Uniwersytet Warszawski
Wydział Matematyki, Informatyki i Mechaniki

Marcin Socha
Nr albumu: 418253

Statystyczne dywergencje w zastosowaniu do predykcji sygnałów stochastycznych

Praca magisterska
na kierunku MATEMATYKA

Praca wykonana pod kierunkiem
dr hab. Jana Karbowskiego

Warszawa, grudzień 2024

Streszczenie

W niniejszej pracy rozpatrzono zagadnienia związane z estymacją statystycznych dywergencji dla różnych sygnałów probabilistycznych zarówno stałych jak i zmiennych w czasie. W szczególności omówiono dywergencje Rényiego i Tsallisa oraz dywergencję Kullbacka-Leiblera. Celem pracy jest wyprowadzenie nierówności dla dywergencji pomiędzy dwoma wielowymiarowymi rozkładami prawdopodobieństwa oraz ich zastosowanie w znanych rozkładach prawdopodobieństwa oraz otrzymane wyniki zilustrowano na przykładzie analizy danych sportowych mających charakter stochastyczny.

Słowa kluczowe

Dywergencja Kullbacka-Leiblera, Dywergencje Tsallisa oraz Renyiego, wielowymiarowe rozkłady lognormalne oraz normalne.

Dziedzina pracy (kody wg programu Socrates-Erasmus)

11.1 Matematyka

Klasyfikacja tematyczna

62P99
94A15
60G15

Tytuł pracy w języku angielskim

Statistical divergences in application to prediction of stochastic signals

Serdeczne podziękowania kieruję do mojego Promotora, profesora Jana Karbowskiego, za poświęcony czas, cenne wskazówki, a także za wsparcie i wyrozumiałość.

Pracę dedykuję Moim Rodzicom.

Spis treści

Rozdział 1

Wprowadzenie

1.1. Historia

W 1960 ukazała się praca autorstwa Alfréda Rényiego, w której zajmował się on teorią informacji [1]. Jedną z najważniejszych koncepcji, którą tam wprowadził były f-dywergencje - funkcje pozwalające zmierzyć "odległość"między dwoma rozkładami prawdopodobieństwa. Oznaczamy je zazwyczaj symbolem $D_f(P||Q)$, gdzie P i Q to pewne dwa rozkłady prawdopodobieństwa, określone na tym samym zbiorze probabilistycznym (jeśli $Q(A) = 0$, to również $P(A) = 0$). Jej definicję, która będzie nam użyteczna w pracy, pokażemy w następnym rozdziale. Jedną z jej odmian jest dywergencja wyżej wymienionego Rényiego czy też dywergencja Tsallisa [2]. Jedną z cech f-dywergencji jest fakt, iż nie są symetryczne oraz nie spełniają nierówności trójkąta, więc nie są miarami. Jednak najważniejszą f-dywergencją, którą zajmiemy się w pracy jest inna funkcja. Co ciekawe, została ona opublikowana przed 1960 rokiem, a więc oryginalną pracą Rényiego.

W 1951 roku Solomon Kullback oraz Richard Leibler na łamach *The Annals of Mathematical Statistics* przedstawili asymetryczną funkcję mierzącą odległość między dwoma miarami probabilistycznymi. Dzisiaj nazywana jest ona dywergencją Kullbacka-Leiblera (KL) i jest jednym z najważnieszych narzędzi statystycznych stosowanych do porównywania dwóch rozkładów prawdopodobieństwa. Kullback i Leibler odwoływali się do wcześniejszych prac dotyczących teorii informacji autorstwa Claude'a Shannona [3], w których badano między innymi entropię. Entropia jest miarą statystycznej niepewności danego układu i spełnia zależność: im większa entropia tym większa niepewność co do układu. Od tamtego czasu dywergencja Kullbacka-Leiblera, którą wymiennie nazywa się też względną entropią, znalazła szerokie zastosowania zarówno w matematyce, jak i statystyce, czy uczeniu maszynowym [4, 5, 6]. Dywergencja KL jest szczególnym przypadkiem dywergencji Rényiego oraz Tsallisa, co zostanie pokazane w dalszej części pracy.

Główną motywacją wprowadzenia dywergencji $D_{KL}(P||Q)$ była próba stworzenia "odległości"między dwoma rozkładami prawdopodobieństwa P i Q. Miała ona umożliwiać ocenę, jak dokładnie rozkład P (zazwyczaj jest to pewne przewidywanie rozkładu) odbiega od Q, który często interpretowany jest jako "prawdziwy"rozkład danych. Niestety dywergencja KL również nie jest metryką. Wynika to z faktu, że między innymi, nie jest symetryczna, co oznacza, że $D_{KL}\left(P||Q\right) \neq D_{KL}\left(Q||P\right)$.

