Divergence pravděpodobnostních distribucí a statistická informace

Igor Vajda*

Úvod

První známou mírou informace v náhodném pozorování se stala Fisherova informace, viz např. Anděl [2] anebo podrobněji Lehmann a Casella [17]. Šlo o informaci ohledně neznámého parametru, na kterém závisí rozdělení tohoto pozorování. Zakladatel teorie informace Shannon [28] přišel s myšlenkou měřit informaci náhodného pozorování ohledně neznámého náhodného parametru poklesem entropie parametru v důsledku tohoto pozorování. Podle něj je tedy informace dána rozdílem apriorní a aposteriorní entropie parametru. Tuto informaci lze ekvivalentně vyjádřit jako jistou logaritmickou míru neshody (divergence) $D(P, Q)$ mezi společnou distribucí P parametru a pozorování a součinem Q distribuce parametru a distribuce pozorování. Víme, že Q může být také společnou distribucí parametru a pozorování, ale jen v případě jejich vzájemné nezávislosti. Shannonova informace je tedy mírou statistické asociace mezi pozorováním a parametrem.

Logaritmická míra divergence $D(P, Q)$ zavedená v teorii informace má v případě diskrétních pozitivních distribucí $P = (p_1, p_2, \ldots, p_k)$, $Q = (q_1, q_2, \ldots, q_k)$ tvar

$$D(P, Q) = \sum_{k=1}^{K} p_k \ln \frac{p_k}{q_k}, \tag{1}$$

kde místo přirozeného logaritmu ln lze užít i logaritmus o jiném základu. Kullback a Leibler [15] interpretovali tuto divergenci jako diskriminační informaci při testování hypotézy $\mathcal{H} : P$ proti alternativě $\mathcal{A} : Q$, zobecnili ji na libovolné distribuce (pravděpodobnostní míry) P, Q na obecném měřitelném pozorovacím prostoru $(\mathcal{X}, \mathcal{S})$ a ukázali, že její zachování charakterizuje postačitelnost transformací pozorovacího prostoru. Pozoruhodné vlastnosti této divergence zkoumala později řada autorů, např. Perez [26], Gelfand a spol. [9] a zejména Kullback [14], který systematicky studoval její statistické aplikace. Mimo jiné ukázal, že Fisherova informace o daném parametru $\theta_0 \in \Theta$ je mírou citlivosti distribuce pozorování P_θ na změnu parametru $\theta \in \Theta$ v okolí θ_0, pokud tuto citlivost měříme divergencí $D(P_\theta, P_{\theta_0})$. Možnost použití jiných divergencí a dokonce možnost zobecnění

*Ústav teorie informace a automatizace AV ČR
Pod Vodárenskou věží 4, 182 08 Praha 8; e-mail: vajda@utia.cas.cz

Fisherovy informace touto cestou byla později studována Kaganem [13] a Vajdou [29]. Z nejnovějších prací ohledně zobecněné Fisherovy informace lze citovat například [12].

V šedesátých letech přispěli do oblasti informace a divergence pravděpodobnostních distribucí významným způsobem Csiszár [4, 5] a De Groot [6, 7]. Csiszár (a nezávisle též Ali a Silvey [1]) si povšimli, že základem žádoucích vlastností divergence (1) je konvexnost funkce $f(t) = t \ln t$ a zavedli obecnější f-divergence $D_f(P, Q)$ pro konvexní funkce $f : (0, \infty) \mapsto \mathbb{R}$, které jsou v bodě 1 ryze konvexní a standardizované do tvaru $f(1) = 0$. Pro P, Q uvažované v (1) jsou tyto zobecněné divergence dané formulí

$$D_f(P, Q) = \sum_{k=1}^{K} q_k \, f\left(\frac{p_k}{q_k}\right). \tag{2}$$

Toto umožnilo vedle diskriminační informace typu (1) pokrýt jednotným způsobem další známé statistické divergence jako například Pearsonovu divergenci, která má pro P, Q uvažované v (1) tvar

$$\chi^2(P, Q) = \sum_{k=1}^{K} \frac{(p_k - q_k)^2}{q_k} \tag{3}$$

nebo Hellingerovu divergenci, která má pro tato P, Q tvar

$$H^2(P, Q) = \sum_{k=1}^{K} \left(\sqrt{p_k} - \sqrt{q_k}\right)^2. \tag{4}$$

Obecný pojem f-divergence dovolil vyložit mnoho zdánlivě různorodých metod statistického rozhodování (odhadování a testování) jako speciální případy univerzální metody založené na divergenci empirického a teoretického rozdělení a jako takové je dále zobecnit. Pod zobecňováním nemáme na mysli jen odvozování alternativních metod pro klasické statistické modely, ale též rozšiřování metod z klasických modelů na modely s pozorováními typu náhodných procesů a náhodných polí. Obecné f-divergence jsou totiž dobře definované i na abstraktních pozorovacích prostorech příslušných takovým pozorováním.

De Groot přispěl v jiném směru než Csiszár, totiž novou definicí statistické informace, odlišné od Fisherovy definice. Jeho informace je obdobou Shannonovy, pouze místo entropie neznámého parametru užívá Bayesovo riziko při rozhodování o tomto parametru (speciálním případem je riziko při ztrátové funkci typu 0 - 1, t.j. Bayesova pravděpodobnost chyby při identifikaci parametru). Tudíž statistickou informací v náhodném pozorování je rozdíl mezi apriorním a aposteriorním Bayesovým rizikem. Na rozdíl od Shannonovy informace motivované komunikační situací, zde máme jasnou statistickou motivaci. Vzájemná souvislost mezi f-divergencemi $D_f(P, Q)$ a Ge Grootovými statistickými informacemi v dichotomickém rozhodovacím modelu s podmíněnými distribucemi P, Q a různými ztrátovými funkcemi prozkoumali Österreicher a Vajda [24]. Dokázali, že každá statistická informace je f-divergence pro některé f a každá f-divergence je statistická informace pro některou ztrátovou funkci. Navíc dokázali, že všechny f-divergence jsou průměrné statistické informace měřené poklesem Bayesovy pravděpodobnosti chyby v důsledku pozorování, kde průměr se bere přes apriorní pravděpodobnosti $\pi \in (0, 1)$ hypotézy $\mathcal{H} : P$. Různé f-divergence se tedy liší jen různou distribucí vah na prostoru apriorních pravděpodobností π, ale v zásadě jde vesměs o průměrné rozdíly mezi apriorními a aposteriorními Bayesovými pravděpodobnostmi chyb.

Studium divergencí pravděpodobnostních distribucí a jejich využití ve statistickém rozhodování představuje jeden z tradičních mezinárodně uznávaných a mezinárodně podporovaných výzkumných směrů ÚTIA AV ČR. Například v monografii [18] byly systematicky prostudovány obecné vlastnosti f-divergencí a vlastnosti některých důležitých parametrických tříd těchto divergencí. Z dlouhé řady prací věnovaných v rámci tohoto výzkumu statistickým aplikacím f-divergencí lze uvést například [20, 31, 21, 11, 3, 10]. Dále v práci [25] bylo dokázáno, že jedinými mírami divergence, jejichž zachování charakterizuje postačitelnost transformací pozorovacího prostoru, jsou f-divergence. V [30] jsou uvedeny podmínky, při kterých posloupnost kvantování pozorovacího prostoru zaručuje asymptoticky nulové ztráty zobecněné Fisherovy informace i f-divergence. V současnosti jsou metody testování a odhadování založené na f-divergencích empirických a teoretických distribucí předmětem výzkumu v ÚTIA v grantu AV ČR: *Nové výsledky v testování dobré shody* a taktéž představují jeden z hlavních směrů výzkumu v rámci Aplikačního Centra ÚTIA: *Data, algoritmy, rozhodování* zřízeného z iniciativy MŠMT ČR.

V této práci, v sekcích 1 a 2, uvádíme do problematiky divergenčních měr v teorii informace a matematické statistice. V sekci 3 definujeme obecnou třídu f-divergencí a v sekcích 4 a 5 jednak odvozujeme základní vlastnosti těchto divergencí a také objasňujeme jejich vztah ke statistickým informacím. I když tato práce pojednává o vesměs známých pojmech a výsledcích, přece jen je v převážné míře původní a nová. Konkrétně, pojednání o nejznámějších vzdálenostech používaných v teorii pravděpodobnosti a matematické statistice a jejich vztahu k postačujícím statistikám v sekci 2 je nové a příklady v této sekci (i některé další v jiných sekcích) jsou původní. Dále, nové a původní jsou všechny důkazy uvedené v práci. Totiž, důkazy základních vlastností f-divergencí se v předchozí literatuře opíraly o Jensenovu nerovnost, zejména její verzi pro podmíněné distribuce, jejíž rigorózní formulace a důkaz nejsou zrovna jednoduché. V této práci uvádíme nové důkazy těchto vlastností bez jakéhokoliv odkazu na Jensenovy nerovnosti. Nové důkazy jsou jednoduché a opírají se jen o obyčejný Taylorův rozvoj funkce f se zbytkem v integrálním tvaru. Daň, která se za tuto jednoduchost platí, je předpoklad dvojí spojité diferencovatelnosti funkce f. Avšak prakticky všechny f-divergence používané ve statistických aplikacích tento předpoklad splňují. Výjimkou je totální variace, pro kterou základní vlastnosti dokazujeme pomocí aproximací f_n-divergencemi s dvakrát spojitě diferencovatelnými funkcemi f_n. Dále, zmíněným Taylorovým rozvojem nově dokazujeme i větu o reprezentaci f-divergencí průměrnými statistickými informacemi. Naše metoda je podstatně jednodušší než předchozí, založená na Jensenových nerovnostech v práci [24]. V závěru se zmiňujeme o možnosti rozšíření postupu použitého v této práci na f-divergence s libovolnými nediferencovatelnými funkcemi f.

1 Divergence pravděpodobnostních distribucí

Uvažujme standardní model matematické statistiky, kde je dán měřitelný pozorovací prostor $(\mathcal{X}, \mathcal{S})$ a náhodné pozorování X popsané pravděpodobnostním prostorem $(\mathcal{X}, \mathcal{S}, P_0)$. Pravděpodobnostní distribuce P_0 (pravděpodobnostní míra) se považuje za neznámou a dva základní problémy matematické statistiky jsou (i) testovat na základě pozorování X

hypotézu $\mathcal{H} : P_0 \in \mathcal{P}$ resp. (ii) najít na základě pozorování X odhad $\hat{P} \in \mathcal{P}$ neznámého P_0, kde v obou případech $\mathcal{P}$ je některá daná třída pravděpodobnostních distribucí na prostoru $(\mathcal{X}, \mathcal{S})$. Obě tyto úlohy se obvykle řeší na základě vhodné číselné míry divergence $D(P, Q)$ definované pro libovolné dvojice distribucí P, Q na $(\mathcal{X}, \mathcal{S})$.

