal{C} = \Big\{\{x : \theta^T x > y\} : \theta \in \mathbb{R}^d, y \in \mathbb{R}\Big\}.$$

Potom $\mathcal{C}$ je VC třída (důkaz lze najít v knize van de Geer (2000), str. 40).

(iii) Nechť $S = \mathbb{R}^2$. Položme následující třídu množin

$$\mathcal{C} = \Big\{C \subset \mathbb{R}^2 \mid C \text{ je uzavřená a konvexní}\Big\}.$$

Potom $\mathcal{V}(\mathcal{C}) = \infty$ (tzn. $\mathcal{C}$ není VC-třídou). To ověříme následovně: rozmístíme body $x_1,x_2,...,x_n$ na hranici jednotkového kruhu S^1. Nechť $A \subseteq \{x_1,x_2,...,x_n\}$, potom

- symbolem $conv(A)$ budeme značit konvexní obal množiny A,
- symbolem $\overline{conv(A)}$ budeme značit uzávěr konvexního obálu množiny A.

Všimněme si, že

$$\overline{conv(A)} \in \mathcal{C} \text{ a } \overline{conv(A)} \subset S^1.$$

Odsud $\overline{conv(A)} \cap \{x_1,x_2,...,x_n\} = A$.

Věta 2.1. *Nechť $\mathcal{C}$ a $\mathcal{D}$ jsou VC třídy, pak platí*

(i) $\mathcal{C}^c := \{C^c : C \in \mathcal{C}\}$ *je VC třída.*

(ii) $\mathcal{C} \cup \mathcal{D} := \{C \cup D : C \in \mathcal{C}, D \in \mathcal{D}\}$ *je VC třída.*

(iii) Jestli $\mathcal{C}$ je VC třída množiny $\mathcal{X}$ a $\mathcal{D}$ je VC třída množiny $\mathcal{Y}$, potom $\mathcal{C} \times \mathcal{D}$ je VC třída množiny $\mathcal{X} \times \mathcal{Y}$.

Tato věta (důkaz viz. van der Vaart a Wellner, 1996, lemma 2.6.17) udává základní vlastnosti VC tříd množin.

Dále budeme používat následující značení: $\Psi_n(\mathcal{C},x_1,x_2,...,x_n)$ je počet podmnožin množiny $\{x_1,x_2,...,x_n\}$ tříštěných třídou $\mathcal{C}$, tzn.

$$\Psi_n(\mathcal{C},x_1,x_2,...,x_n) = \#\{C : C \subseteq \{x_1,...,x_n\}, C \text{ je tříštěna třídou } \mathcal{C}\}.$$

Důkaz následujícího lemmatu najděte v knize van der Vaart a Wellner (1996), lemma 2.6.2.

Lemma 2.2. *Nechť $\{x_1,x_2,...,x_n\}$ je libovolná konečná podmnožina S. Potom počet podmnožin $\Delta_n(\mathcal{C},x_1,x_2,...,x_n)$, které jsou vybrané třídou $\mathcal{C}$ je shora omezen počtem podmnožin množiny $\{x_1,x_2,...,x_n\}$ tříštěných třídou $\mathcal{C}$, tzn.*

$$\Delta_n(\mathcal{C},x_1,x_2,...,x_n) \le \Psi_n(\mathcal{C},x_1,x_2,...,x_n).$$

Důležitou vlastností tříd VC je to, že z množiny velikosti n si mohou vybrat nejvýše $O(n^{\mathcal{V}(\mathcal{C})-1})$ podmnožin, což je podstatně méně než $2^n - 1$. Tento kombinatorický výsledek je znám jako Sauerovo lemma.

Věta 2.3. ***(Sauerovo lemma)*** *Pro VC-třídu $\mathcal{C}$ s indexem $\mathcal{V}(\mathcal{C})$, platí*

$$\max_{\mathrm{x}_1,...,\mathrm{x}_n} \Delta_n(\mathcal{C},x_1,x_2,...,x_n) \le \sum_{j=0}^{\mathcal{V}(\mathcal{C})-1} \binom{n}{j}.$$

Speciálně, výraz na levé straně roste nejvýše polynomiálně řádu $O(n^{\mathcal{V}(\mathcal{C})-1})$ při $n \to \infty$

Důkaz. Víme, že VC-třída netříští žádnou množinu velikosti $\mathcal{V}(\mathcal{C})$. Tedy všechny tříštěné množiny jsou velikosti nejvýše $\mathcal{V}(\mathcal{C}) - 1$. Všimněme si, že

$$\forall x_1, x_2, ..., x_n \in \mathcal{X} : \Psi_n(\mathcal{C},x_1,x_2,...,x_n) \le \sum_{j=0}^{\mathcal{V}(\mathcal{C})-1} \binom{n}{j}.$$

Zbytek věty plyne z předchozího lemmatu.

□

Nechť $\mathcal{C}$ je VC třída množin. Potom symbolem $\mathcal{F}_\mathcal{C}$ značíme následující třídu funkcí

$$\mathcal{F}_\mathcal{C} = \{\mathbb{1}_C : C \in \mathcal{C}\}.$$

Důkaz následující věty je proveden v knize van der Vaart a Wellner (1996) Věta 2.6.4.

Věta 2.4. *Existuje konečná konstanta K taková, že pro libovolnou VC-třídu $\mathcal{C}$, pro každou pravděpodobnostní míru Q na $(S,\mathcal{S})$, pro libovolné $\varepsilon \in (0,1)$,*

$$N\Big(\varepsilon, \mathcal{F}_\mathcal{C}, L_1(Q)\Big) \le K \left(\frac{1}{\varepsilon}\right)^{\mathcal{V}(\mathcal{C})-1}.$$

V dalším textu symbolem $\mathcal{Q}(S)$ budeme značit množinu všech pravděpodobnostních měr na $(S,\mathcal{S})$.

Všimněme si, že podle předchozí věty platí

$$\sup_{Q\in\mathcal{Q}(S)} N\Big(\varepsilon,\mathcal{F}_{\mathcal{C}},L_1(Q)\Big) \le K\left(\frac{1}{\varepsilon}\right)^{\mathcal{V}(\mathcal{C})-1}.$$

To jest pokrývací čísla třídy $\mathcal{F}_{\mathcal{C}}$ jsou omezené stejnoměrně v pravděpodobnostních mírách.

2.2 VC třídy funkcí

Nechť $f : S \to \mathbb{R}$ je funkce pak její *podgrafem (subgraph)* rozumíme následující množinu:

$$\Big\{(x;t)\in S\times\mathbb{R} : t < f(x)\Big\}.$$

Definice 2.2. *Nechť $\mathcal{F}$ je třída měřitelných reálných funkcí na $(S,\mathcal{S})$. Potom $\mathcal{F}$ nazveme VC-třídou funkcí, pokud systém všech jejich podgrafů tvoří VC-třídu (na $S\times\mathbb{R}$). $\mathcal{V}(\mathcal{F})$ bude značit VC-dimenzi třídy podgrafů funkcí v systému $\mathcal{F}$.*

Následující věta ukazuje souvislost mezi VC třídami množin a třídami jejich indikátorů.

