Tadeusz Inglot (Wrocław)

Teoria informacji a statystyka matematyczna*

* Niniejszy artykuł jest rozszerzoną wersją wykładu wygłoszonego na XXXVIII Konferencji Statystyka Matematyczna Wisła 2012.

Streszczenie. W niniejszym artykule przedstawiony jest zarys teorii informacji z probabilistycznego i statystycznego punktu widzenia. Ten nurt teorii informacji rozwijał się intensywnie w ostatnich dziesięcioleciach. Wpłynął też w znaczący sposób na rozwój metod statystycznych. Celem artykułu jest wprowadzenie czytelnika w przystępny sposób w podaną powyżej tematykę, dostarczenie mu pewnych intuicji i przybliżenie specyfiki podejścia teorio-informacyjnego w statystyce matematycznej.

Abstract. In the paper we present an outline of the information theory from the probabilistic and statistical point of view. Such a direction of the information theory has been intensively developed in recent decades and significantly influenced a progress in the statistical methodology. The aim of the article is to introduce the reader into these problems, provide some intuitions and acquaint with a specific information-theoretic approach to the mathematical statistics.

2010 Mathematics Subject Classification: Primary 62B10; Secondary 94A15, 94A17, 60F, 62F, 62G.

Key words and phrases: Entropy, Kullback-Leibler distance, Fisher information, entropy convergence, statistical model and source coding, Stein's Lemma, density estimation.

1. Wstęp

Rozwój teorii informacji zapoczątkowała słynna praca Shannona z 1948 roku *A mathematical theory of communication*. Jej podstawowym celem były zastosowania techniczne związane z kodowaniem i przesyłaniem informacji. Wkrótce dostrzeżono znaczenie wyników Shannona dla teorii prawdopodobieństwa i statystyki matematycznej

i podjęto badania w tym kierunku. Wystarczy wspomnieć choćby monografię Kullbacka (1959), podręcznik Rényiego (1962) czy książkę Chinczyna i in. (1957). Jednak dopiero w ostatnich dziesięcioleciach nastąpił burzliwy rozwój probabilistycznej i statystycznej gałęzi teorii informacji. Znaczny udział mieli w nim między innymi A. Barron i J. Rissanen. Niewiele jest publikacji w języku polskim poświęconym tym zagadnieniom. Stąd próba wypełnienia istniejącej luki i przedstwienia ukierunkowanego przeglądu współczesnej teorii informacji w zwartej i przystępnej formie, co daje też okazję do zaproponowania polskich terminów dla niektórych pojęć, gdyż w dostępnej literaturze nie były dotąd używane. W krótkim szkicu nie jest możliwe uwzględnienie wszystkich ważnych wyników tak obszernego działu matematyki. Dlatego zdecydowałem się na pewien wybór materiału, podyktowany celem wykładu i osobistymi zainteresowaniami, jednakże przedstawiony na możliwie ogólnym tle.

Ostatnio ukazała się monografia Dębowskiego (2013) adresowana głównie do informatyków, w której skupiono się na problemach kodowania, złożoności obliczeniowej i analizie ciągów stacjonarnych w kontekście teorii informacji. Tak więc znaczna jej część dotyczy innych zagadnień niż przedstawione w niniejszym opracowaniu. Uzupełniając się wzajemnie, dają czytelnikowi możliwość zapoznania się z szerokim wachlarzem wyników współczesnej teorii informacji.

Rozdział drugi poświęcony jest wprowadzeniu podstawowych pojęć teorii informacji, omówieniu ich własności i wzajemnych związków. Dalsze dwa dotyczą kolejno zastosowań w teorii prawdopodobieństwa i statystyce. Literatura zwiera jedynie prace cytowane w toku wykładu. Wyczerpujący przegląd literatury można znaleźć np. w podręczniku Covera i Thomasa (2006).

2. Podstawy teorii informacji

2.1. Entropia, entropia względna. Niech A będzie zdarzeniem losowym. Zapytajmy, ile informacji jest zawartej w zdaniu *zaszło zdarzenie* A. Jeśli prawdopodobieństwo $P(A)$ jest duże, to informacji jest niewiele, a jeśli $P(A)$ jest mniejsze, to informacji jest więcej. Ponadto, gdy $P(A) = 1$, to nie ma żadnej informacji. Natomiast, gdy $P(A)$ jest coraz bliższe 0, to ilość informacji powinna wzrastać nieograniczenie. Te naturalne postulaty realizuje wyrażenie $-\log_2 P(A)$. Wybór funkcji logarytmicznej oraz podstawy 2 ma głębsze uzasadnienie, o czym powiemy nieco dalej. W dalszym ciągu będziemy opuszczać podstawę logarytmu i pisać $\log P(A)$ (dla odróżnienia, logarytm

naturalny będziemy oznaczać przez ln, a innych podstaw nie będziemy używać).

Rozważmy zmienną losową X o skończonym nośniku i rozkładzie μ danym równościami

$$P(X = x_i) = p_i, \quad i = 1, ..., r. \tag{2.1}$$

Zapytajmy teraz, ile informacji zawiera podanie wartości X. Ponieważ X wyznacza układ zdarzeń $A_i = \{X = x_i\}$, $i = 1, ..., r$, rozsądnie jest określić tę wielkość jako średnią ilość informacji zawartej w zdarzeniach $A_1, ..., A_r$.

Definicja 2.1. *Entropią (entropią Shannona) zmiennej losowej X o rozkładzie (2.1) nazywamy liczbę*

$$H(X) = H(\mu) = -\sum_{i=1}^{r} p_i \log p_i. \tag{2.2}$$

Dodatkowo w (2.2) przyjmujemy konwencję $0 \log 0 = 0$. Taka definicja została przyjęta przez Shannona w pionierskiej pracy z 1948 roku.

Rozumowanie prowadzące do powyższej definicji pozwala ją natychmiast rozszerzyć na zmienne losowe dwuwymiarowe (X, Y). Mianowicie, jeśli

$$P(X = x_i, Y = y_j) = p_{ij}, \quad i = 1, ..., r, \;\; j = 1, ..., s, \tag{2.3}$$

to

$$H(X, Y) = -\sum_{i,j} p_{ij} \log p_{ij}.$$

Analogicznie określa się entropię $H(X_1, ..., X_k)$ zmiennej wielowymiarowej $X = (X_1, ..., X_k)$.

Jeśli zmienna losowa ε ma rozkład dwupunktowy

$$P(\varepsilon = 1) = p = 1 - P(\varepsilon = 0), \tag{2.4}$$

to entropię ε oznaczamy przez $h(p)$ i nazywamy funkcją entropijną (ang. binary entropy function). Mamy więc

$$h(p) = -p \log p - (1 - p) \log(1 - p), \quad p \in [0, 1].$$

Funkcja entropijna jest ciągła, wklęsła, symetryczna względem $1/2$ oraz $h(1/2) = 1$. Wybór podstawy logarytmu 2 jest wyborem jednostki miary ilości informacji. Zatem podanie wyniku pojedynczego rzutu monetą daje 1 bit informacji. Standardowe rozumowanie prowadzi do nierówności

$$4p(1 - p) \leqslant h(p) \leqslant 2\sqrt{p(1 - p)}, \quad p \in [0, 1].$$

Entropia ma następujące proste i posiadające naturalną interpretację własności:

(A1) $0 \leqslant H(X) \leqslant r$, *przy czym równość* $H(X) = r$ *zachodzi wtedy i tylko wtedy, gdy* X *ma rozkład jednostajny na zbiorze* $\mathcal{X} = \{x_1, ..., x_r\}$;

(A2) $H(X,Y) \leqslant H(X) + H(Y)$ *dla dowolnych zmiennych losowych* X, Y, *przy czym równość zachodzi wtedy i tylko wtedy, gdy* X, Y *są niezależne*;

(A3) $H(p\mu + (1-p)\nu) = h(p) + pH(\mu) + (1-p)H(\nu)$ *dla dowolnych miar* μ, ν *o skończonych, rozłącznych nośnikach oraz dowolnego* $p \in [0,1]$.

Dowody (A1) i (A2) wykorzystują jedynie ścisłą wypukłość funkcji $C(y) = y \log y$ i są charakterystycznym, często powtarzającym się rozumowaniem w teorii informacji. Przytoczymy dowód pierwszej z tych własności.

Dowód (A1). Z wypukłości funkcji $C(y)$ wynika nierówność

$$y \log y \geqslant \tfrac{y-1}{\ln 2}, \quad y > 0,$$

gdyż jej prawa strona jest równaniem stycznej do wykresu $C(y)$ w punkcie $y = 1$. Równość ma miejsce tylko dla $y = 1$. Z powyższej nierówności mamy natychmiast

$$\log r - H(X) = \log r + \textstyle\sum_{i=1}^r p_i \log p_i = \sum_{i=1}^r p_i \log r + \sum_{i=1}^r p_i \log p_i$$
$$= \tfrac{1}{r} \textstyle\sum_{i=1}^r (rp_i) \log(rp_i) \geqslant \tfrac{1}{r} \sum_{i=1}^r \tfrac{rp_i - 1}{\ln 2} = 0,$$

przy czym równość ma miejsce wtedy i tylko wtedy, gdy $rp_i = 1$ dla wszystkich i. To kończy dowód. □

Interpretacja własności (A3) wydaje się mniej oczywista. Aby ją przedstawić przyjmijmy, że X, Y są zmiennymi losowymi o rozkładach μ, ν, a ε zmienną losową niezależną od X, Y o rozkładzie dwupunktowym (2.4). Wówczas zmienna losowa

$$V = \begin{cases} X, & \text{gdy } \varepsilon = 1, \\ Y, & \text{gdy } \varepsilon = 0, \end{cases}$$

ma rozkład $p\mu + (1-p)\nu$. Aby podać wartość V trzeba najpierw podać wartość ε, a następnie wartość X, gdy $\varepsilon = 1$ lub Y, gdy $\varepsilon = 0$. Odpowiednie ilości informacji należy zsumować, co daje prawą stronę równości (A3) zwanej własnością dekompozycji lub grupowania.

Okazuje się, że własności (A1)–(A3) charakteryzują entropię Shannona i w ten sposób uzasadniają wybór funkcji logarytmicznej jako miary ilości informacji.

TWIERDZENIE 2.1. *Niech $G(X) = G(\mu)$ będzie funkcją określoną na rozkładach o skończonych nośnikach spełniającą (A1)–(A3). Ponadto załóżmy, że $G(X)$ jest niezmiennicza na bijekcje (tzn. dla dowolnej bijekcji f mamy $G(f(X)) = G(X)$) oraz funkcja $g(p) = G(\varepsilon)$, $p \in [0,1]$, dla zmiennej ε o rozkładzie dwupunktowym (2.4), jest ciągła na $[0,1]$. Wówczas*

$$G(X) = -\sum_{i=1}^{r} p_i \log_c p_i$$

dla pewnego $c > 1$. Przyjmując dodatkowo $g(1/2) = 1$, otrzymujemy $c = 2$ czyli $G(X) = H(X)$.

Faddejew (1956) podał minimalny zbiór założeń gwarantujących tę samą tezę. Rényi (1961) uogólnił powyższe twierdzenie. Zastępując (A3) przez

(A3') $\quad \psi(G(p\mu + (1-p)\nu)) = p^\alpha \psi(G(\mu)) + (1-p)^\alpha \psi(G(\nu)),$

gdzie $\alpha > 0$, $\alpha \neq 1$, $\psi(y) = 2^{(1-\alpha)y}$, i zachowując wszystkie pozostałe własności, udowodnił, że jedyną funkcją spełniającą te założenia jest tzw. α-entropia Rényiego

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{r} p_i^\alpha \right).$$

Wartość parametru $\alpha = 1$ odpowiada entropii określonej przez (2.2). W ostatnich dekadach α-entropia, $\alpha \neq 1$, znalazła liczne zastosowania m.in. w lingwistyce matematycznej, teorii fraktali i informatyce kwantowej. Pewną wersją α-entropii Rényiego jest tzw. α-entropia Tsallisa (1988)

$$\frac{1}{1-\alpha} \left(\sum_{i=1}^{r} p_i^\alpha - 1 \right).$$

Więcej wiadomości o α-entropii można znaleźć np. w pracy Jizby i Arimitsu (2004).

W dalszym ciągu naszych rozważań pozostaniemy przy definicji entropii (2.2) przyjętej przez Shannona, która dobrze odpowiada zagadnieniom kodowania i przesyłania informacji, a co za tym idzie, także zagadnieniom statystyki matematycznej.

Interpretacja pojęć teorii informacji w języku kodów binarnych jest

często bardziej przekonywująca niż w języku mierzenia ilości informacji. Dotyczy to przede wszystkim samego pojęcia entropii.

Niech $\mathcal{X} = \{x_1, ..., x_r\}$ będzie ustalonym zbiorem (alfabetem). W celu przesyłania tekstów zapisanych w alfabecie $\mathcal{X}$ każdą literę kodujemy w postaci ciągu binarnego (słowa kodowego) $x_i \longrightarrow \alpha_1\alpha_2...\alpha_{k_i}$. Liczby $k_1, ..., k_r$ są długościami słów kodowych. Ograniczamy się do kodów jednoznacznie dekodowalnych tzn. takich, które pozwalają jednoznacznie podzielić przesłany bez zakłóceń ciąg binarny na słowa kodowe (litery). Wśród tych kodów szczególną rolę odgrywają kody przedrostkowe (lub inaczej prefiksowe), w których żadne słowo kodowe nie jest początkiem innego.

Przykładem kodu przedrostkowego dla $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$ może być: $x_1 \to 0,\ x_2 \to 10,\ x_3 \to 110,\ x_4 \to 111$. Ciąg 1100101110010 dekodujemy więc jako $110|0|10|111|0|0|10 \longrightarrow x_3x_1x_2x_4x_1x_1x_2$.

Dla kodów jednoznacznie dekodowalnych prawdziwa jest nierówność Krafta.

Twierdzenie 2.2 (nierówność Krafta, 1949). *Kod jest jednoznacznie dekodowalny wtedy i tylko wtedy, gdy* $\sum_{i=1}^{r} 2^{-k_i} \leqslant 1$.

Nierówność tę nazywa się także twierdzeniem Krafta-McMillana (McMillan, 1956).

Zmienną losową X o rozkładzie (2.1) nazywamy źródłem (ang. source). Rozważmy jednoznacznie dekodowalny kod binarny o słowach długości $k_1, ..., k_r$. Wówczas $\sum_{i=1}^{r} k_i p_i$ jest średnią długością słowa kodowego dla źródła X. Z nierówności Krafta wynika, że optymalnym (realizującym równość) wyborem długości słów jest $k_i = -\log p_i$. Ponieważ k_i muszą być liczbami naturalnymi, konieczne jest zaokrąglenie w górę tj. przyjęcie $k_i = \lceil -\log p_i \rceil$. Wtedy średnia długość słowa optymalnego kodu binarnego leży w przedziale $[H(X), H(X) + 1)$. Zatem entropię możemy interpretować jako średnią długość słowa optymalnego kodu binarnego (z dokładnością do 1 bitu) dla źródła X. Optymalnym kodem przedrostkowym jest kod Huffmana (1952).

Powróćmy do naszego przykładu. Jeśli, $p_1 = 0.5$, $p_2 = 0.25$, $p_3 = p_4 = 0.125$, to poprzednio określony kod jest optymalny, średnia długość słowa kodowego wynosi $1.75 = H(X)$ i jest to kod Huffmana. Natomiast, jeśli $p_1 = p_2 = 0.0625$, $p_3 = 0.375$, $p_4 = 0.5$, to ten kod jest daleki od optymalnego i średnia długość słowa kodowego wynosi ok. 2.81 i jest większa niż $H(X) + 1 \approx 2.53$. W tym przypadku kod Huffmana jest określony przez $x_1 \longrightarrow 110,\ x_2 \longrightarrow 111,$

$x_3 \longrightarrow 10$, $x_4 \longrightarrow 0$, ze średnią długością słowa kodowego równą 1.625.

Niech (X, Y) będzie dwuwymiarową zmienną losową o rozkładzie (2.3). Zapytajmy znów, ile informacji zawiera podanie wartości Y, gdy znamy X. Jeśli wiemy, że zaszło zdarzenie $\{X = x_i\}$, to zdarzenia $\{Y = y_j\}$ mają prawdopodobieństwa (warunkowe) $p_{ij}/p_{i\cdot}$, gdzie $p_{i\cdot} = \sum_{j=1}^{s} p_{ij}$, i ilość informacji potrzebnej wtedy do podania wartości Y wynosi $H(Y|X = x_i) = -\sum_{j=1}^{s} \frac{p_{ij}}{p_{i\cdot}} \log \frac{p_{ij}}{p_{i\cdot}}$. Zatem entropia warunkowa Y przy znanym X wynosi

$$H(Y|X) = \sum_{i=1}^{r} H(Y|X = x_i)\, p_{i\cdot} = -\sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_{i\cdot}}.$$

Powtarzając rozumowanie z dowodu własności (A1), otrzymujemy

$$H(Y|X) \leqslant H(Y),$$

przy czym równość ma miejsce wtedy i tylko wtedy, gdy X, Y są niezależne. Ostatnia nierówność zgadza się z intuicją, że dla X, Y zależnych podanie wartości Y przy znanym X wymaga mniejszej ilości informacji niż $H(Y)$.

Przez bezpośrednie sprawdzenie można udowodnić następujące twierdzenie, którego interpretacja jest oczywista.

Twierdzenie 2.3 (reguła łańcuchowa).

$$H(X, Y) = H(X) + H(Y|X).$$

Ogólniej

$$H(X_1, ..., X_n) = H(X_1) + H(X_2|X_1) + ... + H(X_n|X_1, ..., X_{n-1}).$$

Dodajmy, że dla α-entropii Rényiego reguła łańcuchowa w powyższej postaci nie jest prawdziwa. Używając pojęcia entropii warunkowej, można przyjąć, że ilość informacji o Y zawartej w X wynosi $I(Y, X) = H(Y) - H(Y|X)$. Z reguły łańcuchowej wynika równość $I(Y, X) = H(X) + H(Y) - H(X, Y) = I(X, Y)$ czyli $I(Y, X)$ jest także równa ilości informacji o X zawartej w Y.

Definicja 2.2. *Ilość informacji wzajemnej zmiennych losowych X, Y wynosi*

$$I(X, Y) = H(X) + H(Y) - H(X, Y). \tag{2.5}$$

Dotychczas rozważaliśmy zmienne losowe, których rozkład był znany. Jeśli jednak $p_1, ..., p_r$ nie są znane, to można je przybliżyć prawdo-

podobieństwami $q_1, ..., q_r$ czyli przez pewien rozkład ν na tym samym zbiorze (alfabecie) $\mathcal{X}$. Wtedy ilość informacji potrzebnej do podania wartości X wynosi $-\sum_{i=1}^{r} p_i \log q_i$ i różni się od ilości informacji, gdy znamy p_i o wielkość

$$-\sum_{i=1}^{r} p_i \log q_i + \sum_{i=1}^{r} p_i \log p_i = \sum_{i=1}^{r} p_i \log \frac{p_i}{q_i} \stackrel{ozn}{=} D(\mu||\nu).$$

Powtarzając jeszcze raz rozumowanie analogiczne do dowodu własności (A1), otrzymujemy twierdzenie.

TWIERDZENIE 2.4. *Dla dowolnych rozkładów μ, ν o wspólnym skończonym nośniku $\mathcal{X}$ mamy*

$$D(\mu||\nu) \geqslant 0$$

i równość zachodzi wtedy i tylko wtedy, gdy $\mu = \nu$ tzn. $p_i = q_i$ dla wszystkich i.

A więc zastąpienie nieznanych p_i przez ich przybliżenia q_i powoduje wzrost entropii o $D(\mu||\nu)$. Wielkość $D(\mu||\nu)$ nazywamy entropią względną, odległością Kullbacka-Leiblera lub odległością entropijną μ od ν. Przyjmując nadal konwencję $0 \log 0 = 0$, definicję $D(\mu||\nu)$ możemy rozszerzyć na przypadek, gdy nośnik ν zawiera nośnik μ (rozkład przybliżający ν jest bardziej 'rozmyty'). Uogólnienie pojęcia entropii względnej na szerszą klasę rozkładów omówimy w rozdziale 2.2.

Rozumowanie prowadzące do określenia entropii względnej wygodnie jest zinterpretować w języku kodów binarnych. Jeśli konstruujemy optymalny kod binarny dla źródła X w oparciu o przybliżenia q_i nieznanych prawdopodobieństw p_i, to średnia długość słowa kodowego (z dokładnością do zaokrąglenia w górę) wynosi $-\sum_i p_i \log q_i$. W stosunku do kodu optymalnego opartego na dokładnych prawdopodobieństwach p_i średnie wydłużenie (inny polski termin redundancja, ang. redundancy) słów kodowych wynosi $D(\mu||\nu)$.

Odległość entropijna ma następujące własności, które łatwo sprawdzić bezpośrednim rachunkiem.

TWIERDZENIE 2.5.

(i) $D(\mu||\nu)$ nie jest symetryczną funkcją μ, ν, czyli na ogół $D(\mu||\nu) \neq D(\nu||\mu)$;

(ii) $D(\mu||\lambda) = \log r - H(\mu) = H(\lambda) - H(\mu)$, gdzie λ jest rozkładem jednostajnym na nośniku μ;

(iii) $D(\mu||\mu_1 \times \mu_2) = I(X,Y)$, *gdzie* μ *jest rozkładem dwuwymiarowej zmiennej losowej* (X,Y), *a* μ_1, μ_2 *rozkładami brzegowymi;*

(iv) $$D(\alpha\mu_1 + (1-\alpha)\mu_2||\alpha\nu_1 + (1-\alpha)\nu_2) \leqslant \alpha D(\mu_1||\nu_1) + (1-\alpha)D(\mu_2||\nu_2)$$

dla dowolnych rozkładów μ_1, ν_1 *o tym samym nośniku, dowolnych* μ_2, ν_2 *o tym samym nośniku i* $\alpha \in [0,1]$.