Termín *divergence* obecně označuje rozdílnost, neshodu, odchylku, rozbíhání se. Pod divergencí $D(P, Q)$ tedy budeme rozumět nezápornou míru rozdílnosti, neshody distribucí P a Q. Očekáváme, že divergence bude reflexivní v obvyklém smyslu

$$D(P, Q) = 0 \quad \text{právě když} \quad P = Q, \tag{5}$$

ale nevyžadujeme nutně ani symetrii $D(P, Q) = D(Q, P)$ ani trojúhelníkovou nerovnost. Proto se vyhýbáme běžnějšímu termínu *vzdálenost*. Místo běžných metrických vlastností spíše požadujeme, aby hodnota divergence $D(P, Q)$ charakterizovala jak snadné je v případě $P_0 \in \{P, Q\}$ odlišení distribuce P od distribuce Q. Tato úloha nemusí být symetrická v P, Q. Například je-li $(\mathcal{X}, \mathcal{S})$ borelovská přímka $(\mathbb{R}, \mathcal{B})$ a P, Q jsou rovnoměrné distribuce na intervalech $(0, 10)$ a $(0, 1)$, pak bezchybné zamítnutí hypotézy $\mathcal{H} : P_0 = Q$ na základě pozorování X je mnohem snazší než bezchybné zamítnutí alternativy $\mathcal{A} : P_0 = P$.

Místo symetrie a trojúhelníkové nerovnosti požadujeme od dobré statistické míry divergence jinou vlastnost a sice monotonii vůči transformacím pozorování

$$D(PT^{-1}, QT^{-1}) \leq D(P, Q) \quad \text{pro} \quad T : (\mathcal{X}, \mathcal{S}) \mapsto (\mathcal{Y}, \mathcal{T}). \tag{6}$$

Zde symbol $(\mathcal{X}, \mathcal{S}) \mapsto (\mathcal{Y}, \mathcal{T})$ znamená, že jde o zobrazení $T : \mathcal{X} \mapsto \mathcal{Y}$, přičemž obraz $\mathcal{Y}$ pozorovacího prostoru $\mathcal{X}$ je vybaven σ-algebrou $\mathcal{T}$ a zobrazení T je měřitelné v běžném smyslu $T^{-1}(\mathcal{T}) \subset \mathcal{S}$. V tomto případě transformaci T nazýváme *statistika*. Dále, PT^{-1} a QT^{-1} jsou pravděpodobnostní distribuce (míry) indukované statistikou T na novém pozorovacím prostoru $(\mathcal{Y}, \mathcal{T})$, definované pro všechna $B \in \mathcal{T}$ vztahy

$$PT^{-1}(B) = P(T^{-1}B) \quad \text{a} \quad QT^{-1}(B) = Q(T^{-1}B).$$

Nerovnost (6) představuje velmi srozumitelnou skutečnost, že totiž transformací pozorování (překódováním statistických dat) nelze docílit vyšší rozlišitelnost pravděpodobnostních distribucí. V jednom důležitém speciálním případě by se však rozlišitelnost měla zachovat. Totiž tehdy, když statistika T je v tradičním statistickém smyslu postačující pro dvojici $\{P, Q\}$. Přirozenou doplňkovou podmínkou k (6) je tudíž invariance divergence vůči postačitelným transformacím dat, tj.

$$D(PT^{-1}, QT^{-1}) = D(P, Q) \quad \text{když } T \text{ je postačující pro } \{P, Q\}. \tag{7}$$

Obvykle požadujeme víc a sice, aby netriviální (konečná) divergence byla úplným invariantem postačitelných transformací, tj. aby kromě (7) platilo

$$T \text{ je postačující pro } \{P, Q\}, \text{ když } D(PT^{-1}, QT^{-1}) = D(P, Q) < \infty. \tag{8}$$

V následující sekci uvedeme nekonečnou třídu divergencí splňujících podmínky (5) – (8).

Ve statistických úlohách (i) a (ii) ze začátku této sekce se pozorování X často transformuje měřitelným způsobem na určitou tzv. *empirickou distribuci* (pravděpodobnostní míru) Q na pozorovacím prostoru $(\mathcal{X}, \mathcal{S})$. (Pod měřitelným způsobem se rozumí, že

pravděpodobnosti $Q(A)$ jevů $A \in \mathcal{S}$ jsou měřitelnými funkcemi pozorování X.) Vedle empirické distribuce Q jsou k dispozici na $(\mathcal{X}, \mathcal{S})$ *teoretické distribuce* P z dané hypotetické rodiny $\mathcal{P}$. Hypotéza $\mathcal{H} : P_0 \in \mathcal{P}$ v úloze (i) se obvykle zamítá, když testovací statistika

$$T(X) = \inf_{P \in \mathcal{P}} D(P, Q) \tag{9}$$

přesáhne určitou kritickou hodnotu $c > 0$. (Poznamenejme, že $\mathcal{S}$-měřitelnost pravděpodobností $Q(A)$, $A \in \mathcal{S}$ obvykle postačí k měřitelnosti statistiky $T(X) : (\mathcal{X}, \mathcal{S}) \mapsto (\mathbb{R}, \mathcal{B})$ definované v (9).) Podobně odhad $\hat{P}$ v úloze (ii) se definuje z podmínky

$$\hat{P} = \mathrm{argmin}_{P \in \mathcal{P}} D(P, Q), \tag{10}$$

resp. z podmínky, aby se příliš nelišil od argumentu minima uvažovaného v (10), viz například [31].

Úloha divergence v řešeních (9) a (10) statistických úloh (i) a (ii) často není na první pohled patrná. Pro ilustraci uvažujme jednoduchý příklad, kdy na konečném součinovém pozorovacím prostoru

$$\mathcal{X} = \{1, \ldots, m\}^n$$

je dána parametrická rodina součinových modelů $\mathcal{P} = \{P_\theta^n : \theta \in \Theta\}$, kde

$$P_\theta = (p_\theta(1), \ldots, p_\theta(m))$$

jsou pozitivní diskrétní distribuce na $\{1, \ldots, m\}$ závislé na parametru θ z některé neprázdné množiny $\Theta \subset \mathbb{R}^S$. Budeme řešit statistickou úlohu (ii), kdy na základě pozorování $X = (X_1, X_2, \ldots, X_n)$ s konečným výběrovým pravděpodobnostním prostorem $(\mathcal{X}, P_0)$ máme najít odhad $\hat{P}$ neznámého rozdělení P_0. O tomto rozdělení budeme předpokládat, že patří do rodiny $\mathcal{P}$, tj. že

$$P_0 = P_{\theta_0}^n,$$

kde θ_0 je některá neznámá hodnota parametru $\theta \in \Theta$. Pozorování X definuje na $\mathcal{X}$ součinové empirické rozdělení Q^n, kde

$$Q = (q(1), \ldots, q(m))$$

je diskrétní rozdělení na $\{1, \ldots, m\}$ dané formulí

$$q(j) = \frac{1}{n} \sum_{i=1}^{n} I(X_i = j), \quad 1 \le j \le m$$

přičemž symbol $I(A)$ rezervujeme pro indikátor jevu $A \subset \mathcal{X}$. Pravděpodobnosti $q(j)$ jsou tedy relativní četnosti jevů $j \in \{1, \ldots, m\}$ ve výběru X.

Uvažujme nyní divergenci (1) s tím, že distribuce $Q = (q_1, \ldots, q_K)$ nemusí nutně být pozitívní. Pro případ $q_k = 0$ přijmeme v (1) konvenci $0 \ln 0 = 0$. Pak obdržíme

$$\begin{aligned} D(P_\theta^n, Q^n) &= nD(P_\theta, Q) = n \sum_{j=1}^{m} q(j) \ln \frac{q(j)}{p_\theta(j)} \\ &= -H(Q^n) - L_n(\theta). \end{aligned}$$

V této formuli

$$H(Q^n) = -n \sum_{j=1}^{m} q(j) \ln q(j)$$

označuje entropii empirického rozdělení Q^n a

$$\begin{aligned} L_n(\theta) &= \sum_{j=1}^{m} \sum_{i=1}^{n} I(X_i = j) \ln p_\theta(j) \\ &= \ln \prod_{i=1}^{n} p_\theta(X_i) \\ &= \ln P_\theta^n(X) \end{aligned}$$

označuje logaritmus věrohodnosti rozdělení $P_\theta^n \in \mathcal{P}$ při daném pozorování X. Metodou minimální divergence (1) získáme tudíž odhad

$$\hat{P} = P_{\hat{\theta}}^n \in \mathcal{P}$$

neznámého rozdělení $P_0 = P_{\theta_0}^n$ kde

$$\hat{\theta} = \text{argmax}_{\theta \in \Theta} L_n(\theta) \tag{11}$$

není nic jiného než maximálně věrohodný odhad neznámého parametru $\theta_0 \in \Theta$ v obecném diskrétním modelu $\{P_\theta = (p_\theta(1), \ldots, p_\theta(m)) : \theta \in \Theta\}$ s n nezávislými pozorováními $X = (X_1, \ldots, X_n)$. Jinými slovy, klasický maximálně věrohodný odhad $\hat{\theta}$ dostaneme jako speciální případ odhadu (10) při divergenci (1). Podobně lze ukázat, že klasický věrohodnostní test hypotézy $\theta_0 \in \Theta$ v obecném diskrétním modelu lze získat jako speciální případ testu hypotézy $\mathcal{H} : P_0 \in \mathcal{P}$ založeného na statistice (9) při divergenci (1).

Tyto výsledky lze rozšířit i na maximálně věrohodné odhady a zobecněné věrohodnostní testy v libovolných (ne nutně diskrétních) statistických modelech (podrobněji viz Liese, Vajda [18]). Jestliže použijeme jiné divergence než (1), pak dostaneme jiné odhady a testy. Některé z nich jsou již využívány ve statistice, jiné zatím nikoliv. Jak již bylo řečeno, systematické studium těchto odhadů a testů a jim příslušných divergencí je součástí probíhajícího výzkumu na ÚTIA AV ČR.