Věta 2.5. *$\mathcal{C}$ je VC třída množin právě tehdy když $\mathcal{F}_{\mathcal{C}} := \{\mathbb{1}_C : C\in\mathcal{C}\}$ je VC třída funkcí. Dokonce platí*

$$\mathcal{V}(\mathcal{C}) = \mathcal{V}(\mathcal{F}_{\mathcal{C}}).$$

Důkaz. Podle definice VC-indexu tříd funkcí musíme pracovat s podgrafy funkcí $\mathcal{F}_{\mathcal{C}}$. Tedy

$$\begin{aligned}&\Big\{(x,t) : t < \mathbb{1}_C(x)\Big\} = \Big\{(x,t) : t<0,\ \mathbb{1}_C(x)=0\Big\} \cup \Big\{(x,t) : t<1,\ \mathbb{1}_C(x)=1\Big\}\\ &= \Big\{(x,t) : t<0,\ x\notin C\Big\} \cup \Big\{(x,t) : t<1,\ x\in C\Big\}\\ &= \Big[C^c\times(-\infty,0)\Big] \cup \Big[C\times(-\infty,1)\Big].\end{aligned}$$

Tedy pracujeme s následující třídou množin

$$\mathcal{D} = \Big\{\Big[C^c\times(-\infty,0)\Big] \cup \Big[C\times(-\infty,1)\Big] : C\in\mathcal{C}\Big\}.$$

Předpokládejme, že $\mathcal{D}$ je VC třída a $\mathcal{V}(\mathcal{D}) = k$. Tedy žádná množina

$$\Big\{(x_1,0),...,(x_k,0)\Big\}$$

není tříštěná třídou $\mathcal{D}$. Proto žádná množina $\{x_1,...,x_k\}$ není tříštěná třídou $\mathcal{C}$. To jest $\mathcal{V}(\mathcal{C}) \le \mathcal{V}(\mathcal{D})$.

Teď naopak předpokládejme, že $\mathcal{C}$ je VC třída a $\mathcal{V}(\mathcal{C}) = k$.

Dokážeme, že žádná množina $\Big\{(x_1,t_1),...,(x_k,t_k)\Big\} \subset S\times\mathbb{R}$ není tříštěná třídou $\mathcal{D}$.

Kdyby existovalo $i \in \{1, ..., k\}$ takové, že $t_i < 0$, pak neumíme nevybrat množinu $\{(x_i, t_i)\}$ třídou $\mathcal{D}$.

Tedy můžeme předpokládat, že $\forall i \in \{1, ..., k\} : t_i \in [0, 1)$. Kdyby

$$\Big\{(x_1, t_1), ..., (x_k, t_k)\Big\}$$

byla tříštěna třídou $\mathcal{D}$, pak $\{x_1, ..., x_k\}$ je tříštěná třídou $\mathcal{C}$, což je spor.

To jest žádná množina $\Big\{(x_1, t_1), ..., (x_k, t_k)\Big\} \subset S \times \mathbb{R}$ není tříštěná třídou $\mathcal{D}$. Odsud $\mathcal{V}(\mathcal{D}) \leq \mathcal{V}(\mathcal{C})$.

□

Příklad 2.2. (i) Podle předchozí věty a příkladu 2.1 máme, že

$$\Big\{\mathbb{1}(-\infty, t] \mid t \in \mathbb{R}\Big\}$$

je VC třída funkcí.

(ii) Pokud $\mathcal{F}$ je konečně dimenzionální vektorový prostor obsahující měřitelné funkce $f : (S, \mathcal{S}) \longrightarrow (\mathbb{R}, \mathcal{B})$. Potom $\mathcal{F}$ je VC třída funkcí. Důkaz lze najít v knize van der Vaart a Wellner (1996), lemma 2.6.15.

Následující věta nám ukazuje, že pokrývací čísla libovolné VC třídy funkcí jsou také stejnoměrně omezená.

Věta 2.6. *Nechť $\mathcal{F}$ je VC-třída funkcí, nechť F je měřitelná, obálková funkce pro systém $\mathcal{F}$. Potom existuje $K < \infty$ takové, že pro každou pravděpodobnostní míru Q na $(S, \mathcal{S})$ (takovou, že $||F||_{L_1(Q)} > 0$) a pro každé $\varepsilon \in (0{,}1)$ platí*

$$N\Big(\varepsilon ||F||_{L_1(Q)},\ \mathcal{F},\ L_1(Q)\Big) \leq K \left(\frac{1}{\varepsilon}\right)^{\mathcal{V}(\mathcal{F})-1}.$$

Důkaz je proveden v knize van der Vaart a Wellner (1996) Věta 2.6.7. Je technický a my jej v práci neuvádíme.

Teď máme matematický aparát abychom uvedli vztah mezí VC-třídami a Glivenkovými-Cantelliho třídami.

Věta 2.7. *Nechť $\mathcal{F}$ je VC-třída funkcí, která je P-měřitelná. Nechť F je příslušná měřitelná obálková funkce taková, že $0 < P(F) < \infty$. Pak $\mathcal{F}$ je $P-$Glivenkova-Cantelliho třída.*

Důkaz. Nejdříve si všimněme, že

$$\forall M > 0\ \forall \varepsilon > 0\ \forall n \in \mathbb{N}:\ N\Big(\varepsilon, \mathcal{F}^*_M, L_1(P_n)\Big) \leq N\Big(\varepsilon, \mathcal{F}, L_1(P_n)\Big).$$

Proto podle věty 1.6 stačí ověřit následující podmínku

$$\forall \varepsilon > 0 :\ \log\ N\Big(\varepsilon, \mathcal{F}, L_1(P_n)\Big) = o_{P^*}(n). \tag{2.1}$$

Také si všimněme, že pro libovolnou pravděpodobnostní míru Q na $(S, \mathcal{S})$ platí

$$f(\varepsilon) = N\Big(\varepsilon, \mathcal{F}, L_1(Q)\Big) \text{ je nerostoucí v } \varepsilon.$$

Proto stačí podmínku (2.1) ověřit jen pro malá ε. Z nezápornosti obálkové funkce máme

$$||F||_{L_1(P_n)} = \frac{1}{n}\sum_{i=1}^{n} F(X_i).$$

Uvažujme množiny

$$A_n := \Big\{0 < ||F||_{L_1(P_n)} \leq 2\ P(F)\Big\}.$$

Z měřitelnosti obálkové funkce F jsou množiny A_n měřitelné pro každé $n \in \mathbb{N}$. Podle silného zákonu velkých čísel platí

$$||F||_{L_1(P_n)} \xrightarrow[n\to\infty]{s.j} P(F).$$

Odsud dostáváme, že

$$P(A_n) \xrightarrow[n\to\infty]{} 1.$$

Připomeňme, že podle poznámky 1.4 pro libovolné kladné r pokrývací čísla $N(r, \mathcal{F}, L_1(P_n))$ jsou náhodná zobrazení.

Tedy podle Věty 2.6 máme

$$\exists K > 0\ \exists W > 0\ \forall n \in \mathbb{N}\ \forall \omega \in A_n\ \forall \varepsilon \in (0,1):$$

$$N\Big(2\varepsilon P(F),\ \mathcal{F},\ L_1(P_n)\Big)(\omega) \leq N\Big(\varepsilon ||F||_{L_1(P_n)},\ \mathcal{F},\ L_1(P_n)\Big)(\omega) \leq K\left(\frac{1}{\varepsilon}\right)^W.$$

Tedy máme

$$\forall \varepsilon \in (0,1):\ P^*\left(N\Big(2\varepsilon P(F),\ \mathcal{F},\ L_1(P_n)\Big) \leq K\left(\frac{1}{\varepsilon}\right)^W\right) \xrightarrow[n\to\infty]{} 1.$$

Což lze ekvivalentně přepsat následovně

$$\forall \widetilde{\varepsilon} \in \Big(0,\ 2P(F)\Big):\ P^*\left(N\Big(\widetilde{\varepsilon},\ \mathcal{F},\ L_1(P_n)\Big) \leq K\left(\frac{2P(F)}{\widetilde{\varepsilon}}\right)^W\right) \xrightarrow[n\to\infty]{} 1.$$

Odsud dostaneme, že $\forall \widetilde{\varepsilon} \in \Big(0,\ 2P(F)\Big)\ \exists K_{\widetilde{\varepsilon}} > 0:$

$$P^*\left(\frac{\log N\Big(\widetilde{\varepsilon}, \mathcal{F}, L_1(P_n)\Big)}{n} \leq \frac{K_{\widetilde{\varepsilon}}}{n}\right) \xrightarrow[n\to\infty]{} 1.$$

Zvolme libovolně $\widetilde{\varepsilon} \in \Big(0,\ 2P(F)\Big)$ a $\eta > 0$, potom

$$P^*\left(\frac{\log N\Big(\widetilde{\varepsilon}, \mathcal{F}, L_1(P_n)\Big)}{n} > \eta\right) \leq P^*\left(\frac{\log N\Big(\widetilde{\varepsilon}, \mathcal{F}, L_1(P_n)\Big)}{n} > \eta,\ A_n\right) + P^*(A_n^c)$$

$$\leq P^*\left(\frac{K_{\widetilde{\varepsilon}}}{n} > \eta,\ A_n\right) + P(A_n^c) \xrightarrow[n\to\infty]{} 0.$$

Odsud už snadno plyne, že podmínka (2.1) je splněna. To jest podle věty 1.6 třída funkcí $\mathcal{F}$ je $P-$Glivenkova-Cantelliho.