1.2. Zastosowanie

Jednym z zastosowań dywergencji KL jest porównywanie modeli probabilistycznych. W statystyce, przykładowo, bada się jak dobrze dopasowany jest model do prawdziwych danych. Służy też do porównywania hipotez, czy optymalizacji parametrów w metodach bayesowskich [7]. W szczególności dywergencja KL potrafi przybliżyć nam ile informacji zyskujemy aktualizując wcześniejsze przekonania (a priori) na podstawie nowych danych (posterior). W uczeniu maszynowym dywergencja KL odgrywa kluczową rolę na przykład w generatywnych modelach probabilistycznych, takich jak generative adversarial network (GAN) [8].

Ponadto, dywergencja KL ma istotne zastosowania w analizie dynamicznych systemów probabilistycznych [9]- szczególnie w kontekście modelowania zmian w czasie. W niniejszej pracy jednym z elementów jest również badanie zmian dywergencji jako funkcji zależnej od czasu. Udaje się to dzięki analizie nierówności ograniczających tempo tych zmian. Dodatkowo w dalszej części pracy pokazano również zastosowania praktyczne, między innymi na wielowymiarowych rozkładach Gaussa. Nasze badania mogą mieć znaczenie zarówno teoretyczne, jak i praktyczne, ponieważ umożliwiają zastosowanie dywergencji KL w analizie sygnałów stochastycznych takich jak dane sportowe, co pokazano w ostatnim rozdziale tej pracy.

1.3. Cel pracy

Celem pracy jest przedstawienie podstaw teoretycznych oraz praktycznego użycia dywergencji Kullbacka-Leiblera w porównaniu z innymi dywergencjami jak Rényiego czy Tsallisa. Wyniki, które otrzymujemy mogą posłużyć jako narzędzie do dalszego badania nad sygnałami stochastycznymi.

W niniejszej pracy przeprowadzono szczegółową analizę nierówności dla ograniczeń tempa zmian dywergencji statystycznych zaproponowanych przez J. Karbowskiego w jego pracy z 2024 roku [10], a także praktyczne zastosowania dywergencji Kullbacka-Leiblera. W tym celu wykonano niezbędne obliczenia oraz wizualizacje umożliwiające dalsze badania i rozwój w tej gałęzi teorii informacji oraz statystyce. Ponadto zaprezentowano przykłady zastosowań tych nierówności, szczególnie w kontekście rozkładów Gaussa, aby ukazać ich praktyczne znaczenie i potencjalne zastosowania w modelowaniu, analizie statystycznej lub innych dziedzinach nauki. Poprzez te działania pragniemy stworzyć podstawy dla bardziej zaawansowanych badań i praktycznych zastosowań przedstawionych nierówności.

Rozdział 2

Wstępne definicje oraz nierówności

2.1. Podstawowe definicje

Na początku naszych rozważań dotyczących dywergencji między rozkładami przedstawmy kilka najważniejszych definicji używanych w następnych rozdziałach.

Często będziemy rozważać rozkłady prawdopodobieństwa zarówno dyskretne, jak i ciągłe, jednak zgodnie z właściwościami gęstości prawdopodobieństwa dalsze rozważania przy pomocy całek będą analogiczne dla sum. W związku z tym, w zależności od potrzeb, będziemy przyjmowali różne notacje.

We wstępie powiedzieliśmy o f-dywergencjach. Przedstawimy teraz jej szczegółową definicję. Zarówno tę jak i bardziej ogólną definicję można znaleźć w publikacji [11].

Definicja 2.1. Niech $p(x)$ oraz $q(x)$ będą pewnymi gęstościami na $\mathbb{R}^n$ rozkładów prawdopodobieństwa odpowiednio P i Q . Niech dla każdego zbioru mierzalnego A zachodzi
$(\mathbb{P}(Q \in A) = 0) \Rightarrow (\mathbb{P}(P \in A) = 0)$.
Wówczas, jeśli $f : (0, +\infty) \to \mathbb{R}$ jest funkcją wypukłą, taką, że $f(1) = 0$ oraz $f(0) = \lim_{t\to 0^+} f(t)$, to *$f$-dywergencją między rozkładami P oraz Q* nazywamy

$$D_f(P||Q) := \int_{\mathbb{R}^n} f\left(\frac{p}{q}\right) q dx. \tag{2.1}$$

Definicja 2.2. Niech $a, b > 0$, oraz $\frac{1}{a} + \frac{1}{b} = 1$. Dodatkowo niech $\mathbb{R}^n = X$ oraz $(X, \mathcal{M}, \mu)$ będzie przestrzenią mierzalną probabilistyczną z miarą μ, a f i g będą funkcjami mierzalnymi na X [12]. Wówczas zachodzi *nierówność Holdera:*

$$\int_X fg\, d\mu \leq \left(\int_X f^a\, d\mu\right)^{\frac{1}{a}} \left(\int_X g^b\, d\mu\right)^{\frac{1}{b}}. \tag{2.2}$$

W niektórych przypadkach przybiera ona formę

$$\sum_{i=1}^{\infty} f_i g_i \leq \left(\sum_{i=1}^{\infty} f_i^a\right)^{\frac{1}{a}} \left(\sum_{i=1}^{\infty} g_i^b\right)^{\frac{1}{b}}, \tag{2.3}$$

gdzie $(f_i)_{i\in\mathbb{N}}$ oraz $(g_i)_{i\in\mathbb{N}}$ są pewnymi ciągami liczb rzeczywistych.