Własność (iii) wiąże ilość informacji wzajemnej X, Y z odległością entropijną rozkładu (X,Y) od produktu rozkładów brzegowych. Własność (iv) nazywamy wypukłością odległości entropijnej. W dowodzie (iv) korzysta się z tzw. nierówności logarytmiczno-sumacyjnej (ang. log–sum inequality)

$$(x_1 + x_2)\log\frac{x_1+x_2}{y_1+y_2} \leqslant x_1\log\frac{x_1}{y_1} + x_2\log\frac{x_2}{y_2}, \quad x_1, x_2, y_1, y_2 > 0,$$

która jest bezpośrednią konsekwencją wypukłości funkcji $C(y)$.

2.2. Entropia różniczkowa, odległość Kullbacka-Leiblera. W modelowaniu zjawisk losowych w naturalny sposób pojawiają się rozkłady o nieskończonym, a nawet nieprzeliczalnym nośniku. Dlatego pojęcia wprowadzone w poprzednim paragrafie przeniesiemy na przypadek nieskończonych nośników.

Niech X będzie dyskretną zmienną losową o rozkładzie μ danym równościami

$$P(X = x_i) = p_i, \quad i = 1, 2, ...$$

Rozumowanie prowadzące do definicji entropii (por. (2.2)) stosuje się także w tym przypadku i entropię X określamy wzorem

$$H(X) = -\sum_{i=1}^{\infty} p_i \log p_i. \tag{2.6}$$

Entropia pozostaje wielkością dodatnią, ale może przyjmować wartość $+\infty$ (szereg może być rozbieżny). Dla rozkładów μ, ν o tych samych nośnikach (lub ewentualnie nośniku ν zawierającym nośnik μ) analogicznie określamy entropię względną $D(\mu||\nu) = \sum_{i=1}^{\infty} p_i \log(p_i/q_i)$ i analogicznie wykazujemy, że jest ona nieujemna i równa zeru wtedy i tylko wtedy, gdy $\mu = \nu$.

Jako przykład rozważmy zmienną losową X o rozkładzie geometrycznym γ_m z parametrem $1/m$, $m > 1$, tzn.

$$P(X = i) = \frac{1}{m}\left(1 - \frac{1}{m}\right)^{i-1}, \quad i = 1, 2, ...$$

Bezpośredni rachunek daje $H(X) = m\,h(1/m)$. Łatwo sprawdzić, że $\lim_{m\to\infty} m\,h(1/m) = \infty$. Oznacza to, że jeśli prawdopodobieństwo sukcesu $1/m$ maleje do 0, to ilość informacji potrzebnej do podania wartości X rośnie nieograniczenie. A więc własność (A1) entropii mówiąca, że na zbiorze skończonym $\mathcal{X}$ istnieje rozkład o maksymalnej entropii nie ma odpowiednika w klasie wszystkich rozkładów o nośniku przeliczalnym i skończonej entropii. Można jednak wykazać, że w klasie rozkładów na zbiorze liczb naturalnych o ustalonej wartości oczekiwanej m istnieje rozkład o maksymalnej entropii.

TWIERDZENIE 2.6. *W klasie $\mathcal{M}_m$ rozkładów na zbiorze liczb naturalnych o ustalonej wartości oczekiwanej $m > 1$, maksymalną entropię osiąga rozkład geometryczny γ_m z parametrem $1/m$. Mamy więc*

$$\sup_{\mu\in\mathcal{M}_m} H(\mu) = H(\gamma_m) = m\,h(1/m),$$

przy czym równość ma miejsce wtedy i tylko wtedy, gdy $\mu = \gamma_m$.

Dowód. Dla dowolnego rozkładu $\mu \in \mathcal{M}_m$ czyli takiego, że $\sum_{i=1}^{\infty} ip_i = m$ mamy

$$D(\mu||\gamma_m) = \textstyle\sum_{i=1}^{\infty} p_i \log \frac{p_i\, m^i}{(m-1)^{i-1}}$$

$$= \textstyle\sum_{i=1}^{\infty} p_i \log p_i - (\sum_{i=1}^{\infty} ip_i) \log \frac{m-1}{m} + \log(m-1) = -H(\mu) + H(\gamma_m).$$

Ponieważ $D(\mu||\gamma_m) \geqslant 0$ i równość zachodzi tylko gdy $\mu = \gamma_m$, to $H(\mu) \leqslant H(\gamma_m)$, co kończy dowód. □

Entropia określona wzorem (2.6) zachowuje wszystkie własności entropii rozkładów o skończonym nośniku.

Niech teraz X będzie rzeczywistą zmienną losową o gęstości $p(x)$. Na chwilę załóżmy dodatkowo, że p jest ciągła na $\mathbb{R}$. Podanie dokładnej wartości X wymaga nieskończonej ilości informacji. Podanie wartości X z dokładnością $\delta > 0$ prowadzi do dyskretyzacji X. Mianowicie, niech X_δ oznacza zmienną losową określoną wzorem

$$X_\delta = i\delta, \quad \text{gdy} \quad X \in [i\delta, (i+1)\delta), \quad i \in \mathbb{Z}.$$

Wtedy jej rozkład określają prawdopodobieństwa

$$P(X_\delta = i\delta) = \int_{[i\delta,(i+1)\delta)} p(x)\,dx \;=\; \delta p(x_i^*), \quad i \in \mathbb{Z},$$

dla pewnych $x_i^* \in [i\delta, (i+1)\delta)$ z twierdzenia o wartości średniej dla całki i ciągłości p. Entropia dyskretyzacji X_δ wynosi zatem

$$H(X_\delta) = -\sum_{i\in\mathbb{Z}} \delta p(x_i^*) \log(\delta p(x_i^*))$$
$$= -\sum_{i\in\mathbb{Z}} (p(x_i^*) \log p(x_i^*))\, \delta - \log \delta. \tag{2.7}$$

Pierwszy składnik po prawej stronie wzoru (2.7) jest sumą całkową całki Riemanna $-\int_{\mathbb{R}} p(x) \log p(x)\, dx$. Z ciągłości p wynika następująca równość

$$\lim_{\delta\to 0}(H(X_\delta) + \log \delta) = -\int_{\mathbb{R}} p(x) \log p(x)\, dx. \tag{2.8}$$

Wyrażenie po prawej stronie wzoru (2.8) nazywamy entropią różniczkową zmiennej losowej X (krótko entropią X) i oznaczamy tak jak poprzednio przez $H(X)$. W dalszym ciągu prawą stronę (2.8) przyjmujemy jako definicję entropii dowolnej zmiennej losowej o rozkładzie absolutnie ciągłym i gęstości $p(x)$, dla której ta całka (jako całka Lebesgue'a) istnieje, z dopuszczeniem wartości $+\infty$ oraz $-\infty$. Interpretacja $H(X)$ wynika z (2.8) i jest nieco inna niż poprzednio. Ilość informacji potrzebna do podania wartości X z dokładnością do m miejsc binarnych po przecinku wynosi $H(X) + m$. Bezpośrednio z definicji dostajemy

(i) $H(X + c) = H(X), \quad H(cX) = H(X) + \log|c|, \quad c \in \mathbb{R}$;

(ii) $H(Y|X) = -\int\int_{\mathbb{R}\times\mathbb{R}} p(x, y) \log p(y|x) dx\, dy$, *gdzie* $p(x, y)$ *jest gęstością dwuwymiarowej zmiennej losowej* (X, Y), *a* $p(y|x)$ *gęstością warunkową* Y *przy warunku* $X = x$;

(iii) $I(X, Y) = H(X) + H(Y) - H(X, Y)$;

(iv) $H(X + Y) \geqslant \max\{H(X), H(Y)\}$, *gdy* X, Y *niezależne.* (2.9)

Własność (i) odróżnia entropię różniczkową od entropii zmiennej losowej dyskretnej, gdyż niezmienniczość na bijekcje przestaje obowiązywać. Ilość informacji wzajemnej $I(X, Y)$ zachowuje taką samą interpretację jak w przypadku dyskretnym. Własność (iv) jest konsekwencją relacji $H(X, Y) = H(X, X + Y)$ oraz (A2). Definicja entropii różniczkowej (2.8) przenosi się natychmiast na przypadek wielowymiarowy $X = (X_1, ..., X_k)$ wzorem

$$H(X) = H(X_1, ..., X_k) = -\int_{\mathbb{R}^k} p(x_1, ..., x_k) \log p(x_1, ..., x_k) dx_1 ... dx_k.$$

Analogia między entropią różniczkową i entropią zmiennych dyskretnych prowadzi do pytania o istnienie rozkładu ciągłego o maksymalnej entropii. Odpowiedź jest nieco bardziej złożona niż w przypadku zmiennych dyskretnych. Najpierw rozważmy zmienną losową o nośniku zwartym $[a, b]$. Wówczas mamy następujące twierdzenie,

które jest odpowiednikiem własności (A1).

TWIERDZENIE 2.7. *Jeśli X ma rozkład absolutnie ciągły o nośniku $[a,b]$, to*

$$H(X) \leqslant \log(b-a) = H(U_{[a,b]}),$$

gdzie $U_{[a,b]}$ oznacza zmienną losową o rozkładzie jednostajnym na odcinku $[a,b]$. Równość ma miejsce wtedy i tylko wtedy, gdy $X \stackrel{\mathcal{D}}{=} U_{[a,b]}$.

Dowód. Rozumowanie jest analogiczne do dowodu własności (A1). Niech $p(x)$ będzie gęstością X. Wtedy z wypukłości funkcji $C(y)$ mamy

$$\begin{aligned}\log(b-a) - H(X) &= \textstyle\int_a^b p(x)\log[(b-a)p(x)]dx \\ &= \tfrac{1}{b-a}\textstyle\int_a^b [(b-a)p(x)]\log[(b-a)p(x)]dx \\ &\geqslant \tfrac{1}{b-a}\textstyle\int_a^b \frac{(b-a)p(x)-1}{\ln 2}dx = 0,\end{aligned}$$

przy czym równość zachodzi wtedy i tylko wtedy, gdy $(b-a)p(x) = 1$ prawie wszędzie. To kończy dowód. □

Tak samo dowodzi się następujące twierdzenie.

TWIERDZENIE 2.8.

(1) *Jeśli X ma rozkład absolutnie ciągły o nośniku $[0,\infty)$ oraz $EX = m$, to*

$$H(X) \leqslant \log(e\,m) = H(W_m),$$

gdzie W_m oznacza zmienną losową o rozkładzie wykładniczym z parametrem $1/m$. Równość ma miejsce wtedy i tylko wtedy, gdy $X \stackrel{\mathcal{D}}{=} W_m$;

(2) *jeśli X ma rozkład absolutnie ciągły o nośniku $\mathbb{R}$ oraz $EX = 0$ i $\operatorname{Var} X = t$, to*

$$H(X) \leqslant \frac{1}{2}\log(2\pi\, e\, t) = H(Z_t), \tag{2.10}$$

gdzie Z_t oznacza zmienną losową o rozkładzie normalnym, średniej 0 i wariancji t. Równość ma miejsce wtedy i tylko wtedy, gdy $X \stackrel{\mathcal{D}}{=} Z_t$;

(3) *jeśli X ma rozkład absolutnie ciągły o nośniku $\mathbb{R}^k$, średniej 0 i macierzy kowariancji K, to*

$$H(X) \leqslant \frac{1}{2}\log\left((2\pi\, e)^k \det K\right) = H(Z_K), \tag{2.11}$$

gdzie Z_K oznacza zmienną losową o rozkładzie normalnym $N(0,K)$. Równość ma miejsce wtedy i tylko wtedy, gdy $X \stackrel{\mathcal{D}}{=} Z_K$.

Część (1) twierdzenia 2.8 jest odpowiednikiem twierdzenia 2.6, a (2) i (3) nie mają odpowiednika w klasie rozkładów dyskretnych i zawierają ciekawe i ważne spostrzeżenie, iż rozkład normalny maksymalizuje entropię i pełni rolę rozkładu jednostajnego w klasie rozkładów o nośniku $\mathbb{R}^k$. Ustalenie wartości wariancji jest konieczne, gdyż $\lim_{t\to\infty} H(Z_t) = \lim_{t\to\infty} \log(2\pi\, e\, t) = \infty$.

Definicja entropii różniczkowej (2.8) sugeruje sposób przeniesienia pojęcia entropii względnej na przypadek rozkładów absolutnie ciągłych.

Definicja 2.3. *Niech μ, ν będą rozkładami absolutnie ciągłymi w $\mathbb{R}^k$ o gęstościach, odpowiednio, p i q. Wielkość*

$$D(\mu||\nu) = D(p||q) = \begin{cases} \displaystyle\int_{\mathbb{R}^k} p(x) \log \frac{p(x)}{q(x)} dx, & \text{gdy } \mu \prec\prec \nu, \\ +\infty, & \text{poza tym}, \end{cases} \tag{2.12}$$

nazywamy odległością Kullbacka-Leilblera, entropią względną lub odległością entropijną μ od ν.

Interpretacja odległości entropijnej podana w poprzednim paragrafie pozostaje słuszna w przypadku rozkładów absolutnie ciągłych. Jeśli nieznana gęstość p zmiennej losowej X jest przybliżana przez gęstość q (być może o większym nośniku), to ilość informacji potrzebna do podania X (z ustaloną dokładnością) zwiększa się o $D(p||q)$ w stosunku do sytuacji, gdy p jest znana. Podobnie, optymalny kod binarny oparty na przybliżeniu q nieznanej gęstości p źródła powoduje wydłużenie (redundancję) średniej długości słowa kodowego o $D(p||q)$.

Własności odległości entropijnej omówione w poprzednim rozdziale obowiązują także w obecnej sytuacji. Dla kompletności powtórzymy je i uzupełnimy dalszymi, ważnymi dla rozważań teoretycznych.

Twierdzenie 2.9.

(i) $D(p||q) \geqslant 0$ i równość zachodzi wtedy i tylko wtedy, gdy $p = q$ prawie wszędzie;

(ii) $D(p||q)$ nie jest funkcją symetryczną p, q, czyli na ogół $D(p||q) \neq D(q||p)$;

(iii) $I(X,Y) = D(\mu||\mu_1 \times \mu_2)$, gdzie μ jest rozkładem dwuwymiarowej zmiennej losowej ciągłej (X,Y), a μ_1, μ_2 rozkładami brzegowymi;

(iv) $D(\alpha\mu_1 + (1-\alpha)\mu_2||\alpha\nu_1 + (1-\alpha)\nu_2)$
$$\leqslant \alpha D(\mu_1||\nu_1) + (1-\alpha)D(\mu_2||\nu_2)$$

dla dowolnych rozkładów absolutnie ciągłych $\mu_1 \prec\prec \nu_1$ *i* $\mu_2 \prec\prec \nu_2$ *oraz* $\alpha \in [0,1]$*;*

(v) *jeśli* $p_n \xrightarrow{\lambda} p$ *oraz* $q_n \xrightarrow{\lambda} q$, *to* $\liminf_{n\to\infty} D(p_n||q_n) \geqslant D(p||q)$, *gdzie* λ *oznacza miarę Lebesgue'a;*

(vi) $(\log e)\, d^2_{\mathcal{H}}(p,q) \leqslant D(p||q)$*;*

(vii) $$\frac{\log e}{2}||p-q||^2_{L_1} \leqslant D(p||q). \tag{2.13}$$

Własność (i) nazywa się nierównością Gibbsa, własność (v) oznacza dolną półciągłość odległości entropijnej, $d_{\mathcal{H}}$ jest odległością Hellingera ($d^2_{\mathcal{H}}(p,q) = \int_{\mathbb{R}}(\sqrt{p}-\sqrt{q})^2 d\lambda$), a własność (vii) nazywa się nierównością Pinskera (Csiszár, 1967, Kullback, 1967). Nierówności (vi) i (vii) pokazują, że odległość Kullbacka-Leiblera ma charakter kwadratu 'odległości' rozkładów. Własność (ii) z twierdzenia 2.5, pominięta w twierdzeniu 2.9 stanowi treść twierdzeń 2.7 i 2.8. Dowód trzech ostatnich własności szkicujemy poniżej.

Dowód. Własność (v) wynika bezpośrednio z Lematu Fatou i faktu, że funkcja $C(y) = y\log y$ jest ograniczona z dołu. Dla dowodu (vi) wystarczy zauważyć, że funkcja $\psi(y) = y\ln y + 2\sqrt{y} - 2y$ jest nieujemna na $(0,\infty)$ oraz że $\int_{\mathbb{R}^k} \psi(\frac{p(x)}{q(x)})q(x)dx = \frac{1}{\log e}D(p||q) - d^2_{\mathcal{H}}(p,q)$. Zamiast (vii) udowodnimy tutaj nieco słabszą nierówność tj. ze stałą 4 zamiast 2. Z nierówności Schwarza, własności (vi) oraz faktu, że $d^2_{\mathcal{H}}(p,q) = 2 - 2\int_{\mathbb{R}}\sqrt{pq}\,d\lambda$ mamy

$$\begin{aligned}\Big(\int_{\mathbb{R}^k} |p(x)-q(x)|dx\Big)^2 &\\ \leqslant \int_{\mathbb{R}^k}(\sqrt{p(x)}-\sqrt{q(x)})^2dx \int_{\mathbb{R}^k}(\sqrt{p(x)}+\sqrt{q(x)})^2dx &\\ = (4 - d^2_{\mathcal{H}}(p,q))d^2_{\mathcal{H}}(p,q) \leqslant 4d^2_{\mathcal{H}}(p,q) \leqslant \tfrac{4}{\log e}D(p||q)&\end{aligned}$$

co dowodzi (2.13) z gorszą stałą. □

Na koniec zauważmy, że pojęcia entropii i odległości Kullbacka-Leiblera można przenieść natychmiast na przypadek zmiennych losowych o wartościach w dowolnej przestrzeni miarowej σ-skończonej $(\mathcal{X}, \mathcal{B}, \lambda)$, przepisując definicje (2.8) i (2.12). Prawie wszystkie omówione przez nas własności entropii oraz odległości entropijnej przenoszą się na przypadek ogólny wraz z dowodami. Dołączymy do nich jeszcze

jedną, którą można nazwać własnością kontrakcji odległości Kullbacka-Leiblera.

TWIERDZENIE 2.10 (Kullback, 1959). *Niech μ_1, μ_2 będą rozkładami na $(\mathcal{X}, \mathcal{B}, \lambda)$ o gęstościach $p_1 = \frac{d\mu_1}{d\lambda}, p_2 = \frac{d\mu_2}{d\lambda}$ i niech $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{Y}, \mathcal{C})$ będzie odwzorowaniem mierzalnym. Wówczas*

$$D(\mu_1 T^{-1} || \mu_2 T^{-1}) \leqslant D(\mu_1 || \mu_2)$$

i równość ma miejsce wtedy i tylko wtedy, gdy $\frac{p_1(x)}{p_2(x)} = \frac{q_1(T(x))}{q_2(T(x))}$ $\lambda-$p.w., gdzie q_1, q_2 są gęstościami rozkładów $\mu_1 T^{-1}, \mu_2 T^{-1}$ względem miary indukowanej λT^{-1} na $\mathcal{Y}$.

Dowód. Stosując zamianę zmiennych w całce, otrzymujemy

$$D(\mu_1 T^{-1} || \mu_2 T^{-1}) = \int_{\mathcal{X}} p_1(x) \log \tfrac{q_1(T(x))}{q_2(T(x))} \lambda(dx).$$

Stąd

$$D(\mu_1 || \mu_2) - D(\mu_1 T^{-1} || \mu_2 T^{-1}) = \int_{\mathcal{X}} p_1(x) \log \tfrac{p_1(x) q_2(T(x))}{p_2(x) q_1(T(x))} \lambda(dx).$$

Ponieważ $\int \frac{q_1(T(x))}{q_2(T(x))} p_2(x) \lambda(dx) = 1$, to powtarzając rozumowanie z dowodu własności (A1), pokazujemy, że całka po prawej stronie jest nieujemna i równa 0 wtedy i tylko wtedy, gdy $\frac{p_1(x) q_2(T(x))}{p_2(x) q_1(T(x))} = 1$ prawie wszędzie względem miary λ, co kończy dowód. □

2.3. Rzut informacyjny. Rozważania przeprowadzimy w ogólnym przypadku tj. dla zmiennych losowych o wartościach w dowolnej przestrzeni miarowej σ-skończonej $\mathcal{X}$. Niech $\mathcal{P}$ będzie pewną rodziną rozkładów absolutnie ciągłych na $\mathcal{X}$ i niech $\mu \in \mathcal{P}^c$ będzie ustalonym rozkładem absolutnie ciągłym takim, że $D(\nu||\mu) < \infty$ dla pewnego $\nu \in \mathcal{P}$.

DEFINICJA 2.4. *Rozkład $\mu^* \in \mathcal{P}$ nazywa się rzutem informacyjnym (ang. information projection) rozkładu μ na $\mathcal{P}$, jeśli*

$$D(\mu^*||\mu) = \min_{\nu \in \mathcal{P}} D(\nu||\mu).$$

Piszemy $\mu^* = Iproj_{\mathcal{P}}(\mu)$.

Warunki wystarczające istnienia rzutu informacyjnego podaje następujące twierdzenie.

TWIERDZENIE 2.11 (Csiszár, 1975). *Jeśli rodzina $\mathcal{P}$ jest wypukła i domknięta w normie całkowitego wahania, to dla dowolnego rozkładu absolutnie ciągłego $\mu \in \mathcal{P}^c$ takiego, że $D(\nu||\mu) < \infty$ dla pewnego $\nu \in \mathcal{P}$, rzut informacyjny $Iproj_{\mathcal{P}}(\mu)$ istnieje.*

Dowód. Wybierzmy ciąg rozkładów ν_n o gęstościach $q_n = \frac{d\nu_n}{d\lambda}$, dla którego $\lim_{n\to\infty} D(\nu_n||\mu) = \inf_{\nu\in\mathcal{P}} D(\nu||\mu)$. Z wypukłości $\mathcal{P}$ oraz nierówności Pinskera (2.13) mamy

$$\begin{aligned}\tfrac{\log e}{4}||q_n - q_m||_{L_1}^2 &\leqslant \tfrac{\log e}{2}(||q_n - \tfrac{q_n+q_m}{2}||_{L_1}^2 + ||q_m - \tfrac{q_n+q_m}{2}||_{L_1}^2)\\ &\leqslant D(q_n||\tfrac{q_n+q_m}{2}) + D(q_m||\tfrac{q_n+q_m}{2})\\ &= D(q_n||p) + D(q_m||p) - 2D(\tfrac{q_n+q_m}{2}||p) \to 0,\end{aligned}$$

gdzie p jest gęstością μ. To oznacza, że ciąg (q_n) jest podstawowy w L_1, czyli z domkniętości $\mathcal{P}$ zbieżny w normie L_1 do pewnej gęstości p^* miary $\mu^* \in \mathcal{P}$. Z dolnej półciągłości odległości Kullbacka-Leiblera wynika, że μ^* jest rzutem informacyjnym. □

Csiszár pokazał również ważną własność geometryczną rzutu informacyjnego.