$\square$

Věta 2.8. *Nechť $\mathcal{F}$ je VC třída funkcí. Předpokládejme, že $\mathcal{F}$ je omezená kladnou konstantou R a je $P-$měřitelná pro každé $P \in \mathcal{P}(\Omega)$, potom $\mathcal{F}$ je silně stejnoměrná Glivenkova-Cantelliho.*

Důkaz. Podle věty 1.7 stačí dokázat, že pro $\forall \varepsilon > 0$

$$\sup_{Q \in \mathcal{Q}_n(S)} \log\, N\Big(\varepsilon R, \mathcal{F}, L_1(Q)\Big) = o(n). \tag{2.2}$$

Stejně jako v důkazu věty 2.7, jelikož pokrývací čísla jsou nerostoucí v ε, tak stačí podmínku (2.2) ověřit pro každé $\varepsilon \in (0,1)$. Z věty 2.6 víme, že

$$\exists K > 0\ \forall \varepsilon \in (0,1)\ \forall Q \in \mathcal{Q}(S):\ N\Big(\varepsilon R, \mathcal{F}, L_1(Q)\Big) \leq K \left(\frac{1}{\varepsilon}\right)^{\mathcal{V}(\mathcal{F})-1},$$

kde $\mathcal{Q}(S)$ je množina všech pravděpodobnostních měr na $(S, \mathcal{S})$. Tudíž

$$\sup_{Q \in \mathcal{Q}(S)} \log N\Big(\varepsilon R, \mathcal{F}, L_1(Q)\Big) \leq \log\left(K \left(\frac{1}{\varepsilon}\right)^{\mathcal{V}(\mathcal{F})-1}\right) = o(n).$$

Tedy také platí (2.2), jelikož $\mathcal{Q}_n(S) \subset \mathcal{Q}(S)$ pro $\forall n \in \mathbb{N}$.

□

3. Aplikace v teorii strojového učení

V předchozích kapitolách jsme zavedli pojmy Glivenkových-Cantelliho tříd funkcí a VC tříd funkcí. V této kapitole ukážeme, že tyto třídy funkcí hrají důležitou roli v matematických základech strojového učení. V kapitole 3.1 nejprve popíšeme úlohu strojového učení. V kapitole 3.2 formálně zavedeme pojem „učení“. Hlavní aplikací bude dokázat základní větu statistického učení, což uděláme v kapitole 3.3.

3.1 Struktura úlohy strojového učení

Zhruba řečeno pod *strojovým učením* rozumíme numerické metody využívající zkušenosti k vytváření přesných předpovědí (tady pod zkušenostmi rozumíme informace kterou máme). Hlavním cílem strojového učení je konstrukce efektivních algoritmů pro vytváření předpovědi. Známou klasifikační úlohou strojového učení je tzv. spamový problém. Spamový problém se zabývá automatickou klasifikací emailů na dvě kategorie: SPAM nebo NESPAM. Použijeme spamový problém pro ilustraci základních definic.

- **VSTUP:**
  - *Definiční obor (Domain set):* Je to libovolná množina $\mathcal{X}$ (množina objektů, které chceme označit). V spamovém problému je $\mathcal{X}$ množina emailů. Obvykle prvky definičního oboru jsou reprezentovány vektorem atributů (např. v spamovém problému to může být délka zprávy, jméno odesílatele, přítomnost klíčových slov ve zprávě). Nechť $\mathcal{A}_{\mathcal{X}}$ bude značit σ-algebru množiny $\mathcal{X}$.
  - *Množina označení (Label set).* Tato množina obsahuje kategorie přiřazené k vzorům. V klasifikačních problémech, vzorům jsou přiřazeny specifické kategorie. V našem spamovém problému máme dvě kategorie: SPAM a NESPAM. Nechť $\mathcal{Y}$ bude značit množinu označení a $\mathcal{A}_{\mathcal{Y}}$ bude značit σ-algebru množiny $\mathcal{Y}$.
  - *Trénovací data (Training data).* Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor. Předpokládejme, že
  $$\left(\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}\right)$$
  je náhodný výběr na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$ s hodnotami v $(\mathcal{X} \times \mathcal{Y}, \mathcal{A}_{\mathcal{X}} \otimes \mathcal{A}_{\mathcal{Y}})$. Trénovacími daty nazveme libovolnou realizaci náhodného výběru. Trénovací data budeme značit následovně
  $$S_n^{\omega} = \left(\binom{X_1(\omega)}{Y_1(\omega)}, \ldots, \binom{X_n(\omega)}{Y_n(\omega)}\right), \text{ kde } \omega \in \Omega.$$
- **VÝSTUP:** Na výstupu chceme mít funkci $h : (\mathcal{X},\mathcal{A}_{\mathcal{X}}) \longrightarrow (\mathcal{Y}, \mathcal{A}_{\mathcal{Y}})$. Tuto funkce nazýváme *prediktorem (predictor, hypothesis, classifier).* Prediktor nám dovolí predikovat označení libovolného prvku definičního oboru. V dalším textu $\mathcal{H}$ značí množinu prediktorů.

Všimněme si, že prediktor je měřitelná funkce. To můžeme předpokládat, protože třída měřitelných funkcí (definovaných na $(\mathcal{X},\mathcal{A}_{\mathcal{X}})$ s hodnotami v $(\mathcal{Y},\mathcal{A}_{\mathcal{Y}})$) je dostatečně velká. Proto se stačí pro praktické účely strojového učení omezit na měřitelné funkce.

3.2 Probably Approximately Correct learning

V předchozí kapitole 3.1 jsme popsali úlohu strojového učení, ale neodpověděli jsme na otázku jsme na otázku: „Co znamená učení?“. V této kapitole odpovíme na tuto otázku.

Algoritmem učení budeme nazývat zobrazení, které libovolným trénovacím datům S_n^ω přiřadí nějaký prediktor h_n^ω. To jest

$$ALG(S_n^\omega) = h_n^\omega.$$

Pro dané $h \in \mathcal{H}$ zavedeme *chybu klasifikátoru (error of a classifier, generalization error, risk)* následujícím způsobem

$$L_P(h) := P\Big\{\omega \in \Omega : h(X_1(\omega)) \neq Y_1(\omega)\Big\}.$$

Cílem našeho algoritmu je najít takový prediktor h, který minimalizuje chybu $L_P(h)$. Všimněme si, že jelikož neznáme rozdělení P, tak nemůžeme spočítat chybu klasifikátoru. Můžeme ji však odhadnout tzv. *empirickou chybou (training error, empirical error, empirical risk)*

$$L_n^\omega(h) := \frac{\sum_{i=1}^n \mathbb{1}\{h(X_i(\omega)) \neq Y_i(\omega)\}}{n}.$$

Řekněme, že prediktor h_n^ω minimalizuje empirické riziko, pokud

$$h_n^\omega \in \underset{h\in\mathcal{H}}{argmin}\ L_n^\omega(h).$$

V dalším textu $\widehat{h}_n^\omega$ bude značit prediktor minimalizující empirické riziko. Také řekneme, že algoritmus učení ALG minimalizuje empirické riziko, pokud pro libovolná trénovací data S_n^ω platí

$$ALG(S_n^\omega) \in \underset{h\in\mathcal{H}}{argmin}\ L_n^\omega(h).$$

Hlavní myšlenku procesu „učení“ lze vysvětlit následujícím příkladem. Uvažujme hru s dvěma hráči (Hráč 1, Hráč 2).