2 Vzdálenosti které nejsou divergencemi

V teorii pravděpodobnosti a matematické statistice se již téměř sto let s úspěchem využívají vhodné vzdálenosti $\rho(P, Q)$ pravděpodobnostních distribucí P, Q. Nejznámější z nich probereme z hlediska způsobilosti vyhovět podmínkám (5) – (8). Protože reflexivnost (5) je u těchto vzdáleností samozřejmá, budeme se zajímat jen o monotonii (6) a invarianci (7) resp. (8).

Pro distribuce P, Q na borelovské přímce $(\mathcal{X}, \mathcal{S}) = (\mathbb{R}, \mathcal{B})$ se běžně používá *Kolmogorova vzdálenost*

$$\rho_K(P, Q) = \sup_{x \in \mathbb{R}} |F(x) - G(x)|, \tag{12}$$

kde $F(x) = P((-\infty, x))$ a $G(x) = Q((-\infty, x))$ jsou (zleva spojité) distribuční funkce příslušné P a Q. Jak známo, Kolmogorova vzdálenost je metrika v prostoru všech pravděpodobnostních distribucí na $(\mathbb{R}, \mathcal{B})$. Tato vzdálenost však není monotonní ve smyslu (6) ani invariantní ve smyslu (7). Vhodnou transformací dat $T : \mathbb{R} \mapsto \mathbb{R}$ můžeme totiž tuto vzdálenost zvětšit. Jako příklad můžeme vzít distribuce P, Q s hustotami

$$\begin{aligned} p(x) &= I(-1 < x < -1/2) + I(0 < x < 1/2), \\ q(x) &= I(-1/2 < x < 0) + I(1/2 < x < 1) \end{aligned}$$

a transformaci

$$T(x) = x\left[I(|x| \geq 1/2) - I(|x| < 1/2)\right],$$

přičemž symbolem $I(A)$ označujeme indikátor jevu A. Zde

$$\rho_K(P, Q) = F(-1/2) - G(-1/2) = F(1/2) - G(1/2) = 1/2$$

a distribuce $\tilde{P} = PT^{-1}$, $\tilde{Q} = QT^{-1}$ mají hustoty

$$\tilde{p}(x) = I(-1 < x < 0), \quad \tilde{q}(x) = I(0 < x < 1)$$

takže

$$\rho_K(\tilde{P}, \tilde{Q}) = \tilde{F}(0) - \tilde{G}(0) = 1,$$

což je ve sporu s monotonií (6). Transformace T je přitom jedno-jednoznačná a tudíž postačující pro jakoukoliv dvojici $\{P, Q\}$. Z tohoto důvodu vzdálenost ρ_K nesplňuje ani podmínku (7). Na druhé straně, modifikovaná transformace

$$T(x) = x\left[I(|x| \geq 1/2)\right]$$

není postačující pro $\{P, Q\}$, protože rozdílný průběh hustot p a q na intervalu $-1/2 < x < 1/2$ není touto transformací postižen. Přitom pro příslušné distribuce $\tilde{P} = PT^{-1}$, $\tilde{Q} = QT^{-1}$ resp. distribuční funkce $\tilde{F}$, $\tilde{G}$ a všechna $-1/2 \leq x \leq 1/2$ platí

$$\rho_K(\tilde{P}, \tilde{Q}) = \tilde{F}(x) - \tilde{G}(x) = 1/2 = \rho_K(P, Q).$$

To znamená, že vzdálenost ρ_K taktéž nesplňuje podmínku (8).

Přesto je Kolmogorova vzdálenost vhodná pro některé speciální statistické aplikace. Například za hypotézy $\mathcal{H} : P_0 = P$ se ukázal jako schůdný výpočet asymptotického rozdělení známé Kolmogorov–Smirnovy statistiky $\rho_K(P, Q_n)$, kde Q_n označuje empirické rozdělení

$$Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

na $(\mathbb{R}, \mathcal{B})$ příslušející nezávislému výběru $X = (X_1, X_2, \ldots, X_n)$ při hypotéze $\mathcal{H}$. Rovněž statistické odhady minimalizující statistiku $\rho_K(P, Q_n)$ v podobném smyslu jako odhady definované v (9), vykazují některé žádoucí vlastnosti, podrobněji viz Kůs [16].

Dalším známým pojmem je *Lévyho vzdálenost*

$$\rho_L(P, Q) = \inf\left\{\varepsilon > 0 : F(x - \varepsilon) - \varepsilon \leq G(x) \leq F(x + \varepsilon) + \varepsilon,\ \forall\, x \in \mathbb{R}\right\},$$

kde P, Q a F, G jsou stejné jako v (12). Je známo, že Lévyho vzdálenost je metrika v prostoru pravděpodobnostních distribucí na $(\mathbb{R}, \mathcal{B})$, která metrizuje topologii slabé konvergence. Porovnáním poslední formule s formulí

$$\rho_K(P,Q) = \inf\{\varepsilon > 0 : F(x) - \varepsilon \le G(x) \le F(x) + \varepsilon,\ \forall\, x \in \mathbb{R}\},$$

která je ekvivalentní s (12) vidíme, že $\rho_K(P,Q)$ je maximální vzdálenost mezi grafy distribučních funkcí F a G naměřená ve směru osy y zatímco $\rho_L(P,Q)$ je maximální vzdálenost mezi těmito grafy naměřená ve směru osy $y = -x$. Odsud již snadno nahlédneme, že stejné příklady P, Q a T, které jsme využili k ukázce, že Kolmogorova vzdálenost nesplňuje podmínky (6), (7) a (8), stačí také k důkazu, že ani Lévyho vzdálenost nesplňuje tyto tři podmínky.

Nechť $(\mathcal{X}, \mathcal{S}, \mu)$ je nyní libovolný σ-konečný prostor s mírou a P, Q pravděpodobnostní míry na $(\mathcal{X}, \mathcal{S})$ dominované mírou μ (symbolicky, $\{P, Q\} \ll \mu$). Pak míry (distribuce) P, Q jsou jednoznačně dané mírou μ a Radon–Nikodymovými hustotami

$$p = \frac{\mathrm{d}P}{\mathrm{d}\mu}, \quad q = \frac{\mathrm{d}Q}{\mathrm{d}\mu} \tag{13}$$

na prostoru $\mathcal{X}$. Budeme se zajímat o známou třídu metrických vzdáleností

$$\rho_\alpha(P,Q) = \left(\int_{\mathcal{X}} |p(x) - q(x)|^\alpha \mathrm{d}\mu(x) \right)^{1/\alpha}, \quad \alpha \ge 1 \tag{14}$$

na prostoru distribucí dominovaných danou mírou μ. Je-li $\alpha > 1$ pak vzdálenost $\rho_\alpha(P,Q)$ obecně nesplňuje podmínky (6), (7) a (8). Pro doložení tohoto tvrzení zvolme nejběžnější pozorovací prostor $(\mathcal{X}, \mathcal{S}) = (\mathbb{R}, \mathcal{B})$ s Lebesguovou mírou μ a distribuci P, Q s hustotami

$$p(x) = I(0 < x < 1), \quad q(x) = I(0 < x < 2)/2.$$

Pro tyto hustoty z (14) obdržíme

$$\rho_\alpha(P,Q) = \left[(1/2)^\alpha + (1/2)^\alpha\right]^{1/\alpha} = \left(2^{1-\alpha}\right)^{1/\alpha} < 1 \quad \text{při } \alpha > 1.$$

Přitom lineární transformace $Tx = x/2$ na $\mathbb{R}$ indukuje distribuce $\tilde{P} = PT^{-1}$, $\tilde{Q} = QT^{-1}$ s hustotami

$$\tilde{p}(x) = 2I(0 < x < 1/2), \quad \tilde{q}(x) = I(0 < x < 1).$$

Pro tyto distribuce dostaneme z (14)

$$\rho_a(\tilde{P}, \tilde{Q}) = (1/2 + 1/2)^{1/\alpha} = 1,$$

což přesahuje $\rho_\alpha(P,Q)$. Proto nemůže platit (6) ani (7). Podobně jako v případě Kolmogorovy vzdálenosti se dá snadno ukázat, že nemůže platit ani (8).

Speciální případ $\alpha = 1$ je v mnoha ohledech odlišný od toho, co jsme shledali při $\alpha > 1$. Především v tomto případě integrál v (14) nezávisí na volbě dominující míry μ. Protože každá dvojice P, Q distribucí je dominovaná konečnou mírou $\mu = P + Q$, formule (14) definuje vzdálenost $\rho_1(P,Q)$ pro všechny distribuce P, Q na libovolném měřitelném prostoru $(\mathcal{X}, \mathcal{S})$. Jelikož

$$V(P,Q) = \int |p - q|\, \mathrm{d}\mu \tag{15}$$

je totální variace znaménkové míry $P - Q$ na prostoru $\mathcal{X}$, místo $\rho_1(P, Q)$ se zpravidla používá symbol $V(P, Q)$ a příslušná hodnota se obvykle nazývá totální variace distribucí P a Q. V poslední sekci uvidíme, že totální variace splňuje podmínky (6) a (7) ale nesplňuje (8). K doložení toho, že totální variace skutečně nesplňuje (8) stačí uvažovat diskrétní distribuce

$$P = (1/2, 1/2, 0), \quad Q = (0, 1/2, 1/2)$$

na $\mathcal{X} = \{1, 2, 3\}$ a statistiku $T : \mathcal{X} \mapsto \mathcal{Y} = \{1, 2\}$, kde $T(1) = 1$, $T(2) = T(3) = 2$. Tato statistika není postačující pro $\{P, Q\}$. Přitom indukuje na $\mathcal{Y}$ diskrétní distribuce $\tilde{P} = PT^{-1} = (1/2, 1/2)$ a $\tilde{Q} = QT^{-1} = (0, 1)$ s totální variací

$$V(\tilde{P}, \tilde{Q}) = \sum_{j=1}^{2} |\tilde{p}(j) - \tilde{q}(j)| = \frac{1}{2} + \frac{1}{2}$$

stejnou jako

$$V(P, Q) = \sum_{j=1}^{3} |p(j) - q(j)| = \frac{1}{2} + \frac{1}{2}.$$

3 Nekonečná třída divergencí

V této sekci budeme definovat nekonečnou třídu divergencí pomocí funkcí $f : (0, \infty) \mapsto \mathbb{R}$ standardizovaných ve smyslu $f(1) = 0$, které jsou dvakrát spojitě diferencovatelné na $(0, \infty)$, přičemž druhá derivace f'' je na celém tomto intervalu nezáporná a v okolí bodu 1 je kladná. Třídu všech takových funkcí označíme symbolem $\mathcal{F}$. Funkce $f \in \mathcal{F}$ jsou zřejmě konvexní na intervalu $(0, \infty)$ a ryze konvexní v okolí bodu 1. Dále, s každou f patřící do $\mathcal{F}$ tam zřejmě patří i konjugovaná funkce f^* definovaná vztahem

$$f^*(t) = tf(1/t), \quad t > 0. \tag{16}$$

Všechny funkce $f \in \mathcal{F}$ spojitě rozšiřujeme do bodu 0, tj. klademe

$$f(0) = \lim_{t \downarrow 0} f(t), \tag{17}$$

kde limita $f(0)$ může případně být $+\infty$.