Szkic dowodu: Oznaczmy

$$A = \left(\int_X f^a \, d\mu \right)^{\frac{1}{a}}, \quad B = \left(\int_X g^b \, d\mu \right)^{\frac{1}{b}}. \tag{2.4}$$

Jeśli $A = 0$, to $f = 0$ prawie wszędzie względem μ, a wtedy $fg = 0$ prawie wszędzie względem μ, co dowodzi tezy. Jeśli $A > 0$ i $B = +\infty$, to również mamy tezę.

Możemy więc założyć, że A i B są dodatnie i skończone. Zdefiniujmy nowe funkcje $F = \frac{f}{A}$ i $G = \frac{g}{B}$. Wówczas

$$\int_X F^a \, d\mu = \int_X G^b \, d\mu = 1. \tag{2.5}$$

Jeśli $x \in X$ jest taki, że $F(x)$ i $G(x)$ są dodatnie, to istnieją liczby s i t takie, że $F(x) = e^{\frac{s}{a}}$ i $G(x) = e^{\frac{t}{b}}$. Ponieważ $\frac{1}{a} + \frac{1}{b} = 1$, to korzystając z wypukłości funkcji wykładniczej mamy:

$$e^{\frac{s}{a} + \frac{t}{b}} \leq \frac{1}{a} e^s + \frac{1}{b} e^t. \tag{2.6}$$

Stąd dla dowolnego $x \in X$ otrzymujemy:

$$F(x)G(x) \leq \frac{1}{a} F(x)^a + \frac{1}{b} G(x)^b. \tag{2.7}$$

Całkując powyższą nierówność, dostajemy:

$$\int_X FG \, d\mu \leq \int_X \frac{1}{a} F^a \, d\mu + \int_X \frac{1}{b} G^b \, d\mu = \frac{1}{a} + \frac{1}{b} = 1, \tag{2.8}$$

a stąd:

$$\int_X fg \, d\mu \leq AB = \left(\int_X f^a \, d\mu \right)^{\frac{1}{a}} \left(\int_X g^b \, d\mu \right)^{\frac{1}{b}}. \tag{2.9}$$

□

Dla przestrzeni probabilistycznych nierówność (2.2) przybiera postać:

$$\mathbb{E}\left[fg\right] \leq \left(\mathbb{E}\left[f^2\right]\right)^{1/2} \left(\mathbb{E}\left[g^2\right]\right)^{1/2}, \tag{2.10}$$

gdzie $\mathbb{E}$ oznacza wartość oczekiwaną.

Definicja 2.3. Jeśli $a = b = 2$, to nierówność (2.9) przybiera formę:

$$\int_X fg \, d\mu \leq \left(\int_X f^2 \, d\mu \right)^{\frac{1}{2}} \left(\int_X g^2 \, d\mu \right)^{\frac{1}{2}}. \tag{2.11}$$

Mówimy, że jest to *nierówność Cauchye'ego-Schwartza* [13]. W szczególności dla dwóch ciągów $(x_i)_{i \in \mathbb{N}}$ oraz $(y_i)_{i \in \mathbb{N}}$ zachodzi

$$\sum_{i=1}^{\infty} x_i y_i \leq \left(\sum_{i=1}^{\infty} x_i^2 \right)^{\frac{1}{2}} \cdot \left(\sum_{i=1}^{\infty} y_i^2 \right)^{\frac{1}{2}} \tag{2.12}$$

Rozważmy układ fizyczny, którego stany wewnętrzne są indeksowane przez n i mogą być reprezentowane przez dwa zależne od czasu rozkłady prawdopodobieństwa: $p(n,t)$ oraz $q(n,t)$, gdzie $p(n,t) = \mathbb{P}(X_t = n)$, co implikuje, że $\sum_{n\in\mathbb{N}} p(n,t) = 1$. Podobnie dla $q(n,t)$. Choć nie jest to kluczowe dla dalszej dyskusji, przydatne jest postrzeganie $q(n,t)$ jako prawdziwego (lub referencyjnego) rozkładu prawdopodobieństwa rządzącego stochastyczną dynamiką układu, podczas gdy $p(n,t)$ może być traktowane jako estymacja lub prognoza tego rozkładu. Na początek zdefiniujmy użyteczną wielkość, która posłuży nam jako fundament do innych f-dywergencji