-
  - Předpokládejme, že Hráč 1 zafixuje libovolný interval $[a_0,b_0]$.
  - Pak náhodně (podle nějakého pravidla P) vybere prvky $x_1, ..., x_n \in \mathbb{R}$ a označí prvky x_i, pokud patří do intervalu $[a_0,b_0]$.
  - Ukáže Hráčovi 2 výběr $\{(x_i,y_i)\}_{i=1}^n$, kde $\forall i \in \{1, ..., n\} : y_i \in \{0,1\}$.
- Na základě pozorovacích dat $\{(x_i, y_i)\}_{i=1}^n$ Hráč 2 odhadne interval $[a_0,b_0]$ intervalem $[\hat{a}_n, \hat{b}_n]$.

Předpokládejme, že můžeme spočítat skutečnou pravděpodobnost toho, že interval $[\hat{a}_n, \hat{b}_n]$ nepředpovídá správné označení. Pokud tato pravděpodobnost je malá, pak říkáme, že Hráč 2 vyhrál.

Jestli pro dostatečně velké bude n Hráč 2 vyhrávat s libovolně velkou pravděpodobností (bez ohledu na to, jaký interval $[a_0,b_0]$ a jaké pravidlo P Hráč 1 použije), pak říkáme, že celá hra je Probably Approximately Correct learnable (PAC learnable).

Připomeňme, že symbolem $\mathcal{P}(\Omega)$ značíme množinu všech pravděpodobnostních měr na $(\Omega, \mathcal{A})$.

Definice 3.1. *Množinu prediktorů $\mathcal{H}$ nazveme PAC learnable, pokud platí:*

$$\exists ALG\ \forall \varepsilon > 0 : \lim_{n \longrightarrow \infty} \sup_{P \in \mathcal{P}(\Omega)} P^* \Big\{ \omega \in \Omega : L_P(h_n^\omega) - \inf_{h \in \mathcal{H}} L_P(h) > \varepsilon \Big\} = 0. \tag{3.1}$$

Poznámka 3.1. V literatuře množinu prediktorů $\mathcal{H}$ z definice 3.1 se také někdy označují jako *PAC agnostic learnable* (viz. Shalev-Shwartz a Ben-David, 2014, definice 3.3). Přičemž $\mathcal{H}$ je PAC learnable, pokud $\inf_{h \in \mathcal{H}} L_P(h) = 0$. V naši práci tyto dva případy nebudeme rozlišovat.

Poznámka 3.2. V knize Shalev-Shwartz a Ben-David (2014) definice PAC learnable se uvádí s pravděpodobností P (místo P^*). To odpovídá tomu, že $\forall \varepsilon > 0$ je množina

$$\Big\{ \omega \in \Omega : L_P(h_n^\omega) - \inf_{h \in \mathcal{H}} L_P(h) > \varepsilon \Big\}$$

měřitelná.

Poznámka 3.3. Předpokládejme, že $\mathcal{H}$ je PAC learnable. Zafixujme libovolnou pravděpodobnostní míru $P \in \mathcal{P}(\Omega)$. Potom (3.1) dává

$$L_P(h_n) \xrightarrow[n \longrightarrow \infty]{P^*} \inf_{h \in \mathcal{H}} L_P(h), \tag{3.2}$$

kde připomeňme, že h_n je náhodné neboť závisí na realizaci trénovacích dat S_n^ω, tj. na $\omega \in \Omega$.

V celé kapitole 3 se směřujeme k důkazu základní věty statistického učení. Tato věta je obvykle dokazovaná pro množiny prediktorů, které jsou PAC learnable (viz. Shalev-Shwartz a Ben-David, 2014, věta 6.7). Všimněme si, že definici PAC learnable lze chápat jako konvergence $L_P(h_n)$ ve vnější pravděpodobnosti k $\inf_{h \in \mathcal{H}} L_P(h)$, přičemž tato konvergence je stejnoměrná v $P \in \mathcal{P}(\Omega)$.

V naši práci dokážeme silnější verzi základní věty statistického učení. Toto zobecnění bude spočívat, že místo stejnoměrné konvergence v pravděpodobnosti budeme uvažovat stejnoměrnou konvergenci skoro jistě. Takovou množinu prediktorů budeme v této práci nazývat Almost Sure Approximately Correct learnable (ASAC learnable).

Definice 3.2. *Množinu prediktorů $\mathcal{H}$ nazveme ASAC learnable, pokud platí:*

$$\exists ALG\ \forall \varepsilon > 0 : \lim_{n \longrightarrow \infty} \sup_{P \in \mathcal{P}(\Omega)} P^* \Big\{ \omega \in \Omega : \sup_{m \geq n} \Big(L_P(h_m^\omega) - \inf_{h \in \mathcal{H}} L_P(h) \Big) > \varepsilon \Big\} = 0.$$

Teď si uvedeme příklad množiny prediktorů, která je PAC learnable.

Příklad 3.1. Nechť $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0{,}1\}$. Mějme následující množinu prediktorů

$$\mathcal{H} = \{h_{a,b} : a < b,\ a,\ b \in \mathbb{R}\},$$

kde $h_{a,b}(x) := \mathbb{1}_{[a,b]}(x)$ pro $x \in \mathbb{R}$.

Předpokládejme, že $\forall P \in \mathcal{P}(\Omega)\ \exists a_0, b_0 \in \mathbb{R},\ a_0 < b_0, \forall \omega \in \Omega :$

$$Y_1(\omega) = \mathbb{1}\{X_1(\omega) \in [a_0{,}b_0]\}. \tag{3.3}$$

Jelikož Y_1 je dané determenisticky jako funkce náhodné veličiny X_1, pak stačí uvažovat, že pravděpodobnostní míra $P \in \mathcal{P}(\Omega)$ určuje rozdělení X_1. Dokážeme, že $\mathcal{H}$ je PAC learnable.

Řešení. Všimněme si, že podle předpokladu (3.3) platí, že

$$\forall P \in \mathcal{P}(\Omega)\ \exists a_0, b_0 \in \mathbb{R},\ a_0 < b_0 :\ L_P(h_{a_0,b_0}) = 0.$$

Tedy pro každé $P \in \mathcal{P}(\Omega)$ platí

$$\inf_{h \in \mathcal{H}} L_P(h) = 0.$$

Zvolme libovolné $P \in \mathcal{P}(\Omega)$. Symbolem F_X označíme distribuční funkci náhodné veličiny X_1 při rozdělení P.

V dalším textu pro každé $a \in \mathbb{R}$ použijeme následující značení

$$F_X(a^-) := \lim_{x \longrightarrow a_-} F(x).$$

Předpokládejme, že máme trénovací data S_n^ω. Položme

$$\widehat{a}_n(\omega) = \min\{X_i(\omega) : Y_i(\omega) = 1, i \in \{1, 2, ..., n\}\},$$

$$\widehat{b}_n(\omega) = \max\{X_i(\omega) : Y_i(\omega) = 1, i \in \{1, 2, ..., n\}\}.$$

Všimněme si, že $\widehat{a}_n$ a $\widehat{b}_n$ jsou náhodné veličiny, protože minimum a maximum jsou měřitelné funkce (viz. Cohn, 2013, tvrzení 2.1.4).

