Autoreferat

1. Imię i nazwisko

Łukasz Dębowski

2. Posiadane dyplomy

1999 — **Uniwersytet Warszawski**, Wydział Fizyki

Dyplom ukończenia studiów magisterskich w zakresie **fizyki** specjalność **fizyka teoretyczna** z wynikiem bardzo dobrym (z wyróżnieniem). Praca magisterska pt. *Teoria funkcjonału gęstości dla monowarstw zaadsorbowanych na podłożu krystalicznym* napisana pod kierunkiem prof. dr. hab. Marka Napiórkowskiego.

2005 — **Polska Akademia Nauk**, Instytut Podstaw Informatyki

Dyplom uzyskania stopnia doktora **nauk matematycznych** w zakresie **informatyki** (z wyróżnieniem) uzyskany na podstawie rozprawy doktorskiej pt. *Własności entropii nadwyżkowej dla procesów stochastycznych nad różnymi alfabetami* napisanej pod kierunkiem doc. dr. hab. Jana Mielniczuka.

3. Zatrudnienie w jednostkach naukowych

06.2000–teraz — **Polska Akademia Nauk, Instytut Podstaw Informatyki**, Warszawa

06.2000–11.2005 — zatrudnienie na stanowisku asystenta w wymiarze pełnego etatu

12.2005–teraz — zatrudnienie na stanowisku adiunkta w wymiarze pełnego etatu

01.2008–12.2009 — **Centrum voor Wiskunde en Informatica**, Amsterdam

— zatrudnienie na stanowisku stażysty podoktorskiego w wymiarze pełnego etatu

4. Podstawowe osiągnięcie

Jako podstawowe osiągnięcie przedkładam cykl artykułów pt. *Własności dyskretnych procesów stochastycznych o szybko rosnącej informacji wzajemnej między blokami*, na który składają się następujące prace:

(A) Ł. Dębowski, (2010). Variable-Length Coding of Two-Sided Asymptotically Mean Stationary Measures. *Journal of Theoretical Probability*, **23**:237–256.

(B) Ł. Dębowski, (2011). On the Vocabulary of Grammar-Based Codes and the Logical Consistency of Texts. *IEEE Transactions on Information Theory*, **57**:4589–4599.

(C) Ł. Dębowski, (2012). Mixing, Ergodic, and Nonergodic Processes with Rapidly Growing Information between Blocks. *IEEE Transactions on Information Theory*, **58**:3392-3401.

(D) Ł. Dębowski, (2012). On Hidden Markov Processes with Infinite Excess Entropy. *Journal of Theoretical Probability*, DOI: `10.1007/s10959-012-0468-6`.

Powyższe prace dotyczą teorioinformacyjnych własności dyskretnych procesów stochastycznych z daleką pamięcią, a ich punktem wyjścia i motywacją była refleksja nad statystycznymi własnościami tekstów w języku naturalnym (czyli np. w języku polskim lub angielskim).

Wydaje się intuicyjnie oczywiste, że teksty tworzone przez ludzi nie mogą być tworzone ani w sposób czysto deterministyczny, ani w sposób czysto losowy. Jeżeliby rozpatrywać teksty w języku naturalnym jako realizacje pewnego procesu stochastycznego, to można by przypuszczać, że proces ten ma także własność, jak na razie nieformalnie rozmumianego, niedeterminizmu i dalekiej pamięci. Odpowiednich formalnych pojęć w przypadku procesów o wartościach dyskretnych, a takież wystarczają do modelowania tekstu, dostarcza aparat teorii informacji (Cover i Thomas, 2006; Crutchfield i Feldman, 2003). Warunek niedeterminizmu można sformalizować jako warunek, że intensywność entropii (*entropy rate*, entropia procesu) jest dodatnia, zaś warunek dalekiej pamięci można sformalizować jako warunek, że entropia nadwyżkowa (*excess entropy*, informacja wzajemna między przeszłością a przyszłością procesu) jest nieskończona. (Entropia nadwyżkowa jest skończona dla procesów Markowa i ukrytych procesów Markowa o skończonej liczbie stanów.)

Już w momencie narodzin teorii informacji Shannon (1951) próbował mierzyć entropię języka naturalnego (konkretnie, języka angielskiego). Otrzymał on słynne oszacowanie intensywności entropii jako około 1 bit na literę. Kilka dekad później Hilberg (1990) przerysował wykres entropii warunkowej z pracy Shannona w skali podwójnie logarytmicznej i spostrzegł, że punkty na nim układają się w przybliżeniu na linii prostej. Oznacza to, że informacja wzajemna pomiędzy przylegającymi blokami tekstu długości n rośnie proporcjonalnie do n^β, gdzie $\beta \in (0, 1)$. W konsekwencji entropia nadwyżkowa dla języka naturalnego byłaby istotnie nieskończona.

Badacze badający układy złożone, zainspirowani artykułem Hilberga (1990), zwrócili uwagę, że procesy o nieskończonej entropii nadwyżkowej pojawiają się nie tylko w modelowaniu języka naturalnego (Crutchfield i Young, 1989; Ebeling, 1997; Ebeling i Pöschel, 1994; Gramss, 1994; Bialek *et al.*, 2001; Shalizi i Crutchfield, 2001; Crutchfield i Feldman, 2003; Löhr, 2009). Również z czysto matematycznego punktu widzenia entropia nadwyżkowa jest interesującą miarą zależności dla procesów losowych o wartościach nominalnych, dla których analiza autokorelacji nie jest dobrym narzędziem do opisu zależności procesu.

Pokrótce omawiając wcześniejsze prace, należy wspomnieć, że entropia nadwyżkowa była już badana dla kilku klas procesów stochastycznych, jakkolwiek znane wyniki są dość wycinkowe. Najbardziej klasyczny rezultat dotyczy procesów gaussowskich, dla których Grenander i Szegő (1958, Section 5.5) wyprowadzili przedstawienie całkowe entropii nadwyżkowej, zaś Finch (1960) obliczył wartość tej całki dla procesów autoregresyjnych średniej ruchomej (ARMA). W przypadku procesów ARMA entropia nadwyżkowa jest skończona. Ponadto Dębowski (2005) pokazał, że entropia nadwyżkowa jest skończona dla procesu gaussowskiego wtedy i tylko wtedy, gdy $\sum_{k=1}^{\infty} k\,|\alpha_k|^2 < \infty$, gdzie α_k jest funkcją autokorelacji częściowej. Kilka dalszych prac dotyczy procesów nad alfabetem skończonym o nieskończonej entropii nadwyżkowej. Na przykład, Bradley (1980) skonstruował pierwszy przykład procesu mieszającego mającego tę własność. Gramss (1994) zbadał proces generowany przez częstości zer i jedynek w ciągu króliczym (*rabbit sequence*). Travers i Crutchfield (2011) zbadali kilka ukrytych procesów Markowa o nieskończonej liczbie stanów ukrytych. Były podejmowane także próby uogólnienia entropii nadwyżkowej na przypadek dwuwymiarowych pól losowych (Feldman i Crutchfield, 2003; Bułatek i Kamiński, 2009). Mahoney *et al.* (2009) oraz Ellison *et al.* (2009) wyprowadzili wzór na entropię nadwyżkową w terminach predykcyjnych i wstecznie predykcyjnych ϵ-maszyn, czyli minimalnych unifilarnych ukrytych reprezentacji Markowa procesu (Shalizi i Crutchfield, 2001; Löhr, 2009) (definicję unifilarnych procesów Markowa można znaleźć także w pracy **(D)**).

Po zapoznaniu się z wyżej wymienioną literaturą zainteresowałem się matematycznymi własnościami procesów o potęgowo rosnącej informacji wzajemnej między przyległymi blokami oraz ewentualnymi lingwistycznymi interpretacjami tych własności. Trop ten okazał się być płodny i stanowi motyw przewodni przedkładanego cyklu artykułów. Szczegółowe omawianie moich osiągnieć rozpocznę od pracy **(B)** zawierającej najbardziej interesujący rezultat. W następnej kolejności omówię prace **(C)** i **(A)**, a jako ostatnią zreferuję pracę **(D)**.

Praca (B): W pracy tej przedstawiłem nowe objaśnienie rozkładu słów w języku naturalnym. Aby osiągnąć ten cel, rozpatrzyłem nową klasę kodów opartych na gramatykach (Kieffer i Yang, 2000; Charikar *et al.*, 2005) i zbadałem niektóre teorioinformacyjne własności mocno nieergodycznych procesów stacjonarnych.

Rozkład słów w tekście jest dość dobrze opisywany przez empiryczne prawo Zipfa-Mandelbrota (Zipf, 1965; Mandelbrot, 1954), które głosi, że częstość słowa jest ujemną potęgą jego rangi częstości w zbiorze wszystkich słów w tekście. Trochę wysiłku w stosowanym rachunku prawdopodobieństwa poświęcono wyprowadzaniu tego prawa dla różnych wyidealizowanych modeli. Najbardziej znanym objaśnieniem jest model „małpy przy klawiaturze". W tym objaśnieniu kolejne znaki tekstu są modelowane jako ciąg niezależnych zmiennych losowych o identycznym rozkładzie (IID), które przyjmują wartości zarówno liter jak i spacji, zaś prawo Zipfa-Mandelbrota można wywieść dla ciągów liter oddzielonych spacjami (Mandelbrot, 1954; Miller, 1957). Inne opublikowane objaśnienia bazują np. na procesach multiplikatywnych (Simon, 1955) lub grach (Harremoës i Topsøe, 2001).

W pracy **(B)** rozpatrywałem scałkowaną wersję prawa Zipfa-Mandelbrota, zwyczajowo nazywaną prawem Herdana lub Heapsa w literaturze anglojęzycznej. Prawo to głosi, że liczba różnych słów w tekście jest proporcjonalna do potęgi długości tekstu (Kuraszkiewicz i Łukaszewicz, 1951; Guiraud, 1954; Herdan, 1964; Heaps, 1978). Tezę tę można wywieść z prawa Zipfa-Mandelbrota, zakładając pewną regularność wzrostu tekstu (Khmaladze, 1988; Kornai, 2002).

Objaśnienie prawa Herdana zaproponowane w pracy **(B)** uwzględnia dwie własności języka naturalnego:

1. Słowa, w rozumieniu lingwistycznym, mogą zostać wyróżnione w sposób automatyczny w tekście nawet po usunięciu spacji (Wolff, 1980; de Marcken, 1996; Kit i Wilks, 1999).
2. Teksty, w rozumieniu lingwistycznym, odnoszą się do wielu faktów nie znanych czytelnikowi a priori, ale zwykle odnoszą się do tych faktów w sposób zgodny i powtarzalny.

W pracy **(B)** udowodniłem stwierdzenie, które można wyrazić następująco w sposób nieformalny, zakładając dalej $\beta \in (0, 1)$:

($\mathcal{H}$) Jeżeli tekst długości n opisuje n^β niezależnych faktów w sposób powtarzalny, to tekst ten zawiera co najmniej $n^\beta / \log n$ różnych słów.

Aby przetłumaczyć stwierdzenie ($\mathcal{H}$) na twierdzenie formalne, przyjąłem pewien model matematyczny słów, tekstów i faktów, który jest motywowany lingwistycznie. Podstawowe założenia opisane są poniżej. Zakładam, że $\mathbb{N}$ oznacza zbiór ściśle dodatnich liczb całkowitych. Dla ustalonego przeliczalnego zbioru $\mathbb{X}$, nazywanego alfabetem, oznaczam zbiór niepustych napisów jako $\mathbb{X}^+ := \bigcup_{n\in\mathbb{N}} \mathbb{X}^n$, a zbiór wszystkich napisów jako $\mathbb{X}^* := \mathbb{X}^+ \cup \{\lambda\}$, gdzie λ jest napisem pustym. Długość napisu $w \in \mathbb{X}^*$ będzie oznaczana jako $|w|$.

Liczba różnych słów w tekście: Lingwiści zaobserwowali, że ciągi znaków, które powtarzają się w tekście dostatecznie często, często odpowiadają całym słowom bądź zbitkom takim jak *Nowy Jork*. Szczególnie dobrą odpowiedniość otrzymuje się, gdy granice słów wykrywane są kodem gramatykowym, który minimalizuje długość pewnego kodowania tekstu (Wolff, 1980; de Marcken, 1996; Kit i Wilks, 1999; Nevill-Manning, 1996). Z tego powodu, liczba słów w formalizacji stwierdzenia ($\mathcal{H}$) została zamodelowana przez liczbę różnych symboli nieterminalnych w takim kodowaniu. Poniżej przedstawiam niektóre detale tej konstrukcji.

Kody gramatykowe kompresują teksty, przekształcając je najpierw w specjalne gramatyki, nazywane gramatykami dopuszczalnymi (Kieffer i Yang, 2000), a następnie kodując te gramatyki jako teksty według pewnej prostej metody. Gramatyką *dopuszczalną* nazywa się gramatykę bezkontekstową, która generuje syngieltonowy język $\{w\}$ dla pewnego napisu $w \in \mathbb{X}^*$ (Kieffer i Yang, 2000). Podzbiór takich gramatyk będzie oznaczany jako $\mathcal{G}(w)$, natomiast zbiór gramatyk dopuszczalnych dla wszystkich napisów będzie oznaczany jako $\mathcal{G} := \bigcup_{w\in\mathbb{X}^*} \mathcal{G}(w)$. Jeżeli napis w zawiera powtarzające się podnapisy, to pewna gramatyka $\mathcal{G}(w)$ wynajduje te powtórzenia i reprezentuje napis w w sposób zwięzły.