TWIERDZENIE 2.12 (Csiszár, 1975). *Przy założeniach twierdzenia 2.11 mamy $\mu^* = Iproj_{\mathcal{P}}(\mu)$ wtedy i tylko wtedy, gdy dla każdego rozkładu $\nu \in \mathcal{P}$ zachodzi nierówność*

$$D(\nu||\mu) \geqslant D(\nu||\mu^*) + D(\mu^*||\mu). \tag{2.14}$$

W konsekwencji μ^* jest wyznaczona jednoznacznie.

Kwadrat normy euklidesowej w $\mathbb{R}^k$ ma analogiczną własność do (2.14), gdyż rzut ortogonalny x^* punktu $x \in W^c$ na zbiór wypukły W, punkt x oraz dowolny punkt $y \in W$ tworzą trójkąt rozwartokątny, co skutkuje nierównością $||y-x||^2 \geqslant ||y-x^*||^2 + ||x^*-x||^2$. Jednak odległość Kullbacka-Leiblera ma inną geometrię niż norma euklidesowa. Wiadomo, że jeśli x^* jest rzutem ortogonalnym punktu $x \in \mathbb{R}^k$ na podprzestrzeń afiniczną A, to A jest zawarta w hiperpłaszczyźnie stycznej do kuli o środku x i promieniu $||x^*-x||$ tj. zbiorze wszystkich punktów y, dla których spełnione jest twierdzenie Pitagorasa $||y-x||^2 = ||y-x^*||^2 + ||x^*-x||^2$. Csiszár (1975, str. 152) podał przykład rodziny 'afinicznej' rozkładów, który nie jest zawarty w 'hiperpłaszczyźnie stycznej' $\mathcal{T}$ do D-kuli o środku μ i 'promieniu' $\sqrt{D(\mu^*||\mu)}$ tj. zbiorze wszystkich ν, dla których w (2.14) zachodzi

równość. Pokazał natomiast, że jeśli $\mathcal{P} = \{\alpha\nu_1 + (1-\alpha)\nu_2 : \alpha \in [0,1]\}$ i $\mu^* = Iproj_{\mathcal{P}}(\mu)$ leży wewnątrz odcinka $\mathcal{P}$ (ogólniej, μ^* jest punktem algebraicznie wewnętrznym zbioru wypukłego $\mathcal{P}$), to $\mathcal{P} \subset \mathcal{T}$. Podobnie rodzina $\mathcal{P}$ wyznaczona przez skończoną liczbę warunków całkowych (liniowych) jest zawarta w $\mathcal{T}$.

Dualne pojęcie zwane odwrotnym rzutem informacyjnym (ang. reversed information projection) wprowadził Czencow (1972). Csiszár i Matúš (2003) uogólnili pojęcia rzutu i odwrotnego rzutu informacyjnego i wszechstronnie je przebadali. Niech $\mathcal{P}$ i μ będą jak poprzednio i takie, że $D(\mu||\nu) < \infty$ dla pewnego $\nu \in \mathcal{P}$.

Definicja 2.5. *Rozkład $\mu_* \in \mathcal{P}$ nazywamy odwrotnym rzutem informacyjnym μ na $\mathcal{P}$, jeśli $D(\mu||\mu_*) = \min_{\nu\in\mathcal{P}} D(\mu||\nu)$. Piszemy $\mu_* = RIproj_{\mathcal{P}}(\mu)$.*

Dla odwrotnego rzutu informacyjnego prawdziwe są analogiczne fakty do zawartych w twierdzeniach 2.11 i 2.12. Mówimy, że rodzina $\mathcal{P}$ jest logarytmicznie wypukła, jeśli dla dowolnych $\mu, \nu \in \mathcal{P}$ o gęstościach $p, q,\ p\cdot q \neq 0$, i dowolnego $\alpha \in (0,1)$ rozkład o gęstości $c_\alpha p^\alpha q^{1-\alpha}$ należy do $\mathcal{P}$, gdzie c_α jest stałą normującą. Oczywiście, wykładnicza rodzina rozkładów jest logarytmicznie wypukła.

Twierdzenie 2.13 (Csiszár, Matúš, 2003). *Jeśli rodzina $\mathcal{P}$ jest logarytmicznie wypukła i domknięta w normie całkowitego wahania oraz $D(\mu||\nu) < \infty$ dla pewnego $\nu \in \mathcal{P}$, to istnieje jednoznacznie wyznaczony odwrotny rzut informacyjny μ na $\mathcal{P}$ i $\mu_* = RIproj_{\mathcal{P}}\mu$ wtedy i tylko wtedy, gdy*

$$D(\mu||\nu) \geqslant D(\mu||\mu_*) + D(\mu_*||\nu)$$

dla wszystkich $\nu \in \mathcal{P}$.

Czytelnika zainteresowanego problematyką rzutów informacyjnych, wykładniczych rodzin rozkładów i uogólnionych rodzin wykładniczych odsyłamy do, wspomnianej już, obszernej pracy Csiszára i Matúša (2003) oraz monografii Barndorffa-Nielsena (1978).

2.4. Macierz informacji Fishera. Rozważmy parametryczną rodzinę $\mathcal{P} = \{p_\vartheta : \vartheta \in \Theta,\ \Theta \subset \mathbb{R}^k \text{ otwarty}\}$ gęstości rozkładów na przestrzeni σ-skończonej $(\mathcal{X}, \mathcal{B}, \lambda)$. Zapytajmy, jak zmienia się odległość entropijna $p_{\vartheta'}$ od p_ϑ, gdy ϑ' przebiega małe otoczenie ϑ. Załóżmy, że $\mathcal{P}$ jest dostatecznie regularna, w szczególności p_ϑ mają

ciągłe pochodne cząstkowe rzędu 2 względem ϑ w otoczeniu ϑ (precyzyjne sformułowanie założeń pomijamy). Wtedy ze wzoru Taylora mamy

$$-\log\frac{p_\vartheta}{p_{\vartheta'}} = \log p_{\vartheta'} - \log p_\vartheta = (\nabla \log p_\vartheta)^T(\vartheta' - \vartheta) \tag{2.15}$$

$$+\frac{1}{2}(\vartheta' - \vartheta)^T \frac{\partial^2 \log p_\vartheta}{\partial\vartheta\partial\vartheta^T}(\vartheta' - \vartheta) + o(||\vartheta' - \vartheta||^2),$$

gdzie $\nabla\psi_\vartheta = \frac{\partial\psi_\vartheta}{\partial\vartheta}$ dla funkcji ψ_ϑ, $\vartheta \in \mathbb{R}^k$. Ponieważ $\int_{\mathcal{X}} p_\vartheta \nabla \log p_\vartheta d\lambda = \log e \int_{\mathcal{X}} \nabla p_\vartheta d\lambda = 0$, to, mnożąc (2.15) przez p_ϑ i całkując stronami, dostajemy

$$D(p_\vartheta||p_{\vartheta'}) = \frac{\log e}{2}(\vartheta' - \vartheta)^T \mathbf{J}(\vartheta)(\vartheta' - \vartheta) + o(||\vartheta' - \vartheta||^2),$$

gdzie

$$\mathbf{J}(\vartheta) = -\int_{\mathcal{X}} \frac{\partial^2 \ln p_\vartheta}{\partial\vartheta\partial\vartheta^T} p_\vartheta d\lambda = \int_{\mathcal{X}} (\nabla \ln p_\vartheta)(\nabla \ln p_\vartheta)^T p_\vartheta d\lambda \tag{2.16}$$

nazywamy macierzą informacji Fishera. Jak widać odpowiada ona za lokalną prędkość zmian odległości entropijnej rozkładów względem kwadratu odległości euklidesowej parametrów. Analogicznie dowodzi się relacji

$$D(p_{\vartheta'}||p_\vartheta) = \frac{\log e}{2}(\vartheta' - \vartheta)^T \mathbf{J}(\vartheta)(\vartheta' - \vartheta) + o(||\vartheta' - \vartheta||^2).$$

Prawą stronę wzoru (2.16) przyjmujemy za definicję macierzy informacji Fishera $\mathbf{J}(\vartheta)$ bez wymagania poprawności wyprowadzenia rozwinięcia (2.15) i równości w (2.16), a tylko istnienia całki. $\mathbf{J}(\vartheta)$ jest więc macierzą kowariancji wektora $\ell_\vartheta = \nabla \ln p_\vartheta$, który nazywamy wektorem wynikowym (ang. score vector) i który występuje w rozwinęciu (2.15).

Macierz informacji Fishera można też interpretować nieco inaczej. Jeśli ϑ' jest oszacowaniem nieznanego (prawdziwego) ϑ, to przy tej samej małej odległości entropijnej $D(p_\vartheta||p_{\vartheta'})$ oszacowanie jest tym dokładniejsze (w sensie odległości euklidesowej) im $\mathbf{J}(\vartheta)$ większa.

Dla dwóch macierzy symetrycznych A, B relacje $A > B$ i $A \geqslant B$ oznaczają, że $A - B$ jest, odpowiednio, dodatnio i nieujemnie określona.

Podstawową własnością macierzy informacji Fishera jest nierówność Rao-Craméra.

Twierdzenie 2.14 (nierówność Rao-Craméra). *Niech X będzie zmienną losową o wartościach w $\mathcal{X}$ i gęstości $p_\vartheta \in \mathcal{P}$, gdzie $\mathcal{P}$ jest rodziną gęstości o ciągłych pochodnych cząstkowych względem ϑ taką, że*

macierz informacji Fishera określona przez prawą stronę wzoru (2.16) istnieje i jest ciągłą funkcją ϑ. Niech $T(x) \in \mathbb{R}^k$ będzie funkcją mierzalną na $\mathcal{X}$ taką, że $E\,T(X) = \vartheta$ oraz Cov $T(X) = E(T(X) - \vartheta)(T(X) - \vartheta)^T = K_\vartheta > 0$. Wówczas

$$\mathbf{J}(\vartheta) \geqslant K_\vartheta^{-1}$$

lub w równoważnej postaci, częściej używanej w statystyce matematycznej, $K_\vartheta \geqslant \mathbf{J}(\vartheta)^{-1}$. Równość ma miejsce wtedy i tylko wtedy, gdy $T = \mathbf{J}(\vartheta)^{-1}\nabla \ln p_\vartheta + \vartheta$.

W literaturze statystycznej (np. Lehmann, Casella, 1998) nierówność Rao-Craméra jest nazywana także nierównością informacyjną.

Dowód. Z założenia $E\,T(X) = \vartheta$ i różniczkowalności rodziny $\mathcal{P}$ mamy

$$\int_{\mathcal{X}} (\nabla \ln p_\vartheta) T^T p_\vartheta d\lambda = \int_{\mathcal{X}} (\nabla p_\vartheta) T^T d\lambda = \nabla \int_{\mathcal{X}} p_\vartheta T^T d\lambda = \nabla \vartheta^T = I,$$

gdzie I oznacza macierz jednostkową. W konsekwencji z (2.16)

$$\begin{aligned} 0 &\leqslant E(\nabla \ln p_\vartheta(X) - K_\vartheta^{-1}(T(X) - \vartheta))(\nabla \ln p_\vartheta(X) - K_\vartheta^{-1}(T(X) - \vartheta))^T \\ &= \mathbf{J}(\vartheta) - 2K_\vartheta^{-1} + K_\vartheta^{-1} E(T(X) - \vartheta)(T(X) - \vartheta)^T K_\vartheta^{-1} = \mathbf{J}(\vartheta) - K_\vartheta^{-1} \end{aligned}$$

co kończy dowód. □

W twierdzeniu 2.14 oraz w definicji macierzy informacji Fishera założenia o różniczkowalności punktowej można zastąpić przez wygodniejsze założenie różniczkowalności średniokwadratowej (ang. differentiability in quadratic mean) rodziny $\mathcal{P}$ (por. van der Vaart, 1998, str. 93-97).

2.5. Nieparametryczna informacja Fishera. Ważne znaczenie w teorii informacji ma nieparametryczna macierz informacji Fishera. Słowo 'nieparametryczna' jest tutaj użyte dla odróżnienia od macierzy informacji Fishera w rodzinie parametrycznej omówionej w rozdziale 2.4. Niech X będzie zmienną losową w $\mathbb{R}^k$ z absolutnie ciągłą gęstością p oraz $\mathcal{P} = \{p_\vartheta : p_\vartheta(x) = p(x-\vartheta), \vartheta \in \mathbb{R}^k\}$ rodziną z parametrem przesunięcia. Wektor wynikowy ℓ_ϑ zdefiniowany w poprzednim rozdziale ma teraz postać $\ell_\vartheta = -\frac{\partial \ln p}{\partial x}(\cdot - \vartheta)$. To sugeruje, aby wektor

$$\ell = -\ell_0 = \frac{\partial \ln p}{\partial x} = \nabla \ln p \tag{2.17}$$

także nazywać wektorem wynikowym (teraz ∇ oznacza różniczkowanie względem zmiennej x). Z niezmienniczości miary Lebesgue'a na prze-

sunięcia wynika, że macierz informacji Fishera $\mathbf{J}(\vartheta)$ (o ile istnieje) nie zależy od ϑ, tylko od p i z (2.16) otrzymujemy

$$\mathbf{J} = \mathbf{J}(p) = \mathbf{J}(X) = \int_{\mathcal{X}} \ell\ell^T p \, d\lambda. \tag{2.18}$$

Ponadto liczbę $J(X) = \operatorname{tr} \mathbf{J}(X) = E||\nabla \ln p(X)||^2$ nazywamy informacją Fishera. W przypadku jednowymiarowym macierz informacji Fishera jest liczbą, $\mathbf{J}(X) = J(X) = \int_{\mathbb{R}} \left(\frac{p'(x)}{p(x)}\right)^2 p(x)\, dx$ i słowo *macierz* opuszczamy. Bezpośrednio z definicji (2.17) wynikają następujące własności macierzy $\mathbf{J}$:

(i) $\mathbf{J}(X + c) = \mathbf{J}(X), \quad \mathbf{J}(cX) = \frac{1}{c^2}\mathbf{J}(X)$;

(ii) $\mathbf{J}(AX) = (A^{-1})^T \mathbf{J}(X) A^{-1}$ *dla dowolnej macierzy nieosobliwej* A;

(iii) jeśli K *jest macierzą dodatnio określoną i* Z_K *jest zmienną losową o rozkładzie normalnym* $N(0, K)$ *w* $\mathbb{R}^k$, *to* $\mathbf{J}(Z_K) = K^{-1}$;

(iv) dla dowolnej zmiennej losowej X *w* $\mathbb{R}^k$ *o wartości oczekiwanej 0 i dodatnio określonej macierzy kowariancji* K *mamy* $\mathbf{J}(X) \geqslant K^{-1}$ *oraz* $J(X) \geqslant \operatorname{tr} K^{-1}$, *przy czym równość zachodzi wtedy i tylko wtedy, gdy* X *ma rozkład normalny* $N(0, K)$.

Własność (iv) jest szczególnym przypadkiem nierówności Rao-Craméra. Możemy z niej wyprowadzić interpretację $J(X)$. Zauważmy, że dla dowolnej nieosobliwej, symetrycznej macierzy K stopnia k mamy $\operatorname{tr} K^{-1} \geqslant k/t$, gdzie t jest największą wartością własną macierzy K i równość zachodzi wtedy i tylko wtedy, gdy $K = t\, I$. Zatem dostajemy nierówność $J(X) \geqslant k/t$ i równość zachodzi wtedy i tylko wtedy, gdy X ma rozkład normalny $N(0, t\, I)$. Oznacza to, że wyrażenie $k/J(X)$ jest wariancją k-wymiarowego szumu białego o zadanej informacji Fishera $J(X)$.

Przez analogię do określenia odległości Kullbacka-Leiblera można wprowadzić pojęcie odległości Fishera (ang. Fisher information distance).

Definicja 2.6. *Niech* X, Y *będą zmiennymi losowymi w* $\mathbb{R}^k$ *o rozkładach* μ, ν, $\mu \prec\prec \nu$, *i absolutnie ciągłych gęstościach* p, q. *Wówczas*

$$\mathbf{J}(X||Y) = \mathbf{J}(\mu||\nu) = \mathbf{J}(p||q) = \int_{\mathbb{R}^k} \left(\nabla \ln \frac{p}{q}\right) \left(\nabla \ln \frac{p}{q}\right)^T p \, d\lambda,$$

nazywamy odległością Fishera μ od ν, o ile całka jest skończona.

Oczywiście, $\mathbf{J}(\mu||\nu) \geqslant 0$ i równa zeru wtedy i tylko wtedy, gdy $\mu = \nu$. Podobnie jak dla odległości entropijnej, $\mathbf{J}(\mu||\nu)$ nie jest symetryczna i na ogół $\mathbf{J}(\mu||\nu) \neq \mathbf{J}(\nu||\mu)$. W przypadku $k = 1$ i nieskończonej całki przyjmujemy, że $J(X||Y) = +\infty$. Dla przykładu, jeśli β_{pq} oznacza gęstość rozkładu beta na odcinku $(0, 1)$ z parametrami p, q, to bezpośredni rachunek daje $J(\beta_{21}||\beta_{31}) = +\infty$, $J(\beta_{31}||\beta_{21}) = 3$, $J(\beta_{31}||\beta_{51}) = 12$, a $J(\beta_{51}||\beta_{31}) = 20/3$.

Z nierówności Rao-Craméra (iv) wynika, że rozkład normalny ma minimalną macierz informacji Fishera w klasie rozkładów o tej samej macierzy kowariancji. Dlatego odległość Fishera od rozkładu normalnego odgrywa ważną rolę. Powtarzając dowód nierówności Rao-Craméra, otrzymujemy następujące twierdzenie.

TWIERDZENIE 2.15. *Jeśli K jest macierzą nieosobliwą i Z_K ma rozkład $N(0, K)$ w $\mathbb{R}^k$, to dla dowolnej zmiennej losowej X w $\mathbb{R}^k$ o średniej 0, macierzy kowariancji K i absolutnie ciągłej gęstości mamy*

$$\mathbf{J}(X||Z_K) = \mathbf{J}(X) - K^{-1} = \mathbf{J}(X) - \mathbf{J}(Z_K). \tag{2.19}$$

Przypomnijmy, że z (2.11)

$$D(X||Z_K) = \int_{\mathbb{R}^k} p(x) \log p(x)\, dx + \frac{1}{2} \log[(2\pi)^k \det K] + \frac{\log e}{2} \int_{\mathbb{R}^k} x^T K^{-1} x\, p(x) dx = H(Z_K) - H(X) \tag{2.20}$$

co jest widocznym odpowiednikiem (2.19).

W dalszym ciągu przez φ_t będziemy oznaczać gęstość rozkładu normalnego o średniej 0 i wariancji t. Związek odległości Fishera z innymi odległościami rozkładów podajemy w kolejnym twierdzeniu.

TWIERDZENIE 2.16. *Jeśli X jest rzeczywistą zmienną losową (k=1) o średniej 0, wariancji 1 i absolutnie ciągłej gęstości p, to*

(i) $D(p||\varphi_1) \leqslant \dfrac{\log e}{2} J(p||\varphi_1)$;

(ii) $\sup\limits_{x \in \mathbb{R}} |p(x) - \varphi_1(x)| \leqslant \left(1 + \sqrt{\dfrac{6}{\pi}}\right) \sqrt{J(p||\varphi_1)}$.

Dowód (i) podamy w rozdziale 2.6. Nierówność (ii) udowodnił Shimizu (1975). Obie nierówności pokazują, że odległość Fishera, o ile

jest skończona, majoryzuje inne odległości rozkładów, łącznie z odległością Kullbacka-Leiblera.

2.6. Tożsamość de Bruijna. Maksymalizacja entropii i minimalizacja informacji Fishera przez rozkład normalny łącznie z (2.19) i (2.20) sugerują, że istnieje związek pomiędzy odległością Fishera i odległością entropijną. Mówi o tym kolejne twierdzenie.

TWIERDZENIE 2.17 (tożsamość de Bruijna). *Niech X będzie rzeczywistą zmienną losową o średniej 0, wariancji 1 i gęstości p i niech Z będzie zmienną losową niezależną od X o rozkładzie standardowym normalnym. Wówczas dla każdego $t > 0$*

$$\frac{d}{dt}H(X+\sqrt{t}Z) = \frac{\log e}{2}J(X+\sqrt{t}Z). \tag{2.21}$$

Ponadto

$$D(p||\varphi_1) = \frac{\log e}{2}\int_0^\infty \left[J(X+\sqrt{t}Z) - \frac{1}{1+t}\right] dt. \tag{2.22}$$

Dowód postaci różniczkowej (2.21) tożsamości de Bruijna podał Stam (1959). Opiera się on na fakcie, że φ_t spełnia równanie przewodnictwa cieplnego, z czego wynika, że gęstość $X+\sqrt{t}Z$ także spełnia to równanie. Dalsze rozumowanie sprowadza się do różniczkowania pod znakiem całki określającej $H(X+\sqrt{t}Z)$ i całkowaniu przez części. Szczegóły pomijamy. Postać całkowa (2.22) została udowodniona przez Barrona (1986). Zauważmy, że funkcja podcałkowa w (2.22) jest odległością Fishera $J(X+\sqrt{t}Z)||\sqrt{1+t}Z)$.