Nechť $\widehat{I}_\omega := [\widehat{a}_n(\omega), \widehat{b}_n(\omega)]$. Všimněme si, že $\widehat{I}_\omega$ implicitně závisí na rozsahu výběru n, ale pro jednoduchost značení tuto závislost nebudeme zdůrazňovat. Uvažujme následující algoritmus

$$ALG(S_n^\omega) = h_{\widehat{I}_\omega}.$$

Všimněme si, že podle předpokladu (3.3) platí

$$\{X_i(\omega) : Y_i(\omega) = 1, i \in \{1, 2, ..., n\}\} \subseteq J \text{ pro } \forall\ \omega \in \Omega.$$

To jest $\widehat{I}_\omega \subseteq J$ pro $\forall \omega \in \Omega$. Tedy dostaneme

$$a_0 \leq \widehat{a}_n(\omega) \leq \widehat{b}_n(\omega) \leq b_0 \text{ pro } \forall\ \omega \in \Omega.$$

Položme $A_n^\omega = [a_0, \widehat{a}_n(\omega)]$ a $B_n^\omega = [\widehat{b}_n(\omega), b_0]$. Potom

$$\begin{aligned} L_P(h_{\widehat{I}_\omega}) &= \int_{\{h_{\hat{I}_\omega} \neq \mathbb{1}_{[a_0,b_0]}\}} dP_X(x) = \int_{[a_0, \hat{a}_n(\omega))} dP_X(x) + \int_{(\hat{b}_n(\omega), b_0]} dP_X(x) \\ &\leq F_X(\widehat{a}_n(\omega)) - F_X(a_0^-) + F_X(b_0) - F_X(\widehat{b}_n(\omega)). \end{aligned}$$

Z měřitelnosti distribuční funkce F_X dostaneme

$$\{\omega \in \Omega : F_X(\widehat{a}_n(\omega)) - F_X(a_0^-) + F_X(b_0) - F_X(\widehat{b}_n(\omega)) > \varepsilon\} \in \mathcal{A}.$$

Zvolme libovolné $\varepsilon > 0$, potom

$$\begin{aligned}
&P^*\{\omega \in \Omega : L_P(h_{\widehat{I_\omega}}) > \varepsilon\} \\
&\leq P\{\omega \in \Omega : F_X(\widehat{a}_n(\omega)) - F_X(a_0^-) + F_X(b_0) - F_X(\widehat{b}_n(\omega)) > \varepsilon\} \\
&\leq P\left\{\omega : F_X(\widehat{a}_n(\omega)) - F_X(a_0^-) > \frac{\varepsilon}{2}\right\} + P\left\{\omega : F_X(b_0) - F_X(\widehat{b}_n(\omega)) > \frac{\varepsilon}{2}\right\}.
\end{aligned}$$

Na chvíli předpokládejme, že pro zvolené $\varepsilon > 0$ platí

$$\left\{x \in \mathbb{R} : F_X(x) \geq F_X(a_0^-) + \frac{\varepsilon}{2}\right\} \neq \emptyset. \tag{3.4}$$

V opačném případě

$$\forall P \in \mathcal{P}(\Omega)\ \forall n \in \mathbb{N} :\ P\left\{\omega \in \Omega : F_X(\widehat{a}_n(\omega)) - F_X(a_0^-) > \frac{\varepsilon}{2}\right\} = 0.$$

Tedy položme

$$y := \inf\left\{x \in \mathbb{R} : F_X(x) \geq F_X(a_0^-) + \frac{\varepsilon}{2}\right\}.$$

Jelikož distribuční funkce je zprava spojitá, potom

$$F_X(y) \geq F_X(a_0^-) + \frac{\varepsilon}{2}.$$

Tedy dostaneme

$$\begin{aligned}
&P\left\{F_X(\widehat{a}_n) - F_X(a_0^-) > \frac{\varepsilon}{2}\right\} \leq P\Big\{X_i \notin [a_0, y]\ \forall i \in \{1, ..., n\}\Big\} \\
&\overset{iid}{=} \Big(P\Big\{X_1 \notin [a_0, y]\Big\}\Big)^n = \Big(1 - P\Big\{X_1 \in [a_0, y]\Big\}\Big)^n \leq \left(1 - \frac{\varepsilon}{2}\right)^n.
\end{aligned}$$

Analogicky lze ukázat, že

$$P\left\{F_X(b_0) - F_X(\hat{b}_n) > \frac{\varepsilon}{2}\right\} \leq \left(1 - \frac{\varepsilon}{2}\right)^n.$$

Připomeňme z matematické analýzy, že platí následující nerovnost

$$(1 - x) \leq e^{-x} \text{ pro } x \geq 0.$$

Tedy celkem dostaneme

$$P\{\omega \in \Omega : L_P(h_{\widehat{I_\omega}}) > \varepsilon\} \leq 2e^{-\varepsilon n/2}.$$

Zvolme libovolné $\delta > 0$, potom pro $\forall P \in \mathcal{P}(\Omega)$ a $\forall n \geq \lceil \frac{2}{\varepsilon} \log(\frac{2}{\delta}) \rceil$

$$P\{\omega \in \Omega : L_P(h_{\widehat{I_\omega}}) > \varepsilon\} \leq \delta.$$

To jest $\mathcal{H}$ je PAC learnable.

$\square$

3.3 Základní věta statistického učení

V této kapitole se zaměříme na důkaz základní věty statistického učení (viz. věta 3.2). Zhruba řečeno základní věta statistického učení dává charakterizace úloh, které se „dají učit“.

Všimněme si, že definici 3.1 jsme zavedli pro konkrétní množinu prediktorů $\mathcal{H}$. Obvykle $\mathcal{H}$ odpovídá našim apriorním znalostem o zkoumaném problému. Přirozeně vzniká otázka, jestli existuje universální algoritmus, který je „úspěšný“ nezávislé na zkoumaném problému. Následující věta tvrdí, že podobný algoritmus neexistuje, pokud $|\mathcal{X}|$ je alespoň dva krát větší, něž je rozsah trénovacích dat.

Připomeňme si, že $\mathcal{A}_{\mathcal{X}}$ je $\sigma-$algebra na $\mathcal{X}$ (viz. kapitola 3.1). V dalším textu budeme předpokládat následující

$$\forall x \in \mathcal{X}: \ \{x\} \in \mathcal{A}_{\mathcal{X}}. \tag{3.5}$$

Poznámka 3.4. Všimněme si, že předpoklad (3.5) není omezující, protože chceme uvažovat i všechna diskrétní rozdělení v $\mathcal{P}(\Omega)$.

Věta 3.1. ***(No-Free-Lunch Theorem)*** *Nechť ALG je libovolný algoritmus pro úlohu binární klasifikace. Předpokládejme, že* $n \leq \frac{|\mathcal{X}|}{2}$. *Potom existuje* $P \in \mathcal{P}(\Omega)$ *takové, že*

- $\exists f : \mathcal{X} \longrightarrow \{0,1\} : L_P(f) = 0.$
- $\Big\{\omega \in \Omega : L_P(h_n^\omega) \geq \frac{1}{8}\Big\} \in \mathcal{A}.$
- $P\Big\{\omega \in \Omega : L_P(h_n^\omega) \geq \frac{1}{8}\Big\} \geq \frac{1}{7}.$

Důkaz. Tato věta je dokázána v knize Shalev-Shwartz a Ben-David (2014), jako věta 5.1. Z tohoto důkazu však není na první pohled jasné, proč by měla být množina

$$\left\{\omega \in \Omega : L_P(h_n^\omega) \geq \frac{1}{8}\right\} \tag{3.6}$$

měřitelná. Proto se soustředíme jenom na důkaz této měřitelnosti.