Definicja 2.4. Niech $p(n,t)$ oraz $q(n,t)$ będą dyskretnymi rozkładami prawdopodobieństwa. Wówczas dla $\alpha > 0$ *współczynnikiem Chernoffa* α lub *dywergencją)* nazywamy wyrażenie $C_\alpha(p||q)$ dane wzorem:

$$C_\alpha(p||q) = \mathbb{E}_p\left[\left(\frac{p}{q}\right)^{\alpha-1}\right] = \sum_n p(n,t)\cdot\left(\frac{p(n,t)}{q(n,t)}\right)^{\alpha-1}, \tag{2.13}$$

gdzie α jest daną liczbą rzeczywistą, a $\mathbb{E}_p[\cdot]$ oznacza wartość oczekiwaną względem rozkładu prawdopodobieństwa p. Przyjmujemy oznaczanie, że $C_\alpha := C_\alpha(p||q)$.

Możemy kontrolować jak istotne dla nas będą różnice między p i q poprzez zmiany parametru α. Jeśli chcemy, aby na przykład "promować"jak najmniejsze wartości p/q, to wówczas zwiększymy α.

Możemy rozważać p zarówno jako dyskretny, jak i ciągły rozkład, zgodnie z właściwościami gęstości prawdopodobieństwa. Ważne jest jednak, aby zarówno p, jak i q były tego samego rodzaju (dyskretny lub ciągły). Będzie to użyteczne w dalszej części pracy jeśli zauważymy że:

$$C_\alpha(p||q) = \int_{\mathbb{R}^n}\left(\frac{p(x)}{q(x)}\right)^{\alpha-1} p(x)\,dx = \int_{\mathbb{R}^n}\left(\frac{p(x)}{q(x)}\right)^{\alpha} q(x)\,dx = \mathbb{E}_q\left[\left(\frac{p}{q}\right)^{\alpha}\right], \tag{2.14}$$

co jest również prawdziwe dla rozkładów dyskretnych.

Zauważmy także, że $C_0(p||q) = C_1(p||q) = 1$, ponieważ:

$$C_0(p||q) = \mathbb{E}_p\left[\left(\frac{p}{q}\right)^{-1}\right] = \mathbb{E}_p\left[\frac{q}{p}\right] = \int_{\mathbb{R}^n}\frac{q(x)}{p(x)}p(x)\,dx = \int_{\mathbb{R}^n} q(x)\,dx = 1, \tag{2.15}$$

oraz:

$$C_1(p||q) = \int_{\mathbb{R}^n}\left(\frac{p(x)}{q(x)}\right)^{0} p(x)\,dx = \int_{\mathbb{R}^n} p(x)\,dx = 1. \tag{2.16}$$

Współczynnik Chernoffa będzie niezbędny do wprowadzenia dwóch innych f- dywergencji, o których była mowa we wstępie, czyli dywergencje, które zdefiniujemy między dwoma rozkładami p i q: Tsallisa (T_α) oraz Rényiego (R_α) [14]:

Definicja 2.5. Jeśli p oraz q są pewnymi rozkładami prawdopodobieństwa, to wówczas wyrażenie

$$T_\alpha(p||q) = \frac{C_\alpha(p||q) - 1}{\alpha - 1}, \tag{2.17}$$

nazywamy *dywergencją Tsallisa* zaś:

$$R_\alpha(p||q) = \frac{\ln(C_\alpha(p||q))}{\alpha - 1}. \tag{2.18}$$

nazywamy *dywergencją Rényiego*.

Wynika stąd zależność między tymi dywergencjami, którą otrzymujemy po prostych obliczeniach:

$$T_\alpha = \left(\exp\left((\alpha-1)R_\alpha\right) - 1\right)/(\alpha - 1). \tag{2.19}$$

Mówimy, że jeśli R_α lub T_α są bliskie zeru, to znaczy, że rozkład q dobrze przybliża rozkład p. Dla $\alpha = 2$ dywergencja T_2 bywa nazywana χ^2 [15]. Te dwie dywergencje można uznać za uogólnienie dywergencji Kullbacka-Leiblera, która brzmi:

Definicja 2.6. Analogicznie jak wcześniej *dywergencją Kullbacka-Leiblera między dwoma rozkładami prawdopodobieństwa p i q* nazywamy:

$$D_{KL}(p||q) = \mathbb{E}_p\left[\ln\left(\frac{p}{q}\right)\right] = \sum_{n\in\mathbb{N}} p_n \ln\left(\frac{p_n}{q_n}\right). \tag{2.20}$$

Zakładamy, że jeśli D_{KL} zapiszemy tożsamościowo jako $\mathbb{E}_p\left[\ln(p) - \ln(q)\right]$, to możemy ją interpretować jako średnią liczbę bitów informacji, które tracimy korzystając z danego przybliżenia zamiast prawdziwego rozkładu [16]. W związku z tym jest ona często wykorzystywana przy optymalizacji modeli opartych na pewnych rozkładach prawdopodobieństwa lub wykorzystujących zmienne losowe.