Nechť nyní P, Q jsou libovolné pravděpodobnostní distribuce na nějakém měřitelném prostoru $(\mathcal{X}, \mathcal{S})$ a

$$p = \frac{\mathrm{d}P}{\mathrm{d}\mu}, \quad q = \frac{\mathrm{d}Q}{\mathrm{d}\mu}$$

hustoty těchto distribucí vzhledem k nějaké σ-konečné míře μ na $(\mathcal{X}, \mathcal{S})$. Pro každé $f \in \mathcal{F}$ definujme *f-divergenci* distribucí P, Q jako integrál

$$D_f(P, Q) = \int qf\left(\frac{p}{q}\right) \mathrm{d}\mu, \tag{18}$$

kde zamlčujeme proměnnou $x \in \mathcal{X}$ v hustotách p, q a integrační obor $\mathcal{X}$ pod symbolem pro integrál (srovnej formule (15) a (14)).

Protože hustoty p, q jsou obecně nezáporné, musíme integrand v (18) rozšířit z oblasti $p \geq 0$, $q > 0$ na celý kvadrant $p \geq 0$, $q \geq 0$ v $\mathbb{R}^2$. Toho lze docílit mnoha způsoby. My volíme formuli

$$0f\left(\frac{p}{0}\right) = pf^*(0) \quad \text{s konvencí} \quad 0f^*(0) = 0. \tag{19}$$

Při $p > 0$ se zjevně jedná o spojité rozšíření. Při obecném $p \geq 0$ jde o rozšíření konvexní a zdola polospojité v $\mathbb{R}^2$. Funkce $pf(p/q)$ je totiž konvexní v proměnných $(p, q) \in [0, \infty) \times (0, \infty)$ a jako taková musí zůstat konvexní i po spojitém rozšíření do konvexní oblasti $[0, \infty) \times [0, \infty)$ s vyjmutým bodem $(0, 0)$. Protože při $p > 0$ platí $pf(p/p) = f(1) = 0$, zdola polospojité rozšíření do bodu $(0, 0)$ vyžaduje $0f(0/0) \leq 0$, zatímco konvexní rozšíření vyžaduje $0f(0/0) \geq 0$. Proto (19) představuje jediné konvexní zdola polospojité rozšíření integrandu v (18) na celý uzavřený kvadrant $p \geq 0$, $q \geq 0$.

S využitím konvence uvedené v (19) můžeme (18) přepsat do tvaru

$$D_f(P, q) = \int_{\{q>0\}} qf\left(\frac{p}{q}\right) \mathrm{d}\mu + f^*(0)\, P(q = 0), \tag{20}$$

odkud lze nahlédnout, že hodnota $D_f(P, Q)$ nezávisí na zvolené dominující míře μ.

Prvním důsledkem formule (20) je tento důležitý výsledek.

Věta 1. Pro každé $f \in \mathcal{F}$ a libovolnou dvojici distribucí P, Q platí

$$D_f(P, Q) = D_{f^*}(Q, P),$$

kde $f^* \in \mathcal{F}$ je funkce konjugovaná k f ve smyslu (16).

Důkaz. Zřejmé ze (20) a (16).

Další důsledek formule (20) představuje užitečné pomocné tvrzení.

Lemma 1. Ve formulích (18) nebo (20) lze bez újmy na obecnosti předpokládat, že $f \in \mathcal{F}$ splňuje podmínku

$$f'(1) = 0. \tag{21}$$

Důkaz. Uvažujme funkci $\tilde{f} = f(t) - f'(1)\,(t-1)$, která splňuje (21). Postačí když dokážeme, že $D_{\tilde{f}}(P, Q) = D_f(P, Q)$. Konjugovaná funkce k $\tilde{f}$ je

$$\tilde{f}^*(t) = f^*(t) + f'(1)\,(t - 1).$$

Proto $\tilde{f}^*(0) = f^*(0) - f'(1)$ a tudíž podle (20)

$$\begin{aligned} D_{\tilde{f}}(P, Q) - D_f(P, Q) &= -f'(1)\left[\int_{\{q>0\}} (p - q)\,\mathrm{d}\mu + P(q = 0)\right] \\ &= -f'(1)\left[\int_{\mathcal{X}} (p - q)\,\mathrm{d}\mu\right] = 0, \end{aligned}$$

což bylo dokázat. □

Při analýze diferencovatelných funkcí se často s výhodou využívá Taylorův rozvoj se zbytkem v Lagrangeově nebo Cauchyho tvaru. Při analýze konvexních diferencovatelných funkcí lze výhodně využívat Taylorův rozvoj se zbytkem v integrálním tvaru. Příslušnou formuli uvádíme v následující lemmě.

Lemma 2. Jestliže $f \in \mathcal{F}$ splňuje podmínku (21), pak pro všechna $t > 0$ platí

$$f(t) = \int_t^1 (s-t)\, f''(s)\, \mathrm{d}s.$$

Důkaz. Nechť $f \in \mathcal{K}$. Podle Taylorovy věty se zbytkem v integrálním tvaru (viz např. Fichtengolc [8], str. 147), pro všechna $t > 0$ platí

$$f(t) = f(1) + f'(1)\,(t-1) + \int_1^t (t-s)\, f''(s)\, \mathrm{d}s.$$

Tudíž zbývá využít předpoklady $f(1) = f'(1) = 0$. □

Lemma 3. Jestliže $f \in \mathcal{F}$ splňuje podmínku (21), pak pro všechna $0 \le t_1 < t_2 \le 1$ platí $f(t_1) > f(t_2)$.

Důkaz. Z Lemmy 2 vidíme, že pro každé $0 < t \le 1$ existuje nezáporná funkce $\phi_t(s)$, pro kterou

$$f(t) = \int_0^1 \phi_t(s)\, f''(s)\, \mathrm{d}s.$$

Navíc, při pevných $0 < t_1 < t_2 \le 1$ je funkce $\phi_{t_1}(s) - \phi_{t_2}(s)$ nezáporná pro všechna $0 < s < 1$ a kladná v okolí $s = 1$. Protože $f''(s)$ je podle předpokladu nezáporná pro všechna $0 < s < \infty$ a kladná v okolí $s = 1$, rozdíl $f(t_1) - f(t_2)$ musí být kladný. Tato monotonie se rozšíří z oblasti $0 < t_1 < t_2 \le 1$ do oblasti $0 \le t_1 < t_2 \le 1$ přímo z definice $f(0)$ v (17). □

Z našich tří jednoduchých lemm snadno vyplyne první velmi závažný výsledek ohledně f-divergencí a sice věta o oboru hodnot. Poznamenejme, že pod singularitou $P \perp Q$ rozumíme situaci, kdy distribuce P a Q mají v σ-algebře $\mathcal{S}$ vzájemně disjunktní nosiče. Singularita je tudíž nejvyšší možnou formou nepodobnosti distribucí P a Q.

Věta 2. Každá f-divergence splňuje nerovnost

$$0 \le D_f(P,Q) \le f(0) + f^*(0). \tag{22}$$

Levá rovnost nastává právě když $P = Q$, zatímco pravá nastává když $P \perp Q$. Je-li ovšem horní mez $f(0) + f^*(0)$ konečná, pak pravá rovnost nastává právě když $P \perp Q$.

Důkaz. Z disjunktního rozkladu $\{q > 0\} = \{q > p\} + \{p > q > 0\} + \{p = q > 0\}$, ze (20) a z podmínky $f(1) = 0$ získáme formuli

$$D_f(P,Q) = C_f(P,Q) + C_{f^*}(Q,P), \tag{23}$$

kde

$$C_f(P,Q) = \int_{\{q>p\}} f\left(\frac{p}{q}\right) \mathrm{d}Q.$$

Navíc podle Lemmy 1 můžeme předpokládat $f'(1) = 0$ (přičemž podle definice konjugované funkce $f^* \in \mathcal{F}$ toto již implikuje podobnou rovnost také pro f^*). Tudíž podle Lemmy 3 v oblasti $\{q > p\} \in \mathcal{S}$ platí

$$f(0) \geq f\left(\frac{p}{q}\right) > f(1) = 0,$$

kde rovnost nastane právě když $p = 0$. Integrací těchto nerovností v oblasti $\{q > p\}$ dostaneme

$$f(0)\, Q(\{q > p\}) \geq C_f(P,Q) \geq 0,$$

přičemž pravá rovnost platí právě když $q \leq p$ Q-s. j., což je ekvivalentní $P = Q$. Levá rovnost platí, když $p \geq q$ P-s. j., přičemž pro konečná $f(0)$ lze "když" zaměnit za "právě když". Podobně

$$f^*(0)\, P(\{p < q\}) \geq C_{f^*}(Q,P) \geq 0,$$

kde pravá rovnost je ekvivalentní rovnosti $P = Q$ a levá plyne z podmínky $q \geq p$ Q-s. j. resp. je s touto podmínkou ekvivalentní. Požadovaný výsledek tedy dostáváme z formule (23) a z toho, že

$$p \geq q \ P\text{-s. j.} q \geq p \ Q\text{-s. j.}$$

je ekvivalentní podmínce

$$q = 0 \ P\text{-s. j.} p = 0 \ Q\text{-s. j.},$$

kterou zkráceně zapisujeme jako $P \perp Q$.