W gramatyce dopuszczalnej dla każdego symbolu nieterminalnego istnieje dokładnie jedna reguła i można uporządkować symbole nieterminalne w taki sposób, że symbole przepisywane są na ciągi ściśle następujących po sobie symboli (Kieffer i Yang, 2000; Charikar *et al.*, 2005). Zatem taka gramatyka jest dana przez zbiór reguł produkcji

$$\mathsf{G} = \left\{ \begin{array}{l} A_1 \to \alpha_1, \\ A_2 \to \alpha_2, \\ ..., \\ A_n \to \alpha_n \end{array} \right\}, \tag{1}$$

gdzie A_1 jest symbolem startowym, inne A_i nazywam nieterminalami drugorzędnymi, zaś prawe strony reguł produkcji spełniają $\alpha_i \in (\{A_{i+1}, A_{i+2}, ..., A_n\} \cup \mathbb{X})^*$.

Konkretny przykład gramatyki dopuszczalnej to

$$\left\{ \begin{array}{l} A_1 \to A_2A_2A_4A_5\textbf{dear\_children}A_5A_3\textbf{all.} \\ A_2 \to A_3\textbf{you}A_5 \\ A_3 \to A_4\textbf{\_to\_} \\ A_4 \to \textbf{Good\_morning} \\ A_5 \to \textbf{,\_} \end{array} \right\}. \tag{2}$$

Jeżeli rozpoczniemy wyprowadzenie od symbolu A_1 i będziemy podążać za regułami produkcji, otrzymamy tekst piosenki

Good morning to you,
Good morning to you,
Good morning, dear children,
Good morning to all.

W kompresjach dłuższych tekstów, symbole nieterminalne A_i często odpowiadają całym słowom lub zbitkom słów, zwłaszcza jeżeli narzuca się dodatkowy warunek, żeby drugorzędne symbole nieterminalne były definiowane jako ciągi wyłącznie symboli terminalnych (Kit i Wilks, 1999). Taki rodzaj gramatyk nazywam *gramatykami płaskimi.*

Liczba różnych symboli nieterminalnych w gramatyce (1) będzie nazywana *rozmiarem słownika* gramatyki G i oznaczana jako

$$\mathbf{V}[\mathsf{G}] := \operatorname{card}\{A_1, A_2, ..., A_n\} = n. \tag{3}$$

Z kolei funkcja $\Gamma : \mathbb{X}^+ \to \mathcal{G}$ taka, że $\Gamma(w) \in \mathcal{G}(w)$ dla wszystkich $w \in \mathbb{X}^+$ nazywana jest *transformacją gramatykową* (Kieffer i Yang, 2000). W pracy **(B)**, skonstruowałem *dopuszczalnie minimalne* transformacje gramatykowe, które minimalizują długość pewnego kodowania tekstu. Do rozmiaru słownika tych gramatyk odnosi się formalizacja stwierdzenia ($\mathcal{H}$). Ścisła definicja transformacji dopuszczalnie minimalnych jest zbyt techniczna, aby ją tu przedstawiać — można ją znaleźć w pracy **(B)**. Mogę nadmienić, że transformacje dopuszczalnie minimalne przypominają transformacje rozpatrywane przez lingwistów (de Marcken, 1996; Kit i Wilks, 1999) oraz transformacje, które badali Charikar *et al.* (2005), a które nazywam transformacjami minimalnymi w sensie Yanga-Kieffera. W szczególności istnieją transformacje dopuszczalnie minimalne, których obrazami są wyłącznie gramatyki płaskie.

W drugiej kolejności potrzebujemy sformułować model nieskończonego tekstu, który opisuje losowe fakty w sposób powtarzalny. Zarówno tekst jak i zbiór faktów wielokrotnie opisywanych w tekście będą modelowane jako procesy stochastyczne. Ponieważ wprowadziłem nowy model matematyczny języka naturalnego, poświęcam więcej miejsca na jego umotywowanie.

Model tekstów i faktów: Niech $(X_i)_{i\in\mathbb{Z}}$ będzie procesem stochastycznym na przestrzeni probabilistycznej $(\Omega, \mathfrak{J}, P)$, gdzie zmienne $X_i : \Omega \to \mathbb{X}$ przyjmują wartości z przeliczalnego alfabetu $\mathbb{X}$. Ten proces będzie modelować nieskończenie długi tekst, gdzie X_i są kolejnymi jednostkami tekstu. Możemy sobie wyobrażać, że wartościami X_i są znaki (litery, spacje, lub znaki interpunkcyjne), jeżeli zbiór $\mathbb{X}$ jest skończony, bądź słowa lub zdania, jeżeli zbiór $\mathbb{X}$ jest nieskończony. Z kolei, niech $Z_k : \Omega \to \{0, 1\}$, $k \in \mathbb{N}$, będą niezależnymi binarnymi zmiennymi losowymi o rozkładzie jednostajnym. Będziemy zakładać, że wartości Z_k są nie znane a priori czytelnikowi tekstu, ale można się ich z tekstu nauczyć. Zmienne Z_k będą nazywane faktami. Możemy sobie wyobrażać, że wartościami Z_k są wartości logiczne (1=prawda i 0=fałsz) pewnych systematycznie wyliczanych stwierdzeń, które są logicznie niezależne.

W szczególności będziemy zakładać, że wartość każdego faktu Z_k można wywnioskować z półnieskończonego tekstu, jeżeli rozpoczniemy czytanie od arbitralnej pozycji. Założenie to czynimy, aby uwzględnić postulat, że fakty opisywane są w tekście w sposób powtarzalny. W dalszej kolejności notacja $X_{m:n} := (X_i)_{m\le i\le n}$ będzie używana na oznaczenie ciągów zmiennych X_i, nazywanych także blokami. Bloki $X_{m:n}$ modelują teksty skończone. Następująca defnicja ujmuje potrzebną nam własność:

Definicja 1. *Proces stochastyczny $(X_i)_{i\in\mathbb{Z}}$ nazywa się* mocno nieergodycznym*, jeżeli istnieje ciąg niezależnych zmiennych losowych $(Z_k)_{k\in\mathbb{N}}$ o rozkładzie jednostajnym $P(Z_k = 0) = P(Z_k = 1) = \frac{1}{2}$ oraz funkcje $s_k : \mathbb{X}^* \to \{0, 1\}$, $k \in \mathbb{N}$, takie, że*

$$\lim_{n\to\infty} P(s_k(X_{t+1:t+n}) = Z_k) = 1, \qquad \forall t \in \mathbb{Z}, \forall k \in \mathbb{N}. \tag{4}$$

Definicja ta została po raz pierwszy sformułowana w omawianej dalej pracy **(E)**.

Funkcje s_k motywowane są ideą, że istnieje ustalona metoda interpretowania skończonych tekstów w języku naturalnym, aby wywnioskowywać z nich wartości różnych faktów. W żargonie lingwistycznym metodę tę nazywa się kompetencją językową. Dzięki niej dowolny fakt, który jest wzmiankowany w tekście w sposób powtarzalny, może zostać wywnioskowany przez czytelników z tekstu niezależnie od miejsca, w którym rozpoczną czytanie.

Z probabilistycznego punktu widzenia, proces stacjonarny jest mocno nieergodyczny wtedy i tylko wtedy, gdy istnieje ciągła zmienna losowa $Y : \Omega \to (0, 1)$ mierzalna względem σ-algebry niezmienniczej generowanej przez proces. Taka zmienna stanowi przykładu parametru w statystyce bayesowskiej. Na przykład, wzięcie $Y = \sum_{k\in\mathbb{N}} 2^{-k} Z_k$ odpowiada jednostajnemu rozkładowi a priori na Y. Stąd wynika, że każdy stacjonarny proces mocno nieergodyczny jest nieergodyczny, zob. praca **(E)** omawiana dalej.

Aby zilustrować, jak pojęcie procesu mocno nieergodycznego odpowiada intuicjom dotyczącym ludzkiej komunikacji, rozpatrzmy następujący przykład, który nazwałem procesem Santa Fe. Jest on bardzo prosty, a zarazem odmienny od modeli parametrycznych tradycyjnie rozpatrywanych w statystyce. Na chwilę, połóżmy alfabet $\mathbb{X} = \mathbb{N} \times \{0, 1\}$ i wprowadźmy proces $(X_i)_{i\in\mathbb{Z}}$ postaci

$$X_i := (K_i, Z_{K_i}), \tag{5}$$

gdzie procesy $(Z_k)_{k\in\mathbb{N}}$ i $(K_i)_{i\in\mathbb{Z}}$ są probabilistycznie niezależne, a $(K_i)_{i\in\mathbb{Z}}$ jest takim ergodycznym procesem stacjonarnym, że $P(K_i = k) > 0$ dla każej liczby naturalnej $k \in \mathbb{N}$. Przy takich założeniach można pokazać, że $(X_i)_{i\in\mathbb{Z}}$ jest procesem mocno nieergodycznym (por. praca **(E)** dyskutowana dalej).

Zmiennym $X_i = (K_i, Z_{K_i})$ można przydać następującą interpretację lingwistyczną: Wyobraźmy sobie, że $(X_i)_{i\in\mathbb{Z}}$ jest ciągiem kolejnych stwierdzeń wydobytych z nieskończenie długiego tekstu opisującego nieskończony obiekt losowy $(Z_k)_{k\in\mathbb{N}}$ w sposób niesprzeczny. Każde stwierdzenie $X_i = (k, z)$ komunikuje zarówno adres k losowego bitu obiektu $(Z_k)_{k\in\mathbb{N}}$ jak i jego wartość $Z_k = z$. Logiczna niesprzeczność opisu odzwierciedlona jest w następującej własności: Jeżeli dwa stwierdzenia $X_i = (k, z)$

i $X_j = (k', z')$ opisują bity o tym samym adresie ($k = k'$) to zawsze przypisują im tę samą wartość ($z = z'$). Zwróćmy uwagę, że kolekcja faktów $(Z_k)_{k\in\mathbb{N}}$ może być postrzegana jako obiektywny stan świata, który poprzedza tekst $(X_i)_{i\in\mathbb{Z}}$, bądź jako dziedzictwo historyczne, które jest generowane w trakcie tworzenia tekstu i zapamiętywane do generowania kolejnych zmiennych X_i. Model (5) dopuszcza dowolną z tych dwu interpretacji.

W formalizacji stwierdzenia $(\mathcal{H})$ liczba różnych faktów opisywanych przez skończony tekst $X_{1:n}$ została utożsamiona z liczbą zmiennych Z_i, które można przewidzieć z prawdopodobieństwem co najmniej δ przy danym $X_{1:n}$. To znaczy liczbę tę rozumie się jako moc zbioru

$$U_\delta(n) := \{k \in \mathbb{N} : P\left(s_k\left(X_{1:n}\right) = Z_k\right) \geq \delta\}, \tag{6}$$

gdzie $\delta > \frac{1}{2}$. Jak pokazałem w pracy **(B)**, moc zbioru $U_\delta(n)$ jest rzędu n^β dla procesu Santa Fe (5), jeżeli K_i są niezależnymi zmiennymi losowymi o identycznym rozkładzie potęgowym

$$P(K_i = k) = k^{-1/\beta}/\zeta(\beta^{-1}), \tag{7}$$

gdzie $\beta \in (0,1)$, a $\zeta(x) = \sum_{k=1}^{\infty} k^{-x}$ to funkcja zeta. Dla odmiany, moc zbioru $U_\delta(n)$ jest rzędu $\log n$, jeżeli $(X_i)_{i\in\mathbb{Z}}$ jest procesem Bernoulliego z losowym parametrem $Y = \sum_{k\in\mathbb{N}} 2^{-k} Z_k$. Należy zwrócić uwagę, że moc zbioru $U_\delta(n)$ dla danego procesu zależy od doboru funkcji s_k i faktów Z_k. Rozpatrywana dalej formalizacja stwierdzenia $(\mathcal{H})$ jest jednak prawdziwa dla dowolnego wyboru funkcji s_k i faktów Z_k, jeżeli tylko zachodzi warunek (4).

Teraz mogę już przedstawić główny wynik pracy **(B)**. Niech $\mathbf{E}$ będzie operatorem wartości oczekiwanej, zaś $\operatorname{card} A$ niech oznacza moc zbioru A. Używam także następującego pojęcia, które wprowadził Shields (1997):

Definicja 2. *Proces $(X_i)_{i\in\mathbb{Z}}$ nazywany jest* procesem o skończonej energii, *jeżeli*

$$P(X_{t+|w|+1:t+|wu|} = u | X_{t+1:t+|w|} = w) \leq Kc^{|u|} \tag{8}$$

dla wszystkich $t \in \mathbb{Z}$, wszystkich $u, w \in \mathbb{X}^*$, oraz pewnych stałych $0 < c < 1$ i $K > 0$, pod warunkiem, że $P(X_{t+1:t+|w|} = w) > 0$.

Warunek (8) jest spełniony dla dowolnego procesu stochastycznego zakłóconego przez probabilistycznie niezależny szum IID (Shields, 1997).

Moja formalizacja stwierdzenia $(\mathcal{H})$ ma następującą postać:

Twierdzenie 1. *Niech $(X_i)_{i\in\mathbb{Z}}$ będzie stacjonarnym mocno nieergodycznym procesem o skończonej energii nad skończonym alfabetem $\mathbb{X}$. Załóżmy, że zachodzi nierówność*

$$\liminf_{n\to\infty} \frac{\operatorname{card} U_\delta(n)}{n^\beta} > 0 \tag{9}$$

dla pewnych $\beta \in (0,1)$, $\delta \in (\frac{1}{2}, 1)$ i zbiorów (6), gdzie funkcje s_k spełniają (4). Wówczas

$$\limsup_{n\to\infty} \mathbf{E}\left(\frac{\mathbf{V}[\Gamma(X_{1:n})]}{n^\beta (\log n)^{-1}}\right)^p > 0 \tag{10}$$

dla dowolnego $p > 1$ i dowolnej dopuszczalnie minimalnej transformacji gramatykowej $\Gamma : \mathbb{X}^+ \to \mathcal{G}$.

Jak pokazałem w pracach **(A)** i **(B)**, pewien przykład procesu nad alfabetem skończonym, który spełnia przesłankę twierdzenia 1, może zostać skonstruowany przez stacjonarne kodowanie procesu Santa Fe (5) z K_i spełniającymi (7).