Szkic dowodu (2.22). Oznaczmy przez g_t gęstość zmiennej losowej $X+\sqrt{t}Z$. Ponieważ rozkład normalny maksymalizuje entropię, a entropia sumy niezależnych składników jest nie mniejsza od entropii składników (por. (2.9)), to z (2.10) mamy $\frac{1}{2}\log 2\pi et = H(\varphi_t) \leqslant H(g_t) \leqslant H(\varphi_{1+t}) = \frac{1}{2}\log 2\pi e(t+1)$, skąd dostajemy

$$\lim_{t\to\infty}[H(g_t) - \tfrac{1}{2}\log 2\pi e(1+t)] \longrightarrow 0.$$

W konsekwencji z (2.21) dla $0 < a < b$

$$\begin{aligned}\tfrac{\log e}{2}\textstyle\int_a^b (J(g_t) - \frac{1}{1+t})dt &= H(g_b) - \tfrac{\log(1+b)}{2} - H(g_a) + \tfrac{\log(1+a)}{2} \\ &= H(g_b) - \tfrac{\log(2\pi e(1+b))}{2} + D(g_a||\varphi_{1+a}).\end{aligned} \tag{2.23}$$

Przechodząc w (2.23) z b do nieskończoności dostajemy

$$D(g_a||\varphi_{1+a}) = \tfrac{\log e}{2}\textstyle\int_a^\infty \left(J(g_t) - \frac{1}{1+t}\right) dt$$

dla każdego $a > 0$. Biorąc granicę z obu stron względem $a \to 0^+$ (możliwość przejścia do granicy pod znakiem całki wymaga jeszcze nietrywialnego uzasadnienia) dostajemy (2.22). □

W celu podania jeszcze jednej interpretacji informacji Fishera przyjmijmy, że X jest sygnałem, a $\sqrt{t}Z$ niezależnym od niego szumem gaussowskim o wariancji t, którą interpretujemy jako moc (lub poziom) szumu $\sqrt{t}Z$. Z niezależności X i Z wynika, że ilość informacji wzajemnej $X + \sqrt{t}Z$ i Z ma postać (por. (2.5))

$$I(X + \sqrt{t}Z, Z) = H(X + \sqrt{t}Z) + H(Z) - H(X + \sqrt{t}Z, Z)$$
$$= H(X + \sqrt{t}Z) + H(Z) - H(X) - H(Z) = H(X + \sqrt{t}Z) - H(X).$$

Zatem z (2.21) otrzymujemy

$$\frac{d}{dt} I(X + \sqrt{t}Z, Z) = \frac{\log e}{2} J(X + \sqrt{t}Z) \tag{2.24}$$

dla wszystkich $t > 0$. Riuol (2011) pokazał, że równość (2.24) jest prawdziwa dla $t = 0$ i w konsekwencji mamy

$$I(X + \sqrt{t}Z, Z) = \frac{\log e}{2} J(X)t + o(t)$$

gdy $t \to 0^+$. Ostatnia równość oznacza, że dla małego poziomu szumu ilość informacji wzajemnej sygnału zakłóconego addytywnym szumem gaussowskim i szumu zależy liniowo od mocy szumu t, a $J(X)$ jest współczynnikiem proporcjonalności. Tak więc informację Fishera $J(X)$ można interpretować jako czułość sygnału X na zakłócenie szumem gaussowskim. Najmniejsza czułość jest dla sygnału gaussowskiego, co jest oczywiste.

Tożsamości (2.21) i (2.22) można przenieść na przypadek wielowymiarowy. Przytoczymy tę drugą w postaci udowodnionej przez Johnsona i Suchowa (2001).

Twierdzenie 2.18 (wielowymiarowa tożsamość de Bruijna). *Niech X, Z_K będą niezależnymi zmiennymi losowymi w $\mathbb{R}^k$ o średnich 0, macierzach kowariancji odpowiednio, B i K, $K > 0$, przy czym Z_K ma rozkład normalny $N(0, K)$. Jeśli p jest gęstością X, to*

$$D(p||\varphi_K) = \frac{\log e}{2} \int_0^\infty \left[tr(K\,\mathbf{J}(X + \sqrt{t}Z_K)) - \frac{k}{1+t} \right] dt$$
$$+ \frac{\log e}{2} [tr(K^{-1}B) - k].$$

Gdy $B = K$, drugi człon po prawej stronie znika.

Tożsamość de Bruijna ma liczne, ważne zastosowania oraz służy jako narzędzie w rozważaniach teoretycznych. Na przykład, dowód centralnego twierdzenia granicznego w wersji entropijnej, podany przez Barrona (1986), wykorzystuje tę tożsamość. Omówimy go w rozdziale 3.4.

2.7. Nierówności dla informacji Fishera. Sumowanie niezależnych zmiennych losowych ma fundamentalne znaczenie w rachunku prawdopodobieństwa i statystyce matematycznej. Dlatego ważne jest ustalenie związków entropii sumy niezależnych zmiennych losowych z entropią składników i uzyskanie podobnych związków dla informacji Fishera. W tym rozdziale przyjrzymy się temu drugiemu zagadnieniu, które prowadzi do nierówności dla informacji Fishera (ang. Fisher information inequalities). Rozpoczniemy od pomocniczego lematu Barrona i Johnsona (2004).

LEMAT 2.2. *Jeśli X, Y są niezależnymi, rzeczywistymi ($k = 1$) zmiennymi losowymi o gęstościach p, q, skończonych informacjach Fishera i funkcjach wynikowych (por. (2.17)) $\ell_X = p'/p$, $\ell_Y = q'/q$, to dla każdego $\alpha \in [0, 1]$*

$$\begin{aligned} &E(\ell_{X+Y}(X+Y) - \alpha\ell_X(X) - (1-\alpha)\ell_Y(Y))^2 \\ &\quad = \alpha^2 J(X) + (1-\alpha)^2 J(Y) - J(X+Y). \end{aligned}$$

Szkic dowodu. Najpierw należy pokazać, że

$$\ell_{X+Y}(u) = E(\ell_X(X)|X+Y=u) = E(\ell_Y(Y)|X+Y=u) \quad \text{p.w.},$$

co nie jest całkiem oczywiste. Następnie mnożąc odpowiednio przez α i $1-\alpha$ i dodając stronami, dostajemy

$$\ell_{X+Y}(X+Y) = E([\alpha\ell_X(X) + (1-\alpha)\ell_Y(Y)]|X+Y)$$

prawie na pewno. Ponieważ warunkowa wartość oczekiwana jest rzutem ortogonalnym, to

$$\begin{aligned} E\left(\alpha\ell_X(X) + (1-\alpha)\ell_Y(Y)\right)^2 &= E(\ell_{X+Y}(X+Y))^2 \\ &\quad + E(\ell_{X+Y}(X+Y) - \alpha\ell_X(X) - (1-\alpha)\ell_Y(Y))^2, \end{aligned}$$

co daje prawą stronę dzięki niezależności X, Y i własności $E\ell_X(X) = \int_{\mathbb{R}} \ell_X p\, d\lambda = 0$ funkcji wynikowej. □

Bezpośrednim wnioskiem z lematu 2.2 jest nierówność dla informacji Fishera.

TWIERDZENIE 2.19. *Przy założeniach lematu 2.2 dla każdego* $\alpha \in [0,1]$

$$J(X+Y) \leqslant \alpha^2 J(X) + (1-\alpha)^2 J(Y). \tag{2.25}$$

Równość ma miejsce wtedy i tylko wtedy, gdy X, Y *mają taki sam rozkład normalny.*

Indukcja matematyczna pozwala przenieść nierówność (2.25) na dowolną liczbę składników:

$$J(X_1 + ... + X_n) \leqslant \alpha_1^2 J(X_1) + ... + \alpha_n^2 J(X_n), \tag{2.26}$$

gdzie $X_1, ..., X_n$ są niezależne oraz $\alpha_i \in [0,1]$, $\sum_{i=1}^n \alpha_i = 1$. Nierówność (2.26) udowodnili Stam (1959) oraz Blachman (1965), nie korzystając z lematu 2.2. Nierówność (2.26) można zapisać w kilku równoważnych postaciach:

(i) $$J(\sqrt{\alpha_1}X_1 + ... + \sqrt{\alpha_n}X_n) \leqslant \alpha_1 J(X_1) + ... + \alpha_n J(X_n); \tag{2.27}$$

(ii) $$\frac{1}{J(X_1 + ... + X_n)} \geqslant \frac{1}{J(X_1)} + ... + \frac{1}{J(X_n)}; \tag{2.28}$$

(iii) $$\frac{1}{J(\alpha_1 X_1 + ... + \alpha_n X_n)} \geqslant \frac{\alpha_1^2}{J(X_1)} + ... + \frac{\alpha_n^2}{J(X_n)}.$$

Równoważność (2.26) i (2.27) oraz (ii) i (iii) otrzymuje się przez podstawienie $X_i = \sqrt{\alpha_i} X_i'$ oraz skorzystanie z własności $J(cX) = J(X)/c^2$. Równoważność (2.26) z (ii) wystarczy udowodnić dla $n = 2$ i następnie zastosować indukcję matematyczną. W tym celu podstawiamy w (2.26) $\alpha_1 = J(X_2)/(J(X_1) + J(X_2))$ i $\alpha_2 = J(X_1)/(J(X_1) + J(X_2))$ i dostajemy (ii). Natomiast dowód w przeciwną stronę wynika z elementarnej nierówności arytmetycznej $\frac{uv}{u+v} \leqslant \alpha^2 u + (1-\alpha)^2 v$ prawdziwej dla dowolnych dodatnich u, v oraz $\alpha \in [0,1]$.

Nierówności dla informacji Fishera mają ważne zastosowania w teorii informacji i były intensywnie badane w ostatnich latach m.in. przez Barrona i Johnsona (2004), Barrona i Madimana (2007) czy Rioula (2011). W szczególności wiążą się z problematyką centralnego twierdzenia granicznego w wersji entropijnej. Podstawiając w nierówności (2.27) $\alpha_1 = ... = \alpha_n = 1/\sqrt{n}$ oraz rozważając zmienne o tym samym rozkładzie, dostajemy prostą nierówność

$$J\left(\frac{X_1 + ... + X_n}{\sqrt{n}}\right) \leqslant J(X_1),$$

która oznacza, że sumowanie niezależnych składników prowadzi do zmniejszania się informacji Fishera.

Jako przykład zastosowania nierówności dla informacji Fishera podajemy dowód nierówności (i) z twierdzenia 2.16.

Dowód (i) z twierdzenia 2.16. Z całkowej tożsamości de Bruijna, (2.27) i faktu, że $J(Z) = 1$ mamy

$$\begin{aligned}\tfrac{2}{\log e}D(p||\varphi_1) &= \textstyle\int_0^\infty (J(X+\sqrt{t}Z) - \tfrac{1}{1+t})dt \\ &= \textstyle\int_0^\infty \tfrac{1}{1+t}\left[J(\sqrt{\tfrac{1}{1+t}}X + \sqrt{\tfrac{t}{1+t}}Z) - 1\right]dt \\ &\leqslant \textstyle\int_0^\infty \tfrac{1}{1+t}\left[\tfrac{1}{1+t}J(X) + \tfrac{t}{1+t} - 1\right]dt \\ &= \textstyle\int_0^\infty \tfrac{1}{(1+t)^2}dt\,(J(X)-1) = J(p||\varphi_1)\end{aligned}$$

co kończy dowód. □

2.8. Nierówności dla mocy entropijnej. W tym rozdziale przyjrzymy się nierównościom dla entropii sum niezależnych zmiennych losowych zwanym nierównościami dla mocy entropijnej (ang. entropy power inequalities).

Ze wzoru (2.11) otrzymujemy entropię białego szumu gaussowskiego $Z_t = (Z_{t1}, ..., Z_{tk})$ w $\mathbb{R}^k$ o wariancji (mocy) t jako

$$H(Z_t) = \frac{k}{2}\log 2\pi e t.$$

Stąd dostajemy $t = \frac{1}{2\pi e}2^{2H(Z_t)/k} = N(Z_t)$ czyli wyrażenie określające wariancję (moc) białego szumu gaussowskiego w $\mathbb{R}^k$ o danej entropii $H(Z_t)$. Równość powyższą przyjmujemy jako definicję w przypadku ogólnym.

Definicja 2.7. *Mocą entropijną zmiennej losowej X w $\mathbb{R}^k$ o macierzy kowariancji K nazywamy liczbę*

$$N(X) = \frac{1}{2\pi e}2^{2H(X)/k}.$$

Ponieważ $H(X) \leqslant H(Z_K) = \frac{k}{2}\log 2\pi e + \frac{1}{2}\log \det K$, to

$$N(X) \leqslant (\det K)^{1/k} \leqslant \kappa,$$

gdzie κ jest największą wartością własną K. Równość zachodzi wtedy i tylko wtedy, gdy X jest białym szumem o wariancji (mocy) $t = \kappa$.

Nierówność dla mocy entropijnej sformułował Shannon (1948) w następującej wersji.

TWIERDZENIE 2.20 (Shannon, 1948). *Jeśli X, Y są niezależnymi zmiennymi losowymi w $\mathbb{R}^k$ o rozkładach absolutnie ciągłych, to*

$$2^{2H(X+Y)/k} \geqslant 2^{2H(X)/k} + 2^{2H(Y)/k}$$

to znaczy

$$N(X+Y) \geqslant N(X) + N(Y),$$

przy czym równość ma miejsce wtedy i tylko wtedy, gdy X ma rozkład normalny $N(0,K)$, a Y rozkład normalny $N(0,cK)$ dla pewnego $c>0$.

Przez indukcję dostajemy

$$N(X_1 + ... + X_n) \geqslant N(X_1) + ... + N(X_n) \tag{2.29}$$

i równość zachodzi tylko dla zmiennych normalnych o proporcjonalnych macierzach kowariancji. Dowód (2.29) podali Stam (1959) i Blachman (1965) we wspomnianych już pracach. Nierówność (2.29) nie jest prawdziwa dla zmiennych losowych dyskretnych. Z własności entropii wynika, że dla dowolnego $c>0$ jest $N(cX) = c^2 N(X)$. Korzystając z tej własności i podstawiając w (2.29) $X_i = \sqrt{\alpha_i} X_i'$, gdzie $\alpha_i \geqslant 0$, $\sum_{i=1}^n \alpha_i = 1$, otrzymujemy natychmiast nierówność

$$N(\sqrt{\alpha_1}X_1 + ... + \sqrt{\alpha_n}X_n) \geqslant \alpha_1 N(X_1) + ... + \alpha_n N(X_n),$$

która jest odpowiednikiem (2.27) dla mocy entropijnej, a stąd i z wklęsłości logarytmu

$$H(\sqrt{\alpha_1}X_1 + ... + \sqrt{\alpha_n}X_n) \geqslant \alpha_1 H(X_1) + ... + \alpha_n H(X_n),$$

przy czym równość w obu ostatnich nierównościach zachodzi tylko dla zmiennych losowych normalnych o proporcjonalnych macierzach kowariancji. Można wykazać, że dwie ostatnie nierówności są równoważne. Ponadto są one prawdziwe także dla zmiennych dyskretnych.

Nierówności dla mocy entropijnej i dla samej entropii oraz ich uogólnienia były badane w wielu pracach, ostatnio m.in. przez Barrona i Madimana (2007) czy Rioula (2011). W szczególności Barron i Madiman (2007) pokazali, że dla zmiennych niezależnych o jednakowym rozkładzie ciąg entropii

$$\frac{H(X_1 + ... + X_n)}{\sqrt{n}}$$

jest niemalejący, co oznacza, że sumowanie niezależnych składników prowadzi do wzrostu entropii, odwrotnie niż dla informacji Fishera.

2.9. Przetwarzanie danych. Na koniec krótko wspomnimy o bardzo ważnym, np. w informatyce czy telekomunikacji, zagadnieniu

przetwarzania danych i kompresji danych. Czynimy to dlatego, że ma ono także odniesienia do statystyki matematycznej i będziemy z podanych poniżej wyników korzystać w rozdziałach 3.1 i 4.3.

Najpierw rozszerzymy pojęcie entropii, zdefiniowane w rozdziałach 2.1 i 2.2, na przypadek niektórych rozkładów osobliwych. Niech X będzie zmienną losową o wartościach w dowolnej przestrzeni $(\mathcal{X}, \mathcal{B}, \lambda)$, a T odwzorowaniem mierzalnym na $\mathcal{X}$. Wówczas przyjmujemy $H(X, T(X)) = H(X)$ i w konsekwencji $H(T(X)|X) = 0$. Podobnie przyjmujemy $H(X, Y, T(X)) = H(X, Y)$.

Definicja 2.8 (warunkowa ilość informacji). *Wyrażenie*

$$I(Z, X|Y) = H(Z|Y) - H(Z|X, Y)$$

nazywamy warunkową ilością informacji wzajemnej X, Z przy warunku Y, gdzie X, Y, Z są dowolnymi zmiennymi losowymi, dla których prawa strona jest zdefiniowana.

Bezpośrednie sprawdzenie daje $I(Z, X|Y) = H(Z|Y) + H(X|Y) - H(Z, X|Y)$. Analogiczne rozumowanie jak w dowodzie własności (A1) z rozdziału 2.1 pokazuje, że $I(Z, X|Y) \geqslant 0$. Gdy p jest gęstością łączną X, Y, Z, to równość ma miejsce wtedy i tylko wtedy, gdy $p(z, x|y) = p(z|y)p(x|y)$ p.w. tzn. gdy X, Z są warunkowo niezależne przy warunku Y. Ponadto mamy następującą własność.

Twierdzenie 2.21 (reguła łańcuchowa). *Dla dowolnych X, Y, Z, jak w definicji 2.8, mamy*

$$I((Y, Z), X) = I(Z, X) + I(Y, X|Z). \tag{2.30}$$

Mówimy, że zmienne losowe X, Y, Z tworzą łańcuch Markowa, jeśli

$$I(Z, X|Y) = 0. \tag{2.31}$$

Fakt ten notujemy w postaci $X \to Y \to Z$. Oczywiście $X \to Y \to Z$ wtedy i tylko wtedy, gdy $Z \to Y \to X$. Ponadto, jeśli X, Y, Z mają gęstość łączną p, to definicja (2.31) oznacza, że przy danej 'teraźniejszości' Y, 'przeszłość' X nie zależy od 'przyszłości' Z. Z drugiej strony, jeśli $Z = T(Y)$, to dla dowolnej zmiennej X mamy $X \to Y \to Z$.

Lemat 2.3. *Jeśli $X \to Y \to Z$, to $I(X, Y|Z) = I(X, Y) - I(X, Z)$. W konsekwencji $I(X, Y|Z) \leqslant I(X, Y)$.*

Dowód. Z (2.30) i założenia (2.31) mamy

$$\begin{aligned} I(X,Y|Z) &= I((Y,Z),X) - I(X,Z) \\ &= I(Z,X|Y) + I(X,Y) - I(X,Z) = I(X,Y) - I(X,Z) \end{aligned}$$

co kończy dowód. □

Zauważmy, że nierówność z lematu 2.3 nie musi zachodzić, gdy zmienne nie tworzą łańcucha Markowa (por. przykład w monografii Covera i Thomasa, 2006, str. 35). Bezpośrednim wnioskiem z lematu 2.3 jest następująca ważna i pożyteczna nierówność.

TWIERDZENIE 2.22 (nierówność dla danych przetworzonych). *Jeśli* $X \to Y \to Z$, *to*

$$I(X,Z) \leqslant I(X,Y). \tag{2.32}$$

Równość ma miejsce wtedy i tylko wtedy, gdy $I(X,Y|Z) = 0$.

Nierówność (2.32) oznacza, że dane przetworzone Z nie zawierają więcej informacji o X niż dane wyjściowe Y. W literaturze angielskiej nierówność (2.32) nazywana jest 'data processing inequality'.

3. Twierdzenia graniczne

Druga zasada termodynamiki mówi, że układ odizolowany od otoczenia ewoluuje w kierunku stanów o większej entropii i w nieskończonym horyzoncie czasowym osiąga stan o maksymalnej entropii. Matematycznym odpowiednikiem tego prawa są twierdzenia mówiące o zbieżności entropii. W tym rozdziale przedstawimy kilka najważniejszych twierdzeń tego typu oraz twierdzenia graniczne, które mają odniesienia do pojęć teorii informacji.

3.1. Asymptotyczna zasada ekwipartycji. Podstawową własnością średniej z próby jest jej zbieżność do wartości oczekiwanej pojedynczej obserwacji. Stanowi ona treść prawa wielkich liczb, które można formułować w różnych wersjach. Asymptotyczna zasada ekwipartycji (ang. asymptotic equipartition property) są uogólnieniami praw wielkich liczb, ale wypowiedzianymi w szczególnej sytuacji związanej z ewolucją procesu stochastycznego i jego entropii. Ograniczymy się do standardowej wersji tego twierdzenia, udowodnionej przez Barrona (1985) i niezależnie Oreya (1985) (por. także Cover i Thomas, 2006, str. 644). Dla prostoty podamy wynik Barrona w szczególnym

przypadku.

Rozważmy stacjonarny i ergodyczny ciąg $X_1, X_2, ...$ wektorów losowych w $\mathbb{R}^k$ lub dyskretnych zmiennych losowych o wartościach w ustalonym zbiorze przeliczalnym $\mathcal{X}$, o gęstościach skończenie wymiarowych $p(x_1, ..., x_n)$ (względem miary Lebesgue'a lub miary liczącej). Wówczas przyrost entropii w $n+1$-szym kroku wynosi

$$D_n = E \log p(X_{n+1}|X_1, ..., X_n) = -H(X_{n+1}|X_1, ..., X_n).$$

Załóżmy, że dla pewnego $n_0 \geqslant 1$ jest $D_{n_0} > -\infty$. Wówczas ciąg D_n, $n \geqslant n_0$, jest niemalejący. Istotnie, ze stacjonarności i definicji 2.8 mamy

$$\begin{aligned} D_{n+1} - D_n &= -H(X_{n+2}|X_1, ..., X_{n+1}) + H(X_{n+1}|X_1, ..., X_n) \\ &= -H(X_{n+2}|X_1, ..., X_{n+1}) + H(X_{n+2}|X_2, ..., X_{n+1}) \\ &= I(X_{n+2}, X_1|X_2, ..., X_{n+1}) \geqslant 0. \end{aligned}$$

Zatem istnieje granica $D = \lim_{n\to\infty} D_n$ zwana tempem wzrostu entropii (ang. relative entropy rate lub entropy rate).

Twierdzenie 3.1 (Barron, 1985, Orey, 1985) *Przy powyższych założeniach*

$$\frac{1}{n} \log p(X_1, ..., X_n) \longrightarrow D \tag{3.1}$$

prawie na pewno (p.n.) i w L_1.