Jako první příklad uvedeme divergenci definovanou pro speciální diskrétní distribuce již dříve ve formuli (1). Uvažujme funkci $f(t) = -\ln t$ ze třídy $\mathcal{F}$ s limitou $f(0) = \infty$ v $t = 0$. Konjugovaná funkce je $f^*(t) = t \ln t$ s limitou $f^*(0) = 0$. Příslušná divergence je tedy podle (20) dána předpisem

$$D(P,Q) = \int_{\{p>0\}} q \ln \frac{q}{p} \, \mathrm{d}\mu + \infty Q(p = 0) \tag{24}$$

s konvencemi $0 \ln 0 = 0$ a $\infty \cdot 0 = 0$. Tudíž $\infty Q(p = 0)$ je 0 pokud $Q(p = 0) = 0$, což nastane právě když $Q \ll P$ (absolutní spojitost vzhledem k P). Toho jsme využili v diskrétním modelu uvažovaném v (1). Tam totiž všechny souřadnice P byly kladné, v kterémžto případě $Q \ll P$ platí pro libovolné diskrétní $Q = (q_1, ..., q_K)$. Je zřejmé, že pak integrál z (24) přejde při libovolném $\mu \gg \{P,Q\}$, například při $\mu = P$, do tvaru sumy (1), pokud v ní také respektujeme konvenci $0 \ln 0 = 0$. Z Věty 2 pro divergenci (24) dostaneme obor hodnot

$$0 \leq D(P,Q) \leq \infty. \tag{25}$$

Zde rovnost nule nastane právě když $P = Q$ a rovnost nekonečnu při singularitě $P \perp Q$. Rovnost nekonečnu ovšem může nastat i jindy, dokonce i při absolutní spojitosti $Q \ll P$,

kdy $\infty Q(p=0)=0$. Pro tento účel stačí uvažovat na intervalu $\mathcal{X}=(1,\infty)$ vzájemně absolutně spojité pravděpodobnostní hustoty

$$p(x)=e^{1-x},\quad q(x)=\frac{1}{x^2} \tag{26}$$

pro které platí $D(P,Q)=\infty$. Přitom obrácená divergence $D(Q,P)$ je konečná.

Nyní se budeme věnovat f-divergenci pro funkci $f(t)=t\ln t$ ze třídy $\mathcal{F}$, která má výjimečný význam v matematické statistice i teorii informace. V literatuře se označuje i nazývá různým způsobem. My ji budeme v dalším označovat symbolem $D(P\|Q)$ a nazývat *informační divergence*. Protože konjugovaná funkce je zde $f^*(t)=-\ln t$, z Věty 1 vyplývá rovnost

$$D(P\|Q)=D(Q,P),$$

kde $D(P,Q)$ je divergence (24). Tudíž podle (24)

$$D(P\|Q)=\int_{\{q>0\}} p\ln\frac{p}{q}\,\mathrm{d}\mu+\infty P(q=0) \tag{27}$$

při konvencích $0\ln 0{=}0$ a $\infty\cdot 0=0$. Dále, všechny vlastnosti $D(P,Q)$ uvedené výše automaticky platí i pro informační divergenci $D(Q\|P)$. Tak například nerovnost (25) a podmínky $P=Q$ resp. $P\perp Q$ pro příslušné rovnosti jsou symetrické v P, Q a platí tedy i pro $D(P\|Q)$. Naproti tomu ale obdržíme $D(P\|Q)<\infty$ resp. $D(Q\|P)=\infty$ pro konkrétní distribuce P, Q na intervalu $\mathcal{X}=(1,\infty)$ určené Lebesguovskými hustotami (26).

Výjimečnost informační divergence vyplývá z její aditivnosti

$$D\left(\otimes_{i=1}^n P_i\middle\|\otimes_{i=1}^n Q_i\right)=\sum_{i=1}^n D(P_i\|Q_i), \tag{28}$$

kde $\otimes$ je symbol pro kartézský součin. Má se totiž za to, že aditivnost informace získaná ze statistických pozorování je charakteristická vlastnost nezávislosti těchto pozorování. Přitom ze známého logaritmického řešení Cauchyovy funkcionální rovnice

$$\phi(pq)=\phi(p)+\phi(q)\quad\text{pro}\quad p,\,q>0$$

lze vyvodit, že (27) je jediná f-divergence aditivní ve smyslu (28). Chápeme-li tedy divergenci $D_f(P,Q)$ jako míru informace o $P_0\in\{P,Q\}$ obsažené v pozorování X s výběrovým prostorem $(\mathcal{X},\mathcal{S},P_0)$, pak $D(P\|Q)$ je jediná taková míra, která kumulaci informace z nezávislých pozorování považuje za aditivní proces.

Navíc, informační divergence intimně souvisí se základním pojmem teorie informace a to se Shannonovou informací $I(X;Y)$ ve stochastickém výstupu Y některého pozorovacího nebo komunikačního kanálu o jeho stochastickém vstupu X. Nechť (X,Y) je dvojice náhodných veličin s výběrovým pravděpodobnostním prostorem $(\mathcal{X}\otimes\mathcal{Y},\,\mathcal{S}\otimes\mathcal{T},\,P_{X,Y})$ a nechť P_X resp. P_Y budou příslušné marginální distribuce na vstupním resp. výstupním měřitelném prostoru $(\mathcal{X},\mathcal{S})$ resp. $(\mathcal{X},\mathcal{T})$. Stochastická kanálová transformace $X\mapsto Y$ vstupu na výstup je utajena v simultánním rozdělení $P_{X,Y}$. Kdyby se kanálem nepřenášela žádná informace, pak by vstup X a výstup Y byly stochasticky nezávislé náhodné veličiny,

tj. platilo by $P_{X,Y} = P_X \otimes P_Y$. Shannon definoval informaci $I(X;Y)$ jako pokles entropie zprávy X v důsledku pozorování Y, což je ovšem ekvivalentní vztahu

$$I(X;Y) = D\left(P_{X,Y} \| P_X \otimes P_Y\right).$$

Z něj je zřejmé, že Shannonova informace $I(X;Y)$ není nic jiného než statistická míra asociace náhodných veličin X a Y. Místo informační divergence by bylo možné k podobnému účelu použít libovolnou f-divergenci $D_f(P_{X,Y}, P_X \otimes P_Y)$, ale opět by se vytratila důležitá aditivnost. O f-divergenčních mírách statistické asociace jako první systematicky pojednala Zvárová [33].

Zmiňme se pro úplnost ještě o dvou statisticky významných příkladech f-divergencí. První z nich je *Pearsonova divergence* $\chi^2(P,Q)$ definovaná funkcí $f(t) = (t-1)^2$ s konjugovanou $f^*(t) = (t-1)^2/t$, kde $f^*(0) = \infty$. Podle (20) tedy

$$\chi^2(P,Q) = \int_{\{q>0\}} \frac{(p-q)^2}{q}\,\mathrm{d}\mu + \infty P(q=0) \tag{29}$$

s konvencí $\infty \cdot 0 = 0$. Podobně jako u divergencí $D(P,Q)$ a $D(P\|Q)$, zde tedy $\chi^2(P,Q) = \infty$ když neplatí $P \ll Q$, ale nekonečná Pearsonova divergence je možná i při $P \ll Q$, kdy $P(q=0) = 0$. K důkazu stačí přehodit distribuce s hustotami (26). Na podobném principu lze zkonstruovat i příklad s diskrétním pozorovacím prostorem $\mathcal{X} = \{1, 2, \ldots\}$. Při konečném $\mathcal{X}$ ovšem z absolutní spojitosti $P \ll Q$ plyne konečnost divergence $\chi^2(P,Q)$. Poznamenejme, že f-divergence příslušná funkci $f(t) = (t-1)^2/t$, tj. obrácená Pearsonova divergence, se někdy nazývá Neymanova divergence. Pro tuto divergenci nebudeme zavádět speciální symbol.

Posledním příkladem bude Hellingerova divergence $H^2(P,Q)$ definovaná funkcí $f(t) = (1-\sqrt{t}^2)$ z $\mathcal{F}$. Tato funkce je samokonjugovaná, tj. $f^*(t) = f(t)$. S tím souvisí, že

$$H^2(P,Q) = \int \left(\sqrt{p} - \sqrt{q}\right)^2 \mathrm{d}\mu \tag{30}$$

je symetrická v P, Q. Protože $f(0) = f^*(0) = 1$, z Věty 2 dostaneme pro tuto divergenci obor hodnot

$$0 \le H^2(P,Q) \le 2, \tag{31}$$

kde levá rovnost platí právě když $P = Q$ a pravá právě když $P \perp Q$. Lze snadno nahlédnout, že odmocnina $H(P,Q)$ je metrika v prostoru pravděpodobnostních distribucí P, Q na kterémkoliv měřitelném prostoru $(\mathcal{X}, \mathcal{S})$.

V následující sekci ukážeme, že všechny f-divergence splňují podmínky (6), (7). Ty, které mají f ryze konvexní všude na $(0,\infty)$, splňují také (8). Náš důkaz bude zcela nový, založený na výsledku, který prokazuje vzájemnou příbuznost a společnou jednoduchou reprezentovatelnost všech těchto zdánlivě různorodých divergenčních měr. Jak uvidíme, jde vlastně jen o různým způsobem průměrované informace ve statistickém experimentu

$(\mathcal{X}, \mathcal{S}, P_0)$, kde P_0 je z rodiny $\mathcal{P} = \{P, Q\}$.

4 Statistická informace a f-divergence

Mějme systém, který může být v hypotetickém stavu $\mathcal{H}$ s pravděpodobností $\pi \in (0,1)$ a alternativním stavu $\mathcal{A}$ s komplementární pravděpodobností $1-\pi$. Máme-li rozhodnout o stavu bez jakékoliv další informace, pak Bayesovu apriorní pravděpodobnost chyby

$$B_\pi = \pi \wedge (1-\pi) \stackrel{\triangle}{=} \min\{\pi,\, 1-\pi\}$$

dosáhneme, když se rozhodneme pro stav, který je a priori pravděpodobnější.