Definiujemy $T_\alpha = T_\alpha(p||q)$ i analogicznie dla R_α i D_{KL}. Korzystając z faktu, że każda funkcja pierwotna funkcji ciągłej jest również ciągła, to zachodzi tożsamość

$$\begin{aligned}
T_1 &= \lim_{\alpha\to 1} T_\alpha \\
&= \lim_{\alpha\to 1} \frac{C_\alpha - 1}{\alpha - 1} \\
&= \lim_{\alpha\to 1} \left(\int_{\mathbb{R}} \left(\left(\frac{p(x)}{q(x)}\right)^{\alpha-1} - 1\right) p(x)dx\right)/(\alpha-1) \\
&= \left(\int_{\mathbb{R}} \lim_{\alpha\to 1} \left(\left(\frac{p(x)}{q(x)}\right)^{\alpha-1} - 1\right) \frac{p(x)}{\alpha-1}\ dx\right) \\
&= \int_{\mathbb{R}} \lim_{\alpha\to 1} \left(\ln\left(\frac{p(x)}{q(x)}\right)\left(\frac{p(x)}{q(x)}\right)^{\alpha-1}\right) p(x)\ dx.
\end{aligned} \tag{2.21}$$

Następnie używając reguły de l'Hospitala

$$\begin{aligned}
T_1 &= \int_{\mathbb{R}} \ln\left(\frac{p(x)}{q(x)}\right) p(x)\ dx \\
&= \mathbb{E}_p\left[\ln\left(\frac{p}{q}\right)\right] \\
&= D_{KL}(p||q).
\end{aligned} \tag{2.22}$$

W analogiczny sposób można obliczyć, że $R_1 = \lim_{\alpha\to 1} R_\alpha = T_1 = D_{KL}$, więc D_{KL} jest szczególnym przypadkiem dywergencji Tsallisa oraz Renyiego. Łatwo zauważyć, że jeśli T_α, R_α lub D_{KL} są bliskie zeru, to p jest bardzo dobrą estymacją q.

2.2. Nierówności między dywergencjami

W tym podrozdziale przedstawimy kilka nierówności [15] między podanymi wyżej dywergencjami, aby lepiej zrozumieć zależności między nimi. Zauważmy, że dla $0 < \alpha < \beta < 1$ korzystając z twierdzenia Jensena zachodzi

$$\begin{aligned}(C_\alpha)^{(\beta-1)/(\alpha-1)} &= \left(\mathbb{E}_p\left[\left(\frac{p}{q}\right)^{\alpha-1}\right]\right)^{(\beta-1)/(\alpha-1)} \\ &\leq \mathbb{E}_p\left[\left(\frac{p}{q}\right)^{(\alpha-1)(\beta-1)/(\alpha-1)}\right] \\ &= C_\beta. \end{aligned} \tag{2.23}$$

Uzyskujemy to korzystając z faktu, że dla $0 < \alpha < \beta < 1$ przekształcenie $x \mapsto x^{(\beta-1)/(\alpha-1)}$ jest przekształceniem wypukłym. Stąd oraz korzystając z definicji (2.5) dostajemy dla $0 < \alpha < \beta < 1$:

$$R_\alpha \leq R_\beta. \tag{2.24}$$

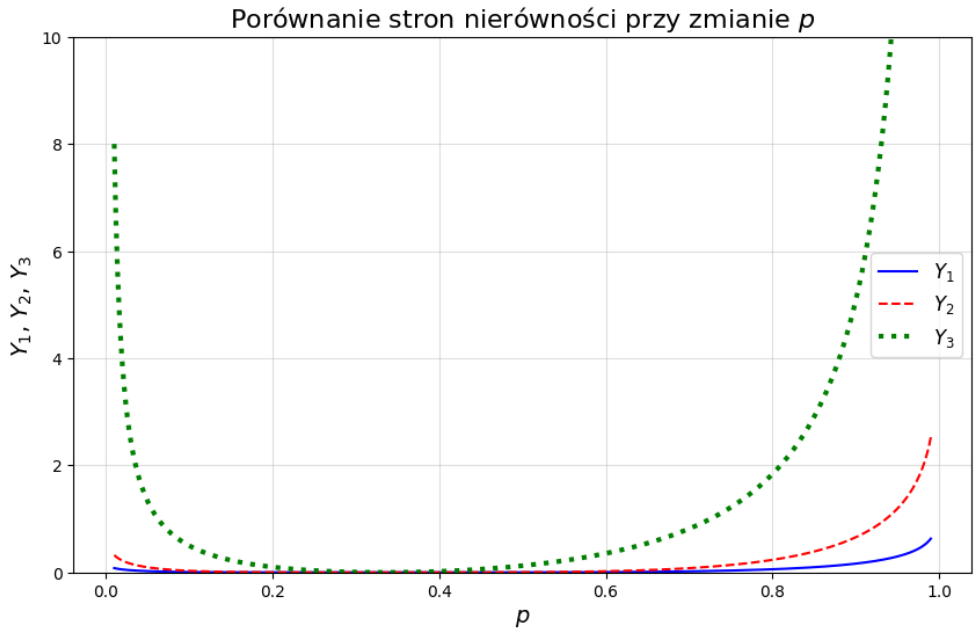
Analogicznie dla $1 < \alpha < \beta$ można udowodnić, że $R_\alpha \leq R_\beta$. Innymi użytecznymi nierównościami dla $0 < \alpha < 1$ są [15]:

$$R_\alpha(P||Q)R_\alpha(Q||P) \leq D_{KL}(P||Q)D_{KL}(Q||P) \leq \frac{1}{4}\left[\exp\left(R_2(P||Q) + R_2(Q||P)\right) - 1\right] \tag{2.25}$$

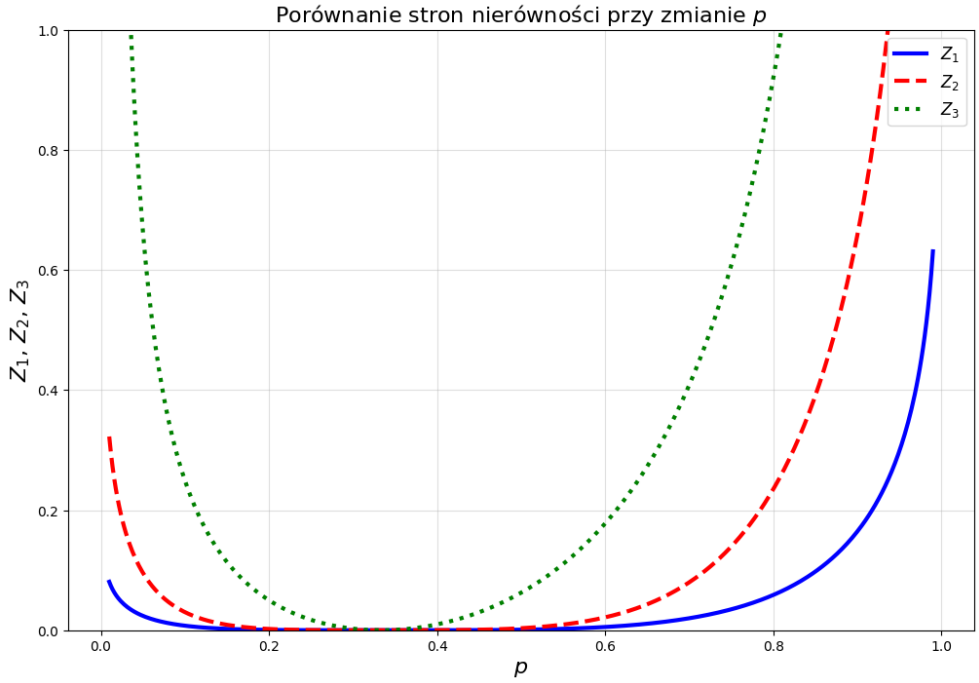
oraz

$$\begin{aligned}R_\alpha(P||Q) + R_\alpha(Q||P) &\leq D_{KL}(P||Q) + D_{KL}(Q||P) \\ &\leq \frac{1}{2}\left[\left(\exp\left(R_2(P||Q)\right) + \exp\left(R_2(Q||P)\right)\right) - 1\right]. \end{aligned} \tag{2.26}$$

Co pokazuje nam, że nawet jeśli $R_\alpha(P||Q) > D_{KL}(P||Q)$, to wtedy na pewno zachodzi $R_\alpha(Q||P) < D_{KL}(Q||P)$. Na rysunkach (2.1) oraz (2.2) przedstawiono kolejno nierówności (2.25) oraz(2.26) w zależności od parametru p (oś x) przy założeniu, że Q i P to rozkłady dwupunktowe, takie że $Q(0) = \frac{1}{3}, Q(1) = \frac{2}{3}$ oraz $P(0) = p, P(1) = 1 - p$, oraz dla $\alpha = 0.5$.