Zvolme $C \subset \mathcal{X}$ takové, že $|C| = 2n$. Existuje $T = 2^{2n}$ možných funkcí z C do $\{0,1\}$. Označme je $f_1, ..., f_T$. Všimněme si, že

$$\forall x \in \mathcal{X}\ \forall y \in \{0,1\} : \ \Big\{X_1 = x, Y_1 = y\Big\} \in \mathcal{A} \text{ (to plyne z předpokladu 3.5).}$$

Pro každé $i \in \{1, ..., T\}$ zadefinujme pravděpodobnostní míru P_i na $(\Omega, \mathcal{A})$ následujícím způsobem

$$P_i\Big(X_1 = x, Y_1 = y\Big) = \begin{cases} \frac{1}{2n}, & \text{jestli } y = f_i(x). \\ 0, & \text{jinak.} \end{cases}$$

Existuje $K = (2n)^n$ různých konečných posloupností n příkladů z C. Označíme tyto posloupnosti jako $S_1, ..., S_K$.

Jestliže $S_j = (x_1, ..., x_n)$, potom použijeme následující značení

$$S_j^i := \Big(\Big(x_1, f_i(x_1)\Big), ..., \Big(x_n, f_i(x_n)\Big)\Big).$$

Pokud je tedy skutečné rozdělení P_i, pak jako trénovací data S_n^ω obdržíme jednu z množin $S_1^i, ..., S_K^i$. Tedy máme

$$L_{P_i}(h_n^\omega) = L_{P_i}(ALG(S_n^\omega)) = \sum_{j=1}^{K} L_{P_i}\Big(ALG(S_j^i)\Big)\mathbb{1}\Big[S_n^\omega = S_j^i\Big]. \tag{3.7}$$

Všimněme si, že $\forall j \in \{1, ..., K\}$ $\forall i \in \{1, ..., T\}$ platí

$$\Big\{\omega : S_n^\omega = S_j^i\Big\} = \bigcap_{r=1}^{n} \Big\{\omega : X_r(\omega) = x_r\Big\} \in \mathcal{A}.$$

Proto funkce z (3.7) je měřitelná $\forall i \in \{1, ..., T\}$. Podle důkazu z knihy Shalev-Shwartz a Ben-David (2014) hledané rozdělení P je jedno z $P_1, ..., P_T$. Tedy množina (3.6) je měřitelná.

$\square$

Věta 3.1 tvrdí, že pro libovolný algoritmus binární klasifikace a pro dostatečně velké $\mathcal{X}$ existuje takové rozdělení na kterém tento algoritmus „selže“. Tedy, pokud $|\mathcal{X}| = \infty$ a

$$\mathcal{H} = \Big\{\text{všechny funkce z } \mathcal{X} \text{ do } \{0, 1\}\Big\}. \tag{3.8}$$

Potom podle věty 3.1 platí

$$\forall n \in \mathbb{N}\ \forall ALG\ \exists P \in \mathcal{P}(\Omega) :\ P\left\{\omega \in \Omega : L_P(h_n^\omega) \geq \frac{1}{8}\right\} \geq \frac{1}{7}$$

a zároveň $\inf\limits_{h\in\mathcal{H}} L_P(h) = 0$. Celkem dostáváme, že

$$L_P(h_n) \underset{n\longrightarrow\infty}{\overset{P}{\not\longrightarrow}} 0.$$

Tudíž $\mathcal{H}$ (tj. množina všech možných prediktorů) není PAC learnable.

Věta 3.2. ***(Základní věta statistického učení)*** *Nechť $\mathcal{H}$ je množina prediktorů z $\mathcal{X}$ do $\{0,1\}$. Označme*

$$\mathcal{F}_{\mathcal{H}} := \Big\{\mathbb{1}_{\{h(x)\neq y\}}(x,y) : h \in \mathcal{H}\Big\}.$$

Předpokládejme, že $\mathcal{F}_{\mathcal{H}}$ je $P-$měřitelná (viz. definice 1.8) $\forall P \in \mathcal{P}(\Omega)$. Pak následující tvrzení jsou ekvivalentní

(i) $\mathcal{F}_{\mathcal{H}}$ je silně stejnoměrná Glivenkova-Cantelliho třída funkcí.

(ii) $\mathcal{H}$ je ASAC learnable.

(iii) $\mathcal{H}$ je VC-třída funkcí.

Důkaz. Nejprve dokážeme $(i) \implies (ii)$. Za tímto účelem si všimněme, že

$$\sup_{f\in\mathcal{F}_{\mathcal{H}}} |P_n(f)(\omega) - P(f)| = \sup_{h\in\mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{h(X_i(\omega)) \neq Y_i(\omega)\} - \mathbb{E}(\mathbb{1}\{h(X_1) \neq Y_1\})\right|$$
$$= \sup_{h\in\mathcal{H}} |L_n^\omega(h) - L_P(h)|. \tag{3.9}$$

Z vlastnosti infima platí, že

$$\forall n \in \mathbb{N}\ \exists h_n^* \in \mathcal{H}:\ L_P(h_n^*) - \inf_{h\in\mathcal{H}} L_P(h) < \frac{1}{n}.$$

Nechť $\widehat{h}_n^\omega \in \operatorname*{argmin}_{h\in\mathcal{H}} L_n^\omega(h)$, potom

$$\begin{aligned}&L_P(\widehat{h}_n^\omega) - L_P(h_n^*) = L_P(\widehat{h}_n^\omega) - L_n^\omega(\widehat{h}_n^\omega) + L_n^\omega(\widehat{h}_n^\omega) - L_P(h_n^*)\\ &\le L_P(\widehat{h}_n^\omega) - L_n^\omega(\widehat{h}_n^\omega) + L_n^\omega(h_n^*) - L_P(h_n^*).\end{aligned}$$

Tedy s využitím (3.9)

$$\begin{aligned}&L_P(\widehat{h}_n^\omega) - L_P(h_n^*) \le |L_P(\widehat{h}_n^\omega) - L_n^\omega(\widehat{h}_n^\omega)| + |L_n^\omega(h_n^*) - L_P(h_n^*)|\\ &\le 2\sup_{h\in\mathcal{H}}|L_n^\omega(h) - L_P(h)| = 2\sup_{f\in\mathcal{F}_\mathcal{H}}|P_n(f)(\omega) - P(f)|.\end{aligned}$$

Celkem dostaneme následující

$$L_P(\widehat{h}_n^\omega) - L_P(h_n^*) \le 2\sup_{f\in\mathcal{F}_\mathcal{H}}|P_n(f)(\omega) - P(f)|.$$

Tedy dostaneme

$$\begin{aligned}&L_P(\widehat{h}_n^\omega) - \inf_{h\in\mathcal{H}} L_P(h) \le L_P(\widehat{h}_n^\omega) - L_P(h_n^*) + L_P(h_n^*) - \inf_{h\in\mathcal{H}} L_P(h)\\ &\le 2\sup_{f\in\mathcal{F}_\mathcal{H}}|P_n(f)(\omega) - P(f)| + \frac{1}{n}.\end{aligned}$$

Dále dostáváme, že $\forall P \in \mathcal{P}(\Omega)$ platí

$$\begin{aligned}&P^*\Big\{\omega\in\Omega : \sup_{m\ge n}\Big(L_P(\widehat{h}_m^\omega) - \inf_{h\in\mathcal{H}} L_P(h)\Big) > \varepsilon\Big\} \le P^*\Big(2\sup_{m\ge n}||P_m - P||_{\mathcal{F}_\mathcal{H}} + \frac{1}{n} > \varepsilon\Big)\\ &= P^*\left(\sup_{m\ge n}||P_m - P||_{\mathcal{F}_\mathcal{H}} > \frac{\varepsilon - \frac{1}{n}}{2}\right)\end{aligned}$$