Twierdzenie powyższe jest uogólnieniem słynnego twierdzenia Shannona-McMillana-Breimana dla procesów dyskretnych na skończonym alfabecie. Shannon (1948) rozważał ciąg niezależnych zmiennych o jednakowym rozkładzie. McMillan (1953) udowodnił zbieżność w L_1, a Breiman (1957) zbieżność prawie na pewno dla ciągów ergodycznych. Znacznie ogólniejsze, niż podana przez nas, wersje tego twierdzenia można znaleźć w monografii Graya (1990). Zauważmy, że lewą stronę (3.1) możemy napisać w postaci średniej

$$\frac{1}{n} \sum_{i=2}^{n} \log p(X_i|X_1, ..., X_{i-1}) + \frac{1}{n} \log p(X_1),$$

która w przypadku jednorodnego łańcucha Markowa jest średnią (na ogół zależnych) zmiennych o jednakowym rozkładzie i $D = D_n = -H(X_2|X_1)$, o ile $H(X_2|X_1) < \infty$. W przypadku ciągu niezależnych zmiennych o jednakowym rozkładzie, gęstości p i $H(X_1) < \infty$, otrzymujemy mocne prawo wielkich liczb w postaci

$$\frac{1}{n}\sum_{i=1}^{n}\log p(X_i) \longrightarrow -H(X_1) = E\log p(X_1) \quad \text{p.n.}$$

W tym przypadku przyrosty entropii w każdym kroku są takie same i wynoszą $-H(X_1)$ czyli tempo wzrostu entropii także jest równe $-H(X_1)$.

Asymptotyczna zasada ekwipartycji ma prostą interpretację. Ciąg $(x_1, ..., x_n)$ nazywamy typowym, jeśli

$$p(x_1, ..., x_n) \in \left(e^{nD-n\varepsilon}, e^{nD+n\varepsilon}\right)$$

dla każdego $\varepsilon > 0$. Definicja oznacza więc, że dla ciągów typowych gęstość $p(x_1, ..., x_n)$ jest asymptotycznie stała. Natomiast twierdzenie mówi, że prawie na pewno dla dostatecznie dużych n ciąg $(X_1, ..., X_n)$ jest typowy. To spostrzeżenie wykorzystuje się do kompresji danych (por. Cover i Thomas, 2006, str. 78).

3.2. Rozkład empiryczny i zasada wielkich odchyleń. Niech $\mathcal{X} = \{x_1, ..., x_r\}$ zbiorem skończonym, a $X = (X_1, ..., X_n)$ będzie próbą prostą z rozkładu μ na $\mathcal{X}$ zadanego przez wektor prawdopodobieństw $p = (p_1, ..., p_r)$, gdzie $p_i = P(X_1 = x_i)$, $i = 1, ..., r$. Oznaczmy przez μ_n rozkład empiryczny próby X, a przez Υ_n zbiór wszystkich możliwych rozkładów empirycznych próby n-elementowej. Zatem $\nu \in \Upsilon_n$, gdy $n\nu(\{x_i\})$ jest liczbą całkowitą. Oczywiście, mamy $P(\mu_n \in \Upsilon_n) = 1$.

Lemat 3.1. *Dla dowolnego n i dowolnego zbioru Γ_n rozkładów na $\mathcal{X}$ prawdziwe są oszacowania*

$$P(\mu_n \in \Gamma_n) \geqslant c_1(r)n^{-r/2}\,2^{-nM}, \tag{3.2}$$

$$P(\mu_n \in \Gamma_n) \leqslant c_2(r)(n+r)^{r-1}\,2^{-nM}, \tag{3.3}$$

gdzie $M = \min\{D(\nu||\mu) : \nu \in \Gamma_n \cap \Upsilon_n\}$.

Dowód. Niech $\nu \in \Upsilon_n$ będzie dowolnym rozkładem empirycznym z wektorem prawdopodobieństw $q = (q_1, ..., q_r)$. Wtedy

$$\begin{aligned} P(\mu_n = \nu) &= N_\nu \textstyle\prod_{i=1}^{r} p_i^{nq_i} = N_\nu 2^{\sum nq_i \log p_i} \\ &= N_\nu 2^{\sum nq_i \log q_i - \sum nq_i \log(q_i/p_i)} = N_\nu 2^{-nH(\nu) - nD(\nu||\mu)}, \end{aligned}$$

gdzie N_ν jest liczbą realizacji próby X, dających ten sam rozkład empiryczny ν. Zatem $N_\nu = n!/((nq_1)!...(nq_r)!)$. Ze wzoru Stirlinga po prostych przekształceniach otrzymujemy

$$c_1(r)n^{-r/2}2^{nH(\nu)} \leqslant N_\nu \leqslant c_2(r)2^{nH(\nu)}.$$

Wstawiając oszacowanie do poprzedniego wyrażenia, dostajemy

$$c_1(r)n^{-r/2}2^{-nD(\nu||\mu)} \leqslant P(\mu_n = \nu) \leqslant c_2(r)2^{-nD(\nu||\mu)}$$

dla dowolnego $\nu \in \Upsilon_n$. Zatem

$$P(\mu_n \in \Gamma_n) = P(\mu_n \in \Gamma_n \cap \Upsilon_n) \geqslant c_1(r)\, n^{-r/2}\, 2^{-nM},$$

co daje (3.2). Z drugiej strony

$$P(\mu_n \in \Gamma_n) \leqslant c_2(r)(\overline{\overline{\Gamma_n \cap \Upsilon_n}})\, 2^{-nM},$$

gdzie $\overline{\overline{A}}$ oznacza moc zbioru skończonego A. Ponieważ $\overline{\overline{\Gamma_n \cap \Upsilon_n}} \leqslant \overline{\overline{\Upsilon_n}} = \begin{pmatrix} n+r-1 \\ r-1 \end{pmatrix} < (n+r)^{r-1}$, to (3.3) zachodzi. □

Bezpośrednim wnioskiem z lematu 3.1 jest następujące twierdzenie o zbieżności entropijnej rozkładów empirycznych do rozkładu μ.

TWIERDZENIE 3.2. *Przy powyższych oznaczeniach i założeniach*

$$\frac{n}{(\log n)(\log\log n)} D(\mu_n||\mu) \to 0 \quad \text{p.n.}$$

W szczególności $D(\mu_n||\mu) \to 0$ p.n.

Dowód. Dla dowolnego $\varepsilon > 0$ połóżmy w lemacie 3.1 $\Gamma_n = \Gamma = \{\nu : D(\nu||\mu) \geqslant \varepsilon\}$. Wtedy z (3.3) dostajemy

$$P(D(\mu_n||\mu) \geqslant \varepsilon) \leqslant c_2(r)(n+r)^{r-1}2^{-n\varepsilon}.$$

Stąd

$$\sum_{n=1}^{\infty} P\left(\tfrac{n}{(\log n)(\log\log n)} D(\mu_n||\mu) \geqslant \varepsilon\right) \leqslant \sum_{n=1}^{\infty} c_2(r)\tfrac{(n+r)^{r-1}}{n^{\varepsilon \log\log n}} < \infty.$$

Z dowolności ε i z lematu Borela-Cantellego wynika teza. □

Kolejnym wnioskiem z lematu 3.1 jest twierdzenie Sanowa dla rozkładów na zbiorze skończonym.

TWIERDZENIE 3.3 (twierdzenie Sanowa dla źródeł o skończonym alfabecie). *Jeśli* Γ *jest zbiorem rozkładów na* $\mathcal{X}$ *o niepustym wnętrzu i* $\Gamma \subset \mathrm{cl}(\mathrm{int}\,\Gamma)$*, to*

$$\frac{1}{n}\log P(\mu_n \in \Gamma) \longrightarrow -\inf_{\nu\in\Gamma} D(\nu||\mu). \tag{3.4}$$

Dowód. Logarytmując obustronnie (3.2) i (3.3), dostajemy

$$\tfrac{1}{n}\log P(\mu_n \in \Gamma) \geqslant -\min_{\nu\in\Gamma\cap\Upsilon_n} D(\nu||\mu) + \tfrac{\log c_1(r) - (r/2)\log n}{n}$$

oraz

$\frac{1}{n}\log P(\mu_n \in \Gamma) \leqslant -\min_{\nu\in\Gamma\cap\Upsilon_n} D(\nu||\mu) + \frac{\log c_2(r)+(r-1)\log n}{n}$.

Stąd natychmiast wynika (3.4). □

Twierdzenie Sanowa (1957) jest prawdziwe w znacznie ogólniejszej sytuacji. Przytoczymy tutaj standardową wersję tego twierdzenia.

TWIERDZENIE 3.4. *Niech $X_1, X_2, ...$ będzie ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie μ na przestrzeni polskiej $\mathcal{X}$ i niech μ_n będzie rozkładem empirycznym próby prostej $X = (X_1, ..., X_n)$. Wówczas dla dowolnego zbioru Γ rozkładów na $\mathcal{X}$ mamy*

$$\liminf_n \frac{1}{n}\log P(\mu_n \in \Gamma) \geqslant -\inf_{\nu\in\operatorname{int}\Gamma} D(\nu||\mu),$$

$$\limsup_n \frac{1}{n}\log P(\mu_n \in \Gamma) \leqslant -\inf_{\nu\in\operatorname{cl}\Gamma} D(\nu||\mu),$$

gdzie wnętrze i domknięcie Γ jest w topologii słabej zbieżności miar.

Twierdzenie Sanowa jest szczególnym przypadkiem zasady wielkich odchyleń (ang. large deviation principle) zastosowanej do rozkładów empirycznych. Wynika z twierdzenia Craméra o wielkich odchyleniach, gdyż rozkład empiryczny jest średnią niezależnych losowych miar punktowych o jednakowym rozkładzie $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. Przy dość słabych założeniach twierdzenie Craméra orzeka, że dla ciągu $X_1, X_2, ...$ niezależnych zmiennych losowych o jednakowym rozkładzie, o wartościach w lokalnie wypukłej przestrzeni liniowo-topologicznej $\mathcal{Y}$ i mających skończoną funkcję tworzącą momenty, zachodzą relacje

$$\liminf_n \frac{1}{n}\ln P\left(\frac{X_1+...+X_n}{n} \in A\right) \geqslant -\inf_{y\in\operatorname{int} A} \Lambda^*(y),$$

$$\limsup_n \frac{1}{n}\ln P\left(\frac{X_1+...+X_n}{n} \in A\right) \leqslant -\inf_{y\in\operatorname{cl} A} \Lambda^*(y)$$

dla dowolnego zbioru borelowskiego $A \subset \mathcal{Y}$, gdzie $\Lambda^*(y) = \sup_{y^*}(y^*(y) - \Lambda(y^*))$ jest transformatą Legendre'a-Fenchela logarytmu funkcji tworzącej momenty $\Lambda(y^*) = \ln\int e^{y^*}d\mu$, a y^* jest funkcjonałem liniowym ciągłym na $\mathcal{Y}$. Funkcja Λ^* nazywa się funkcją tempa wielkich odchyleń (ang. rate function). Precyzyjne sformułowanie twierdzenia Craméra można znaleźć np. w artykule Dembo i Zeitouni (2002), str. 399. My ograniczymy się tutaj do identyfikacji funkcji Λ^* w warunkach twierdzenia 3.4. Wtedy $\mathcal{Y}$ jest przestrzenią miar skończonych znakowanych na $\mathcal{X}$ z topologią słabej zbieżności, a $\Lambda(f) = \log E\, 2^{f(X_1)}$, gdzie f

jest funkcją ciągłą ograniczoną na $\mathcal{X}$. Jeśli μ ma gęstość p, to dla dowolnego rozkładu ν o gęstości q z nierówności Jensena wynika

$$\begin{aligned}\int f d\nu - \Lambda(f) &= \int f\, q\, d\lambda - \log \int 2^f \tfrac{p}{q} q d\lambda \\ &\leqslant \int f\, q\, d\lambda - \int (f - \log \tfrac{q}{p}) q\, d\lambda = D(\nu||\mu)\end{aligned}$$

co dowodzi nierówności $\Lambda^*(\nu) \leqslant D(\nu||\mu)$. Z drugiej strony wartość $D(\nu||\mu)$ jest osiągana na ciągu funkcji ciągłych i ograniczonych, zbieżnym do funkcji $\log \frac{q}{p}$. To pokazuje, że istotnie $D(\nu||\mu)$ jest funkcją tempa wielkich odchyleń w przypadku miar empirycznych.

3.3. Łańcuchy Markowa. W tym rozdziale przedstawimy klasyczne twierdzenie Rényiego (1961) o zbieżności entropijnej rozkładów łańcucha Markowa do rozkładu ergodycznego. Niech więc P będzie macierzą przejścia jednorodnego łańcucha Markowa o skończonym zbiorze stanów $\mathcal{X} = \{x_1, ..., x_r\}$, p rozkładem początkowym łańcucha, a p^* rozkładem stacjonarnym tzn. spełniającym równanie $p^*P = p^*$, gdzie p i p^* oznaczają macierze wierszowe, a p^*P jest iloczynem macierzy.

Twierdzenie 3.6 (Rényi, 1961). *Przy powyższych oznaczeniach i założeniach mamy*

$$D(pP^n||p^*) \longrightarrow 0. \tag{3.5}$$

Dowód. Podstawowym krokiem dowodu jest wykazanie monotoniczności ciągu odległości entropijnych tzn. relacji $D(pP||p^*) \leqslant D(p||p^*)$, w której równość zachodzi wtedy i tylko wtedy, gdy $p = p^*$. Istotnie, oznaczmy $q_j = \sum_i p_i p_{ij}$. Naśladując dowód własności (A1) i korzystając ze stacjonarności p^* tzn. relacji $\sum_i p_i^* p_{ij} = p_j^*$ dla wszystkich j oraz z relacji $\sum_j q_j = 1$, mamy

$$\begin{aligned}D(p||p^*) - D(pP||p^*) &= \textstyle\sum_{ij} p_i p_{ij} \log \frac{p_i}{p_i^*} - \sum_{ij} p_i p_{ij} \log \frac{q_j}{p_j^*} \\ &= \textstyle\sum_{ij} \frac{p_i^* p_{ij} q_j}{p_j^*} \frac{p_i p_j^*}{p_i^* q_j} \log \frac{p_i p_j^*}{p_i^* q_j} \geqslant \frac{1}{\ln 2} \sum_{ij} \frac{p_i^* p_{ij} q_j}{p_j^*} (\frac{p_i p_j^*}{p_i^* q_j} - 1) = 0,\end{aligned}$$

przy czym równość zachodzi wtedy i tylko wtedy, gdy $p_i p_j^* = p_i^* q_j$ dla wszystkich i, j. Sumując tę równość obustronnie względem j dostajemy $p_i = p_i^*$ dla wszystkich i.

Z monotoniczności wynika, że ciąg $D(pP^n||p^*)$ jest zbieżny do pewnej granicy nieujemnej D_p. Ponieważ zbiór stanów jest skończony, to istnieje podciąg potęg macierzy przejścia $P^{n'}$ zbieżny do pewnej macierzy P_0, co oznacza, że $D_p = D(pP_0||p^*)$. Korzystając jeszcze raz z

monotoniczności, dostajemy $D(pP^{n'+1}||p^*) \longrightarrow D(pP_0P||p^*) = D_p = D(pP_0||p^*)$. Ale równość ma miejsce tylko wtedy, gdy $pP_0 = p^*$, co daje $D_p = 0$ i kończy dowód (3.5). □

Powyższe elemntarne twierdzenie zostało rozszerzone na przypadek przeliczalnej liczby stanów i czasu ciągłego przez Kendalla (1963). Zauważmy jeszcze, że pP^n jest rozkładem łańcucha w chwili n. Twierdzenie mówi więc, że rozkład łańcucha w chwili n dąży monotonicznie w sensie odległości entropijnej do rozkładu stacjonarnego (ergodycznego). Można potraktować ten fakt jako odpowiednik drugiej zasady termodynamiki dla przypadku łańcuchów Markowa. Jest to bardziej widoczne, gdy macierz przejścia P jest podwójnie stochastyczna tzn. p^* jest jednostajny i wtedy $D(pP^n||p^*) = H(p^*) - H(pP^n) \searrow 0$.

3.4. Centralne twierdzenie graniczne. Niech $X_1, X_2, ...$ będzie ciągiem niezależnych rzeczywistych zmiennych losowych o tym samym rozkładzie, wartości oczekiwanej 0 i wariancji 1. Oznaczmy przez $S_n = (X_1 + ... + X_n)/\sqrt{n}$ ciąg unormowanych sum częściowych. Jak wspomnieliśmy w rozdziale 2.8, Barron i Madiman (2007) pokazali monotoniczność ciągu $H(S_n)$, co odpowiada monotoniczności odległości entropijnej rozkładów łańcucha Markowa od rozkładu stacjonarnego. Można więc spodziewać się podobnego twierdzenia, mówiącego o zbieżności monotonicznej entropii $H(S_n)$ do maksymalnej entropii (czyli entropii rozkładu normalnego). Istotnie, zbieżność odległości entropijnej jest prostym wnioskiem z nierówności dla informacji Fishera i jest treścią dwóch kolejnych twierdzeń.

Twierdzenie 3.7 (zbieżność informacji Fishera). *Jeśli* $J(S_{n_0}) < \infty$ *dla pewnego* $n_0 \geqslant 1$, *to*

$$J(S_n) \longrightarrow J \geqslant 1. \tag{3.6}$$

Dowód. Niech $J_0 = \inf_n J(S_n)$. Dla dowolnego ustalonego $\varepsilon > 0$ wybierzmy liczbę $k_\varepsilon = k > n_0$ tak, aby $J(S_k) < J_0 + \varepsilon$. Wówczas dla $n = ik+m$, gdzie m jest resztą z dzielenia n przez k, sumę S_n możemy podzielić na bloki długości k i otrzymujemy

$$S_n = \sqrt{\tfrac{k}{n}}\tfrac{X_1+...+X_k}{\sqrt{k}} + \sqrt{\tfrac{k}{n}}\tfrac{X_{k+1}+...+X_{2k}}{\sqrt{k}} + ...+$$
$$+\sqrt{\tfrac{k}{n}}\tfrac{X_{(i-2)k+1}+...+X_{(i-1)k}}{\sqrt{k}} + \sqrt{\tfrac{k+m}{n}}\tfrac{X_{(i-1)k+1}+...+X_n}{\sqrt{k+m}}.$$

Z nierówności dla informacji Fishera dostajemy

$$J(S_n) \leqslant \tfrac{k}{n} J(S_k)\,(i-1) + \tfrac{k+m}{n} J(S_{k+m}) \leqslant J_0 + \varepsilon + \tfrac{k+m}{n} J(S_{k+m}),$$

co pociąga $\limsup_{n\to\infty} J(S_n) \leqslant J_0 + \varepsilon$ i z dowolności ε tezę (3.6) z $J = J_0$. □

Natychmiastowym wnioskiem z twierdzenia 3.7 jest zbieżność entropijna sum S_n.

TWIERDZENIE 3.8 (zbieżność entropijna S_n). *Jeśli $D(S_{n_0}||Z) < \infty$ dla pewnego $n_0 \geqslant 1$, gdzie Z jest zmienną losową o rozkładzie standardowym normalnym, to*

$$D(S_n||Z) \longrightarrow D \geqslant 0. \tag{3.7}$$

Dowód. Z tożsamości de Bruijna mamy dla $n \geqslant n_0$

$$\begin{aligned} D(S_n||Z) &= \int_0^\infty \left[J(S_n + \sqrt{t}Z) - \tfrac{1}{1+t}\right] dt \\ &= \int_0^\infty \left[J\left(\tfrac{(X_1+\sqrt{t}Z_1)+\ldots+(X_n+\sqrt{t}Z_n)}{\sqrt{n}\sqrt{1+t}}\right) - 1\right] \tfrac{dt}{1+t}, \end{aligned} \tag{3.8}$$

gdzie $Z_1, ..., Z_n$ są niezależnymi kopiami Z. Z twierdzenia 3.7 wynika, że funkcja podcałkowa w (3.8) jest zbieżna dla każdego t. Z drugiej strony nierówność (2.28) daje $J(S_n + \sqrt{t}Z) \leqslant J(\sqrt{t}Z) = 1/t$ co oznacza, że w przedziale $[1, \infty)$ funkcja podcałkowa jest majoryzowana przez funkcję całkowalną $1/t - 1/(t+1)$. Podobnie, nierówności (2.27) i (2.28) pozwalają oszacować z góry funkcję podcałkową w (3.8) przez $J(X_1 + \sqrt{t}Z_1) - \frac{1}{1+t} \leqslant J(X_1) - \frac{1}{1+t}$, która jest całkowalną na $[0, 1]$. Zatem twierdzenie Lebesgue'a o zbieżności ograniczonej zastosowane do całki (3.8) daje (3.7). □

Pozostaje jednak trudny problem wykazania, że w istocie mamy zbieżność do rozkładu standardowego normalnego tzn. w (3.7) jest $D = 0$, a w (3.6) jest $J = 1$. Poniższe twierdzenia podają pozytywne rozwiązanie tego problemu.

TWIERDZENIE 3.9 (Barron, 1986). *Przy założeniach twierdzenia 3.8 mamy*

$$J(S_n + \sqrt{t}Z) \longrightarrow \frac{1}{1+t}$$

dla każdego $t > 0$ i w konsekwencji z tożsamości de Bruijna

$$D(S_n||Z) = H(Z) - H(S_n) \longrightarrow 0. \tag{3.9}$$

Zbieżność (3.9) oznacza, że entropia S_n osiąga w granicy maksymalną

wartość $H(Z)$. Stanowi więc kolejny odpowiednik drugiej zasady termodynamiki.

TWIERDZENIE 3.10 (Barron i Johnson, 2004). *Przy założeniach twierdzenia 3.7 mamy*

$$J(S_n) \longrightarrow 1 \tag{3.10}$$

czyli równoważnie $J(S_n||Z) = J(S_n) - J(Z) \to 0$.

Barron i Johnson (2004) otrzymali nie tylko zbieżność informacji Fishera (3.10), ale, przy dość silnych założeniach wykazali, że szybkość zbieżności w (3.10) jest rzędu $O(1/n)$. Bobkov i in. (2013) przebadali szczegółowo szybkość zbieżności entropijnej w (3.9). W najprostszej postaci ich wynik jest następujący.