Twierdzenie 1 jest ściśle związane z dwoma stwierdzeniami odnoszącymi się do informacji wzajemnej miedzy przyległymi blokami. Dla stacjonarnego procesu $(X_i)_{i\in\mathbb{Z}}$ o wartościach dyskretnych, zdefiniujmy entropię blokową

$$H(n) := H(X_{t+1:t+n}) = -\mathbf{E} \log P(X_{t+1:t+n}), \tag{11}$$

gdzie log jest logarytmem naturalnym. Oznaczmy informację wzajemną między przyległymi blokami jako

$$E(n) := 2H(n) - H(2n) = I(X_{1:n}; X_{n+1:2n}), \tag{12}$$

którą dyskutowali m.in. Crutchfield i Feldman (2003). Moje zainteresowania oryginalnie dotyczyły procesów, dla których $E(n)$ rośnie jak potęga n. I w istocie, w pracy **(B)** udało mi się uzyskać następujące wyniki dotyczące takiego zachowania $E(n)$:

Twierdzenie 2. *Niech $(X_i)_{i\in\mathbb{Z}}$ będzie stacjonarnym procesem mocno nieergodycznym nad skończonym alfabetem $\mathbb{X}$. Załóżmy, że nierówność (9) zachodzi dla pewnych $\beta \in (0,1)$, $\delta \in (\frac{1}{2}, 1)$ i zbiorów (6), gdzie funkcje s_k spełniają (4). Wówczas*

$$\limsup_{n\to\infty} \frac{E(n)}{n^\beta} > 0. \tag{13}$$

Twierdzenie 3. *Niech $(X_i)_{i\in\mathbb{Z}}$ będzie stacjonarnym procesem o skończonej energii nad skończonym alfabetem $\mathbb{X}$. Załóżmy, że nierówność*

$$\liminf_{n\to\infty} \frac{E(n)}{n^\beta} > 0 \tag{14}$$

zachodzi dla pewnego $\beta \in (0,1)$. Wówczas nierówność (10) zachodzi dla dowolnej dopuszczalnie minimalnej transformacji gramatykowej $\Gamma : \mathbb{X}^+ \to \mathcal{G}$.

Chociaż twierdzenie 1 nie wynika z koniunkcji twierdzeń 2 i 3, dowód wszystkich trzech faktów jest prawie jednoczesny. Oprócz pewnych własności rozkładu ergodycznego procesu stacjonanego oraz pewnych własności dopuszczalnie minimalnych transformacji gramatykowych dowód wykorzystuje lemat:

Lemat 1. *Rozpatrzmy funkcję $G : \mathbb{N} \to \mathbb{R}$ taką, że $\lim_{k\to\infty} G(k)/k = 0$ i $G(n) \geq 0$ dla wszystkich oprócz skończenie wielu n. Dla nieskończenie wielu n zachodzi $2G(n) - G(2n) \geq 0$.*

Lemat ten jest źrodłem asymetrii: $\liminf_n$ po stronie przesłanek, względnie $\limsup_n$ po stronie tez. Pomimo wielu prób nie udało mi się tego stwierdzenia wzmocnić. Z kolei założenie procesu o skończonej energii jest potrzebne, aby ograniczyć długość najdłuższego powtórzenia w bloku $X_{1:n}$. Ograniczenie to jest źródłem wyrazu $\log n$ we wzorze (10), zaś źródłem wykładnika $p > 1$ w tym samym wzorze jest nierówność Höldera.

Należy tu także przypomnieć, że hipotezę $E(n) \propto n^\beta$ dla języka naturalnego sformułował Hilberg (1990). Taka była jego interpretacja wykresu entropii warunkowej ze słynnej pracy Shannona (1951) i przypuścił on, że $\beta \approx \frac{1}{2}$. Moim zdaniem twierdzenie 2 wskazuje, że hipotezę Hilberga można umotywować racjonalnie, zaś twierdzenie 3 pokazuje, że hipoteza ta implikuje pewne empiryczne zależności. W istocie wyniki moich wstępnych eksperymentów wskazują, że rozmiar słownika dopuszczalnie minimalnych transformacji gramatykowych jest znacznie większy dla tekstów w języku naturalnym niż dla procesów IID (Dębowski, 2007).

Wyniki osiągnięte w pracy **(B)** są istotne z dwóch powodów. Po pierwsze dostarczają one nowego, bardziej satysfakcjonującego wytłumaczenia zjawiska lingwistycznego, będącego przedmiotem intensywnych badań już od kilkudziesięciu lat. Mianowicie, udało mi się wywieść prawo Herdana przy bardzo ogólnym założeniach mówiących, że teksty opisują pewne fakty w sposób losowy lecz zgodny i że gramatykowa kompresja danych dostarcza algorytmu wyróżniania słów w tekście. Po drugie wyniki z pracy **(B)** rzucają światło na własności dyskretnych procesów stochastycznych z daleką pamięcią. Mianowicie, to samo prawo Herdana zachodzi dla dowolnego stacjonarnego procesu mocno nieergodycznego o skończonej energii spełniającego warunek (9) lub (14). Szczególnym przypadkiem takich procesów są procesy Santa Fe, nowy model procesu z daleką pamięcią. Tu warto określić rolę tych procesów w modelowaniu języka naturalnego. Konstrukcja procesów Santa Fe jest po części motywowana lingwistycznie, lecz daleko im do realistycznych modeli języka naturalnego. O ile można sobie wyobrażać, że tekst w języku naturalnym da się sprowadzić do ciągu kolejnych stwierdzeń postaci (5), o tyle warunek, że zmienne K_i są probabilistycznie niezależne, jest daleko posuniętą idealizacją. Można jednak przypuszczać, że dla zależnych zmiennych K_i warunki (9) lub (14) mogą być także spełnione.

Praca (C): Praca ta dotyczy własności uogólnionych procesów Santa Fe oraz pewnego ich kodowania stacjonarnego nad alfabetem skończonym. Uogólnionym procesem Santa Fe nazywam proces $(X_i)_{i\in\mathbb{Z}}$ postaci

$$X_i = (K_i, Z_{i,K_i}), \tag{15}$$

gdzie procesy $(K_i)_{i\in\mathbb{Z}}$ i $(Z_{ik})_{i\in\mathbb{Z}}$, gdzie $k \in \mathbb{N}$, są niezależne i mają następujący rozkład. Po pierwsze, zmienne K_i są niezależne o identycznym rozkładzie (7). Po drugie, każdy z procesów $(Z_{ik})_{i\in\mathbb{Z}}$ jest stacjonarnym łańcuchem Markowa o rozkładzie brzegowym

$$P(Z_{ik} = 0) = P(Z_{ik} = 1) = 1/2 \tag{16}$$

i prawdopodobieństwach przejścia

$$P(Z_{ik} = 0|Z_{i-1,k} = 1) = P(Z_{ik} = 1|Z_{i-1,k} = 0) = p_k. \tag{17}$$

Szczególnym przypadkiem procesu (15) jest proces Santa Fe (5), który otrzymujemy dla $p_k = 0$ i $(Z_k)_{k\in\mathbb{N}} = (Z_{ik})_{k\in\mathbb{N}}$. Uogólniając proces Santa Fe na przypadek $p_k > 0$ możemy się posiłkować analogią lingwistyczną w następujący sposób. Fakty opisywane w tekstach w sposób powtarzalny można podzielić z grubsza na dwie kategorie: a) fakty o obiektach, które nie zmieniają się w czasie, (jak np. stałe matematyczne lub fizyczne) oraz b) fakty o obiektach, które ewoluują z różną prędkością (jak np. kultura, język, geografia). Dla $p_k = 0$ obiekt $(Z_k)_{k\in\mathbb{N}}$ opisywany przez proces Santa Fe nie ewoluuje bądź, równoważnie, żaden bit Z_k nie jest zapominany po ujawnieniu. Z kolei dla $p_k > 0$ obiekt $(Z_{ik})_{k\in\mathbb{N}}$ opisywany przez proces Santa Fe jest funkcją chwili i, a prawdopodobieństwo, że k-ty bit zmieni wartość, wynosi właśnie p_k.

W pracy **(C)** wykazałem po pierwsze, że uogólniony proces Santa Fe jest mieszający w sensie Cornfeld *et al.* (1982, rozdział 1.§6) dla prawdopodobieństw przejścia różnych od 0 lub 1.

Twierdzenie 4. *Uogólniony proces Santa Fe $(X_i)_{i\in\mathbb{Z}}$ dany wzorem (15) jest mieszający, jeżeli wszystkie $p_k \in (0, 1)$.*

Dowód polega na spostrzeżeniu, że nieskończone produkty bezpośrednie procesów mieszających są mieszające. Jest to proste uogólnienie dobrze znanego faktu dla produktów skończonych (Cornfeld *et al.*, 1982, rozdział 10.§1).

Pokazałem także, że dla uogólnionego procesu Santa Fe informacja wzajemna między przyległymi blokami rośnie potęgowo. Oznaczmy $\mu = P((X_i)_{i\in\mathbb{Z}} \in \cdot)$ i $E_\mu(n) = I(X_{1:n}; X_{n+1:2n})$. Wówczas mamy:

Twierdzenie 5. *Informacja wzajemna między przyległymi blokami $E_\mu(n)$ dla ugólnionego procesu Santa Fe $(X_i)_{i\in\mathbb{Z}}$ danego wzorem (15) spełnia*

$$\limsup_{n\to\infty} \frac{E_\mu(n)}{n^\beta} \le \frac{(2-2^\beta)\Gamma(1-\beta)}{[\zeta(\beta^{-1})]^\beta}. \tag{18}$$

Ograniczenia dolnych granic w szczególnych przypadkach są następujące:

1. *Jeżeli $p_k \le P(K_i = k)$, to*

$$\liminf_{n\to\infty} \frac{E_\mu(n)}{n^\beta} \ge \frac{A(\beta)}{[\zeta(\beta^{-1})]^\beta}, \tag{19}$$

gdzie

$$A(\beta) := \sup_{\delta\in(1/2,1)} (1-\eta(\delta))^\beta \int_{\sqrt{\delta}}^1 \frac{(1-u)^2\, du}{u\,(-\ln u)^{\beta+1}}, \tag{20}$$

a $\eta(\delta)$ jest entropią rozkładu binarnego $(\delta, 1-\delta)$,

$$\eta(\delta) := -\delta\log\delta - (1-\delta)\log(1-\delta).$$

2. *Jeżeli $\lim_{k\to\infty} p_k/P(K_i = k) = 0$, to $E_\mu(n)$ spełnia*

$$\lim_{n\to\infty} \frac{E_\mu(n)}{n^\beta} = \frac{(2-2^\beta)\Gamma(1-\beta)}{[\zeta(\beta^{-1})]^\beta}. \tag{21}$$

Zgodnie z twierdzeniami 4 i 5.1–2 istnieją procesy ergodyczne, dla których warunek (14) jest spełniony. Procesy te jednak nie spełniają pełnej przesłanki twierdzenia 3, gdyż nie są procesami nad alfabetem skończonym.

W następnej kolejności wprowadziłem podobne procesy ergodyczne nad alfabetem skończonym. W tym celu użyłem pewnej transformacji, która przekształca procesy nad alfabetem nieskończonym w procesy nad alfabetem skończonym, a która zachowuje stacjonarność i (nie)ergodyczność procesu i nie zaburza nadmiernie entropii. Nazywam tę transformację kodowaniem stacjonarnym (zmiennej długości). Podobne, acz mniej ogólne konstrukcje rozpatrywali Cariolaro i Pierobon (1977); Gray i Kieffer (1980); Timo *et al.* (2007). Kodowanie stacjonarne zmiennej dlugości składa się ona z dwóch operacji.

Po pierwsze, niech funkcja $f : \mathbb{X} \to \mathbb{Y}^*$, zwana funkcją kodującą, przekształca symbole z alfabetu $\mathbb{X}$ w skończone ciągi nad innym alfabetem $\mathbb{Y}$. Definiujemy rozszerzenie tej funkcji na ciągi podwójnie nieskończone $f^\mathbb{Z} : \mathbb{X}^\mathbb{Z} \to \mathbb{Y}^\mathbb{Z} \cup (\mathbb{Y}^* \times \mathbb{Y}^*)$ jako

$$f^\mathbb{Z}((x_i)_{i\in\mathbb{Z}}) := ...f(x_{-1})f(x_0)\boldsymbol{.}f(x_1)f(x_2)..., \tag{22}$$

gdzie $x_i \in \mathbb{X}$, a wytłuszczona kropka rozdziela symbole zerowy i pierwszy. Wówczas dla procesu stacjonarnego $(X_i)_{i\in\mathbb{Z}}$ na $(\Omega, \mathcal{J}, P)$, gdzie zmienne X_i przyjmują wartości z przestrzeni $(\mathbb{X}, \mathcal{X})$, wprowadzamy proces

$$(Y_i)_{i\in\mathbb{Z}} := f^\mathbb{Z}((X_i)_{i\in\mathbb{Z}}), \tag{23}$$

gdzie zmienne Y_i przyjmują wartości z przestrzeni $(\mathbb{Y}, \mathcal{Y})$, pod warunkiem, że prawa strona jest ciągiem podwójnie nieskończonym.

Druga operacja jest następująca. Transformacja (23) w ogólności nie zachowuje stacjonarności, ale proces $(Y_i)_{i\in\mathbb{Z}}$ jest asymptotycznie średnio stacjonarny (AMS) przy łagodnych warunkach, które są spełnione w dyskutowanym dalej przypadku. Proces $(Y_i)_{i\in\mathbb{Z}}$ nazywa się AMS, gdy dla rozkładu

$$P((Y_i)_{i\in\mathbb{Z}} \in \cdot) = \nu \tag{24}$$

i operacji przesunięcia $T((y_i)_{i\in\mathbb{Z}}) := (y_{i+1})_{i\in\mathbb{Z}}$ istnieje miara stacjonarna

$$\bar{\nu}(A) := \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \nu \circ T^{-i}(A), \tag{25}$$

nazywana średnią stacjonarną miary ν (Gray i Kieffer, 1980). Wygodnie jest założyć, że przestrzeń $(\Omega, \mathcal{J}, P)$ jest dostatecznie bogata, aby istniał na niej proces $(\bar{Y}_i)_{i\in\mathbb{Z}}$ o rozkładzie

$$P((\bar{Y}_i)_{i\in\mathbb{Z}} \in \cdot) = \bar{\nu}. \tag{26}$$

Pomimo że proces $(Y_i)_{i\in\mathbb{Z}}$ nie musi być stacjonarny, proces $(\bar{Y}_i)_{i\in\mathbb{Z}}$ jest stacjonarny i będzie nazywany kodowaniem stacjonarnym (zmiennej długości) procesu $(X_i)_{i\in\mathbb{Z}}$.