Rysunek 2.1: Wykres nierówności (2.25) dla $Y_1 = R_\alpha(P||Q)R_\alpha(Q||P)$, $Y_2 = D_{KL}(P||Q)D_{KL}(Q||P)$, $Y_3 = \frac{1}{4}\left[\exp\left(R_2(P||Q) + R_2(Q||P)\right) - 1\right]$



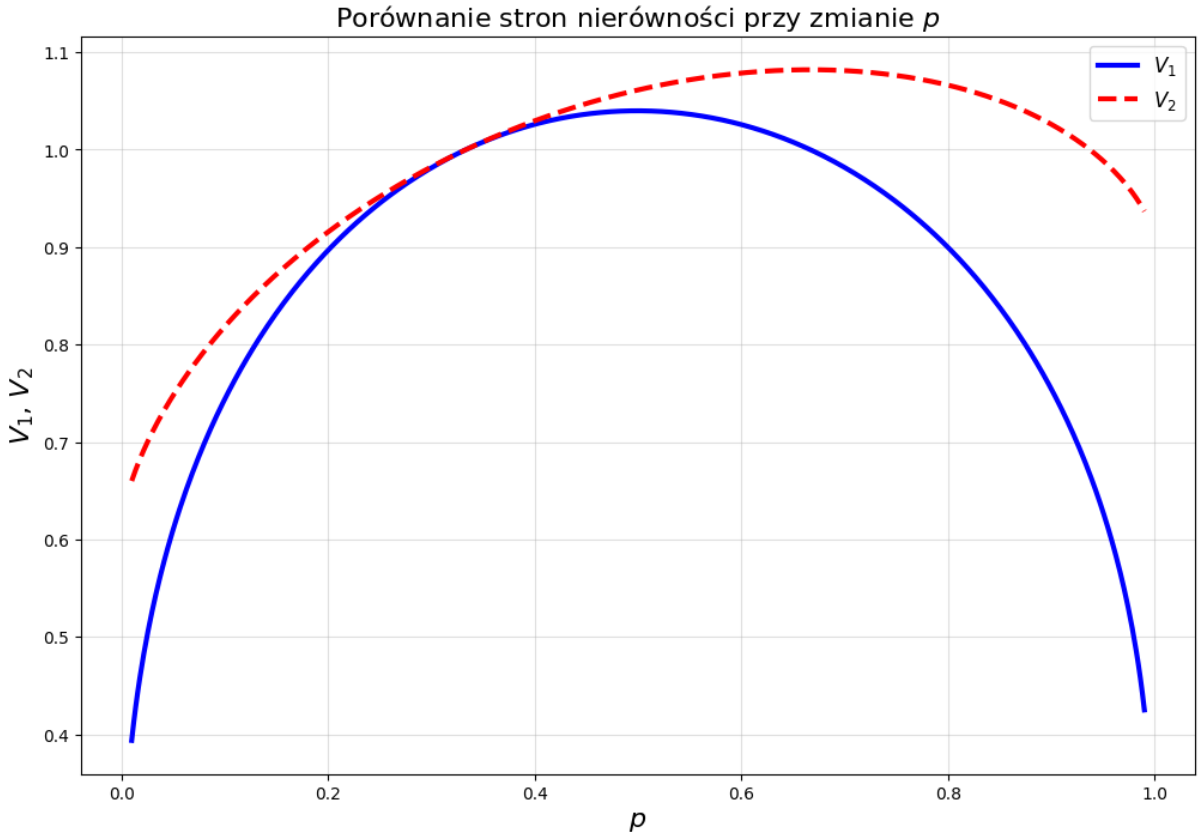
Rysunek 2.2: Wykres nierówności (2.26) dla $Z_1 = R_\alpha(P||Q) + R_\alpha(Q||P)$, $Z_2 = D_{KL}(P||Q) + D_{KL}(Q||P)$, $Z_3 = \frac{1}{2}\left[\exp\left(R_2(P||Q) + R_2(Q||P)\right) - 1\right]$

Niezależnie od nierówności (2.25), (2.26) bada się w literaturze następującą nierówność

[17]

$$\frac{D_{KL}(P||Q)}{D_{KL}(Q||P)} \leq \frac{1}{2}\frac{T_2(P||Q)}{D_{KL}(P||Q)}, \tag{2.27}$$

więc widzimy, że D_{KL} podniesione do kwadratu podzielone przez odwrotną dywergencję jest mniejsze niż T_2. Poniższy rysunek (2.3) przedstawia nierówność (2.27) dla rozkładów Q i P przy założeniu, że Q i P to rozkłady dwupunktowe, takie że $Q(0) = \frac{1}{3}, Q(1) = \frac{2}{3}$ oraz $P(0) = p, P(1) = 1 - p$, oraz dla $\alpha = 0.5$.



Rysunek 2.3: Wykres nierówności (2.27) dla $V_1 = \frac{D_{KL}(P||Q)}{D_{KL}(Q||P)}$, $V_2 = \frac{1}{2}\frac{T_2(P||Q)}{D_{KL}(P||Q)}$

Widzimy, że w każdej z tych nierówności zachodzi równość wtedy i tylko wtedy, gdy $p = q = \frac{1}{3}$ oraz, że obie strony nierówności (2.49) są wtedy równe zero. Można stąd wysnuć wniosek, że są one ostre wtedy i tylko wtedy gdy $P \neq Q$ poza zbiorami miary 0.