TWIERDZENIE 3.11 (Bobkov i in., 2013). *Jeśli* $D(S_{n_0}||Z) < \infty$ *dla pewnego* $n_0 \geqslant 1$ *oraz* $EX_1^4 < \infty$, *to*

$$D(S_n||Z) = \frac{1}{12\,n}(EX_1^3)^2 + o\left(\frac{1}{n\log n}\right).$$

Johnson (2004) rozszerzył twierdzenie 3.10 m.in. na przypadek zmiennych losowych o niejednakowych rozkładach, zmiennych losowych wielowymiarowych oraz ciągów mieszających.

4. Wnioskowanie statystyczne

Ten główny rozdział niniejszego opracowania poświęcimy wybranym, ale, w odczuciu autora, ważnym, zagadnieniom statystyki matematycznej, które wiążą się z pojęciami i twierdzeniami teorii informacji oraz które można badać metodami teorii informacji.

Niech $X = (X_1, ..., X_n) \in \mathcal{X}^n$ będzie próbą (niekoniecznie prostą) o wartościach w n-krotnym produkcie przestrzeni miarowej σ-skończonej $(\mathcal{X}, \mathcal{B}, \lambda)$ i niech

$$\mathcal{P} = \{p_{n\vartheta} : \vartheta \in \Theta\} \tag{4.1}$$

będzie pewną rodziną gęstości rozkładów na $\mathcal{X}^n$, gdzie Θ jest dowolnym zbiorem parametrów. Rozkład próby X może być elementem $\mathcal{P}$ (model prawdziwy) lub nie.

4.1. Estymatory największej wiarogodności. Estymatorem największej wiarogodności parametru ϑ w rodzinie $\mathcal{P}$ danej przez (4.1)

opartym na próbie X nazywamy statystykę $\widehat{\vartheta}$ określoną przez relację

$$\widehat{\vartheta} = \widehat{\vartheta}(X) = \text{argmax}_{\vartheta\in\Theta}\, p_{n\vartheta}(X)$$

(o ile maksimum jest osiągane na Θ).

Wydaje się intuicyjnie oczywiste, że estymator największej wiarogodności powinien odpowiadać najlepszemu wyborowi ϑ, w sensie ilości informacji, czy też wyborowi rozkładu z $\mathcal{P}$ najbliższego, w sensie odległości entropijnej, od prawdziwego rozkładu próby. Dla potwierdzenia tych intuicji przytoczymy dwa rezultaty.

Niech $\mathcal{X} = \{x_1, ..., x_r\}$ będzie skończonym alfabetem, $X = (X_1, ..., X_n)$ próbą prostą z pewnego rozkładu μ na $\mathcal{X}$, μ_n rozkładem empirycznym tej próby (będącym także rozkładem na $\mathcal{X}$), a

$$\mathcal{F} = \{p_\vartheta : p_\vartheta = (p_{\vartheta 1}, ..., p_{\vartheta r}),\ \vartheta \in \Theta \subset \mathbb{R}^k\}$$

rodziną parametryczną rozkładów (pojedynczych obserwacji) na $\mathcal{X}$. Okazuje się, że estymator największej wiarogodności $\widehat{\vartheta}$ odpowiada odwrotnemu rzutowi informacyjnemu (por. definicja 2.5) rozkładu empirycznego na rodzinę $\mathcal{F}$.

TWIERDZENIE 4.1. *Dla skończonego alfabetu $\mathcal{X}$ i przy powyższych oznaczeniach estymator największej wiarogodności $\widehat{\vartheta} \in \Theta$ istnieje wtedy i tylko wtedy, gdy istnieje odwrotny rzut informacyjny μ_n na $\mathcal{F}$. Ponadto*

$$p_{\widehat{\vartheta}} = RIproj_{\mathcal{F}}\mu_n.$$

Dowód. Mamy

$$\begin{aligned}\text{argmax}_\vartheta \log \textstyle\prod_{i=1}^n p_{\vartheta X_i} &= \text{argmin}_\vartheta \textstyle\sum_{i=1}^n \log \frac{1}{p_{\vartheta X_i}} \\ &= \text{argmin}_\vartheta \textstyle\sum_{j=1}^r n\mu_n(\{x_j\}) \log \frac{1}{p_{\vartheta j}} \\ &= \text{argmin}_\vartheta \textstyle\sum_{j=1}^r \mu_n(\{x_j\}) \log \frac{\mu_n(\{x_j\})}{p_{\vartheta j}} = \text{argmin}_\vartheta D(\mu_n||p_\vartheta),\end{aligned}$$

co kończy dowód. □

Csiszár podał inny związek estymatora największej wiarogodności z rzutem informacyjnym. Niech μ będzie pewnym rozkładem na dowolnej (niekoniecznie skończonej) przestrzeni $(\mathcal{X}, \mathcal{B}, \lambda)$, $T = (T_1, ..., T_k)$ wektorem liniowo niezależnych funkcji mierzalnych i $a \in \mathbb{R}^k$ ustalonym wektorem. Rozważmy rodzinę rozkładów na $\mathcal{X}$, dla których wartość oczekiwana T wynosi a. Dokładniej

$$\mathcal{L}_a = \{\nu :\ \nu \prec\prec \mu,\ \textstyle\int T\, d\nu = \int T \frac{d\nu}{d\mu}\, d\mu = a\}.$$

Oczywiście, $\mathcal{L}_a$ jest wypukła i domknięta w normie całkowitego wahania (por. twierdzenie 2.11). Poniższe twierdzenie odpowiada twierdzeniu 3.1 w pracy Csiszára (1975).

TWIERDZENIE 4.2. *Jeśli istnieje $\vartheta \in \mathbb{R}^k$ takie,że funkcja $c_\vartheta e^{\vartheta \circ T}$ jest gęstością (względem μ) rozkładu wykładniczego γ_ϑ z odpowiednio dobraną stałą normującą c_ϑ oraz $\gamma_\vartheta \in \mathcal{L}_a$ tzn.*

$$\int T d\gamma_\vartheta = a,$$

to

$$\gamma_\vartheta = Iproj_{\mathcal{L}_a} \mu.$$

Ponadto zachodzi odpowiednik twierdzenia Pitagorasa

$$D(\nu||\mu) = D(\nu||\gamma_\vartheta) + D(\gamma_\vartheta||\mu)$$

dla wszystkich $\nu \in \mathcal{L}_a$ tzn. $\mathcal{L}_a$ jest zawarta w 'hiperpłaszczyźnie stycznej' do D-kuli o środku μ i 'promieniu' $\sqrt{D(\gamma_\vartheta||\mu)}$ w punkcie γ_ϑ.

Zastosowanie powyższego twierdzenia w szczególnej sytuacji rodzin wykładniczych pozwala na pewną interpretację estymatorów największej wiarogodności. Niech więc μ i T będą jak wyżej, μ^n n-krotnym produktem μ, a

$$\mathcal{P} = \{\gamma_\vartheta^n : \ p_{n\vartheta} = \frac{d\gamma_\vartheta^n}{d\mu^n} = c_\vartheta^n \, e^{n\vartheta \circ \overline{T}}, \ \vartheta \in \Theta \subset \mathbb{R}^k\}$$

rodziną wykładniczą rozkładów produktowych na $\mathcal{X}^n$, gdzie $\overline{T}(x) = \frac{1}{n}\sum_{i=1}^n T(x_i)$. Niech $X = (X_1, ..., X_n)$ będzie próbą prostą z rozkładu μ, a $\overline{T}(X)$ średnią z tej próby. Rozważmy rodzinę rozkładów

$$\mathcal{L}_{\overline{T}(X)} = \{\nu : \nu \prec\prec \mu^n, \ \int \overline{T} \frac{d\nu}{d\mu^n} d\mu^n = \overline{T}(X)\}.$$

Zauważmy, że na ogół μ^n nie należy do $\mathcal{L}_{\overline{T}(X)}$, gdyż średnia teoretyczna $ET = \int \overline{T} d\mu^n$ nie jest na ogół równa średniej empirycznej. Oznaczmy przez $\widehat{\vartheta} \in \Theta$ estymator największej wiarogodności w rodzinie $\mathcal{P}$ oparty na próbie X. Wtedy

$$\gamma_{\widehat{\vartheta}}^n \in \mathcal{L}_{\overline{T}(X)}. \tag{4.2}$$

Istotnie, z własności rodziny wykładniczej dla dowolnego $\vartheta \in \Theta$ mamy

$$0 = \nabla \int p_{n\vartheta} d\mu^n = \int \nabla p_{n\vartheta} d\mu^n = \int (n\overline{T} + n\nabla \ln c_\vartheta) p_{n\vartheta} d\mu^n,$$

gdzie ∇ oznacza różniczkowanie względem $\vartheta \in \Theta$. Stąd $\int \overline{T} p_{n\widehat{\vartheta}} d\mu^n =$

$-\nabla \ln c_\vartheta|_{\vartheta=\widehat{\vartheta}}$. Z drugiej strony, z równania wiarogodności wynika $\nabla \ln c_\vartheta|_{\vartheta=\widehat{\vartheta}} = -\overline{T}(X)$, co uzasadnia stwierdzenie (4.2). Z niego oraz z twierdzenia 4.2 wynika, że $\gamma^n_{\widehat{\vartheta}} = Iproj_{\mathcal{L}_{\overline{T}(X)}}\mu^n$. Oznacza to, że rozkład wykładniczy $\gamma^n_{\widehat{\vartheta}}$, gdzie $\widehat{\vartheta}$ jest estymatorem największej wiarogodności, jest najbliższym μ^n (w sensie odległości entropijnej) rozkładem w $\mathcal{L}_{\overline{T}(X)}$.

4.2. Model statystyczny. Podstawowym obiektem rozważań w statystyce matematycznej jest model statystyczny.

Rozważmy rodzinę gęstości $\mathcal{P}$ jak w (4.1) i model statystyczny $(\mathcal{X}^n, \mathcal{B}^n, \lambda^n, \mathcal{P})$ i załóżmy, że rozkład próby X należy do $\mathcal{P}$. Aby mierzyć ilość informacji o parametrze ϑ, potrzebna jest struktura miarowa na zbiorze Θ. Odpowiada to podejściu bayesowskiemu w statystyce matematycznej. Niech więc $(\Theta, \mathcal{S}, \sigma)$ będzie przestrzenią miarową σ-skończoną. Parametr ϑ będziemy traktować jako wartość zmiennej losowej θ o gęstości a priori $w(\vartheta)$ (względem σ). Wtedy $p_{n\vartheta}$ jest gęstością warunkową próby X przy warunku $\theta = \vartheta$, a iloczyn $p_{n\vartheta}\, w(\vartheta)$ gęstością łączną zmiennej losowej (θ, X). W konsekwencji gęstość (brzegowa) próby X ma postać $f_n = \int_\Theta p_{n\vartheta}\, w(\vartheta) d\sigma$, a gęstość a posteriori zmiennej θ wyraża się wzorem $p_{n\vartheta}\, w(\vartheta)/f_n$.

Ten ogólny schemat i oznaczenia będziemy przyjmować w dalszym ciągu rozdziału 4.

4.3. Statystyki dostateczne. Niech $\mathcal{P}$ będzie rodziną gęstości próby X jak w (4.1), a $T(X)$ statystyką. Wtedy $\theta, X, T(X)$ tworzą łańcuch Markowa (por. rozdział 2.9) tzn. $\theta \to X \to T(X)$. Zatem z nierówności dla danych przetworzonych mamy $I(\theta, T(X)) \leqslant I(\theta, X)$ tzn. żadna statystyka $T(X)$ nie zawiera więcej informacji o parametrze θ niż próba X. Ale może zachodzić równość.

Definicja 4.1. *Statystyka $T(X)$ nazywa się dostateczna dla parametru θ, jeśli*

$$I(\theta, T(X)) = I(\theta, X) \tag{4.3}$$

tzn. $\theta \to T(X) \to X$.

Inaczej mówiąc statystyka dostateczna stanowi kompresję danych bez straty informacji o parametrze θ. Równość (4.3) można wyrazić równoważnie w postaci $H(\theta|X) = H(\theta|T(X))$. Powyższa definicja dostateczności statystyki (wypowiedziana w języku teorii informacji) jest równoważna z definicją klasyczną. Istotnie, z (iii) w twierdzeniu 2.9 mamy $I(\theta, X) = D(p_{n\vartheta}\, w || f_n\, w)$. Stosując twierdzenie 2.10 do

odwzorowania $(\theta, X) \to (\theta, T(X))$, równość (4.3) zachodzi wtedy i tylko wtedy, gdy

$$\frac{p_{n\vartheta}(x)w(\vartheta)}{f_n(x)w(\vartheta)} = \frac{q_{n\vartheta}(T(x))w(\vartheta)}{g_n(T(x))w(\vartheta)} \quad \sigma \times \lambda\text{−p.w.},$$

gdzie $q_{n\vartheta}w$ i $g_n\, w$ są gęstościami rozkładów indukowanych przez powyższe odwzorowanie. To oznacza, że dla prawie wszystkich ϑ

$$p_{n\vartheta}(x) = \frac{q_{n\vartheta}(T(x))}{g_n(T(x))} f_n(x) \quad \lambda\text{−p.w.}$$

Z kryterium faktoryzacji Neymana (1935) (por. Bartoszewicz, 1981, wniosek 3.5) wynika dostateczność statystyki T w sensie Fishera.

Idąc dalej w definicji 4.1 możemy powiedzieć, że statystyka $T^*(X)$ jest minimalną statystyką dostateczną, gdy jest funkcją każdej innej statystyki dostatecznej tzn. dla dowolnej statystyki dostatecznej $T(X)$ istnieje funkcja mierzalna ψ_T taka, że $T^*(X) = \psi_T(T(X))$. Oznacza to, że dla dowolnej statystyki dostatecznej mamy $\theta \to T^*(X) \to T(X) \to X$. Zatem minimalna statystyka dostateczna stanowi maksymalną kompresję danych bez straty informacji o parametrze θ.

4.4. Rozwinięcie Barrona-Clarke'a. W tym rozdziale przedstawimy rozwinięcie podane przez Clarke'a i Barrona (1990). Najpierw sformułujemy odpowiednie założenia regularności rodziny gęstości $\mathcal{P}$ jak w (4.1).

Niech $X = (X_1, ..., X_n)$ będzie próbą prostą, gdzie X_i przyjmują wartości w przestrzeni polskiej $\mathcal{X}$. Ponadto, załóżmy, że spełnione są następujące warunki.

(BC1) $\Theta \subset \mathbb{R}^k$ *ma niepuste wnętrze;*

(BC2) *gęstości p_ϑ pojedynczych obserwacji X_i mają pochodne cząstkowe rzędu 2 w pewnym otoczeniu U_{ϑ_0} punktu $\vartheta_0 \in \operatorname{int}\Theta$ (na wspólnym zbiorze $\mathcal{X}_0 \subset \mathcal{X}$ miary pełnej) i są ciągłe w punkcie ϑ_0 oraz*

$$\sum_{r,s=1}^{k} \int_{\mathcal{X}} \sup_{\vartheta \in U_{\vartheta_0}} \left(\frac{\partial^2 \ln p_\vartheta(x)}{\partial\vartheta_r \partial\vartheta_s} \right)^2 p_{\vartheta_0}(x)\lambda(dx) < \infty.$$

Macierz informacji Fishera $\mathbf{J}(\vartheta_0)$ jest nieosobliwa i równa macierzy $-\int \frac{\partial^2 \ln p_{\vartheta_0}}{\partial\vartheta\partial\vartheta^T} p_{\vartheta_0} d\lambda$;

(BC3) *$\vartheta \to \vartheta_0$ wtedy i tylko wtedy, gdy P_ϑ jest słabo zbieżny do P_{ϑ_0}, gdzie P_ϑ oznacza rozkład o gęstości p_ϑ;*

(BC4) *gęstość a priori $w(\vartheta)$ względem miary Lebesgue'a na Θ jest ciągła i dodatnia w punkcie ϑ_0.*

Założenie (BC2) gwarantuje istnienie macierzy informacji Fishera, a nawet skończoność całek $\int \sup_{\vartheta \in U_{\vartheta_0}} (\ell_{\vartheta r})^2 p_{\vartheta_0} d\lambda$, $r = 1, ..., k,$, gdzie $\ell_{\vartheta r} = \frac{\partial \ln p_\vartheta}{\partial \vartheta_r}$ są składowymi wektora wynikowego ℓ_ϑ. Zauważmy, że rodzina wykładnicza rozkładów wyznaczona przez wektor $T = (T_1, ..., T_k)$ funkcji mierzalnych, liniowo niezależnych spełnia założenia (BC2) i (BC3).

TWIERDZENIE 4.3 (Clarke i Barron, 1990). *Przy założeniach (BC1) – (BC4) mamy następujące rozwinięcie*

$$\ln \frac{p_{n\vartheta_0}(X)}{f_n(X)} = -\frac{1}{2} N_{k\vartheta_0}(X) + \frac{k}{2} \ln \frac{n}{2\pi} + \frac{1}{2} \ln \frac{\det \mathbf{J}(\vartheta_0)}{w^2(\vartheta_0)} + o_{L_1}(1), \quad (4.4)$$

gdzie

$$N_{k\vartheta_0}(X) = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell_{\vartheta_0}(X_i) \right)^T \mathbf{J}^{-1}(\vartheta_0) \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell_{\vartheta_0}(X_i) \right)$$

jest statystyką wynikową Neymana-Rao.

Rozwinięcie (4.4) ma dwa natychmiastowe wnioski. Obliczając wartość oczekiwaną obu stron (4.4) względem rozkładu $P^n_{\vartheta_0}$, dostajemy następujące rozwinięcie.

WNIOSEK 4.1 (Clarke i Barron, 1990). *Przy założeniach twierdzenia 4.3 mamy*

$$D(p_{n\vartheta_0} || f_n) = \frac{k}{2} \log \frac{n}{2\pi\, e} + \frac{1}{2} \log \frac{\det \mathbf{J}(\vartheta_0)}{w^2(\vartheta_0)} + o(1). \quad (4.5)$$

Z kolei odejmując (4.4) od (4.5) i korzystając z asymptotycznej normalności wektora losowego $(\sum_{i=1}^n \ell_{\vartheta_0}(X_i))/\sqrt{n}$, otrzymujemy drugi wniosek.

WNIOSEK 4.2 (Clarke i Barron, 1990). *Przy założeniach twierdzenia 4.3 mamy*

$$\ln \frac{f_n(X)}{p_{n\vartheta_0}(X)} + (\ln 2)\, D(p_{n\vartheta_0} || f_n) + \frac{k}{2} = \frac{1}{2} N_{k\vartheta_0}(X) + o_{L_1}(1) \xrightarrow{\mathcal{D}} \frac{1}{2} \chi^2_k, \quad (4.6)$$

gdzie χ^2_k jest zmienną losową o rozkładzie chi-kwadrat z k stopniami swobody.

Rozwinięcie Barrona-Clarke'a ma rozmaite zastosowania w staty-

styce matematycznej. Niektóre pokażemy w następnych paragrafach.

4.5. Kodowanie źródła, ilość informacji o θ zawartej w próbie. Problem kodowania źródła (ang. source coding) jest ważne zarówno w technice przesyłania informacji, jak i dla wnioskowania statystycznego.

Niech $X = (X_1, ..., X_n)$ będzie próbą (źródłem), niekoniecznie prostą, o gęstości $p_{n\vartheta} \in \mathcal{P}$, przy czym ϑ jest nieznane, a q_n dowolną gęstością używaną jako przybliżenie nieznanej gęstości próby X. Wówczas wydłużenie średniej długości słowa kodowego optymalnego kodu binarnego opartego na q_n wynosi $D(p_{n\vartheta}||q_n)$. Wielkość

$$R_n = \inf_{q_n} \int_\Theta D(p_{n\vartheta}||q_n)\, w(\vartheta)d\sigma, \tag{4.7}$$

gdzie q_n przebiega wszystkie gęstości próby, nazywamy minimalnym wydłużeniem bayesowskim, a wielkość

$$R_n^+ = \inf_{q_n} \sup_{\vartheta} D(p_{n\vartheta}||q_n)$$

wydłużeniem minimaksowym tzn. minimalnym przy najmniej korzystnym $p_{n\vartheta}$. Łatwo zauważyć następujący fakt.

TWIERDZENIE 4.4. *Minimalne wydłużenie bayesowskie R_n realizuje kod bayesowski oparty na $f_n(x) = \int p_{n\vartheta}\, w(\vartheta)d\sigma$ czyli o słowach długości $\lceil \log(1/f_n(x)) \rceil$. Zatem*

$$R_n = \int D(p_{n\vartheta}||f_n)\, w(\vartheta)d\sigma.$$

Dowód. Dla dowolnego q_n mamy

$$\begin{aligned}\int_\Theta [D(p_{n\vartheta}||q_n) - D(p_{n\vartheta}||f_n)]w(\vartheta)d\sigma &= \int_\Theta \int_{\mathcal{X}^n} p_{n\vartheta} \log \tfrac{f_n}{q_n} w(\vartheta)d\sigma\, d\lambda^n \\ &= \int_{\mathcal{X}^n} \left(\int_\Theta p_{n\vartheta} w(\vartheta)d\sigma\right) \log \tfrac{f_n}{q_n}\, d\lambda^n = D(f_n||q_n) \geqslant 0\end{aligned}$$

co kończy dowód. □

Definicja (4.7) sugeruje określenie przepustowości (ang. capacity) kanału informacyjnego wzorem

$$C(\Theta, X) = \sup_w R_n, \tag{4.8}$$

gdzie supremum jest względem wszystkich gęstości a priori na przestrzeni parametrów Θ. Bezpośrednio z definicji (4.8) wynika relacja $R_n^+ \geqslant C(\Theta, X)$. Istotnie, korzystając z faktu, że $w(\vartheta)$ jest gęstością rozkładu, mamy

$$R_n^+ = \inf_{q_n} \sup_\vartheta D(p_{n\vartheta}||q_n) = \sup_w \inf_{q_n} [\sup_\vartheta D(p_{n\vartheta}||q_n)] \int w(\vartheta)d\sigma$$
$$\geqslant \sup_w \inf_{q_n} \int D(p_{n\vartheta}||q_n)w(\vartheta)d\sigma = \sup_w R_n = C(\Theta, X).$$

Haussler (1997) wykazał, że ma miejsce równość tj. $R_n^+ = C(\Theta, X)$.