Procesy $(X_i)_{i\in\mathbb{Z}}$, $(Y_i)_{i\in\mathbb{Z}}$ oraz $(\bar{Y}_i)_{i\in\mathbb{Z}}$ mają izomorficzne algebry niezmiennicze dla pewnych dogodnych funkcji kodujących, które nazwałem iniekcjami synchronizowalnymi (zob. definicja w pkt. 2 omówienia pracy **(A)**). Na przykład dla nieskończonego alfabetu $\mathbb{X} = \mathbb{N} \times \{0, 1\}$ rozpatrzmy alfabet trójkowy $\mathbb{Y} = \{0, 1, 2\}$ i funkcję kodującą

$$f(k, z) = b(k)z2, \tag{27}$$

gdzie $b(k) \in \{0, 1\}^+$ jest rozwinięciem binarnym liczby naturalnej k pozbawionym początkowej cyfry 1. Funkcja kodująca (27) jest przykładem iniekcji synchronizowalnej. Stąd mamy następujący fakt:

Twierdzenie 6. *Niech $(\bar{Y}_i)_{i\in\mathbb{Z}}$ będzie kodowaniem stacjonarnym otrzymanym przez zastosowanie funkcji kodującej (27) do uogólnionego procesu Santa Fe (15). Proces $(\bar{Y}_i)_{i\in\mathbb{Z}}$ jest ergodyczny, jeżeli wszystkie $p_k \in (0, 1)$, a nieergodyczny, jeżeli któreś $p_k = 0$.*

Rozpatrzmy teraz informację wzajemną między przyległymi blokami dla stacjonarnego kodowania uogólnionego procesu Santa Fe. Przypomnijmy, że $\bar{\nu} = P((\bar{Y}_i)_{i\in\mathbb{Z}} \in \cdot)$ i $E_{\bar{\nu}}(m) = I\left(\bar{Y}_{1:m}; \bar{Y}_{m+1:2m}\right)$. Jako ostatni wynik w pracy **(C)** wykazałem następujący fakt:

Twierdzenie 7. *Niech $(\bar{Y}_i)_{i\in\mathbb{Z}}$ będzie kodowaniem stacjonarnym otrzymanym przez zastosowanie funkcji kodującej (27) do uogólnionego procesu Santa Fe (15). Zdefiniujmy tempo ekspansji $L := \mathbf{E}\,|f(X_i)|$. Informacja wzajemna między przyległymi blokami $E_{\bar{\nu}}(m)$ dla procesu $(\bar{Y}_i)_{i\in\mathbb{Z}}$ spełnia*

$$\limsup_{m\to\infty} \frac{E_{\bar{\nu}}(m)}{m^\beta} \le \frac{1}{L^\beta} \frac{(2 - 2^\beta)\Gamma(1 - \beta)}{[\zeta(\beta^{-1})]^\beta}. \tag{28}$$

Ograniczenia dolnych granice w szczególnych przypadkach są następujące:

1. *Jeżeli $p_k \le P(K_i = k)$, to*

$$\liminf_{m\to\infty} \frac{E_{\bar{\nu}}(m)}{m^\beta} \ge \frac{1}{L^\beta} \frac{A(\beta)}{[\zeta(\beta^{-1})]^\beta}, \tag{29}$$

gdzie $A(\beta)$ jest zdefiniowane w (20).

2. *Jeżeli $\lim_{k\to\infty} p_k/P(K_i = k) = 0$, to*

$$\lim_{m\to\infty} \frac{E_{\bar{\nu}}(m)}{m^\beta} = \frac{1}{L^\beta} \frac{(2 - 2^\beta)\Gamma(1 - \beta)}{[\zeta(\beta^{-1})]^\beta}. \tag{30}$$

Twierdzenie 7 wynika z twierdzenia 5 poprzez ograniczenia Chernoffa.

Zgodnie z twierdzeniem 7 istnieją procesy, dla których przesłanka twierdzenia 3 z pracy **(B)** jest spełniona. Z kolei z twierdzenia 6 procesy te nie spełniają przesłanki twierdzenia 1 z pracy **(B)**. Wyniki te są istotne, gdyż dowodzą, że istnieją procesy o potęgowo rosnącej informacji między przyległymi blokami, które są ergodyczne (a więc nie są mocno nieergodyczne), a których konstrukcja jest także motywowana lingwistycznie.

Praca (A): Praca ta zawiera wyniki pomocnicze, na które powoływałem się później w pracach **(B)** i **(C)**. W pracy tej rozpatrywałem problem kodowania stacjonarnego zmiennej długości określonego wzorami (23) i (26), które w mniej ogólnych sformułowaniach badali wcześniej Cariolaro i Pierobon (1977), Gray i Kieffer (1980) oraz Timo *et al.* (2007). Motywującym zastosowaniem było pokazanie, że istnieje proces, dla którego przesłanka twierdzenia 1 z pracy **(B)** jest spełniona.

Główne wyniki pracy **(A)** są następujące:

1. Przypomnijmy, że miara ν na $(\mathbb{Y}^\mathbb{Z}, \mathcal{Y}^\mathbb{Z})$ nazywana jest *asymptotycznie średnio stacjonarną (AMS)*, jeżeli istnieją granice (25) dla każdego $A \in \mathcal{Y}^\mathbb{Z}$, zob. Gray i Kieffer (1980); Krengel (1985). Pokazałem, że miara $\mu \circ \left(f^\mathbb{Z}\right)^{-1}$ jest AMS dla miary AMS μ, jeżeli tempo ekspansji $\bar{l}(x^\mathbb{Z}) := \lim_n n^{-1} \sum_{i=1}^n |f(x_i)|$ należy do przedziału $(0, \infty)$ μ-prawie wszędzie. Ten wynik ugólnia wynik Graya i Kieffera (1980, przykład 6), gdzie pokazano, że $\mu \circ \left(f^\mathbb{Z}\right)^{-1}$ jest AMS pod warunkiem, że miara μ jest stacjonarna, $f^\mathbb{Z}$ jest iniekcją i wartość oczekiwana $\bar{l}$ względem μ jest skończona.
2. Funkcję $\pi : \mathbb{X}^\mathbb{Z} \to \mathbb{Y}^\mathbb{Z}$ nazywam *synchronizowalną iniekcją*, jeżeli π jest iniekcją i $T^i\pi(x^\mathbb{Z}) = \pi(b^\mathbb{Z})$ dla $i \in \mathbb{Z}$ implikuje $T^j x^\mathbb{Z} = b^\mathbb{Z}$ dla pewnego $j \in \mathbb{Z}$. Pokazałem, że jeżeli $f^\mathbb{Z}$ jest synchronizowalną iniekcją, to algebry niezmiennicze procesów $(X_i)_{i\in\mathbb{Z}}$ i $f^\mathbb{Z}((X_i)_{i\in\mathbb{Z}})$ pozostają w zależności wzajemnie jednoznacznej i rozkłady tych procesów na tych algebrach są identyczne.
3. Zbiór ciągów $\mathcal{L} \subset \mathbb{Y}^*$ nazywany jest (i) *bezprzedrostkowym*, jeżeli $w \neq zs$ dla $w, z \in \mathcal{L}$ oraz $s \in \mathbb{Y}^+$, (ii) *bezprzyrostkowym*, jeżeli $w \neq sz$ dla $w, z \in \mathcal{L}$ oraz $s \in \mathbb{Y}^+$, (iii) *bezrostkowym*, jeżeli jest zarówno bezprzedrostkowy jak bezprzyrostkowy, (iv) *zupełnym*, jeżeli spełnia równość Krafta $\sum_{w\in\mathcal{L}} (\operatorname{card} \mathbb{Y})^{-|w|} = 1$, gdzie $\operatorname{card} \mathbb{Y}$ jest mocą zbioru $\mathbb{Y}$. Funkcja $f : \mathbb{X} \to \mathbb{Y}^*$ nazywana jest *(zupełną) bezprzedrostkową/bezprzyrostkową/bezrostkową*, jeżeli f jest iniekcją, a obraz $f(\mathbb{X})$ jest odpowiednio (zupełny) bezprzedrostkowy/bezprzyrostkowy/bezrostkowy. Dla skończonego $f(\mathbb{X})$, f nazywana jest *skończoną*. Pokazałem, że miara $\nu \circ f^\mathbb{Z}$ jest stacjonarna względnie AMS dla stacjonarnej względnie AMS miary ν, jeżeli f jest zupełna bezrostkowa.
4. Przypomnijmy definicję miary o skończonej energii—definicja 2 z pracy **(B)**. Pokazałem, że średnia stacjonarna $\bar{\mu}$ ma skończoną energię, jeżeli μ ma skończoną energię. Ponadto $\mu \circ \left(f^\mathbb{Z}\right)^{-1}$ ma skończoną energię, jeżeli miara μ ma skończoną energię, a funkcja kodująca f jest skończona bezprzedrostkowa.
5. Oznaczmy zbiory cylindryczne jako $[u] := \left\{x^\mathbb{Z} : x^{|u|} = u\right\}$. Entropia blokowa miary μ na $(\mathbb{X}^\mathbb{Z}, \mathcal{X}^\mathbb{Z})$, to funkcja
$$H_\mu(i; n) := -\sum_{u\in\mathbb{X}^n} \mu(T^{-i}[u]) \log \mu(T^{-i}[u]). \tag{31}$$
Używam także skrótu $H_\mu(n) := H_\mu(0; n)$. Pokazałem, że dla iniekcji o stałej długości $f : \mathbb{X} \to \mathbb{Y}^K$, skończonego $\mathbb{X}$ oraz $\nu = \mu \circ \left(f^\mathbb{Z}\right)^{-1}$, entropie blokowe $H_{\bar{\mu}}(n)$ i $H_{\bar{\nu}}(nK)$ średnich stacjonarnych nie różnią więcej niż o stałą.

W następnej kolejności zastosowałem niektóre z tych wyników, aby pokazać, że stacjonarne kodowanie uzyskane z zastosowania funkcji kodującej (27) do procesu Santa Fe (5) spełnia przesłankę twierdzenia 1. Innymi słowy, dowiodłem następującego faktu:

Twierdzenie 8. *Niech $\mu = P((X_i)_{i\in\mathbb{Z}} \in \cdot)$ będzie rozkładem procesu (5) spełniającego (7), gdzie $\zeta(\beta^{-1}) > 4$. Rozpatrzmy funkcję kodującą (27). Proces $(\bar{Y}_i)_{i\in\mathbb{Z}}$ o rozkładzie równym średniej stacjonarnej $\bar{P}((\bar{Y}_i)_{i\in\mathbb{Z}} \in \cdot) = \overline{\mu \circ (f^\mathbb{Z})^{-1}}$ spełnia następujące warunki:*

1. *$(\bar{Y}_i)_{i\in\mathbb{Z}}$ jest procesem nad skończonym alfabetem $\mathbb{Y} = \{0, 1, 2\}$.*

2. $(\bar{Y}_i)_{i\in\mathbb{Z}}$ *jest stacjonarny.*
3. $(\bar{Y}_i)_{i\in\mathbb{Z}}$ *ma własność skonczonej energii.*
4. *Istnieją niezależne zmienne losowe* $(\bar{Z}_k)_{k\in\mathbb{N}}$ *o identycznym rozkładzie* $\bar{P}(\bar{Z}_k = z) = 1/2$, $z \in \{0,1\}$ *mierzalne względem* σ*-algebry niezmienniczej procesu* $(\bar{Y}_i)_{i\in\mathbb{Z}}$ *takie, że*

$$\liminf_{n\to\infty} \frac{\operatorname{card} \bar{U}_\delta(n)}{n^\beta} > 0 \tag{32}$$

zachodzi dla $\delta \in (1/2, 1)$ *i zbiorów* $\bar{U}_\delta(n) := \left\{k \in \mathbb{N} : \bar{P}\left(\bar{s}_k\left(\bar{Y}_{1:n}\right) = \bar{Z}_k\right) \geq \delta\right\}$, *gdzie funkcje* $\bar{s}_k$ *spelniają*

$$\lim_{n\to\infty} \bar{P}\left(\bar{s}_k\left(\bar{Y}_{i+1:i+n}\right) = \bar{Z}_k\right) = 1, \quad \forall i \in \mathbb{Z}. \tag{33}$$

W istocie zmienne $\bar{Z}_k$ można skonstruować jako $\bar{Z}_k = \bar{s}_k((\bar{Y}_i)_{i\in\mathbb{Z}})$, gdzie

$$\bar{s}_k(w) := \begin{cases} 0, & \text{jeżeli } 2b(k)02 \sqsubseteq w \text{ i } 2b(k)12 \not\sqsubseteq w, \\ 1, & \text{jeżeli } 2b(k)12 \sqsubseteq w \text{ i } 2b(k)02 \not\sqsubseteq w, \\ 2, & \text{w pozostałych przypadkach} \end{cases} \tag{34}$$

dla $w \in \mathbb{Y}^{\mathbb{Z}} \cup \mathbb{Y}^*$, gdzie piszę $a \sqsubseteq b$, jeżeli a jest podciągiem b.

Wyniki z pracy **(A)** są istotne, gdyż pokazują szereg intuicyjnych własności spełnionych przez kodowanie stacjonarne zmiennej długości. Własności te są przydatne w wykazaniu, że istnieje proces, dla którego przesłanka twierdzenia 1 z pracy **(B)** jest spełniona.