Situace se zkomplikuje, když vedle apriorních pravděpodobností π, $1-\pi$ budeme mít k dispozici také aposteriorní informaci a sice pozorování X s výběrovým prostorem $(\mathcal{X}, \mathcal{S}, P_0)$, kde $P_0 \in \{P, Q\}$, přičemž $P_0 = P$ když systém je ve stavu $\mathcal{H}$ a $P_0 = Q$, když je ve stavu $\mathcal{A}$. Rozhodování o stavu můžeme představit jako měřitelnou funkci $\varphi : \mathcal{X} \mapsto [0,1]$, kde $\varphi(X)$ je pravděpodobnost, že zamítneme $\mathcal{H}$. Jsou-li p, q hustoty příslušné distribucím P, Q jak byly uvažovány v (18) – (20), pak integrál

$$\int (\pi p\varphi + (1-\pi)\, q(1-\varphi))\, \mathrm{d}\mu$$

představuje pravděpodobnost chyby příslušnou rozhodovací funkci φ a

$$B_\pi(P,Q) = \inf_\varphi \int [\pi p\varphi + (1-\pi)\, q(1-\varphi)]\, \mathrm{d}\mu \tag{32}$$

je Bayesova aposteriorní pravděpodobnost chyby. Tato pravděpodobnost se dosáhne například při Bayesově rozhodovací funkci

$$\varphi_B(x) = I(\pi p(x) \le (1-\pi)\, q(x)) \tag{33}$$

pro kterou z (32) dostaneme formuli

$$B_\pi(P,Q) = \int \pi p \wedge (1-\pi)\, q\, \mathrm{d}\mu. \tag{34}$$

Speciálně tedy pro všechna $s > 0$ platí

$$\int p \wedge qs\, \mathrm{d}\mu = (1+s)\, B_{\frac{1}{1+s}}(P,Q). \tag{35}$$

Apriorní Bayesova pravděpodobnost splňuje podmínku

$$B_\pi = \int [\pi p\varphi_A + (1-\pi)\, q(1-\varphi_A)]\, \mathrm{d}\mu$$

pro

$$\varphi_A(x) = I(\pi \le 1-\pi).$$

Porovnáním s (32) zjistíme, že Bayesovy pravděpodobnosti splňují nerovnost $B_\pi \geq B_\pi(P,Q)$.

Rozdíl

$$\mathcal{I}_\pi(P,Q) = B_\pi - B_\pi(P,Q) \tag{36}$$

bude tím větší, čím více informace relevantní pro daný problém (tj. pro rozhodnutí mezi stavy systému $\mathcal{H} : P_0 = P$ a $\mathcal{A} : P_0 = Q$) obsahuje veličina X. De Groot [6, 7] nazval veličinu $\mathcal{I}_\pi(P,Q)$ *statistickou informací* v pozorování X.

Ukážeme, že každá f-divergence $D_f(P,Q)$ je ve skutečnosti vážená informace $\mathcal{I}_\pi(P,Q)$, kde váha $w_f(\pi) > 0$ je distribuovaná na pravděpodobnostech $\pi \in (0,1)$ a závisí na $f \in \mathcal{F}$. Za tímto účelem se nám hodí dvě jednoduché lemmy.

Lemma 4. Pro libovolné konstanty $p \geq 0$, $q > 0$ a funkci $\phi : (0,\infty) \mapsto [0,\infty)$ integrovatelnou na konečných intervalech platí

$$\int_1^{p/q} (p - qs)\,\phi(s)\,\mathrm{d}s = \int_0^1 (qs - p \wedge qs)\,\phi(s)\,\mathrm{d}s + \int_0^\infty (p - p \wedge qs)\,\phi(s)\,\mathrm{d}s.$$

Důkaz. Je-li $p/q \geq 1$, pak

$$p \wedge qs = \begin{cases} qs & \text{pro} \quad 0 < s \leq p/q \\ p & \text{pro} \quad s > p/q \end{cases}$$

a lemma tudíž platí. Je-li $p/q < 1$ pak

$$p \wedge qs = \begin{cases} qs & \text{pro} \quad 0 < s < p/q \\ p & \text{pro} \quad s > p/q \end{cases}$$

a lemma opět platí. □

Lemma 5. Pro $f \in \mathcal{F}$ platí

$$f^*(0) = \int_0^\infty f''(s)\,\mathrm{d}s.$$

Důkaz. Podle definice a Lemmy 2

$$f^*(0) = \lim_{t\to\infty} \frac{f(t)}{t} = \lim_{t\to\infty} \int_1^t \frac{t-s}{t}\, f''(s)\,\mathrm{d}s.$$

Tvrzení plyne tedy z věty o monotonní konvergenci integrálů. □

Věta 3. Každá f-divergence je vážená statistická informace v tom smyslu, že pro libovolné distribuce P, Q platí

$$D_f(P,Q) = \int_0^1 \mathcal{I}_\pi(P,Q)\, w_f(\pi)\,\mathrm{d}\pi,$$

kde

$$w_f(\pi) = f''\left(\frac{1-\pi}{\pi}\right) \cdot \frac{1}{\pi^3} \quad \text{pro } \pi \in (0,1).$$

Důkaz. Podle (20), Lemmy 4 a (35), rozdíl $D_f(P,Q) - f^*(0)\,P(q=0)$ lze vyjádřit jako následující integrály

$$\begin{aligned}
&\int_{\{q>0\}} \left[\int_1^{p/q} (p - qs)\, f''(s)\,\mathrm{d}s\right] \mathrm{d}\mu \\
= &\int_{\{q>0\}} \left[\int_0^1 (qs - p\wedge qs)\, f''(s)\,\mathrm{d}s + \int_1^\infty (p - p\wedge qs)\, f''(s)\,\mathrm{d}s\right] \mathrm{d}\mu \\
= &\int_0^1 \Big[s - (1+s)\, B_{\frac{1}{s}}(P,Q)\Big]\, f''(s)\,\mathrm{d}s + \int_1^\infty \Big[P(q>0) - (1+s)\, B_{\frac{1}{1+s}}(P,Q)\Big]\, f''(s)\,\mathrm{d}s.
\end{aligned}$$

Dále podle Lemmy 5

$$\begin{aligned}
&f^*(0)\,P(q=0) + \int_1^\infty \Big[P(q>0) - (1+s)\, B_{\frac{1}{1+s}}(P,Q)\Big]\, f''(s)\,\mathrm{d}s \\
= &\int_1^\infty \Big[1 - (1+s)\, B_{\frac{1}{1+s}}(P,Q)\Big]\, f''(s)\,\mathrm{d}s.
\end{aligned}$$

Tudíž podle (36)

$$\begin{aligned}
D_f(P,Q) &= \int_0^1 (1+s)\left[\frac{1}{1+s} \wedge \frac{s}{1+s} - B_{\frac{1}{1+s}}(P,Q)\right] f''(s)\,\mathrm{d}s \\
&= \int_0^\infty (1+s)\, \mathcal{I}_{\frac{1}{1+s}}(P,Q)\, f''(s)\,\mathrm{d}s.
\end{aligned}$$

Nyní zbývá přejít k substituci $\frac{1}{1+s} \mapsto \pi$. □

Pro příklady f-divergencí uvedené na konci předchozí sekce dostaneme z Věty 3 tyto reprezentace

$$D(P\|Q) = \int_0^1 \frac{\mathcal{I}_\pi(P,Q)}{(1-\pi)\,\pi^2}\,\mathrm{d}\pi, \tag{37}$$

$$\chi^2(P,Q) = \int_0^1 \frac{\mathcal{I}_\pi(P,Q)}{\pi^3}\,\mathrm{d}\pi, \tag{38}$$

$$H^2(P,Q) = \int_0^1 \frac{\mathcal{I}_\pi(P,Q)}{\sqrt{(1-\pi)^3\pi^3}}\,\mathrm{d}\pi. \tag{39}$$

Pro obrácenou informační divergenci $D(P,Q) = D(Q\|P)$ a obrácenou Pearsonovu divergenci dostaneme po dosazení příslušných funkcí f do formule pro $w_f(\pi)$ ve Větě 3

$$\begin{aligned}
D(Q\|P) &= \int_0^1 \frac{\mathcal{I}_\pi(P,Q)}{\pi(1-\pi)^2}\,\mathrm{d}\pi, \\
\chi^2(Q,P) &= \int_0^1 \frac{\mathcal{I}_\pi(P,Q)}{(1-\pi)^3}\,\mathrm{d}\pi.
\end{aligned}$$

Tyto výsledky dostaneme též z (37), (38) pokud využijeme zřejmou rovnost $\mathcal{I}_\pi(Q,P) = \mathcal{I}_{1-\pi}(P,Q)$ a nasadíme substituci $1-\pi \mapsto \pi$. Ze všech těchto příkladů je patrné, že různé f-divergence se vzájemně liší jen v důrazu, který se klade na různé apriorní pravděpodobnosti $\pi \in (0,1)$ ve statistickém rozhodování problému, který s divergencí distribucí P, Q spojujeme.

Hlavním výsledkem této sekce a celé práce je následující věta, která se vyjadřuje k podmínkám (6) – (8). Protože podmínky (7) a (8) se týkají postačujících statistik $T : (\mathcal{X},\mathcal{S}) \mapsto (\mathcal{Y},\mathcal{T})$, připomeňme klasickou tzv. faktorizační podmínku postačitelnosti. Podle ní je měřitelné zobrazení $T : \mathcal{X} \mapsto \mathcal{Y}$ postačující pro dvojici pravděpodobnostních distribucí $\{P,Q\}$ na $(\mathcal{X},\mathcal{S})$ s hustotami $p = \mathrm{d}P/\mathrm{d}\mu$, $q = \mathrm{d}Q/\mathrm{d}\mu$, když existují měřitelná zobrazení α, $\beta : \mathcal{Y} \to [0,\infty)$ a $h : \mathcal{X} \mapsto [0,\infty)$ spl

ující μ-s. j. na $\mathcal{X}$ podmínku

$$p(x) = \alpha(Tx)\, h(x), \quad q(x) = \beta(Tx)\, h(x).$$

Je známo, že toto nastane právě když hustota

$$\frac{\mathrm{d}P}{\mathrm{d}(P+Q)} = \frac{p}{p+q}\, I(p+q>0) \tag{40}$$

bude měřitelná vzhledem k pod-σ-algebře $\tilde{\mathcal{S}} \subset \mathcal{S}$ generované pod-σ-algebrou $T^{-1}\mathcal{T} \subset \mathcal{S}$ a μ-nulovými množinami z $\mathcal{S}$. K tomu je nutná a postačující $\tilde{\mathcal{S}}$-měřitelnost množiny $\{\pi p < (1-\pi)\, q\}$ pro každé $\pi \in (0,1)$.

Věta 4. Všechny f-divergence $D_f(P,Q)$ splňují podmínky (6) a (7). Je-li navíc druhá derivace f'' kladná všude na $(0,\infty)$, tj. je-li f na tomto intervalu ryze konvexní, pak f-divergence splňuje také (8).