Popatrzmy teraz na postawione zagadnienie estymacji ϑ nieco inaczej i zapytajmy ile informacji o θ jest zawartej w próbie X. Zgodnie z definicją (2.6) jest ona wyrażona przez $I(\theta, X) = H(\theta) - H(\theta|X)$. Korzystając z przyjętych oznaczeń rozkładu łącznego zmiennych θ i X, mamy stąd

$$\begin{aligned} I(\theta, X) &= -\int_\Theta w \log w d\sigma + \int_{\mathcal{X}^n} \int_\Theta p_{n\vartheta}\, w \log \tfrac{p_{n\vartheta}\, w}{f_n} d\sigma d\lambda^n \\ &= \int_{\mathcal{X}^n} \int_\Theta f_n \tfrac{p_{n\vartheta}\, w}{f_n} \log \tfrac{p_{n\vartheta}\, w}{f_n\, w} d\sigma d\lambda^n = \int_{\mathcal{X}^n} D(\tfrac{p_{n\vartheta}\, w}{f_n}||w) f_n d\lambda^n \\ &= \int_\Theta D(p_{n\vartheta}||f_n) w(\vartheta) d\sigma = R_n. \end{aligned}$$

Z powyższego rachunku wynika, że ilość informacji o θ zawartej w próbie X jest równa średniej (względem rozkładu próby) odległości entropijnej rozkładu a posteriori od rozkładu a priori i równocześnie średniej (względem rozkładu a priori) odległości entropijnej rozkładu warunkowego próby X od rozkładu bezwarunkowego, a ponadto z twierdzenia 4.4 równa minimalnemu wydłużeniu bayesowskiemu. Możemy zatem napisać

$$C(\Theta, X) = \sup_w R_n = \sup_w I(\theta, X). \tag{4.9}$$

Równość (4.9) pozwala na podanie następującej interpretacji przepustowości kanału informacyjnego. Przypuśćmy, że 'przesyłana jest wiadomość' X o gęstości $p_{n\vartheta}$, przy czym ϑ jest nieznane odbiorcy, ale model $\mathcal{P}$ jest znany. Odbiorca (statystyk) chce na podstawie przesłanej informacji zidentyfikować parametr ϑ. Przepustowość jest zatem maksymalną ilością informacji o θ uzyskaną przez odbiorcę w kanale (modelu statystycznym) $\mathcal{P}$.

Przy dodatkowych założeniach można powiedzieć więcej. Przyjmijmy założenia rozwinięcia Barrona i Clarke'a. Wtedy przez obustronne scałkowanie (4.5) względem ϑ dostajemy następujące twierdzenie.

TWIERDZENIE 4.5 (Clarke i Barron, 1994). *Jeśli założenia twierdzenia 4.3 są spełnione dla każdego ϑ (w szczególności X jest próbą prostą), $\Theta \subset \mathbb{R}^k$ jest zbiorem zwartym, $\int |\ell_{\vartheta r}|^{2+\varepsilon} p_\vartheta d\lambda < \infty$ dla $r = 1, ..., k$ i pewnego $\varepsilon > 0$ i są ciągłe względem ϑ oraz $p_\vartheta \neq p_{\vartheta'}$ na zbiorze miary dodatniej, gdy $\vartheta \neq \vartheta'$, to dla dowolnego rozkładu a priori o gęstości $w(\vartheta)$ mamy*

$$I(\theta, X) = R_n = \int_\Theta D(p_{n\vartheta}||f_n)w(\vartheta)d\vartheta$$
$$= \frac{k}{2}\log\frac{n}{2\pi\, e} + \int_\Theta w(\vartheta)\log\frac{\sqrt{\det \mathbf{J}(\vartheta)}}{w(\vartheta)}d\vartheta + o(1). \qquad (4.10)$$

Ponieważ funkcja $\sqrt{\det \mathbf{J}(\vartheta)}$ jest całkowalna, to po jej unormowaniu stałą $c(\mathbf{J}) = \int \sqrt{\det \mathbf{J}(\vartheta)}d\vartheta$ staje się gęstością $w_0(\vartheta)$ rozkładu zwanego rozkładem Jeffreysa (1946). Zatem (4.10) przyjmuje postać

$$I(\theta, X) = R_n = \frac{k}{2}\log\frac{n}{2\pi\, e} - D(w||w_0) + \log c(\mathbf{J}) + o(1).$$

To oznacza, że z dokładnością do członu $o(1)$ przepustowość jest osiągana (wykorzystana jest pełna informacja o θ zawarta w próbie X), gdy rozkład a priori pokrywa się z rozkładem Jeffreysa i wynosi asymptotycznie (gdy $n \to \infty$)

$$C(\Theta, X) = \sup_w I(\theta, X) = \frac{k}{2}\log\frac{n}{2\pi e} + \log c(\mathbf{J}). \qquad (4.11)$$

Rozważania powyższe dotyczą sytuacji, gdy nieznany parametr jest istotnie k-wymiarowy ($\Theta \subset \mathbb{R}^k$ o niepustym wnętrzu). Jak widać dla modeli wystarczająco regularnych i próby prostej n-elementowej przepustowość wyraża się przez logarytm z n. Powstaje pytanie, jak jest dla mniej regularnych modeli. Częściową odpowiedź daje następujące twierdzenie Rissanena.

TWIERDZENIE 4.6 (Rissanen, 1986). *Jeśli w rodzinie $\mathcal{P}$, danej przez (4.1), zbiór $\Theta \subset \mathbb{R}^k$ jest zwarty o niepustym wnętrzu oraz dla prawie wszystkich ϑ (względem miary Lebesgue'a) istnieje asymptotycznie normalny estymator $\widehat{\vartheta}$ parametru ϑ taki, że*

$$\sum_n \sup_\vartheta P_{n\vartheta}(||\sqrt{n}(\widehat{\vartheta} - \vartheta)|| \geqslant \log n) < \infty,$$

gdzie $||\cdot||$ oznacza normę euklidesową, to dla dowolnej gęstości q_n próby X (niekoniecznie prostej) mamy

$$\liminf_n \frac{D(p_{n\vartheta}||q_n)}{\log n} \geqslant \frac{k}{2} \qquad (4.12)$$

dla prawie wszystkich ϑ.

Z nierówności (4.12) wynika, że asymptotycznie dla dowolnie małego $\varepsilon > 0$ przepustowość nie może być mniejsza niż $(\frac{k}{2} - \varepsilon)\log n$.

Twierdzenia 4.5 i 4.6 dotyczą przypadku, gdy przestrzeń para-

metrów Θ jest nieskończona. Przypadek skończonego zbioru badał Rényi, który rozważał $\Theta = \{\vartheta_0, \vartheta_1\}$ i udowodnił następujące twierdzenie.

TWIERDZENIE 4.7 (Rényi, 1967). *Niech* $\mathcal{P} = \{p_{n\vartheta_0}, p_{n\vartheta_1}\}$ $p_{n\vartheta_0} \neq p_{n\vartheta_1}$ *będzie dwuelementową rodziną rozkładów, a* $w = (w_0, w_1)$ *rozkładem a priori. Wówczas*

$$I(\theta, X) \geqslant H(\theta) - \frac{18}{11}(\log e)\sqrt{w_0 w_1} \int \sqrt{p_{n\vartheta_0} p_{n\vartheta_1}}\, d\lambda^n$$

oraz

$$I(\theta, X) \leqslant H(\theta) - \frac{1}{2} w_0 w_1 \left(\int \sqrt{p_{n\vartheta_0} p_{n\vartheta_1}}\, d\lambda^n \right)^2.$$

Z twierdzenia 4.7 wynika, że jeśli $\liminf_n d^2_{\mathcal{H}}(p_{n\vartheta_0}, p_{n\vartheta_1}) > 0$, to dla dwupunktowego zbioru Θ przepustowość jest asymptotycznie równa 1 i jest osiągana dla rozkładu jednostajnego $w_0 = w_1 = 1/2$. Dokładniej

$$1 - \frac{9 \log e}{11} \int \sqrt{p_{n\vartheta_0} p_{n\vartheta_1}}\, d\lambda^n \leqslant C(\Theta, X) \leqslant 1 - \frac{1}{8} \left(\int \sqrt{p_{n\vartheta_0} p_{n\vartheta_1}}\, d\lambda^n \right)^2.$$

Jeśli próba X jest prosta, to z niezależności obserwacji wynika $\int \sqrt{p_{n\vartheta_0} p_{n\vartheta_1}}\, d\lambda^n = \left(\int \sqrt{p_{\vartheta_0} p_{\vartheta_1}}\, d\lambda \right)^n$, gdzie $p_{\vartheta_0}, p_{\vartheta_1}$ są różnymi gęstościami pojedynczych obserwacji, a to oznacza, że, dla każdego rozkładu a priori, $I(\theta, X)$ dąży wykładniczo do $H(\theta)$ przy $n \to \infty$, a przepustowość dąży wykładniczo do 1. Wynik Rényiego przenosi się bez trudu na dowolny skończony zbiór $\Theta = \{\vartheta_1, ..., \vartheta_k\}$. W szczególności można udowodnić nierówność

$$I(\theta, X) \geqslant H(\theta) - \frac{9}{11} \log e \left(\sum_{i \neq j} \sqrt{w_i w_j} \right) \max_{i \neq j} \int \sqrt{p_{n\vartheta_i}\, p_{n\vartheta_j}}\, d\lambda^n,$$

co również oznacza wykładniczą zbieżność $I(\theta, X)$ do $H(\theta)$ w przypadku próby prostej i zbieżność wykładniczą przepustowości do $\log k$.

4.6. Estymacja gęstości. Rozważmy próbę prostą $X = (X_1, ..., X_n)$ o nieznanej gęstości pojedynczej obserwacji p i niech $\widehat{p}_n$ będzie estymatorem p. Odległość entropijna $D(p||\widehat{p}_n)$ jest rozsądną funkcją straty, gdyż, z nierówności Pinskera, majoryzuje odległość w normie L_1. Jawne oszacowania z góry i z dołu na ryzyko $ED(p||\widehat{p}_n)$ estymatora $\widehat{p}_n$ podali Yang i Barron.

TWIERDZENIE 4.8 (Yang i Barron, 1999). *Załóżmy, że nieznana gęstość* p *pochodzi z rodziny* $\mathcal{F}$ *gęstości jednostajnie ograniczonych na*

przestrzeni miarowej σ-skończonej $(\mathcal{X}, \mathcal{B}, \lambda)$. Jeśli entropia metryczna $M_{\mathcal{H}}^{\mathcal{F}}(\varepsilon)$ rodziny $\mathcal{F}$ względem metryki Hellingera $d_{\mathcal{H}}$ jest rzędu $\varepsilon^{-1/r}$ dla pewnego $r > 0$, to

$$(n \ln n)^{-2r/(2r+1)} \leqslant \min_{\widehat{p}_n} \max_{p \in \mathcal{F}} E_p d_{\mathcal{H}}^2(p, \widehat{p}_n) \leqslant (\ln 2) \min_{\widehat{p}_n} \max_{p \in \mathcal{F}} E_p D(p||\widehat{p}_n)$$
$$\leqslant n^{-2r/(2r+1)} (\ln n)^{1/(2r+1)}, \tag{4.13}$$

gdzie estymator $\widehat{p}_n$ przebiega zbiór wszystkich gęstości.

Przypomnijmy, że entropia metryczna $M(\varepsilon)$ zbioru $\mathcal{F}$ jest równa logarytmowi minimalnej liczby kul w danej metryce, o promieniu nie większym niż ε, pokrywających $\mathcal{F}$. Twierdzenie powyższe podaje dokładny rząd potęgowy malenia ryzyka optymalnego (w sensie minimaksowym) estymatora w klasie wszystkich estymatorów.

Jako przykład zastosowania twierdzenie 4.8 rozważmy rodzinę gęstości gładkich na odcinku $[0, 1]$

$$\mathcal{F} = \{p : \log p \in W_2^r[0, 1]\},$$

gdzie $W_2^r[0, 1]$ oznacza przestrzeń Sobolewa funkcji r-krotnie różniczkowalnych z r-tą pochodną całkowalną z kwadratem. Rozważmy rodzinę wykładniczą $\{p_\vartheta = c_\vartheta e^{\vartheta \circ \Phi},\ \vartheta \in \mathbb{R}^k\}$ gęstości pojedynczych obserwacji na odcinku $[0, 1]$, gdzie Φ jest układem ortonormalnym k wielomianów stopnia co najwyżej k lub ciągiem k pierwszych funkcji układu trygonometrycznego. Niech $\widehat{\vartheta} = \widehat{\vartheta}(X)$ będzie estymatorem największej wiarogodności w rodzinie wykładniczej gęstości (produktowych) $p_{n\vartheta}$ dla próby prostej X i niech $\widehat{p} = p_{\widehat{\vartheta}}$ będzie estymatorem nieznanej gęstości pojedynczej obserwacji $p \in \mathcal{F}$. Barron i Sheu udowodnili następujące twierdzenie.

TWIERDZENIE 4.9 (Barron i Sheu, 1991). *Jeśli $r \geqslant 2$ oraz wymiar $k = k(n)$ rodziny wykładniczej jest rzędu $n^{1/(2r+1)}$, to dla każdego $p \in \mathcal{F}$ mamy*

$$D(p||p_{\widehat{\vartheta}}) = O_p(n^{-2r/(2r+1)}).$$

W przypadku układu trygonometrycznego potrzebne są jeszcze pewne warunki brzegowe dla $\log p$.

Zauważmy, że entropia metryczna rozważanej tutaj rodziny $\mathcal{F}$ względem metryki Hellingera jest rzędu $\varepsilon^{-1/r}$. Oznacza to, że estymator gęstości określony w twierdzeniu 4.9 osiąga optymalny (potęgowy) rząd zbieżności ryzyka.

Rozważmy teraz przypadek, gdy nieznana gęstość p należy do pa-

rametrycznej rodziny gęstości $\mathcal{F} = \{p_\vartheta : \vartheta \in \Theta \subset \mathbb{R}^k\}$ i załóżmy, że nieznana gęstość pojedynczej obserwacji należy do tej rodziny tzn. ma postać p_{ϑ_0}, gdzie ϑ_0 jest nieznane. Dla próby prostej X rozważmy estymator bayesowski gęstości p_{ϑ_0} postaci

$$\widehat{p}_n(\cdot) = \int p_\vartheta(\cdot)\frac{p_{n\vartheta}(X)w(\vartheta)}{f_n(X)}d\vartheta, \tag{4.14}$$

gdzie $w(\vartheta)$ jest gęstością a priori parametru ϑ (względem miary Lebesgue'a), a $p_{n\vartheta}$ jest gęstością produktową. Zauważmy, że estymator $\widehat{p}_n$ minimalizuje średnią stratę a posteriori. Istotnie, dla dowolnej gęstości q mamy

$$\begin{aligned}&\textstyle\int_\Theta D(p_\vartheta||q)\frac{p_{n\vartheta}(X)w(\vartheta)}{f_n(X)}d\vartheta - \int_\Theta D(p_\vartheta||\widehat{p}_n)\frac{p_{n\vartheta}(X)w(\vartheta)}{f_n(X)}d\vartheta\\ &\quad = \textstyle\int_{\mathcal{X}} \log\frac{\widehat{p}_n}{q}\int_\Theta p_\vartheta\frac{p_{n\vartheta}(X)w(\vartheta)}{f_n(X)}d\vartheta\, d\lambda^n = \int_{\mathcal{X}}\widehat{p}_n\log\frac{\widehat{p}_n}{q}\,d\lambda^n \geqslant 0.\end{aligned}$$

Clarke i Barron (1990) udowodnili następujący fakt.

TWIERDZENIE 4.10. *Przy założeniach twierdzenia 4.3 mamy*

$$\frac{1}{n}\sum_{j=0}^{n} E_{\vartheta_0}D(p_{\vartheta_0}||\widehat{p}_j) = \frac{k}{2}\frac{\log n}{n} + O\left(\frac{1}{n}\right), \tag{4.15}$$

gdzie $\widehat{p}_0 = \int p_\vartheta w(\vartheta)d\vartheta$, *a* $\widehat{p}_j$ *jest oparty na* $(X_1, ..., X_j)$ *i dany wzorem (4.14).*

Formuła (4.15) oznacza, że w przybliżeniu ryzyko estymatora bayesowskiego $\widehat{p}_n$ jest rzędu $\frac{k}{2}\frac{\log n}{n}$. Ponieważ w modelu k-wymiarowym entropia metryczna jest rzędu $k\log(1/\varepsilon)$, co z grubsza odpowiada $r = \infty$ w twierdzeniu 4.8, to rząd potęgowej zbieżności ryzyka estymatora bayesowskiego jest zgodny z rzędem n^{-1} wynikającym z (4.13).

4.7. Testowanie hipotez statystycznych. Podobnie jak w poprzednim rozdziale ograniczymy się tutaj do prób prostych $X = (X_1, ..., X_n)$ i niech p będzie gęstością pojedynczej obserwacji. Najpierw rozpatrzymy problem testowania hipotez prostych:

$$H_0 : p = p_0 \quad \text{przeciwko} \quad H_1 : p = p_1,$$

gdzie p_0, p_1 są ustalonymi różnymi gęstościami. Chernoff (1952) udowodnił następujące twierdzenie zwane lematem Steina.

TWIERDZENIE 4.11 (Lemat Steina). *Niech* $\alpha \in (0, 1)$ *będzie dowolnym ustalonym poziomem istotności, a* β_n *prawdopodobieństwem błędu II rodzaju testu Neymana-Pearsona hipotezy* H_0 *przeciwko* H_1.

Wówczas

$$\frac{1}{n}\log\beta_n \longrightarrow -D(p_0||p_1). \tag{4.16}$$

Relacja (4.16) mówi, że dla ustalonego poziomu istotności moc testu najmocniejszego dąży wykładniczo do 1 i odległość entropijna $D(p_0||p_1)$ określa tempo tej wykładniczej zbieżności. Lemat Steina jest szczególnym przypadkiem twierdzenia Craméra o wielkich odchyleniach. Jednak jego dowód można przeprowadzić bezpośrednio, korzystając jedynie z klasycznego prawa wielkich liczb. Przytoczymy go, opuszczając łatwe szczegóły.

Dowód. Logarytm statystyki Neymana-Pearsona ma postać

$$V_n = \log\tfrac{p_{n1}}{p_{n0}}(X) = \textstyle\sum_{i=1}^n \log\frac{p_1(X_i)}{p_0(X_i)}$$

czyli jest sumą niezależnych zmiennych losowych o tym samym rozkładzie. Zatem z prawa wielkich liczb Chinczyna mamy

$$\tfrac{1}{n}V_n \xrightarrow{P_0} E_0V_n = -D(p_0||p_1). \tag{4.17}$$

Wybierzmy liczbę $v_{n\alpha}$ tak, aby

$$P_0(V_n \geqslant -nD(p_0||p_1) + v_{n\,\alpha}) = \alpha.$$

Wtedy $v_{n\,\alpha}/n \to 0$. Stąd i z nierówności Markowa dostajemy

$$\begin{aligned}\beta_n &= P_1(\tfrac{p_{n1}}{p_{n0}} \leqslant 2^{-nD(p_0||p_1)+v_{n\,\alpha}}) = P_1(\tfrac{p_{n0}}{p_{n1}} \geqslant 2^{nD(p_0||p_1)-v_{n\,\alpha}})\\ &\leqslant (E_1\tfrac{p_{n0}}{p_{n1}})2^{-nD(p_0||p_1)+v_{n\,\alpha}} = 2^{-nD(p_0||p_1)+v_{n\,\alpha}},\end{aligned}$$

co po prostych przekształceniach daje

$$\tfrac{1}{n}\log\beta_n \leqslant -D(p_0||p_1) + \tfrac{v_{n\,\alpha}}{n}. \tag{4.18}$$

Z drugiej strony z (4.17) mamy $P_0(V_n \geqslant -nD(p_0||p_1) - n\delta) = 1 - \eta_n \to 1$ dla dowolnego $\delta > 0$, skąd

$$\begin{aligned}\beta_n &\geqslant P_1(2^{-nD(p_0||p_1)-n\delta} \leqslant \tfrac{p_{n1}}{p_{n0}} \leqslant 2^{-nD(p_0||p_1)+v_{n\,\alpha}})\\ &= E_0\left(\tfrac{p_{n1}}{p_{n0}}\mathbf{1}(2^{-nD(p_0||p_1)-n\delta} \leqslant \tfrac{p_{n1}}{p_{n0}} \leqslant 2^{-nD(p_0||p_1)+v_{n\,\alpha}})\right)\\ &\geqslant 2^{-nD(p_0||p_1)-n\delta}P_0(2^{-nD(p_0||p_1)-n\delta} \leqslant \tfrac{p_{n1}}{p_{n0}} \leqslant 2^{-nD(p_0||p_1)+v_{n\,\alpha}})\\ &= 2^{-nD(p_0||p_1)-n\delta}(1-\alpha-\eta_n).\end{aligned}$$

W konsekwencji otrzymujemy

$$\tfrac{1}{n}\log\beta_n \geqslant -D(p_0||p_1) - \delta + \tfrac{1}{n}\log(1-\alpha-\eta_n).$$

Z dowolności δ oraz (4.18) wynika (4.16). $\square$

Role poziomu istotności α i prawdopodobieństwa błędu II rodzaju β można zamienić. Prowadzi to do lematu Steina w postaci dualnej.