Praca (D): Praca ta dotyczy ukrytych procesów Markowa o nieskończonej entropii nadwyżkowej. Zagadnienie to wcześniej badali Travers i Crutchfield (2011). Preliminaria są następujące. Po pierwsze, niech $(Y_i)_{i\in\mathbb{Z}}$ będzie stacjonarnym procesem Markowa (rzędu 1) na przestrzeni $(\Omega, \mathcal{J}, P)$, gdzie zmienne $Y_i : \Omega \to \mathbb{Y}$ przyjmują wartości z przeliczalnie nieskończonego alfabetu $\mathbb{Y}$. Ten proces nazywany będzie procesem ukrytym. Następnie, dla funkcji $f : \mathbb{Y} \to \mathbb{X}$, gdzie alfabet $\mathbb{X} = \{0, 1, ..., D-1\}$ jest skończony, konstruuję proces $(X_i)_{i\in\mathbb{Z}}$, gdzie

$$X_i = f(Y_i). \tag{35}$$

Proces $(X_i)_{i\in\mathbb{Z}}$ będzie nazywany procesem obserwowalnym. W następnej kolejności rozpatrzmy informację wzajemną między przyległymi blokami dla procesu obserwowalnego,

$$E(n) := I(X_{-n+1:0}; X_{1:n}). \tag{36}$$

Interesowały mnie procesy dla których entropia nadwyżkowa $E = \lim_{n\to\infty} E(n)$ jest nieskończona, zaś $E(n)$ rozbiega potęgowo. Chciałem pokazać, że taki efekt jest możliwy dla bardzo prostych ukrytych procesów Markowa. Zwróćmy uwagę, że z nierówności przetwarzania informacji dla procesu Markowa $(Y_i)_{i\in\mathbb{Z}}$ otrzymujemy

$$E(n) \leq I(Y_{-n+1:0}; Y_{1:n}) = I(Y_0; Y_1) \leq H(Y_0). \tag{37}$$

Zatem informacja wzajemna między przyległymi blokami $E(n)$ może rozbiegać jedynie, gdy entropia stanu ukrytego jest nieskończona. Aby osiągnąć ten efekt, zmienna ukryta Y_0 musi przyjmować nieskończoną liczbę wartości.

Wprowadzę teraz zaproponowaną przeze mnie klasę przykładów. Załóżmy, że stany ukryte σ_{nk} można pogrupować w poziomy

$$T_n := \{\sigma_{nk}\}_{1\leq k\leq r(n)}, \tag{38}$$

które zawierają jednakowo prawdopodobne wartości. Ponadto załóżmy, że wskaźnik poziomu

$$N_i := n \iff Y_i \in T_n \tag{39}$$

ma rozkład

$$P(N_i = n) = \frac{C}{n \log^\alpha n}. \tag{40}$$

Dla $\alpha \in (1, 2]$ entropia $H(N_i)$ jest nieskończona i taka sama jest entropia $H(Y_i) \geq H(N_i)$, gdyż N_i jest funkcją zmiennej Y_i. W dalszej kolejności będziemy rozpatrywać ten szczególny rozkład Y_i.

Jak pokazałem w pracy **(D)**, tempo wzrostu informacji między blokami $E(n)$ jest ograniczone w terminach wykładnika α z równości (40). Piszmy $f(n) = O(g(n))$, jeżeli $f(n) \leq Kg(n)$ dla $K > 0$, oraz $f(n) = \Theta(g(n))$, jeżeli $K_1 g(n) \leq f(n) \leq K_2 g(n)$ dla $K_1, K_2 > 0$.

Twierdzenie 9. *Przypuśćmy, że $\mathbb{Y} = \{\sigma_{nk}\}_{1\leq k\leq r(n), n\geq 2}$, gdzie funkcja $r(n)$ spełnia $r(n) = O(n^p)$ dla $p \in \mathbb{N}$. Ponadto załóżmy, że*

$$P(Y_i = \sigma_{nk}) = \frac{1}{r(n)} \cdot \frac{C}{n \log^\alpha n}, \tag{41}$$

gdzie $\alpha \in (1, 2]$ i $C^{-1} = \sum_{n=2}^{\infty} (n \log^\alpha n)^{-1}$. Wówczas

$$E(n) = \begin{cases} O\left(n^{2-\alpha}\right), & \alpha \in (1, 2), \\ O\left(\log n\right), & \alpha = 2. \end{cases} \tag{42}$$

Interesujące jest pytanie, czy istnieją ukryte procesy Markowa, które osiągają górne ograniczenie wyprowadzone w twierdzeniu 9. Jeżeli tak, to czy te procesy mogą być ergodyczne. Odpowiedź na oba pytania jest twierdząca i przedstawiłem kilka przykładów takich procesów.

Pierwszy przykład, który przedstawiłem, jest nieergodyczny, a informacja wzajemna rozbiega wolniej niż byłoby to oczekiwane z twierdzenia 9.

Przykład 1. (Heavy Tailed Periodic Mixture I) *Przykład ten wprowadzili Travers i Crutchfield (2011). Połóżmy $\mathbb{Y} = \{\sigma_{nk}\}_{1\leq k\leq r(n), n\geq 2}$, gdzie $r(n) = n$. Następnie definiujemy prawdopodobieństwa przejścia*

$$P(Y_{i+1} = \sigma_{nk} | Y_i = \sigma_{ml}) = \begin{cases} \mathbf{1}\{n = m, k = l + 1\}, & 1 \leq l \leq m - 1, \\ \mathbf{1}\{n = m, k = 1\}, & l = m. \end{cases} \tag{43}$$

Widać, że graf przejścia procesu $(Y_i)_{i\in\mathbb{Z}}$ składa się z rozłącznych cykli na poziomach T_n. Rozkład stacjonarny procesu Markowa nie jest jednoznaczny, a proces jest nieergodyczny, jeżeli więcej niż jeden cykl ma dodatnie prawdopodobieństwo. Załóżmy dalej, że rozkład cykli dany jest wzorem (40), a więc rozkład brzegowy zmiennych Y_i równy jest (41). Ponadto definiujemy proces obserwowalny jako

$$X_i = \begin{cases} 0, & Y_i = \sigma_{nk},\ 1 \leq k \leq n - 1, \\ 1, & Y_i = \sigma_{nn}. \end{cases} \tag{44}$$

W powyższym przykładzie wskaźnik poziomu N_i ma nieskończoną entropię i jest mierzalny względem algebry niezmienniczej procesu obserwowalnego $(X_i)_{i\in\mathbb{Z}}$. Stąd $E(n)$ dąży do nieskończoności z rozkładu ergodycznego entropii nadwyżkowej (dyskutowanego w omawianej dalej pracy **(E)**). Bardziej dokładne oszacowanie informacji wzajemnej między blokami podaję poniżej.

Twierdzenie 10. *Dla przykładu 1 zachodzi*

$$E(n) = \begin{cases} \Theta(\log^{2-\alpha} n), & \alpha \in (1,2), \\ \Theta(\log\log n), & \alpha = 2. \end{cases} \tag{45}$$

Następny przykład jest także nieergodyczny, ale tempo wzrostu informacji wzajemnej osiąga górne ograniczenie. Wydaje się, że dzieje się tak dlatego, że informacja o stanie ukrytym kodowana jest w procesie obserwowalnym w sposób bardziej zwięzły.

Przykład 2. (Heavy Tailed Periodic Mixture II) *Połóżmy* $\mathbb{Y} = \{\sigma_{nk}\}_{1\le k\le r(n), n\ge 2}$, *gdzie* $r(n) = s(n)$ *jest długością rozwinięcia binarnego liczby* n. *Następnie definiujemy prawdopodobieństwa przejścia*

$$P(Y_{i+1} = \sigma_{nk}|Y_i = \sigma_{ml}) = \begin{cases} \mathbf{1}\{n = m, k = l+1\}, & 1 \le l \le s(m) - 1, \\ \mathbf{1}\{n = m, k = 1\}, & l = s(m). \end{cases} \tag{46}$$

Jak poprzednio graf przejścia procesu $(Y_i)_{i\in\mathbb{Z}}$ *składa się z rozłącznych cykli na poziomach* T_n. *Jak poprzednio zakładamy rozkład cykli (40) i rozkład brzegowy (41). Ponadto niech* $b(n,k)$ *będzie* k*-tą cyfrą rozwinięcia binarnego liczby* n. *(Mamy* $b(n,1) = 1$.*) Proces obserwowalny definiujemy jako*

$$X_i = \begin{cases} 2, & Y_i = \sigma_{n1}, \\ b(n,k), & Y_i = \sigma_{nk}, 2 \le k \le s(n). \end{cases} \tag{47}$$

Twierdzenie 11. *Dla przykładu 2 zachodzi*

$$E(n) = \begin{cases} \Theta(n^{2-\alpha}), & \alpha \in (1,2), \\ \Theta(\log n), & \alpha = 2. \end{cases} \tag{48}$$

W trzecim przykładzie tempo wzrostu informacji wzajemnej również ograniczenie górne, lecz proces jest dodatkowo nieergodyczny. Proces ten przypomina proces Branching Copy (BC), który wprowadzili Travers i Crutchfield (2011). Pomiędzy procesem BC a procesem zaproponowanym przeze mnie zachodzą trzy istotne różnice. Po pierwsze, dyskutuję prostszą nieunifilarną prezentację procesu, a nie bardziej skomplikowaną unifilarną. Po drugie, używam ciągu separatorów $(s(m)+1) \times 3$ w procesie obserwowalnym. Po trzecie, kładę nieco inne prawdopodobieństwa przejścia, aby otrzymać prostszy rozkład stacjonarny. Wszystkie te zmiany prowadzą do prostszych oszacowań informacji wzajemnej.

Przykład 3. (Heavy Tailed Mixing Copy) *Połóżmy* $\mathbb{Y} = \{\sigma_{nk}\}_{1\le k\le r(n), n\ge 2}$, *gdzie* $r(n) = 3s(n)$ *a* $s(n)$ *jest długością rozwinięcia binarnego liczby* n. *Następnie definiujemy prawdopodobieństwa przejścia*

$$P(Y_{i+1} = \sigma_{nk}|Y_i = \sigma_{ml}) = \begin{cases} \mathbf{1}\{n = m, k = l+1\}, & 1 \le l \le r(m) - 1, \\ p(n)\mathbf{1}\{k = 1\}, & l = r(m), \end{cases} \tag{49}$$

gdzie

$$p(n) = \frac{1}{r(n)} \cdot \frac{D}{n \log^\alpha n}, \tag{50}$$

a $D^{-1} = \sum_{n=2}^\infty (r(n) \cdot n \log^\alpha n)^{-1}$. *Tym razem poziomy* T_n *komunikują się przez przejścia* $\sigma_{mr(m)} \to \sigma_{n1}$, *które zdarzają się z prawdopodobieństwami* $p(n)$. *Graf przejścia procesu* $(Y_i)_{i\in\mathbb{Z}}$ *jest mocno spójny i* *istnieje jednoznaczny rozkład stacjonarny. A zatem proces jest ergodyczny. Łatwo sprawdzić, że rozkładem stacjonarnym jest (41), a więc poziomy mają rozkład dany przez (40). Jak poprzednio, niech $b(n, k)$ oznacza k-tą cyfrę rozwinięcia binarnego liczby n. Proces obserwowalny definiujemy jako*

$$X_i = \begin{cases} 2, & Y_i = \sigma_{n1}, \\ b(n,k), & Y_i = \sigma_{nk},\ 2 \le k \le s(n), \\ 3, & Y_i = \sigma_{nk},\ s(n)+1 \le k \le 2s(n)+1, \\ b(n, k-2s(n)), & Y_i = \sigma_{nk},\ 2s(n)+2 \le k \le 3s(n). \end{cases} \tag{51}$$

Twierdzenie 12. *Dla przykładu 3, $E(n)$ spełnia (48).*

Podsumowanie: Podsumowując moje wyniki, chciałbym je następująco skomentować. Potęgowy wzrost informacji wzajemnej między przyległymi blokami był poprzednio uznawany za charakterystyczną cechę procesów stochastycznych, które modelują układy złożone, takie jak teksty w języku naturalnym (Hilberg, 1990; Bialek *et al.*, 2001; Crutchfield i Feldman, 2003). Jednakże przykłady ukrytych procesów Markowa skonstruowane w pracy **(D)** cechują się dość prostymi prawdopodobieństwami przejścia. W konsekwencji można wątpić, czy potęgowy wzrost informacji wzajemnej jest dostatecznym powodem, aby nazywać dany proces stochastyczny modelem układu złożonego, nawet jeżeli ograniczyć się do procesów nad alfabetem skończonym. Bazując na moim doświadczeniu z innymi procesami o szybko rosnącej informacji wzajemnej między przyległymi blokami, takimi jak procesy Santa Fe (5) i (15), które są bardziej motywowane lingwistycznie, sądzę, że nieskończona entropia nadwyżkowa jest tylko jednym z koniecznych warunków. Rozpoznanie innych warunków identyfikujących modele stochastyczne układów złożonych jest kwestią dalszych interdyscyplinarnych badań. Przypuszczam, że warunki te zależą od konkretnego modelowanego systemu.

5. Pozostałe osiągnięcia

Moje zainteresowania naukowe spina zainteresowanie losowymi aspektami języka. Obejmują one szeroki obszar: od lingwistyki komputerowej i kwantytatywnej (czyli badania statystycznych własności tekstów w języku naturalnym) przez teorię informacji do teorii procesów stochastycznych. W niniejszym autoreferacie skupiam się na moich osiągnięciach matematycznych, więc jako pozostałe osiągnięcia przedkładam następujące prace dodatkowe:

(E) Ł. Dębowski, (2009). A general definition of conditional information and its application to ergodic decomposition. *Statistics & Probability Letters*, **79**:1260–1268.