Je zřejmé, že existuje n_ε takové, že pro každé $n \ge n_\varepsilon$ platí

$$\frac{\varepsilon - \frac{1}{n}}{2} > \frac{\varepsilon}{4}.$$

Tedy pro každé $n \ge n_\varepsilon$ dostáváme, že

$$P^*\Big\{\omega\in\Omega : \sup_{m\ge n}\Big(L_P(\widehat{h}_m^\omega) - \inf_{h\in\mathcal{H}} L_P(h)\Big) > \varepsilon\Big\} \le P^*\left(\sup_{m\ge n}||P_m - P||_{\mathcal{F}_\mathcal{H}} > \frac{\varepsilon}{4}\right).$$

Podle předpokladu $\mathcal{F}_\mathcal{H}$ je silně stejnoměrná Gl.-Cant. třída funkcí, proto

$$\forall\varepsilon > 0 : \lim_{n\longrightarrow\infty}\ \sup_{P\in\mathcal{P}(\Omega)} P^*\Big\{\omega\in\Omega : \sup_{m\ge n}\Big(L_P(\widehat{h}_m^\omega) - \inf_{h\in\mathcal{H}} L_P(h)\Big) > \varepsilon\Big\} = 0.$$

To jest $\mathcal{H}$ je ASAC learnable.

Implikace $(ii) \implies (iii)$ dokážeme sporem. Pro spor předpokládejme, že $\mathcal{H}$ není VC-třída. To jest $\mathcal{V}(\mathcal{H}) = \infty$. Jelikož $\mathcal{H}$ je množina funkcí z $\mathcal{X}$ do $\{0,1\}$. Tedy ji můžeme přepsat následujícím způsobem

$$\mathcal{H} = \Big\{\mathbb{1}\{h = 1\} : h \in \mathcal{H}\Big\}.$$

Podle věty 2.5 platí

$$\mathcal{V}(\mathcal{C}) = \mathcal{V}(\mathcal{H}) = \infty, \text{ kde } \mathcal{C} = \Big\{\{x : h(x) = 1\} : h \in \mathcal{H}\Big\}.$$

Jelikož $\mathcal{H}$ je ASAC learnable, proto i PAC learnable.
Tedy $\exists ALG\ \forall\varepsilon > 0\ \forall\delta > 0\ \exists n^{\varepsilon}_{\delta} \in \mathbb{N}\ \forall n \geq n^{\varepsilon}_{\delta}$:

$$\sup_{P\in\mathcal{P}(\Omega)} P^*\Big\{\omega \in \Omega : L_P(h^{\omega}_n) - \inf_{h\in\mathcal{H}} L_P(h) > \varepsilon\Big\} < \delta.$$

Zvolme $\varepsilon < \frac{1}{8}$, $\delta < \frac{1}{7}$ nalezneme příslušné n^{ε}_{δ}. Zvolme $n \geq n^{\varepsilon}_{\delta}$, potom existují $x_1, ..., x_{2n} \in \mathcal{X}$ takové, že $\mathcal{C}$ tříští $\{x_1, ..., x_{2n}\}$.

Z důkazu věty 3.1 (No-Free-Lunch Theorem) víme, že pro každý algoritmus lze sestrojit $P \in \mathcal{P}(\Omega)$ takové, že

$$\sum_{i=1}^{2n} P\Big\{X_1 = x_i\Big\} = 1 \text{ a } P\left\{\omega \in \Omega : L_P(h^{\omega}_n) \geq \frac{1}{8}\right\} \geq \frac{1}{7}.$$

Jelikož $\mathcal{C}$ tříští $\{x_1, ..., x_{2n}\}$, pak

$$\mathcal{H}_{|\{x_1,...,x_{2n}\}} = \Big\{h_{|\{x_1,...,x_{2n}\}} : h \in \mathcal{H}\Big\}$$

obsahuje všechny funkce z $\{x_1, ..., x_{2n}\}$ do $\{0, 1\}$. Proto existuje $h \in \mathcal{H}$ takové, že $L_P(h) = 0$.

Tedy dostaneme

$$\frac{1}{7} > \delta > P\Big\{\omega \in \Omega : L_P(h^{\omega}_n) \geq \varepsilon\Big\} \geq \frac{1}{7}, \text{ což je spor.}$$

Teď dokážeme, že $(iii) \implies (i)$. Nejprve si ukážeme, že $\mathcal{F}_{\mathcal{H}}$ je VC třída funkcí. Jelikož $\mathcal{F}_{\mathcal{H}}$ je třída indikátorových funkcí, tedy stačí dokázat, že

$$\mathcal{C}_{\mathcal{H}} = \left\{\Big\{(x,y) \in \mathcal{X} \times \{0,1\} : h(x) \neq y\Big\} : h \in \mathcal{H}\right\} \text{ je VC třída množin.}$$

Všimněme si, že

$$\begin{aligned}
&\Big\{(x,y) : h(x) \neq y\Big\} = \Big\{(x,1) : h(x) \neq 1\Big\} \cup \Big\{(x,0) : h(x) \neq 0\Big\} \\
&= \Big\{(x,1) : h(x) = 0\Big\} \cup \Big\{(x,0) : h(x) = 1\Big\} \\
&= \left(\Big\{x : h(x) = 0\Big\} \times \Big\{1\Big\}\right) \cup \left(\Big\{x : h(x) = 1\Big\} \times \Big\{0\Big\}\right).
\end{aligned}$$

Víme, že $\left\{\Big\{x : h(x) = 1\Big\} : h \in \mathcal{H}\right\}$ je VC třída množiny $\mathcal{X}$ (to plyne z věty 2.5).

Podle věty 2.1$(i.)$ platí, že $\left\{\Big\{x : h(x) = 0\Big\} : h \in \mathcal{H}\right\}$ je také VC třída množin.

Podle věty 2.1($iii.$) platí, že systémy množin $\Big\{\Big\{x : h(x) = 1\Big\} \times \Big\{0\Big\} : h \in \mathcal{H}\Big\}$ a $\Big\{\Big\{x : h(x) = 0\Big\} \times \Big\{1\Big\} : h \in \mathcal{H}\Big\}$ jsou také VC třídy množiny $\mathcal{X} \times \{0,1\}$.

Podle věty 2.1($ii.$) platí, že

$$\Bigg\{\Big(\Big\{x : h_1(x) = 1\Big\} \times \Big\{0\Big\}\Big) \cup \Big(\Big\{x : h_2(x) = 0\Big\} \times \Big\{1\Big\}\Big) : h_1, h_2 \in \mathcal{H}\Bigg\}$$

je VC třída množiny $\mathcal{X} \times \{0,1\}$. Tedy podle poznámky 2.2 dostaneme, že $\mathcal{C}_{\mathcal{H}}$ je VC třída (to jest $\mathcal{F}_{\mathcal{H}}$ je VC třída funkcí). Je zřejmé, že $\mathcal{F}_{\mathcal{H}}$ je omezená, tudíž podle věty 2.8 je $\mathcal{F}_{\mathcal{H}}$ silně stejnoměrná Glivenkova-Cantelliho.

□

Poznámka 3.5. Všimněme si, že v důkazu implikace $(i) \implies (ii)$ jsme použili algoritmus, který minimalizuje empirické riziko. To nemusí být ideální v praktických problémech protože může nastat tzv. *overfitting* (viz. Shalev-Shwartz a Ben-David, 2014, str.35).

Poznámka 3.6. Také si všimněme, že pokud $\mathcal{H}$ je spočetná, potom $\mathcal{F}_{\mathcal{H}}$ je $P-$měřitelná pro každé $P \in \mathcal{P}(\Omega)$.