Důkaz. (I) Napřed dokážeme, že všechny f-divergence splňují (6) a (7). Distribuce $\tilde{P} = PT^{-1}$, $\tilde{Q} = QT^{-1}$ jsou dominovány na $(\mathcal{Y},\mathcal{T})$ σ-konečnou mírou $\tilde{\mu} = \mu T^{-1}$. Položme

$$\tilde{p} = \frac{\mathrm{d}\tilde{P}}{\mathrm{d}\tilde{\mu}}, \quad \tilde{q} = \frac{\mathrm{d}\tilde{Q}}{\mathrm{d}\tilde{\mu}}.$$

Podle definice Radon–Nikodymovy hustoty, pro každé $B \in \mathcal{T}$ platí

$$\tilde{P}(B) = \int_B \tilde{p}(y)\, \mathrm{d}\tilde{\mu}(y) = \int_{T^{-1}(B)} \tilde{p}(Tx)\, \mathrm{d}\mu(x)$$

a

$$P(T^{-1}B) = \int_{T^{-1}B} p(x)\, \mathrm{d}\mu(x).$$

Dále podle definice $\tilde{P}$ platí $\tilde{P}(B) = P(T^{-1}B)$, takže

$$\tilde{p}(Tx) = p(x) \quad \mu\text{-s. j. na } \mathcal{X}. \tag{41}$$

Nyní podle (32) pro měřitelné funkce $\tilde{\varphi} : \mathcal{Y} \mapsto [0,1]$ platí

$$\begin{aligned} B_\pi(\tilde{P},\tilde{Q}) &= \inf_{\tilde{\varphi}} \int_{\mathcal{Y}} \left[\pi\tilde{p}\tilde{\varphi} + (1-\pi)\,\tilde{q}(1-\tilde{\varphi})\right] \mathrm{d}\mu T^{-1} \\ &= \inf_{\tilde{\varphi}} \int_{\mathcal{X}} \left[\pi\tilde{p}(T)\,\tilde{\varphi}(T) + (1-\pi)\,\tilde{q}(T)\,(1-\tilde{\varphi}(T))\right] \mathrm{d}\mu \\ &= \inf_{\tilde{\varphi}} \int_{\mathcal{X}} \left[\pi p\tilde{\varphi}(T) + (1-\pi)\,q(1-\tilde{\varphi}(T))\right] \mathrm{d}\mu \quad \text{(viz (41))} \\ &\geq \inf_{\varphi} \int_{\mathcal{X}} \left[\pi p\varphi + (1-\pi)\,q(1-\varphi)\right] \mathrm{d}\mu \;=\; B_\pi(P,Q). \end{aligned}$$

Tudíž monotonie (6) pro $D(P,Q) = D_f(P,Q)$ plyne z věty 3. Dále z definice Bayesovy rozhodovací funkce (33), která vede na Bayesovu chybu (34) vidíme, že při T postačujícím pro $\{P,Q\}$ dosáhneme při každém $\pi \in (0,1)$ rovnost $B_\pi(\tilde{P},\tilde{Q}) = B_\pi(P,Q)$ a tudíž také rovnost

$$\mathcal{I}_\pi(\tilde{P},\tilde{Q}) = \mathcal{I}_\pi(P,Q).$$

Proto podmínku (7) pro divergence $D(P,Q) = D_f(P,Q)$ dostaneme rovněž z věty 3.

(II) Zůstaňme u označení z předchozí části důkazu, ale omezme se na $f \in \mathcal{F}$ s druhými derivacemi $f''(s) > 0$ pro všechna $s \in (0,\infty)$. Pak ve Větě 3 máme $w_f(\pi) > 0$ pro všechna $\pi \in (0,1)$. Tudíž podle již dokázané monotonie z předpokladu $D_f(\tilde{P},\tilde{Q}) = D_f(P,Q)$ plyne pro skoro všechna $\pi \in (0,1)$ rovnost

$$B_\pi(\tilde{P},\tilde{Q}) = B_\pi(P,Q). \tag{42}$$

Vzhledem ke spojitosti obou těchto funkcí v proměnné $\pi \in (0,1)$ můžeme upřesnit, že množina $\Pi_1 \subset (0,1)$ na které rovnost (42) neplatí je nejvýše spočetná. Uvažujme nyní libovolné pevné $\pi \in (0,1)$ a Bayesovy rozhodovací funkce φ, $\tilde{\varphi} : \mathcal{X} \mapsto \{0,1\}$, při kterých se dosahují $B_\pi(P,Q)$ a $B_\pi(\tilde{P},\tilde{Q})$. Není-li π v nejvýše spočetné množině

$$\Pi_2 = \{\pi \in (0,1) : \mu(\{\pi p = (1-\pi)\,q\}) > 0\},$$

pak nulovost rozdílu

$$B_\pi(\tilde{P},\tilde{Q}) - B_\pi(P,Q) = \int \left[\tilde{\varphi}(Tx) - \varphi(x)\right]\left[\pi p(x) - (1-\pi)\,q(x)\right] \mathrm{d}\mu(x)$$

má za následek

$$\mu\left(\{\tilde{\varphi}(Tx) \neq I(\pi p(x) < (1-\pi)\,q(x))\}\right) = 0. \tag{43}$$

Tato netriviální skutečnost plyne z toho, že při každém $\pi \in (0,1)$ platí μ-s. j.

$$\tilde{\varphi}(Tx) - \varphi(x) = \tilde{\varphi}(Tx) - 1 \leq 0 \quad \text{pokud} \quad \pi p(x) < (1-\pi)\,q(x),$$

při $\pi \notin \Pi_2$ navíc

$$\tilde{\varphi}(Tx) - \varphi(x) = \tilde{\varphi}(Tx) \geq 0 \quad \text{pokud} \quad \pi p(x) > (1-\pi)\,q(x)$$

a

$$\mu\left(\{\pi p(x) = (1-\pi)\,q(x)\}\right) = 0.$$

Integrand v posledním integrálu je μ-s. j. nezáporný a integrál je nula právě když μ-s. j. na $\mathcal{X}$

$$\tilde{\varphi}(Tx) = I(\pi p(x) < (1-\pi)\, q(x)).$$

Ze (43) již plyne $\tilde{\mathcal{S}}$-měřitelnost podmnožin

$$A = \{\pi p(x) < (1-\pi)\, q(x)\}\,, \quad A^c = \mathcal{X} - A_0$$

množiny $\mathcal{X}$. Vskutku, pro

$$\tilde{A} = \{\tilde{\varphi}(Tx) = 0\} \in T^{-1}\mathcal{T} \quad \text{a doplněk} \quad \tilde{A}^c = \{\tilde{\varphi}(Tx) = 1\}$$

platí podle (43)

$$\mu\left(\tilde{A} \cap A\right) + \mu\left(A^c \cap \tilde{A}^c\right) = 0$$

tj. také podmínka

$$\mu\left(\tilde{A} - A^c\right) + \mu\left(A^c - \tilde{A}\right) = 0$$

pro $\tilde{\mathcal{S}}$-měřitelnost množiny A^c a tudíž i množiny A. Dokázali jsme tedy, že rovnosti $D_f(\tilde{P}, \tilde{Q}) = D_f(P, Q)$ plyne $\tilde{\mathcal{S}}$-měřitelnost množiny A pro každé $\pi \in (0, 1)$ nepatřící do nejvýše spočetné množiny $\Pi = \Pi_1 \cup \Pi_2$. To ovšem znamená, že hustota (40) je $\tilde{\mathcal{S}}$-měřitelná, tj. že pro uvažované f-divergence podmínka (8) platí. □

5 Zobecnění

Uvažujme sigmoidální funkci

$$\varphi(y) = \frac{e^y - 1}{e^y + 1}, \quad y \in \mathbb{R}$$

známou z teorie neuronových sítí (viz např. [23]). Funkce

$$\phi_n(y) = \varphi(n(y-1)), \quad n \geq 1$$

jsou rostoucí na $\mathbb{R}$ a antisymetrické kolem $y = 1$, přičemž posloupnost funkcí ϕ_n je rostoucí na polopřímce $(1, \infty)$ s konstantní limitou 1 a klesající na polopřímce $(-\infty, 1)$ s konstantní limitou -1. Z toho plyne, že funkce f_n definované vztahem

$$f_n(t) = \int_1^t \phi_n(y)\, \mathrm{d}y, \quad t \in \mathbb{R}$$

jsou konvexní a jejich posloupnost konverguje monotónně zdola k funkci

$$f(t) = |t-1|, \quad t \in \mathbb{R}. \tag{44}$$

Tato funkce je konvexní na $\mathbb{R}$, ale ryze konvexní je jen v bodě $t = 1$. Je dvakrát spojitě diferencovatelná na $\mathbb{R}$ s výjimkou bodu $t = 1$, kde není diferencovatelná vůbec. Proto

nepatří do třídy $\mathcal{F}$ studované v předchozích sekcích. Nicméně můžeme na ni aplikovat základní definice práce. Z (16) zjistíme, že konjugovaná funkce je

$$f^*(t) = f(t) = |t-1|,$$

z (17) vyplyne $f(0) = f^*(0) = 1$ a z (18) dostáváme f-divergenci totožnou s totální variací (15), tj.

$$D_f(P,Q) = V(P,Q) = \int |p-q| \,\mathrm{d}\mu. \tag{45}$$

Protože nezáporné hustoty p, q splňují nerovnosti $0 \le |p-q| \le p+q$, platí

$$0 \le V(P,Q) \le 2. \tag{46}$$

Nutnou a postačující podmínkou pro levou rovnost je zjevně $p = q$ μ-s. j., tj. $P = Q$ zatím co podobnou podmínkou pro pravou rovnost je $q = 0$ P-s. j. a $p = 0$ Q-s. j., tj. $P \perp Q$. Jinými slovy, totální variace splňuje větu 2 přesto, že jde o f-divergenci pro $f \notin \mathcal{F}$.