(F) Ł. Dębowski, (2012). On Bounded Redundancy of Universal Codes. *Statistics and Probability Letters*, **82**:2068-2071.

(G) Ł. Dębowski, (2009). Computable Bayesian Compression for Uniformly Discretizable Statistical Models. [w:] R. Gavalda et al., eds., *Proceedings of the 20th International Conference on Algorithmic Learning Theory, ALT 2009, Porto, Portugal, October 3-5, Lecture Notes in Artificial Intelligence 5809.* (53–67)

(H) Ł. Dębowski, (2013). Information Theory and Statistics. Institute of Computer Science, Polish Academy of Sciences.

Prace te będę omawiać w kolejności ich wymienienia (niechronologicznej).

Praca (E): Praca ta poświęcona jest uogólnieniu pojęcia warunkowej informacji wzajemnej na przypadek dowolnych σ-ciał oraz zastosowaniu tej definicji do rozkładu ergodycznego procesu stacjonarnego. Definicję warunkowej informacji wzajemnej dla regularnego rozkładu warunkowego podali

Dobrušin (1959) i Pinsker (1960), a na przypadek dowolnych σ-ciał uogólnił ją Wyner (1978). Nieświadom pracy Wynera, w pracy **(E)** zaproponowałem definicję identyczną z jego definicją i wykazałem jej różne własności. Zaproponowana definicja ma postać:

Definicja 3. *Dla skończonych ciał $\mathcal{A}'$ i $\mathcal{B}'$ na przestrzeni zdarzeń Ω i miary prawdopodobieństwa P na $\mathcal{A}' \vee \mathcal{B}'$, określmy informację wzajemną jako*

$$I_P(\mathcal{A}'; \mathcal{B}') := \sum_{i=1}^{I} \sum_{j=1}^{J} P(A_i \cap B_j) \log \frac{P(A_i \cap B_j)}{P(A_i)P(B_j)},$$

zakładając, że ciała $\mathcal{A}'$ i $\mathcal{B}'$ są generowane odpowiednio przez rozbicia $\{A_i\}_{i=1}^{I}$ i $\{B_j\}_{j=1}^{J}$ zbioru Ω.

Następnie rozpatrzmy przestrzeń probabilistyczną $(\Omega, \mathcal{J}, P)$. Dla dowolnego ciała $\mathcal{C}$ i skończonych ciał $\mathcal{A}'$ i $\mathcal{B}'$, gdzie $\mathcal{A}', \mathcal{B}', \mathcal{C} \subset \mathcal{J}$, definiujemy warunkową informację wzajemną

$$I(\mathcal{A}'; \mathcal{B}'||\mathcal{C}) := I_{P(\cdot||\mathcal{C})}(\mathcal{A}'; \mathcal{B}'),$$

gdzie $P(E||\mathcal{C})$ jest warunkowym prawdopodobieństwem zdarzenia $E \in \mathcal{J}$ względem najmniejszego σ-ciała zawierającego $\mathcal{C}$.

Średnią warunkową informację wzajemną (w skrócie: CMI) pomiędzy dowolnymi ciałami $\mathcal{A}$ i $\mathcal{B}$ względem ciała $\mathcal{C}$ definiujemy jako

$$I(\mathcal{A}; \mathcal{B}|\mathcal{C}) := \sup \mathbf{E}\, I(\mathcal{A}'; \mathcal{B}'||\mathcal{C}), \tag{52}$$

gdzie kres górny jest wzięty po wszystkich skończonych ciałach $\mathcal{A}' \subset \mathcal{A}$ i $\mathcal{B}' \subset \mathcal{B}$.

Następnie wykazałem różne własności CMI. Piszmy $\mathcal{B}_n \uparrow \mathcal{B}$ dla ciągu ciał $(\mathcal{B}_n)_{n\in\mathbb{N}}$ takich, że $\mathcal{B}_1 \subset \mathcal{B}_2 \subset ... \subset \mathcal{B}$ i $\bigcup_{n\in\mathbb{N}} \mathcal{B}_n = \mathcal{B}$. ($\mathcal{B}$ nie musi być σ-ciałem.) Punkt 2. poniższego twierdzenia, udowodnionego w pracy **(E)**, udowodnił także Wyner (1978):

Twierdzenie 13. *Niech $\mathcal{A}$, $\mathcal{B}$, $\mathcal{B}_n$ oraz $\mathcal{C}$ będą podciałami $\mathcal{J}$.*

1. $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) = I(\mathcal{B}; \mathcal{A}|\mathcal{C})$;
2. $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) \ge 0$, *gdzie równość zachodzi wtedy i tylko wtedy, gdy* $P(A \cap B||\mathcal{C}) = P(A||\mathcal{C})P(B||\mathcal{C})$ *prawie na pewno dla wszystkich* $A \in \mathcal{A}$ *i* $B \in \mathcal{B}$;
3. $I(\mathcal{A}; \mathcal{B}_1|\mathcal{C}) \le I(\mathcal{A}; \mathcal{B}_2|\mathcal{C})$, *jeżeli* $\mathcal{B}_1 \subset \mathcal{B}_2$;
4. $I(\mathcal{A}; \mathcal{B}_n|\mathcal{C}) \uparrow I(\mathcal{A}; \mathcal{B}|\mathcal{C})$ *dla* $\mathcal{B}_n \uparrow \mathcal{B}$.

Ważną własnością definicji (52) jest to, że wartość CMI nie zmienia się, kiedy ciała są rozszerzane do σ-ciał (bądź jakichkolwiek pośrednich ciał). Ciało nazywamy *zupełnym* jeżeli zawiera wszystkie zbiory zewnętrznej P-miary 0. Niech $\sigma(\mathcal{A})$ oznacza przecięcie wszystkich zupełnych σ-ciał zawierających $\mathcal{A}$. Kolejny wynik z pracy **(E)** to:

Twierdzenie 14. *Niech $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ i $\mathcal{D}$ będą podciałami $\mathcal{J}$.*

1. $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) = I(\mathcal{A}; \sigma(\mathcal{B})|\mathcal{C})$ *oraz* $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) = I(\mathcal{A}; \mathcal{B}|\sigma(\mathcal{C}))$;
2. $I(\mathcal{A}; \mathcal{B} \vee \mathcal{C}|\mathcal{D}) = I(\mathcal{A}; \mathcal{C}|\mathcal{D}) + I(\mathcal{A}; \mathcal{B}|\mathcal{C} \vee \mathcal{D})$.

Punkt 2. powyższego twierdzenia także udowodnił Wyner (1978).

W pracy **(E)** przedyskutowałem także kilka innych wyników, które wykraczają poza zakres pracy Wynera (1978). W szczególności wyprowadziłem wzór na rozkład ergodyczny entropii nadwyżkowej. Preliminaria są następujące. Proces nazywa się przeliczalnie generowanym, jeżeli σ-ciało procesu jest generowane przez zbiór przeliczalny. Załóżmy, że proces $(X_k)_{k\in\mathbb{Z}}$ jest przeliczalnie generowany i stacjonarny. Wówczas entropia nadwyżkowa to informacja wzajemna $E = \lim_{n\to\infty} I(X_{-n:0}; X_{1:n})$, zaś intensywność entropii to entropia warunkowa $h := \lim_{n\to\infty} H(X_1|X_{-n:0})$. Dla rozkładu procesu $\mu = P((X_k)_{k\in\mathbb{Z}} \in \cdot)$ wprowadzam jawną parametryzację $E_\mu := E$ i $h_\mu := h$. Ponadto zgodnie z twierdzenie o rozkładzie ergodycznym istnieje losowa miara ergodyczna

$$F := \mu(\cdot||\mathcal{I})((X_k)_{k\in\mathbb{Z}}),$$

gdzie $\mathcal{I}$ jest σ-algebrą niezmienniczą (Kallenberg, 1997, twierdzenie 9.10). Niech $\mathcal{F} \subset \mathcal{J}$ będzie najmniejszym σ-ciałem, względem którego F jest mierzalna. Definiuję entropię tego σ-ciała jako samo-informację: $H(\mathcal{F}) := I(\mathcal{F}; \mathcal{F})$. Rozkład ergodyczny entropii nadwyżkowej ma postać

$$E = H(\mathcal{F}) + \mathbf{E}\, E_F. \tag{53}$$

Analogicznego rozkładu ergodycznego intensywności entropii

$$h = \mathbf{E}\, h_F, \tag{54}$$

dla procesów o skończonej liczbie wartości dowiedli Gray i Davisson (1974).

Ostatnia grupa wyników pracy **(E)** dotyczy entropii zdefiniowanej jako samo-informacja, tzn. $H(\mathcal{A}) := I(\mathcal{A}; \mathcal{A})$. Od R. Bradleya uzyskałem dowód, że $H(\mathcal{A}) = \infty$ chyba, że ciało $\mathcal{A}$ jest czysto atomowe. Ponadto wyprowadziłem entropijny analogon twierdzenia 13. Analogicznie do notacji $\mathcal{B}_n \uparrow \mathcal{B}$, będę używał notacji $\mathcal{B}_n \downarrow \mathcal{B}$ w przypadku, gdy $\mathcal{B}_1 \supset \mathcal{B}_2 \supset ... \supset \mathcal{B}$ i $\bigcap_{n\in\mathbb{N}} \mathcal{B}_n = \mathcal{B}$. Twierdzenie dla entropii ma postać:

Twierdzenie 15. *Niech $\mathcal{A}$, $\mathcal{B}$ i $\mathcal{B}_n$ będą podciałami $\mathcal{J}$.*

1. $H(\mathcal{A}) = 0$ *wtedy i tylko wtedy, gdy $\mathcal{A}$ is* trywialne*, tzn., gdy $P(A) \in \{0, 1\}$ dla wszystkich $A \in \mathcal{A}$;*
2. $H(\mathcal{A}|\mathcal{B}_1) \geq H(\mathcal{A}|\mathcal{B}_2)$*, jeżeli* $\mathcal{B}_1 \subset \mathcal{B}_2$*;*
3. $H(\mathcal{A}|\mathcal{B}_n) \downarrow H(\mathcal{A}|\mathcal{B})$ *dla* $\mathcal{B}_n \uparrow \mathcal{B}$ *i skończonego* $\mathcal{A}$*;*
4. $H(\mathcal{A}|\mathcal{B}_n) \uparrow H(\mathcal{A}|\mathcal{B})$ *dla* $\mathcal{B}_n \downarrow \mathcal{B}$*;*
5. $H(\mathcal{A}|\mathcal{B}) = 0$ *wtedy i tylko wtedy, gdy* $\mathcal{A} \subset \sigma(\mathcal{B})$*.*

Omawiając entropię σ-ciała niezmienniczego, wprowadziłem także pojęcie procesu mocno nieergodycznego (definicja 1 z pracy **(B)**) — pod niezbyt fortunną nazwą „procesu nieprzeliczalnego opisu" (*uncountable description process*). Pokazałem, że proces Santa Fe (5) jest mocno nieergodyczny i udowodniłem, że proces stacjonarny jest mocno nieergodyczny wtedy i tylko wtedy, gdy σ-ciało niezmiennicze procesu zawiera bezatomowe pod-σ-ciało. A zatem procesy mocno nieergodyczne są nieergodyczne.

Praca (F): Praca ta jest rozwinięciem wyniku z mojego doktoratu, komunikowanego w pracy **(E)** bez dowodu. Niech $\mathbb{X}$ będzie skończonym alfabetem. Dla stacjonarnej miary μ na przestrzeni mierzalnej $(\mathbb{X}^\mathbb{Z}, \mathcal{X}^\mathbb{Z})$ i zmiennych losowych $X_i((x_k)_{k\in\mathbb{Z}}) := x_i$ rozpatrzmy entropię blokową

$$H_\mu(n) := H_\mu(X_{i+1:i+n}) := \mathbf{E}_\mu \left[-\log \mu(X_{i+1:i+n} = \cdot)\right],$$

gdzie $\mathbf{E}_\mu$ oznacza wartość oczekiwaną względem μ, zaś log to logarytm naturalny. Ponadto będziemy rozpatrywać drugi skończony alfabet $\mathbb{Y} = \{0, 1, ..., D_Y - 1\}$ i kod $C : \mathbb{X}^+ \to \mathbb{Y}^+$, gdzie $\mathbb{X}^+ = \bigcup_{n=1}^\infty \mathbb{X}^n$. Oznaczmy oczekiwaną długość kodu jako

$$H_\mu^C(n) := \mathbf{E}_\mu |C(X_{1:n})| \log D_Y.$$

Kod C nazywamy jednoznacznie dekodowalnym, jeżeli jego rozszerzenie $C^*(u_1, ..., u_k) := C(u_1)...C(u_k)$, $u_i \in \mathbb{X}^n$, $k \in \mathbb{N}$, jest iniekcją dla dowolnego n. Jak wiadomo, zob. np. Cover i Thomas (2006, twierdzenie 5.3.1), własność ta implikuje nierówność kodowania źródła

$$H_\mu^C(n) \geq H_\mu(n). \tag{55}$$

Kod jednoznacznie dekodowalny C nazywa się uniwersalnym, jeżeli

$$\lim_{n\to\infty} \frac{H^C_\mu(n)}{n} = \lim_{n\to\infty} \frac{H_\mu(n)}{n} \tag{56}$$

dla dowolnej stacjonarnej miary μ.

Główny wynik pracy **(F)** to następujące twierdzenie:

Twierdzenie 16. *Niech C będzie kodem jednoznacznie dekodowalnym. Wówczas*

$$\operatorname{card}\left\{\mu : \mu \text{ jest ergodyczna i } \limsup_{n\to\infty} \left[H^C_\mu(n) - H_\mu(n)\right] \le K\right\} \le \exp(K). \tag{57}$$

Twierdzenie to wynika z rozkładu ergodycznego. Używając kodu Shannona-Fano można także pokazać, że istnieje kod, który spełnia warunek (57) prawie z równością. Mianowicie, niech $\exp(K) = M \in \mathbb{N}$, zaś C niech będzie kodem Shannona-Fano dla miary $P = M^{-1}\sum_{i=1}^M \mu_i$, gdzie μ_i są M dowolnymi miarami ergodycznymi. Wówczas $\limsup_n \left[H^C_{\mu_i}(n) - H_{\mu_i}(n)\right] \le K + \log D_Y$ dla wszystkich $i = 1, ..., M$.