Poznámka 3.7. Základní vetu statistického učení jsme dokázali pro množinu prediktorů, která je dokonce ASAC learnable. Obvykle je tato věta formulovaná v literatuře tak, že množina prediktorů je PAC learnable (viz. Shalev-Shwartz a Ben-David, 2014, věta 6.7).

Závěr

V této práci jsme podrobně dokázali zobecněnou Glivenkovu-Cantelliho větu pomocí pokrývacích čísel a vyslovili jsme stejnoměrný zákon velkých čísel. Pak jsme zadefinovali VC třídy funkcí a uvedli jsme základní vztahy mezi VC třídami funkcí a Glivenkovými-Cantelliho třídami funkcí. Hlavní aplikací bylo dokázat základní větu statistického učení.

Tuto práci lze rozšířit alternativním způsobem důkazu zobecněné Glivenkovy-Cantelliho věty pomocí tzv. *bracketing numbers*. Bracketing numbers představují jiný způsob jak měřit velikost třídy funkcí $\mathcal{F}$.

V naši práci jsme zobecnili zákon velkých čísel. Dalším krokem by mohlo být zobecnění centrální limitní věty (viz. van der Vaart a Wellner, 1996, kapitola 2.8.2).

Je třeba si všimnout, že jsme dokázali základní větu statistického učení jen pro úlohu binární klasifikace. Když množina označení není dvouprvková, pak věta 3.2 nemusí platit. Ale existují další nutné a postačující podmínky, aby množina prediktorů byla PAC learnable (viz. Shalev-Shwartz a kol., 2010).

A. Apendix

Lemma A.1. ***(Čebyševova nerovnost)*** *Nechť X je náhodná veličina z L^2 a $a > 0$. Pak platí*

$$P(|X - \mathbb{E}X| \geq a) \leq \frac{var(X)}{a^2}.$$

Důkaz je proveden v knize Georgii (2013), tvrzení 5.5.

Definice A.1. *Nechť $T \neq \emptyset, X_t \in L, t \in T$. Řekněme, že $\{X_t\}_{t\in T}$ jsou stejnoměrně integrovatelné (SMI) náhodné veličiny, když*

$$\lim_{c\to\infty} \sup_{t\in T} \mathbb{E}\Big(|X_t|\mathbb{1}\Big\{|X_t| \geq c\Big\}\Big) = 0.$$

Věta A.2. ***(Postačující podmínky pro SMI)*** *Nechť $X_t \in L, t \in T$. Postačující podmínky pro SMI systému $\{X_t\}_{t\in T}$ jsou*

- $\exists Y \in L_1 \; \forall t \in T : |X_t| \leq Y$

Důkaz lze najít v knize Lachout (2004), lemma 5.12.

Věta A.3. ***(Charakterizace konvergenci v L_p)*** *Buď $X, X_n \in L$, $n \in \mathbb{N}$. Nechť $1 \leq p \leq \infty$. Potom $X_n, X \in L_p$, $n \in \mathbb{N}$ a $X_n \xrightarrow{L^p} X$ právě tehdy když $X_n \xrightarrow{P} X$ a $|X_n|^p$, $n \in \mathbb{N}$ jsou SMI.*

Důkaz je proveden v Lachout (2004), věta 6.10.

Definice A.2. ***(Inverzní Filtrace)*** *Inverzní filtraci nazýváme klesající posloupnost σ-algeber $\Sigma_0 \supseteq \Sigma_1 \supseteq \Sigma_2 \supseteq \cdots$. Pro konzistenci budeme používat zápornou indexaci $\mathcal{F}_{-n} = \Sigma_n, n \geq 0$. To jest $\cdots \subset \mathcal{F}_{-n} \subset \cdots \subset \mathcal{F}_{-2} \subset \mathcal{F}_{-1} \subset \mathcal{F}_0$ je rostoucí posloupnost σ-algeber.*

Definice A.3. ***(Inverzní Sub-Martingal)*** *Nechť $(M_{-n})_{n\geq 0}$ je posloupnost L^1 náhodných veličin. Tuto posloupnost nazveme inverzním sub-martingalem vzhledem k filtraci $(\mathcal{F}_{-n})_{n\geq 0}$ jestli*

- $M_{-n} \in L^1(\mathcal{F}_{-n})$, $\forall n \geq 0$,
- $\mathbb{E}(M_{-n}|\mathcal{F}_{-n-1}) \geq M_{-n-1}$, $\forall n \geq 0$.

Věta A.4. ***(o konvergenci inverzního sub-martingalu)*** *Nechť $\mathcal{F}$ je třída měřitelných funkcí, F je obálková funkce pro systém $\mathcal{F}$, přičemž*

$$P^*(F) := \mathbb{E}^*\left[F(X_i)\right] < \infty.$$

Definujme filtraci $(\Sigma_n)_{n\geq 0}$, kde Σ_n je σ-algebra generovaná měřitelnými funkcemi $h : \mathcal{X}^\infty \to \mathbb{R}$, které jsou symetrické v prvních n-argumentech (kde $\mathcal{X}^\infty = \mathcal{X} \times ... \times \mathcal{X} \times ...$). Potom

$$\mathbb{E}\left(||P_n - P||^*_{\mathcal{F}} \mid \Sigma_n\right) \geq ||P_{n+1} - P||^*_{\mathcal{F}}, \text{ s.j.}$$

Mimoto $(||P_n - P||^*_{\mathcal{F}})_{n\in\mathbb{N}}$ tvoří inverzní sub-martingal a konverguje s.j. ke konstantě.

Viz. van der Vaart a Wellner (1996), Lemma 2.4.5.

Lemma A.5. *Nechť* $(\Omega,\mathcal{A},P)$ *je pravděpodobnostní prostor a* $(\mathbb{D},d)$ *je metrický prostor. Předpokládejme, že* X *je borelovsky měřitelná funkce na* $(\mathbb{D},d)$ *a* $\{X_n\}_{n=1}^{\infty}$ *je posloupnost náhodných zobrazení z* Ω *do* $\mathbb{D}$*. Pak následující tvrzení jsou ekvivalentní:*

(i) $X_n \xrightarrow{s.j^*} X$.

(ii) $\sup\limits_{m\geq n} d(X_m,X) \xrightarrow{P^*} 0$.

Důkaz je proveden v knize van der Vaart a Wellner (1996), Lemma 1.9.2., Lemma 1.9.3.

Seznam použité literatury

Cohn, D. L. (2013). *Measure Theory.* Birkhäuser, Boston. ISBN 978-1-4614-6955-1.

Dudley, R. M., Giné, E. a Zinn, J. (1991). Uniform and universal Glivenko-Cantelli classes. *Journal of Theoretical Probability*, **4**(3), 485–510.

Georgii, H.-O. (2013). *Stochastics introduction to probability and statistics.* De Gruyter, Berlin. ISBN 978-3-11-029254-1.

Ghorpade, S. R. a Limaye, B. V. (2006). *A Course in Calculus and Real Analysis.* Springer, New York. ISBN 0-387-30530-0.

Giné, E. a Zinn, J. (1984). Some Limit Theorems for Empirical Processes. *The Annals of probability*, **12**(4), 929–989.

Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, **58**(301), 13–30.

Lachout, P. (2004). *Teorie Pravděpodobnosti.* Karolinum, Praha. ISBN 80-246-0872-3.

Shalev-Shwartz, S. a Ben-David, S. (2014). *Understanding Machine Learning From Theory to Algorithms.* Cambridge University Press, New York. ISBN 978-1-107-05713-5.

Shalev-Shwartz, S., Shamir, O., Srebro, N. a Sridharan, K. (2010). Learnability, Stability and Uniform Convergence. *Journal of Machine Learning Research*, **11**, 2635–2670.

van de Geer, S. (2000). *Empirical Processes in M-Estimation.* Cambridge University Press, New York. ISBN 978-0-521-65002-1.

van der Vaart, A. W. a Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York. ISBN 978-1-4757-2547-6.