Postačitelnost podmínek $P = Q$ resp. $P \perp Q$ pro rovnost $V(P,Q) = 0$ resp. $V(P,Q) = 2$ dostaneme též z monotonní konvergence

$$V_n(P,Q) \uparrow V(P,Q) \quad \text{pro } n \to \infty, \tag{47}$$

kterou pro f_n-divergence

$$V_n(P,Q) = D_{f_n}(P,Q) = \int q\, f_n\left(\frac{p}{q}\right) \mathrm{d}\mu \tag{48}$$

lze vyvodit z monotonní konvergence $f_n \uparrow f$ dokázané výše. Funkce f_n patří do $\mathcal{F}$ protože jsou dvakrát spojitě diferencovatelné na $\mathbb{R}$. Navíc

$$f_n(0) = -\int_0^1 \phi_n(y) \,\mathrm{d}y \uparrow 1 \quad \text{pro } n \to \infty,$$

protože $\phi_n \downarrow -1$ na intervalu $(-\infty, 1)$. Podobně

$$f_n^*(0) = \lim_{f\to\infty} \frac{1}{t} \int_1^t \phi_n(y) \,\mathrm{d}y \uparrow 1 \quad \text{pro } n \to \infty$$

se dá vyvodit z konvergence $\phi_n \uparrow 1$ na intervalu $(1, \infty)$. Přitom druhé derivace

$$f_n''(t) = \phi_n'(t) = \frac{2n\, e^{n(t-1)}}{\left(e^{n(t-1)} + 1\right)^2}$$

jsou kladné všude na $\mathbb{R}$. Tudíž Věta 2 zaručuje pro f_n-divergence $V_n(P,Q)$ obory hodnot

$$0 \le V_n(P,Q) \le f_n(0) + f_n^*(0) < 2$$

s tím, že $V_n(P,Q) = 0$ právě když $P = Q$ a $V_n(P,Q) = f_n(0) + f_n^*(0)$ právě když $P \perp Q$.

Z konvergence (47) je vidět, že totální variace splňuje i větu 4 přesto, že jde o f-divergenci s $f \notin \mathcal{F}$. Skutečně, budiž T statistika uvažovaná v (6) – (8). Podle věty 4 platí pro všechny f_n-divergence

$$V_n(PT^{-1}, QT^{-1}) \le V_n(P,Q),$$

kde rovnost nastane právě když T je postačující pro $\{P, Q\}$. Tudíž ze (47) je jasné, že totální variace splňuje (6) a (7). Z rovnosti limit $V(PT^{-1}, QT^{-1}) = V(P, Q)$ však ještě neplyne rovnost $V_n(PT^{-1}, QT^{-1}) = V_n(P, Q)$ pro některé $n \geq 1$ a tudíž ani postačitelnost T pro $\{P, Q\}$. K doložení toho, že totální variace skutečně nesplňuje (8) stačí uvažovat diskrétní distribuce

$$P = (1/2, 1/2, 0), \quad Q = (0, 1/2, 1/2)$$

na $\mathcal{X} = \{1, 2, 3\}$ a statistiku $T : \mathcal{X} \mapsto \mathcal{Y} = \{1, 2\}$, kde $T(1) = 1$, $T(2) = T(3) = 2$. Tato statistika není postačující pro $\{P, Q\}$. Přitom indukuje na $\mathcal{Y}$ diskrétní distribuce $\tilde{P} = PT^{-1} = (1/2, 1/2)$ a $\tilde{Q} = QT^{-1} = (0, 1)$ s totální variací

$$V(\tilde{P}, \tilde{Q}) = \sum_{j=1}^{2} |\tilde{p}(j) - \tilde{q}(j)| = \frac{1}{2} + \frac{1}{2}$$

stejnou jako

$$V(P, Q) = \sum_{j=1}^{3} |p(j) - q(j)| = \frac{1}{2} + \frac{1}{2}.$$

V připravované práci Lieseho a Vajdy [19] je odvozené zobecnění klasického Taylorova rozvoje (viz důkaz lemmy 2) na všechny konvexní funkce $f : (0, \infty) \mapsto \mathbb{R}$. Pomocí tohoto zobecnění je možné Věty 2–4 a jejich důkazy rozšířit na f-divergence pro všechny konvexní funkce $f : (0, \infty) \mapsto \mathbb{R}$ standardizované do tvaru $f(1) = 0$, které jsou ryze konvexní v bodě 1 (tj. které nejsou lineární v žádném otevřeném okolí tohoto bodu). Většina takových funkcí nebude zřejmě ve třídě $\mathcal{F}$.

Příkladem může být funkce (44), pro kterou $D_f(P, Q)$ je totální variace $V(P, Q)$. Ze zobecněných verzí vět 2 a 4 vyplynou pro totální variaci automaticky obě tvrzení dokazovaná v této sekci přímo, resp. s využitím f_n-divergencí pro posloupnosti $f_n \in \mathcal{F}$. Ze zobecněné verze věty 3 vyplyne rovnost

$$V(P, Q) = 4\mathcal{I}_{1/2}(P, Q) = 4\left[B_{1/2} - B_{1/2}(P, Q)\right], \tag{49}$$

kde $B_{1/2} = 1/2$ a $B_{1/2}(P, Q)$ jsou Bayesovy pravděpodobnosti chyb při stejně pravděpodobné hypotéze a alternativě. Tuto rovnost můžeme nezávisle ověřit integrací vztahu

$$|p - q| = p + q - 2\min\{p, q\}$$

mezi hustotami $p = \mathrm{d}P/\mathrm{d}\mu$ a $q = \mathrm{d}Q/\mathrm{d}\mu$. Totální variace je tedy (až na jednotku) statistickou informací při stejně pravděpodobné hypotéze a alternativě.

Poděkování. Tato práce vznikla v letech 2005-2006 za podpory grantových projektů AV ČR A1075403 a MŠMT 1M0512.

Reference

[1] ALI M. S, SILVEY S. D.: *A general class of coefficients of divergence of one distribution from another.* Journal of Royal Statistical Society, Series B, 28 (1966), 131-140.

[2] ANDĚL J.: *Matematická statistika.* SNTL, Praha 1978.

[3] BEIRLANT J., DEVROYE L., GYÖRFI L, VAJDA I.: *Large deviations of divergence measures on partitions.* Journal of Statistical Planning and Inference 93 (2001), 1–16.

[4] CSISZÁR I.: *Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten.* Publications of the Mathematical Institute of Hungarian Academy of Sciences, Series A, 8 (1963), 85-108.

[5] CSISZÁR I.: *Information-type measures of difference of probability distributions and indirect observations.* Studia scientiarum Mathematicarum Hungarica 2 (1967), 299-318.

[6] DE GROOT M. H.: *Uncertainty, information and sequential experiments.* Annals of Mathematical Statistics 33 (1962), 404–419.

[7] DE GROOT M. H.: *Optimal Statistical Decisions.* McGraw Hill, New York 1970.

[8] FICHTENGOLC G. M.: *Kurs diferencialnogo i integralnogo isčislenija* (Rusky), 2. díl, Fizmatgiz, Moskva 1962.

[9] GELFAND I. M., KOLMOGOROV A. N., YAGLOM A. M.: *On the general definition of the amount of information.* Doklady Akademii Nauk SSSR 111 (1956), 745–748.

[10] GYÖRFI L., VAJDA I.: *Asymptotic distributions for goodness-of-fit statistics in a sequence of multinomial models.* Statistics & Probability Letters 56 (2002), 57–67.

[11] GYÖRFI L., VAJDA I., VAN DER MEULEN E. C. *Minimum Hellinger distance point estimates consistent under weak family regularity.* Mathematical Methods of Statistics 3 (1994), 25–45.

[12] HOBZA T., MOLINA I., VAJDA I.: *On convergence of Fisher informations in continuous models with quantized observations.* Test 14 (2005), 151–179.

[13] KAGAN A. M.: *On the theory of Fisher's amount of information.* Doklady Akademii Nauk SSSR 151 (1963), 277–278.

[14] KULLBACK S.: *Information Theory and Statistics.* Wiley, New York 1959.

[15] KULLBACK S., LEIBLER R.: *On information and sufficiency.* Annals of Mathematical Statistics 22 (1951), 79–86.

[16] KŮS V.: Nonparametric density estimates consistent of the order of $n^{-1/2}$ in the L_1-norm. Metrika 60 (2004), 1–14.

[17] LEHMANN E. L., CASELLA G.: *Theory of Point Estimation* (2nd Edition). Springer, New York 1998.

[18] LIESE F., VAJDA I.: *Convex Statistical Distances*. Teubner, Leipzig 1987.

[19] LIESE F., VAJDA I.: *On divergences and informations in statistics and information theory*, IEEE Transactions on Information Theory 52 (2006), v tisku.

[20] LUSCHGY H., RUKHIN A., VAJDA I.: *Adaptive tests for stochastic processes*. Stochastic Processes and their Applications 45 (1993), 45–49.

[21] MORALES D., PARDO L., VAJDA I: *Some new statistics for testing hypotheses in parametric models*. Journal of Multivariate Analysis 62 (1997), 137–168.

[22] MORALES D., PARDO L., PARDO M. C., VAJDA I.: *Limit laws for disparities of spacings*. Nonparametric Statistics 15 (2003), 325–342.

[23] MÜLLER G., REINHARDT J., STICKLAND M. T.: *Neural Networks. An Introduction* (2. vydání). Springer, Berlin.

[24] ÖSTERREICHER F., VAJDA I.: *Statistical information and discrimination*. IEEE Transactions on Information Theory 39 (1993), 1036–1039.

[25] PARDO M. C., VAJDA I.: *About distances of discrete distributions satisfying the data processing theorem of information theory*. IEEE Transactions on Information theory 43 (1997), 1288–1293.

[26] PEREZ A.: *Notions généralisées d´incertitude, d´entropie et d´information du point de vue de la theorie de martingales*. Transactions of the First Prague Conference on Information Theory, ..., 183-208, Academia, Praha 1957.

[27] RUKHIN A., VAJDA I.: *Adaptive decision making for stochastic processes*. Journal of Statistical Planning and Inference 45 (1995), 313–329.

[28] SHANNON C. E.: *A mathematical theory of communication*. Bell System Technical Journal 27 (1948), 379–423, 623–656.

[29] VAJDA I.: *χ^α-divergence and generalized Fisher's information*. Transactions of the Sixth Prague Conference on Information Theory, 873–886, Academia, Praha 1973.

[30] VAJDA I.: *On convergence of information contained in quantized observations*. IEEE Transactions on Information Theory 48 (2002), 2163–2172.

[31] VAJDA I., JANŽURA M.: *On asymptotically optimal estimates for general observations*. Stochastic Processes and their Applications 72 (1997), 27–45.

[32] VAJDA I., VAN DER MEULEN E. C.: *On minimum divergence adaptation of discrete bivariate distributions to given marginals*. IEEE Transactions on Information Theory 52 (2005), 313–320.

[33] ZVÁROVÁ J.: *On measures of statistical dependence*. Časopis pro pěstování matematiky 99 (1974), 15-29.