Oznaczmy informację wzajemną $E_\mu(n) := 2H_\mu(n) - H_\mu(2n)$ oraz oczekiwaną algorytmiczną informację wzajemną $E^C_\mu(n) := 2H^C_\mu(n) - H^C_\mu(2n)$. Z twierdzenia 16 wywiodłem analogiczny wynik dla tychże wielkości, a mianowicie:

Twierdzenie 17. *Niech C będzie kodem uniwersalnym. Wówczas*

$$\operatorname{card}\left\{\mu : \mu \text{ jest ergodyczna i } \limsup_{n\to\infty} \left[E^C_\mu(n) - E_\mu(n)\right] \le K\right\} \le \exp(K).$$

Wynik w moim doktoracie był słabszą formą twierdzenia 17, niezawierającą wyrażenia $-E_\mu(n)$.

Praca (G): Praca ta dotyczy pogranicza algorytmicznej teorii informacji i statystyki bayesowskiej. Algorytmiczna teoria informacji sugeruje atrakcyjną interpretację wnioskowania bayesowskiego. Mianowicie, jeżeli wierzymy, że parametr jest typowy dla jakiegoś rozkładu a priori, może to oznaczać, że wierzymy, że parametr ów jest algorytmicznie losowy względem rozkładu a priori. Istotny wkład do tejże interpretacji wnieśli Vovk i V'yugin (1993, 1994) i Takahashi (2008, 2011). W pracy **(G)** także chciałem pokazać, że interpretacja ta jest uzasadniona i ma ważkie konsekwencje dla optymalności wnioskowania bayesowskiego.

Niech $\boldsymbol{Y}$ będzie rekursywną (tzn. obliczalną) miarą prawdopodobieństwa na jednostronnie nieskończonych sekwencjach cyfr (bądź liczbach rzeczywistych z przedziału $[0, 1]$). (Definicję funkcji i miary rekursywnej można znaleźć w Li i Vitányi (2008, definicja 1.7.4 i przykład 1.7.8).) Dla sekwencji $\boldsymbol{y} = (y_i)_{i\in\mathbb{N}}$ i napisu $x^n_1 = x_1...x_n$ będziemy pisać $\boldsymbol{Y}(x^n_1) = \boldsymbol{Y}(\{\boldsymbol{y} : y^n_1 = x^n_1\})$. Nieskończony ciąg cyfr $\boldsymbol{x} = (x_i)_{i\in\mathbb{N}}$ nazywamy algorytmicznie losowym (w sensie Martina-Löfa) względem rekursywnej miary $\boldsymbol{Y}$ jeżeli

$$\inf_{n\in\mathbb{N}} [K(x^n_1) + \log \boldsymbol{Y}(x^n_1)] > -\infty, \tag{58}$$

gdzie $K(x^n_1)$ to bezprzedrostkowa złożoność Kołmogorowa napisu x^n_1 (tzn. długość najkrótszego programu dla bezprzedrostkowego uniwersalnego komputera, którego wynikiem jest x^n_1) (Li i Vitányi, 2008, rozdział 3.1). Zbiór ciągów algorytmicznie losowych będzie oznaczany $\mathcal{L}_{\boldsymbol{Y}}$. Ma on dwie ważne własności. Po pierwsze jest to zbiór miary pełnej, tzn. $\boldsymbol{Y}(\mathcal{L}_{\boldsymbol{Y}}) = 1$. Po drugie, można pokazać, że ciąg $\boldsymbol{x}$ należy do $\mathcal{L}_{\boldsymbol{Y}}$ wtedy i tylko wtedy, gdy

$$\sup_{n\in\mathbb{N}} \frac{\boldsymbol{U}(x^n_1)}{\boldsymbol{Y}(x^n_1)} < \infty \tag{59}$$

dla dowolnej wyliczalnej semimiary $\boldsymbol{U}$ (Vovk i V'yugin, 1994, twierdzenie 1 i lemat 3). Innymi słowy, miara $\boldsymbol{Y}$ zadaje najlepszą wyliczalną kompresję (z dokładnością do stałej multiplikatywnej) ciągów, które są algorytmicznie losowe względem $\boldsymbol{Y}$.

Związek tych wyników z wnioskowaniem bayesowskim jest następujący. Dane są rekursywne jądro $\boldsymbol{P} : \mathbb{X}^* \times \boldsymbol{\Theta} \ni (x, \boldsymbol{\theta}) \mapsto \boldsymbol{P_\theta}(x) \in [0, 1]$, tzn. $\boldsymbol{P}$ jest rekursywną funkcją argumentów $(x, \boldsymbol{\theta})$, $\boldsymbol{P_\theta} : \mathbb{X}^* \to [0, 1]$ jest miarą prawdopodobieństwa dla każdego $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ i odwzorowanie $\boldsymbol{\theta} \mapsto \boldsymbol{P_\theta}$ jest mierzalne, oraz rekursywny rozkład a priori $\boldsymbol{Q} : \mathbb{Y}^* \to [0, 1]$, tzn. rekursywna miara prawdopodobieństwa na $\boldsymbol{\Theta}$, $\boldsymbol{Q}(\boldsymbol{\Theta}) = 1$. Należy podkreślić, że mimo że jądro $\boldsymbol{P}$ jest rekursywną funkcją pary $(x, \boldsymbol{\theta})$, miara $\boldsymbol{P_\theta}$ nie musi być rekursywną funkcją danych x dla ustalonej wartości parametru $\boldsymbol{\theta}$. Taki przypadek może się zdarzyć, gdy $\boldsymbol{\theta} = (\theta_i)_{i \in \mathbb{N}}$ jest algorytmicznie losowe. Zatem aby kompresować dane, które są typowe dla miary $\boldsymbol{P_\theta}$, możemy wykorzystać procedurę bayesowską, w której pojawiają się miara $\boldsymbol{T} : \mathbb{X}^* \times \mathbb{Y}^* \to [0, 1]$, gdzie

$$\boldsymbol{T}(x, y) := \int_{A(y)} \boldsymbol{P_\theta}(x) d\boldsymbol{Q}(\boldsymbol{\theta}), \tag{60}$$

zaś $A(y) := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : y \text{ jest przedrostkiem } \boldsymbol{\theta}\}$, oraz miara

$$\boldsymbol{Y}(x) := \boldsymbol{T}(x, \lambda) = \int \boldsymbol{P_\theta}(x) d\boldsymbol{Q}(\boldsymbol{\theta}), \tag{61}$$

gdzie λ jest napisem pustym. Można pokazać, że miary $\boldsymbol{T}$ i $\boldsymbol{Y}$ są rekursywne. Ponadto $\boldsymbol{Y}(\mathcal{L}_{\boldsymbol{Y}}) = 1$ implikuje następujący fakt:

Twierdzenie 18. $\boldsymbol{P_\theta}(\mathcal{L}_{\boldsymbol{Y}}) = 1$ *dla* $\boldsymbol{Q}$*-prawie wszystkich* $\boldsymbol{\theta}$*.*

Innymi słowy miara $\boldsymbol{Y}$ zadaje najlepszą wyliczalną kompresję danych (ciągów), które są typowe względem miar $\boldsymbol{P_\theta}$ dla $\boldsymbol{Q}$-prawie wszystkich parametrów $\boldsymbol{\theta}$. Fakt ten możemy postrzegać jako pewne uzasadnienie procedur bayesowskich.

Do dalszych rozważań pomocne jest wykorzystanie wyników algorytmicznej teorii informacji. Bylibyśmy zainteresowani takim wzmocnieniem twierdzenia 18, które by orzekło, dla których parametrów $\boldsymbol{\theta}$ dokładnie zachodzi $\boldsymbol{P_\theta}(\mathcal{L}_{\boldsymbol{Y}}) = 1$, a dla których nie. Niech $\mathcal{L}_{\boldsymbol{P}|\boldsymbol{\theta}}$ będą zbiorami ciągów, które są warunkowo algorytmicznie losowe (w sensie Martina-Löfa) względem miar $\boldsymbol{P_\theta}$ (Vovk i V'yugin, 1993; Takahashi, 2008). Mamy $\boldsymbol{P_\theta}(\mathcal{L}_{\boldsymbol{P}|\boldsymbol{\theta}}) = 1$ dla wszystkich $\boldsymbol{\theta}$ oraz

$$\mathcal{L}_{\boldsymbol{Y}} = \bigcup_{\boldsymbol{\theta} \in \mathcal{L}_{\boldsymbol{Q}}} \mathcal{L}_{\boldsymbol{P}|\boldsymbol{\theta}}, \tag{62}$$

jak dowiódł Takahashi (2008, wniosek 4.3 i twierdzenie 5.3). Stąd wynika:

Twierdzenie 19. $\boldsymbol{P_\theta}(\mathcal{L}_{\boldsymbol{Y}}) = 1$*, jeżeli* $\boldsymbol{\theta} \in \mathcal{L}_{\boldsymbol{Q}}$*.*

Innymi słowy miara $\boldsymbol{Y}$ zadaje najlepszą wyliczalną kompresję danych, które są typowe względem miar $\boldsymbol{P_\theta}$ dla parametrów $\boldsymbol{\theta}$, które są algorytmicznie losowe względem rozkładu a priori $\boldsymbol{Q}$. Fakt ten możemy postrzegać jako mocniejsze uzasadnienie procedur bayesowskich.

Dalszy tok myśli jest również naturalny. W rozkładzie (62) pojawiają się wyłącznie parametry $\boldsymbol{\theta}$, które są algorytmicznie losowe względem rozkładu a priori $\boldsymbol{Q}$. Zatem, gdyby zbiory $\mathcal{L}_{\boldsymbol{P}|\boldsymbol{\theta}}$ były rozłączne dla różnych $\boldsymbol{\theta}$, otrzymalibyśmy

$$\boldsymbol{P_\theta}(\mathcal{L}_{\boldsymbol{Y}}) = 0, \text{ jeżeli } \boldsymbol{\theta} \not\in \mathcal{L}_{\boldsymbol{Q}}. \tag{63}$$

Innymi słowy miara $\boldsymbol{Y}$ nie zadaje najlepszej wyliczalnej kompresji danych, które są typowe względem miar $\boldsymbol{P_\theta}$ dla parametrów $\boldsymbol{\theta}$, które nie są algorytmicznie losowe względem rozkładu a priori $\boldsymbol{Q}$. Fakt ten można postrzegać jako ostrzeżenie, że kompresja bayesowska jest suboptymalna, jeżeli poczynimy błędne założenie co do rozkładu a priori, względem którego parametr miałby być typowy. Ostrzeżenie to jest ważkim przyczynkiem algorytmicznej teorii informacji do analizy wnioskowania bayesowskiego.

Możemy zaobserwować, że zbiory $\mathcal{L}_{\boldsymbol{P}|\boldsymbol{\theta}}$ są rozłączne i w konsekwencji zachodzi warunek (63), jeżeli parametr może być zidentyfikowany na podstawie danych. Wówczas dwa ciągi danych, które są typowe dla dwóch różnych wartości parametru muszą być różne, ponieważ różne wartości parametru mogą być wywnioskowane na podstawie każdego z tych ciągów. W oczywisty sposób fakt ten możemy ująć następująco:

Twierdzenie 20. *Niech $\boldsymbol{P}$ będzie rekursywnym jądrem. Następujące stwierdzenia są równoważne:*

1. $\mathcal{L}_{\boldsymbol{P}|\boldsymbol{\theta}} \cap \mathcal{L}_{\boldsymbol{P}|\boldsymbol{\theta}'} = \emptyset$, *jeżeli* $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ *dla* $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$.
2. *Istnieje funkcja* $f : \bigcup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{L}_{\boldsymbol{P}|\boldsymbol{\theta}} \to \boldsymbol{\Theta}$ *taka, że* $f(\boldsymbol{x}) = \boldsymbol{\theta}$ *dla* $\boldsymbol{x} \in \mathcal{L}_{\boldsymbol{P}|\boldsymbol{\theta}}$ *i każdego* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Stwierdzenie to, w słabszej formie, udowodnił także Takahashi (2011, twierdzenie 6.1).

Nasuwa się pytanie, czy funkcję f można skonstruować dla zadanego jądra (parametrycznej rodziny rozkładów). Następujące użyteczne pojęcie zaproponował V'yugin (2007):

Definicja 4. *Estymator $T(x_1^n)$ nazywa się efektywnie ściśle zgodnym, jeżeli istnieje rekursywna funkcja $N(\epsilon, \delta)$ taka, że dla każdego $\boldsymbol{\theta}$ i dla wszystkich ϵ i δ zachodzi*

$$\boldsymbol{P}_{\boldsymbol{\theta}}\left(\left\{\boldsymbol{x} : \sup_{n\geq N(\epsilon,\delta)} |T(x_1^n) - \boldsymbol{\theta}| > \epsilon\right\}\right) \leq \delta. \tag{64}$$

Efektywnie ściśle zgodny estymator istnieje na przykład dla rodziny rozkładów Bernoulliego. W dalszej kolejności V'yugin dowiódł następującego faktu:

Twierdzenie 21. *Jeżeli jądro $\boldsymbol{P}$ dopuszcza efektywnie ściśle zgodny estymator $T(x_1^n)$, to*

$$\lim_{n\to\infty} T(x_1^n) = \boldsymbol{\theta} \tag{65}$$

dla każdego $\boldsymbol{x} \in \mathcal{L}_{\boldsymbol{P}|\boldsymbol{\theta}}$.

Jeżeli zachodzi warunek (65), to możemy położyć $f(\boldsymbol{x}) = \lim_{n\to\infty} T(x_1^n)$. Stąd zbiory $\mathcal{L}_{\boldsymbol{P}|\boldsymbol{\theta}}$ są rozłączne i w konsekwencji zachodzi warunek (63).