TWIERDZENIE 4.12 (Dualny lemat Steina). *Niech $\beta \in (0,1)$ będzie dowolnym ustalonym prawdopodobieństwem błędu II rodzaju, a α_n odpowiadającym mu poziomem istotności testu Neymana-Pearsona hipotezy H_0 przeciwko H_1. Wówczas*

$$\frac{1}{n}\log\alpha_n \longrightarrow -D(p_1||p_0).$$

Stosując podejście bayesowskie, można połączyć obie wersje lematu Steina. Niech $w = (w_0, w_1)$ będzie rozkładem a priori na zbiorze dwupunktowym $\Theta = \{0,1\}$, a p_{n0}, p_{n1} gęstościami rozkładów produktowych. Wówczas $\frac{p_{n0}w_0}{p_{n0}w_0+p_{n1}w_1}$ i $\frac{p_{n1}w_1}{p_{n0}w_0+p_{n1}w_1}$ są prawdopodobieństwami a posteriori. Rényi (1967) zaproponował tzw. test standardowy H_0 przeciwko H_1, który odrzuca H_0, gdy $\frac{p_{n1}w_1}{p_{n0}w_0+p_{n1}w_1} \geqslant \frac{p_{n0}w_0}{p_{n0}w_0+p_{n1}w_1}$ co jest równoważne warunkowi $\frac{p_{n1}}{p_{n0}} \geqslant \frac{w_0}{w_1}$. Dla tak określonego testu oznaczmy α_n, β_n odpowiednio, poziom istotności i prawdopodobieństwo błędu II rodzaju oraz prawdopodobieństwo błędnej decyzji $\epsilon_n = w_0\alpha_n + w_1\beta_n$. Gęstości hipotetyczne p_0, p_1 (zakładamy, że $p_0 \cdot p_1 \neq 0$) połączmy odcinkiem logarytmicznie wypukłym tzn. rozważmy jednoparametrową rodzinę gęstości $\{p_t = c_t p_0^t p_1^{1-t} : t \in [0,1]\}$, gdzie c_t jest stałą normującą (por. twierdzenie 2.12). Wówczas funkcja $D(p_t||p_0)$ jest rosnąca względem t, a funkcja $D(p_t||p_1)$ malejąca względem t. Zatem istnieje liczba $t^* \in (0,1)$ taka, że zachodzi równość tzn. $D(p_{t^*}||p_0) = D(p_{t^*}||p_1)$. Chernoff (1952) udowodnił twierdzenie, które można nazwać lematem Steina w wersji bayesowskiej.

TWIERDZENIE 4.13 (Chernoff, 1952). *Przy powyższych założeniach i oznaczeniach mamy*

$$\frac{1}{n}\log\epsilon_n \longrightarrow -D(p_{t^*}||p_0).$$

Oznacza to, że prawdopodobieństwo błędnej decyzji dąży wykładniczo do zera. Liczbę $D(p_{t^*}||p_0)$ nazywamy informacją Chernoffa. Rényi (1967) udowodnił nierówność $h^{-1}(H(\theta|X)) \leqslant \epsilon_n \leqslant 2\,H(\theta|X)$, gdzie $h(p)$ jest funkcją entropijną, co także pociąga wykładniczą zbieżność ϵ_n do zera.

Rozważmy teraz przypadek ogólny, gdy zbiór parametrów Θ jest dowolny. Niech $\mathcal{F} = \{p_\vartheta : \vartheta \in \Theta\}$ będzie rodziną gęstości pojedyn-

czych obserwacji na dowolnej przestrzeni $(\mathcal{X}, \mathcal{B}, \lambda)$, $X = (X_1, ..., X_n)$ będzie próbą prostą z rozkładu P_ϑ o gęstości $p_\vartheta \in \mathcal{F}$ oraz $\Theta_0 \subset \Theta$. Rozważmy problem testowania

$$H_0 : \vartheta \in \Theta_0 \quad \text{przeciwko} \quad H_1 : \vartheta \in \Theta \setminus \Theta_0. \tag{4.19}$$

Bahadur (1967) uogólnił dualny lemat Steina na przypadek testowania dowolnych hipotez złożonych. Jego wynik przedstawimy w wersji podanej przez Raghavachari (1970).

TWIERDZENIE 4.15 (Bahadur, 1967, Raghavachari, 1970) *Niech T_n będzie statystyką testową testu prawostronnego hipotezy H_0, jak w (4.19), $\vartheta \in \Theta \setminus \Theta_0$ będzie dowolnym parametrem, a $\beta \in (0,1)$ dowolnym ustalonym błędem II rodzaju. Jeśli $t_n = t_n(\beta, \vartheta)$ jest takie, że $P^n_\vartheta(T_n \geqslant t_n) \to \beta$ gdy $n \to \infty$, to*

$$\liminf_{n\to\infty} \frac{1}{n} \log \sup_{\vartheta_0 \in \Theta_0} P^n_{\vartheta_0}(T_n \geqslant t_n) \geqslant - \inf_{\vartheta_0 \in \Theta_0} D(p_\vartheta || p_{\vartheta_0}). \tag{4.20}$$

Ponadto dla statystyki ilorazu wiarogodności

$$\widehat{T}_n = \widehat{T}_n(X) = -\frac{1}{n} \log \frac{\sup_{\vartheta_0 \in \Theta_0} p_{n\vartheta_0}(X)}{\sup_{\vartheta \in \Theta} p_{n\vartheta}(X)}$$

istnieje granica w (4.20) i jest równa prawej stronie (4.20).

Jeśli istnieje granica lewej strony w (4.20) i nie zależy od β, to oznaczmy ją przez $-\frac{1}{2}c_T(\vartheta)$. Liczbę $c_T(\vartheta)$ nazywamy dokładnym nachyleniem w sensie Bahadura testu opartego na statystyce T_n. Zatem test ilorazu wiarogodności ma największe możliwe nachylenie wynoszące $2 \inf_{\vartheta_0 \in \Theta_0} D(p_\vartheta || p_{\vartheta_0})$. Efektywnością Bahadura testu $\widetilde{T}_n$ względem testu T_n nazywamy iloraz $c_{\widetilde{T}}(\vartheta)/c_T(\vartheta)$ dokładnych nachyleń w sensie Bahadura. Zatem test ilorazu wiarogodności jest optymalny w tym sensie, że efektywność Bahadura dowolnego innego testu względem niego nie przekracza 1. Istnienie granicy w (4.20) wiąże się z twierdzeniem o wielkich odchyleniach dla statystyki T_n. Problematyce wielkich odchyleń i efektywności Bahadura testów nieparametrycznych jest poświęcona monografia Nikitina (1995). Z efektywnością Bahadura wiążą się inne pojęcia efektywności np. efektywności Kallenberga (lub inaczej pośredniej) (por. Inglot i Ledwina, 1996), którą daje się najczęściej dużo łatwiej wyznaczyć i która ma tę samą interpretację w języku ilorazu rozmiarów prób, przy których (mówiąc w uproszczeniu) oba testy mają tę samą moc asymptotyczną i ten sam poziom istotności. Bliższe omówienie tej problematyki przekracza ramy niniejszego opracowania.

Clarke i Barron (1990) uogólnili lemat Steina w inny sposób. Niech w (4.19) $\Theta_0 = \{\vartheta_0\}$ oraz $\Theta \subset \mathbb{R}^k$. Wówczas $T_n^* = \ln(f_n/p_{n\vartheta_0})$ jest statystyką optymalnego testu bayesowskiego (prawostronnego), który dla każdego ustalonego poziomu istotności $\alpha \in (0,1)$ minimalizuje średnie prawdopodobieństwo błędu II rodzaju $\overline{\beta}_n$ określone wzorem

$$\overline{\beta}_n = \int_\Theta \beta_n(\vartheta) w(\vartheta) d\vartheta = \int_\Theta \int_{\mathcal{X}^n} p_{n\vartheta} \mathbf{1}(T_n^* \geqslant c_{n\,\alpha}) w(\vartheta) d\vartheta d\lambda^n.$$

Z wniosku 4.2 wynika następujące twierdzenie.

Twierdzenie 4.14 (Clarke i Barron, 1990). *Przy założeniach twierdzenia 4.3 asymptotyczna wartość krytyczna testu bayesowskiego T_n^* ma postać*

$$-(\ln 2) D(p_{n\vartheta_0} || f_n) + \frac{1}{2}\chi_{k,1-\alpha} - \frac{k}{2},$$

gdzie $\chi_{k,\tau}$ jest kwantylem rzędu τ rozkładu chi-kwadrat z k stopniami swobody oraz

$$\liminf_n (\ln \overline{\beta}_n + (\ln 2) D(p_{n\vartheta_0} || f_n)) \geqslant \frac{1}{2}\chi_{k,(1-\alpha)/2} - \frac{k}{2} + \log \frac{1-\alpha}{2}$$

$$\limsup_n (\ln \overline{\beta}_n + (\ln 2) D(p_{n\vartheta_0} || f_n)) \leqslant \frac{1}{2}\chi_{k,1-\alpha}.$$

W konsekwencji

$$\frac{1}{n} \log \overline{\beta}_n = -\frac{1}{n} D(p_{n\vartheta_0} || f_n) + O\left(\frac{1}{n}\right). \tag{4.21}$$

Zatem relację (4.21) można nazwać uogólnionym lematem Steina.

4.8. Informacyjne kryteria wyboru modelu. Dotychczas w naszych rozważaniach model statystyczny był ustalony. W rzeczywistości obserwowane zjawisko można opisać za pomocą wielu modeli. Jest oczywiste, że im model jest bogatszy, tym lepiej, bardziej dokładnie, pozwala przybliżyć prawdziwy rozkład próby. Z drugiej strony powiększanie modelu skutkuje większym błędem oszacowania parametru ϑ w tym modelu. Potrzebny jest więc kompromis. To rodzi problem wyboru takiego modelu, z pewnej z góry ustalonej listy konkurujących modeli, który, w jakimś sensie, najlepiej odpowiada danym tj. próbie X. W czterech minionych dekadach zaproponowano wiele kryteriów wyboru modelu m.in. kryterium informacyjne Akaike (1974) i kryterium informacyjne Schwarza (1978). Dużą popularność zyskały kryteria minimalnej długości opisu (ang. minimal

description length, w skrócie MDL). Idea konstrukcji tych kryteriów pochodzi od Rissanena (1978). Omówimy krótko trzy kryteria MDL. Pierwsze, zwane dwustopniowym MDL wprowadził Rissanen (1983). Niech $\mathcal{P}_s = \{p_{n\vartheta} : \vartheta \in \Theta_s \subset \mathbb{R}^{k(s)}\}$, $s \in S$, będzie listą modeli parametrycznych. W pierwszym kroku dla każdego s wybieramy $\sqrt{n}$-zgodny estymator $\widehat{\vartheta}$ i rozważamy jego dyskretyzację $\widehat{\vartheta}^*$ z ziarnem rzędu $1/\sqrt{n}$ oraz gęstość a priori w_s (względem miary Lebesgue'a). Wtedy ilość informacji (długość optymalnego kodu binarnego) potrzebna do podania $\widehat{\vartheta}^*$ wynosi $H(w_s) + \frac{k(s)}{2} \log n$ z dokładnością do członu $o(1)$ (por. z interpretacją entropii różniczkowej w rozdziale 2.2). W drugim kroku kodujemy zmienną o gęstości $p_{n\widehat{\vartheta}^*}$. Ilość informacji (długość optymalnego kodu binarnego) potrzebna do podania jej wartości wynosi w przybliżeniu $-\log p_{n\widehat{\vartheta}^*}$. Jeśli model jest 'gładki', to tę wartość można zastąpić przez $-\log p_{n\widehat{\vartheta}}$ kosztem wyrażenia $o(1)$. Sumując ilości informacji z obu kroków dostajemy $\mathcal{L}(s) = -\log p_{n\widehat{\vartheta}} + \frac{k(s)}{2} \log n$ po opuszczeniu wyrażeń rzędu $o(1)$. Dwustopniowe MDL Rissanena każe wybrać ten model z listy, dla którego $\mathcal{L}(s)$ jest minimalne. Łatwo widać, że z dokładnością do wyrażeń rzędu $O(1)$ kryterium to sprowadza się do reguły BIC Schwarza (1978). Barron i Clarke (1990) zaproponowali stochastyczne kryterium informacyjne (ang. stochastic information criterion). Dla każdego s rozważamy gęstość bezwarunkową próby f_n i optymalny kod binarny oparty na f_n. Wybierany jest więc ten model z listy, dla którego wyrażenie $-\log f_n$ jest minimalne. Przy założeniach twierdzenia 4.3 mamy z (4.4) z dokładnością do wyrażenia $O(1)$

$$-\log f_n = -\log p_{n\widehat{\vartheta}} + \frac{k(s)}{2} \log n,$$

co oznacza, że asymptotycznie dwustopniowe MDL i stochastyczne kryterium informacyjne są równoważne. Rissanen (1986) zaproponował jeszcze jedno kryterium, szczególnie użyteczne dla szeregów czasowych. Ponieważ

$$-\log p_{n\vartheta} = -\sum_{i=1}^{n} \log p_\vartheta(x_i|x_1, ..., x_{i-1}),$$

to, zastępując w kolejnych składnikach ϑ przez estymator $\widehat{\vartheta}_i$ oparty na obserwacjach do momentu i, wybierany jest model, dla którego wyrażenie

$$-\sum_{i=1}^{n} \log p_{\widehat{\vartheta}_i}(x_i|x_1, ..., x_{i-1})$$

jest minimalne. To kryterium nazywa się predykcyjnym MDL. Hansen i Yu (2001), w obszernym artykule przeglądowym, przedstawili różne koncepcje kryteriów wyboru modelu opartych na zasadzie minimalnej długości opisu, w tym trzy kryteria zasygnalizowane powyżej, oraz szereg przykładów ich zastosowań. Metodzie MDL poświęcona jest monografia Grünwalda (2007).

Podziękowania: Dziękuję Pani prof. T. Ledwinie za przeczytanie wstępnej wersji artykułu i wiele cennych wskazówek.

Literatura

[1] H. Akaike, *A new look at the statistical model identification*, IEEE Trans. Autom. Control **19** (1974), 716–723.

[2] R. R. Bahadur, *An optimal property of the likelihood ratio statistic*, Proc. Fifth Berkeley Symp. Math. Statist. Probab. **1** (1967), 13-26.

[3] O. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory*, New York: Wiley, 1978.

[4] A. R. Barron, *The strong ergodic theorem for densities: generalized Shannon-McMillan-Breiman theorem*, Ann. Probab. **13** (1985), 1292–1303.

[5] A. R. Barron, *Entropy and the central limit theorem*, Ann. Probab. **14** (1986), 336–342.

[6] A. R. Barron, C. Sheu, *Approximation of density functions by sequences of exponential families*, Ann. Statist. **19** (1991), 1347–1369.

[7] J. Bartoszewicz, *Wykłady ze statystyki matematycznej*, Wydawnictwa Uniwersytetu Wrocławskiego, Wrocław, 1981.

[8] N. Blachman, *The convolution inequality for entropy powers*, IEEE Trans. Infrom. Theory **11** (1965), 267–271.

[9] S. G. Bobkov, G. P. Chistyakov, F. Götze, *Rate of convergence and Edgeworth-type expansion in the entropic central limit theorem*, Ann. Probab. **41** (2013), 2479–2512.

[10] L. Breiman, *The individual ergodic theorems of information theory*, Ann. Math. Statist. **28** (1957), 809–811.

[11] N. N. Chentsov, *Statistical Decision Rules and Optimal Inference* (w j. rosyjskim), Nauka, Moskwa, 1972.

[12] H. Chernoff, *A measure of the asymptotic efficiency of tests of a hypothesis based on a sum of observations*, Ann. Math. Statist. **23** (1952), 493–507.

[13] A. J. Chintschin, D. K. Faddejew, A. N. Kolmogoroff, A. Rényi, J. Balatoni, *Arbeiten zur Informationstheorie. I*, Mathematische Forschungsberichte, **4**, VEB Deutscher Verlag der Wissenschaften, Berlin, 1957.

[14] B. S. Clarke, A. R. Barron, *Information-theoretic asymptotics of Bayes methods*, IEEE Trans. Infrom. Theory **36** (1990), 453–471.

[15] B. S. Clarke, A. R. Barron, *Jeffreys' prior is asymptotically least favorable under entropy risk*, J. Statist. Planning Inference **41** (1994), 37-60.

[16] T. M. Cover, J. A. Thomas, *Elements of Information Theory* 2nd ed., John Wiley, New York, 2006.

[17] I. Csiszár, *Information-type measures of difference of probability distributions and indirect observations*, Studia Sci. Math. Hungar. **2** (1967), 299–318.

[18] I. Csiszár, *I-divergence geometry of probability distributions and minimization problems*, Ann. Probab. **3** (1975), 146–158.

[19] I. Csiszár, F. Matúš, *Infromation projections revisited*, IEEE Trans. Infrom. Theory **49** (2003), 1474–1490.

[20] A. Dembo, O. Zeitouni, *Large deviations and applications*, Chapter 6 in: Handbook of Stochastic Analysis and Applications, D. Kannan and V. Lakshmikanthan, eds., Marcel-Dekker, 2002.

[21] Ł. Dębowski, *Information Theory and Statistics*, Institute of Computer Science Polish Academy of Science, Warsaw, 2013.

[22] D. K. Faddeev, *On the concept of entropy of a finite probabilistic scheme*, Uspekhi Mat. Nauk **11** (1956), 227–231.

[23] R. M. Gray, *Entropy and Information Theory*, Springer, New York, 1990.

[24] P. D. Grünwald, *The Minimum Description Length Principle*, MIT Press, 2007.

[25] M. H. Hansen, B. Yu, *Model selection and principle of minimum description length*, J. Amer. Math. Soc. **96** (2001), 746–774.

[26] D. Haussler, *A general minimax result for relative entropy*, IEEE Trans. Inform. Theory **43** (1997), 1276–1280.

[27] D. A. Huffman, *A method of the construction of minimum redundancy codes*, Proc. IRE **40** (1952), 1098–1101.

[28] T. Inglot, T. Ledwina, *Asymptotic optimality of data driven Neyman's tests for uniformity*, Ann. Statist., **24** (1996), 1982-2019.

[29] H. Jeffreys, *An invariant form for the prior probability in estimation problems*, Proc. Royal Soc. London, Ser. A, Math. Physical Sci. **186** (1946), 453–461.

[30] P. Jizba, T. Arimitsu, *The world according to Rényi: thermodynamics of multifractal system*, Ann. Physics **312** (2004), 17–59.

[31] O. Johnson, *Information Theory and The Central Limit Theorem*, London, U.K.: Imperial College Press, 2004.

[32] O. Johnson, A. R. Barron, *Fisher information inequalities and the central limit theorem*, Probab. Theory Relat. Fields **129** (2004), 391–409.

[33] O. Johnson, Y. Suhov, *Entropy and random vectors*, J. Statist. Physics **104** (2001), 145–165.

[34] D. Kendall, *Information theory and the limit theorem for Markov Chains and processes with a countable infinty of states*, Ann. Inst. Statist. Math. **15** (1963), 137–143.

[35] L. G. Kraft, *Advice for quantizing, grouping and coding amplitude modulated pulses*, Master's thesis, Department of Electrical Engineering, MIT, Cambridge, MA, 1949.

[36] S. Kullback, *Information Theory and Statistics*, Wiley, New York, 1959.

[37] S. Kullback, *A lower bound for discrimination information in terms of variation*, IEEE Trans. Inform. Theory **13** (1967), 126–127.

[38] E. L. Lehmann, G. Casella, *Theory of Point Estimation*, 2nd ed. New York: Springer, 1998.

[39] M. Madiman, A. R. Barron, *Generalized entropy powr inequalities and monotonicity properties of information*, IEEE Trans. Infrom. Theory **53** (2007), 2317–2329.

[40] B. McMillan, *The basic theorems of information theory*, Ann. Math. Statist. **24** (1953), 196–219.

[41] B. McMillan, *Two inequalities implied by unique decipherability*, IEEE Trans. Inform. Theory **2** (1956), 115–116.

[42] J. Neyman, *Sur un teorema concernente le cosidette statistiche sufficiente*, Giorn. Ist. Ital. Att. **6** (1935), 320–334.

[43] Ya. Yu. Nikitin, *Asymptotic Efficiency of Nonparametric Tests*, Cambridge University Press, New York, 1995.

[44] S. Orey, *On the Shannon-Peres-Moy theorem*, Contemp. Math. **41** (1985), 319–327.

[45] M. Raghavachari, *On a theorem of Bahadur on the rate of the convergence of test statistics*, Ann. Math. Statist., **41** (1970), 1695-1699.

[46] A. Rényi, *On measures of entropy and information*, in Neyman J., editor, Proceedings of the 4th Berkeley Conference on Mathematical Statistics and Probability, 547–561, Berkeley, University of California Press, 1961.

[47] A. Rényi, *Wahrscheinlichkeitsrechnung. Mit einem Anhang über Informationstheorie*, Hochschulbücher für Mathematik, **54**, VEB Deutscher Verlag der Wissenschaften, Berlin, 1962.

[48] A. Rényi, *On some basic problems of statistics from the point of view of information theory*, In: Proceedings of the 5th Berkeley Conference on Mathematical Statistics and Probability, Vol. 1, 531–543, Berkeley, University of California Press, 1967.

[49] J Rissanen, *Modelling by shortest data description*, Automatica, **14** (1978), 465–471.

[50] J. Rissanen, *A universal prior for integers and estimation by minimum description length*, Ann. Statist. **11** (1983), 416–431.

[51] J. Rissanen, *Stochastic complexity and modelling*, Ann. Statist. **14** (1986), 1080–1100.

[52] O. Rioul, *Information theoretic proofs of entropy power inequalities*, IEEE Trans. Infrom. Theory **57** (2011), 33–55.

[53] I. N. Sanov, *On the probability of large deviations of random variables*, Mat. Sbornik **42** (1957), 11–44.

[54] G. Schwarz, *Estimating the dimension of a model*, Ann. Statist. **6** (1978), 461–464.

[55] C. Shannon, *A mathematical theory of communication*, Bell System Tech. J. **27** (1948), 379–423, 623–656.

[56] R. Shimizu, *On Fisher's ammount of information for location family*, In G. P. Patil et al., editor, Statistical Distributions in Scientific Work, Vol. 3, 305–312, Reidel, 1975.

[57] A. Stam, *Some inequalities satisfied by the quantities of information of Fisher and Shannon*, Information and Control **2** (1959), 101–112.

[58] C. Tsallis, *Possible generalization of Boltzmann-Gibbs statistics*, J. Statist. Physics **52** (1988), 479–487.

[59] A. W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, 1998.

[60] Y. Yang, A. R. Barron, *Information-theoretic determination of minimum rates of convergence*, Ann. Statist. **27** (1999), 1564–1599.

Tadeusz Inglot
Instytut Matematyki i Informatyki
Politechniki Wrocławskiej
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
E-mail: `Tadeusz.Inglot@pwr.wroc.pl`