W pracy **(G)** sformalizowałem pojęcie parametru efektywnie identyfikowalnego w sposób odmienny:

Definicja 5. *Niech $\boldsymbol{P}$ będzie jądrem, zaś $\boldsymbol{Q}$ rozkładem a priori. Zdefiniujmy miary $\boldsymbol{T}$ i $\boldsymbol{Y}$ wzorami (60) i (61). Bayesowski model statystyczny $(\boldsymbol{P}, \boldsymbol{Q})$ nazywam τ-wyuczalnym, jeżeli dla funkcji $\tau : \mathbb{Y}^* \to \mathbb{N}$, wszystkich $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\boldsymbol{P}_{\boldsymbol{\theta}}$-prawie wszystkich $\boldsymbol{x}$ i $n = \tau(\theta_1^m)$ zachodzi*

$$\lim_{m\to\infty} \boldsymbol{T}(x_1^n, \theta_1^m)/\boldsymbol{Y}(x_1^n) = 1. \tag{66}$$

Warunek (66) zapewnia, że parametr zdyskretyzowany do m cyfr można wywnioskować dla wszystkich prócz skończenie wielu m, mając dane x_1^n długości $n = \tau(\theta_1^m)$. Warunek (66) wydaje się być słabszy niż (64), gdyż w warunku (66) ciągi x_1^m nie muszą mieć jednostajnie ograniczonej długości jak w warunku (64). W pracy **(G)** wykazałem, że rodziny rozkładów wykładniczych oraz procesy Santa Fe są przykładami τ-wyuczalnych bayesowskich modeli statystycznych. Ponadto wykazałem następujący fakt:

Twierdzenie 22. *Niech $(\boldsymbol{P}, \boldsymbol{Q})$ będzie τ-wyuczalnym bayesowskim modelem statystycznym z rekursywnym jądrem $\boldsymbol{P}$, rekursywnym rozkładem a priori $\boldsymbol{Q}$ i rekursywną funkcją τ. Wówczas zachodzi warunek (63).*

Przedstawiony dowód tego faktu nie odwołuje się do rozłączności zbiorów $\mathcal{L}_{\boldsymbol{P}|\boldsymbol{\theta}}$, której nie udało mi się wykazać.

Monografia (H): Książka ta stanowi przegląd wybranych zagadnień z teorii informacji, statystyki i informatyki teoretycznej, które powiązane są kwestią kwantyfikacji informacji. Wybór dyskutowanych problemów jest oryginalny, a książka obejmuje pewne zagadnienia nie omawiane gdzie indziej, takie jak entropia nadwyżkowa procesów stacjonarnych i obliczeniowa optymalność wnioskowania bayesowskiego.

Pojęcie entropii nadwyżkowej wprowadzono w rozdziale 4. Jest ona zdefiniowana jako informacja wzajemna między przeszłością a przyszłością procesu stacjonarnego

$$E = \lim_{n\to\infty} I(X_{-n+1}^0; X_1^n), \tag{67}$$

gdzie $I(X_{-n+1}^0; X_1^n)$ to informacja wzajemna między przyległymi blokami długości n (Crutchfield i Feldman, 2003). Następnie pokazano, że entropia nadwyżkowa jest granicznym odchyleniem entropii blokowej od jej asymptotycznie liniowego wzrostu, tzn.

$$E = \lim_{n\to\infty} [H(X_1^n) - nh], \tag{68}$$

gdzie $H(X_1^n)$ to entropia blokowa, a $h = \lim_{n\to\infty} H(X_1^n)/n$ to intensywność entropii. Ponadto, w rozdziale 5. podano wzór na entropię nadwyżkową procesu gaussowskiego. Mianowicie, przy odpowiednich warunkach zachodzi

$$E = \frac{1}{2\pi} \int_{|z|<1} \left| \frac{d}{dz} \log \psi(z) \right|^2 d(\operatorname{Re} z) d(\operatorname{Im} z), \tag{69}$$

gdzie funkcja ψ dana jest wzorem

$$\psi(z) = \exp\left[\frac{1}{2} h(z)\right], \quad h(z) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{e^{i\omega} + z}{e^{i\omega} - z} \log f(\omega) d\omega \tag{70}$$

dla $|z| < 1$, a f to gęstość spektralna procesu. Wynik ten jest konsekwencją twierdzeń, które podali Grenander i Szegő (1958, sekcje 1.1, 1.13, 1.14, 5.5).

Obliczeniowa optymalność wnioskowania bayesowskiego omawiana jest w rozdziale 15. Zagadnienie to zostało dokładniej przedstawione w niniejszym autoreferacie w omówieniu pracy **(G)**, więc nie omawiam go tutaj.

Pozostałe zagadnienia omawiane w monografii **(H)** obejmują problemy dyskutowane w innych książkach, takie jak: wprowadzenie do teorii informacji, intensywność entropii, twierdzenie ergodyczne, twierdzenie Shannona-McMillana-Breimana, uniwersalność kodu Lempela-Ziva, wprowadzenie do statystyki, rodziny rozkładów wykładniczych, własności estymatora największej wiarogodności, informacja Fishera, wnioskowanie bayesowskiej, algorytm EM, modelowanie maksimum entropii, złożoność Kołmogorowa i jej związki i analogie z entropią oraz różne charakteryzacje ciągów losowych w sensie Martina-Löfa.

Wybrane zagadnienia z książki **(H)** były wykorzystane jako treść wykładu monograficznego dla studentów matematyki stosowanej. Z tego powodu każdy rozdział zawiera listę ćwiczeń, a rozwiązania niektórych z nich podano na końcu monografii.

Literatura

W. Bialek, I. Nemenman, N. Tishby (2001) *Complexity through nonextensivity*, Physica A, t. 302, s. 89–99.

R. C. Bradley (1980) *On the strong mixing and weak Bernoulli conditions*, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, t. 50, s. 49–54.

W. Bułatek, B. Kamiński (2009) *On Excess Entropies for Stationary Random Fields*, Probability and Mathematical Statistics, t. 29, s. 353–367.

G. Cariolaro, G. Pierobon (1977) *Stationary symbol sequences from variable-length word sequences*, IEEE Transactions on Information Theory, t. 23, s. 243–253.

M. Charikar, E. Lehman, A. Lehman, D. Liu, R. Panigrahy, M. Prabhakaran, A. Sahai, A. Shelat (2005) *The Smallest Grammar Problem*, IEEE Transactions on Information Theory, t. 51, s. 2554–2576.

I. P. Cornfeld, S. V. Fomin, Y. G. Sinai (1982) *Ergodic Theory*, New York: Springer.

T. M. Cover, J. A. Thomas (2006) *Elements of Information Theory, 2nd ed.*, New York: John Wiley.

J. P. Crutchfield, D. P. Feldman (2003) *Regularities unseen, randomness observed: The entropy convergence hierarchy*, Chaos, t. 15, s. 25–54.

J. P. Crutchfield, K. Young (1989) *Inferring statistical complexity*, Physical Review Letters, t. 63, s. 105–108.

Ł. Dębowski (2005) *Własności entropii nadwyżkowej dla procesów stochastycznych nad różnymi alfabetami*, Rozprawa doktorska, Instytut Podstaw Informatyki PAN.

— (2007) *Menzerath's law for the smallest grammars*, w: P. Grzybek, R. Köhler (red.), *Exact Methods in the Study of Language and Text*, s. 77–85, Berlin: Mouton de Gruyter.

R. L. Dobrušin (1959) *Obščaja formulirovka osnovnoj teoremy Šennona v teorii informacii*, Uspehi Matematičeskih Nauk, t. 14(6), s. 3–104.

W. Ebeling (1997) *Prediction and entropy of nonlinear dynamical systems and symbolic sequences with LRO*, Physica D, t. 109, s. 42–45.

W. Ebeling, T. Pöschel (1994) *Entropy and long-range correlations in literary English*, Europhysics Letters, t. 26, s. 241–246.

C. J. Ellison, J. R. Mahoney, J. P. Crutchfield (2009) *Prediction, Retrodiction, and the Amount of Information Stored in the Present*, Journal of Statistical Physics, t. 136, s. 1005–1034.

D. P. Feldman, J. P. Crutchfield (2003) *Structural Information in Two-Dimensional Patterns: Entropy Convergence and Excess Entropy*, Physical Review E, t. 67, s. 051 104.

P. D. Finch (1960) *On the covariance determinants of autoregressive and moving average models*, Biometrika, t. 47, s. 194–211.

T. Gramss (1994) *Entropy of the symbolic sequence for critical circle maps*, Physical Review E, t. 50, s. 2616–2620.

R. M. Gray, L. D. Davisson (1974) *The ergodic decomposition of stationary discrete random processses*, IEEE Transactions on Information Theory, t. 20, s. 625–636.

R. M. Gray, J. C. Kieffer (1980) *Asymptotically mean stationary measures*, The Annals of Probability, t. 8, s. 962–973.

U. Grenander, G. Szegő (1958) *Toeplitz Forms and Their Applications*, Berkeley: University of California Press.

P. Guiraud (1954) *Les caractères statistiques du vocabulaire*, Paris: Presses Universitaires de France.

P. Harremoës, F. Topsøe (2001) *Maximum Entropy fundamentals*, Entropy, t. 3, s. 191–226.

H. S. Heaps (1978) *Information Retrieval—Computational and Theoretical Aspects*, New York: Academic Press.

G. Herdan (1964) *Quantitative Linguistics*, London: Butterworths.

W. Hilberg (1990) *Der bekannte Grenzwert der redundanzfreien Information in Texten — eine Fehlinterpretation der Shannonschen Experimente?*, Frequenz, t. 44, s. 243–248.

O. Kallenberg (1997) *Foundations of Modern Probability*, New York: Springer.

E. Khmaladze (1988) *The statistical analysis of large number of rare events.* Technical Report MS-R8804. Centrum voor Wiskunde en Informatica, Amsterdam.

J. C. Kieffer, E. Yang (2000) *Grammar-based codes: A new class of universal lossless source codes*, IEEE Transactions on Information Theory, t. 46, s. 737–754.

C. Kit, Y. Wilks (1999) *Unsupervised Learning of Word Boundary with Description Length Gain*, w: M. Osborne, E. T. K. Sang (red.), *Proceedings of the Computational Natural Language Learning ACL Workshop, Bergen*, s. 1–6.

A. Kornai (2002) *How many words are there?*, Glottometrics, t. 4, s. 61–86.

U. Krengel (1985) *Ergodic theorems*, Berlin: Walter de Gruyter.

W. Kuraszkiewicz, J. Łukaszewicz (1951) *Ilość różnych wyrazów w zależności od długości tekstu*, Pamiętnik Literacki, t. 42(1), s. 168–182.

M. Li, P. M. B. Vitányi (2008) *An Introduction to Kolmogorov Complexity and Its Applications, 3rd ed.*, New York: Springer.

W. Löhr (2009) *Properties of the Statistical Complexity Functional and Partially Deterministic HMMs*, Entropy, t. 11, s. 385–401.

J. R. Mahoney, C. J. Ellison, J. P. Crutchfield (2009) *Information accessibility and cryptic processes*, Journal of Physics A, t. 42, s. 362 002.

B. Mandelbrot (1954) *Structure formelle des textes et communication*, Word, t. 10, s. 1–27.

C. G. de Marcken (1996) *Unsupervised Language Acquisition*, Rozprawa doktorska, Massachussetts Institute of Technology.

G. A. Miller (1957) *Some effects of intermittent silence*, American Journal of Psychology, t. 70, s. 311–314.

C. G. Nevill-Manning (1996) *Inferring Sequential Structure*, Rozprawa doktorska, University of Waikato.

M. S. Pinsker (1960) *Informacija i informacionnaja ustojnost' slučajnyh veličin i processov*, Moskva: Izdatelstvo Akademii Nauk SSSR.

C. R. Shalizi, J. P. Crutchfield (2001) *Computational Mechanics: Pattern and Prediction, Structure and Simplicity*, Journal of Statistical Physics, t. 104, s. 819–881.

C. Shannon (1951) *Prediction and entropy of printed English*, Bell System Technical Journal, t. 30, s. 50–64.

P. C. Shields (1997) *String matching bounds via coding*, The Annals of Probability, t. 25, s. 329–336.

H. A. Simon (1955) *On a class of skew distribution functions*, Biometrika, t. 42, s. 425–440.

H. Takahashi (2008) *On a definition of random sequences with respect to conditional probability*, Information and Computation, t. 206, s. 1375–1382.

— (2011) *Algorithmic randomness and monotone complexity on product space*, Information and Computation, t. 209, s. 183–197.

R. Timo, K. Blackmore, L. Hanlen (2007) *On the entropy rate of word-valued sources*, w: *Proceedings of the Telecommunication Networks and Applications Conference, ATNAC 2007*, s. 377–382.

N. F. Travers, J. P. Crutchfield (2011) *Infinite Excess Entropy Processes with Countable-State Generators.* `http://arxiv.org/abs/1111.3393`.

V. G. Vovk, V. V. V'yugin (1993) *On the Empirical Validity of the Bayesian Method*, Journal of the Royal Statistical Society, series B, t. 55, s. 253–266.

— (1994) *Prequential Level of Impossibility with Some Applications*, Journal of the Royal Statistical Society, series B, t. 56, s. 115–123.

V. V. V'yugin (2007) *On Empirical Meaning of Randomness with Respect to a Real Parameter*, w: *Computer Science — Theory and Applications*, s. 387–396, New York: Springer.

J. G. Wolff (1980) *Language acquisition and the discovery of phrase structure*, Language and Speech, t. 23, s. 255–269.

A. D. Wyner (1978) *A definition of conditional mutual information for arbitrary ensembles*, Information and Control, t. 38, s. 51–59.

G. K. Zipf (1965) *The Psycho-Biology of Language: An Introduction to Dynamic Philology, 2nd ed.*, Cambridge, MA: The MIT Press.

Łukasz Dębowski