2.3. Ograniczenia dla prędkości zmian dywergencji

W związku z tym, że rozkłady prawdopodobieństwa P i Q mogą zmieniać się w czasie, to warto rozważyć szybkość zmian statystycznych dywergencji ze względu na czas [10]. Ze względu na związki ustalone we wcześniejszych równaniach, prędkości zmian dywergencji Tsallisa i Rényiego mogą być wyrażone przez prędkość zmiany współczynnika α. Pozwala to skupić się na czasowej zmianie C_α oraz jej ograniczeniach, ponieważ obliczenia dla C_α są nieco prostsze.

W konsekwencji ograniczenia dla prędkości zmian T_α i R_α można bezpośrednio wyprowadzić z ograniczeń na dC_α/dt.

Teraz obliczmy wartość bezwzględną pochodnej czasowej $\left|\frac{dC_\alpha}{dt}\right|$, wykorzystując regułę całkowania Leibniza:

$$\begin{aligned}\left|\frac{dC_\alpha}{dt}\right| &= \left|\frac{d}{dt}\int_{\mathbb{R}} p(x,t)\cdot\left(\frac{p(x,t)}{q(x,t)}\right)^{\alpha-1} dx\right| \\ &= \left|\int_{\mathbb{R}} \frac{d}{dt}\left(\frac{p(x,t)^\alpha}{q(x,t)^{\alpha-1}}\right) dx\right| \\ &= \left|\int_{\mathbb{R}} \left(\frac{\alpha p^{\alpha-1}\dot{p}q^{\alpha-1} - (\alpha-1)q^{\alpha-2}\dot{q}p^\alpha}{q^{2\alpha-2}}\right) dx\right| \\ &= \left|\int_{\mathbb{R}} \alpha\dot{p}\left(\frac{p}{q}\right)^{\alpha-1} - (\alpha-1)\dot{q}\left(\frac{p}{q}\right)^{\alpha} dx\right|.\end{aligned} \tag{2.28}$$

Zdefiniujmy notację $\dot{f} = \frac{df}{dt}$. Wówczas z (2.28) oraz nierówności trójkąta otrzymujemy

$$\begin{aligned}\int_{\mathbb{R}} p(x) &= 1 \\ \int_{\mathbb{R}} \frac{d}{dt}p(x) &= \frac{d}{dt}\cdot 1 = 0 \\ \int_{\mathbb{R}} \dot{p}dx &= 0\end{aligned} \tag{2.29}$$

dalej mamy

$$\begin{aligned}\left|\frac{dC_\alpha}{dt}\right| &= \left|\int_{\mathbb{R}} \alpha\dot{p}\left(\left(\frac{p}{q}\right)^{\alpha-1} - C_\alpha\right) - (\alpha-1)\dot{q}\left(\left(\frac{p}{q}\right)^{\alpha} - C_\alpha\right) dx\right| \\ &\leq |\alpha|\left|\mathbb{E}_p\left[\frac{\dot{p}}{p}\left(\left(\frac{p}{q}\right)^{\alpha-1} - C_\alpha\right)\right]\right| + |\alpha-1|\left|\mathbb{E}_q\left[\frac{\dot{q}}{q}\left(\left(\frac{p}{q}\right)^{\alpha} - C_\alpha\right)\right]\right|.\end{aligned} \tag{2.30}$$

Skoro zachodzą (2.28), (2.30) i (2.21) możemy zapisać następujące górne ograniczenie

$$\begin{aligned}\left|\frac{dD_{KL}}{dt}\right| &= \left|\frac{d}{dt}\lim_{\alpha\to 1} T_\alpha\right| \\ &= \left|\frac{d}{dt}\lim_{\alpha\to 1}\frac{C_\alpha - 1}{\alpha-1}\right| \\ &\leq \left|\lim_{\alpha\to 1}\frac{|\alpha|\mathbb{E}_p\left[\frac{\dot{p}}{p}\left(\left(\frac{p}{q}\right)^{\alpha-1} - C_\alpha\right)\right]}{\alpha-1} + \mathbb{E}_q\left[\frac{\dot{q}}{q}\left(\left(\frac{p}{q}\right)^{\alpha} - C_\alpha\right)\right]\right| \\ &\leq \lim_{\alpha\to 1}\left|\left|\alpha|\mathbb{E}_p\left[\frac{\dot{p}}{p}\left(\left(\frac{p}{q}\right)^{\alpha-1} - C_\alpha\right)/(\alpha-1)\right]\right| + \left|\mathbb{E}_q\left[\frac{\dot{q}}{q}\left(\left(\frac{p}{q}\right)^{\alpha} - C_\alpha\right)\right]\right|.\end{aligned} \tag{2.31}$$

Skoro $\lim_{\alpha\to 1}\left(\left(\frac{p}{q}\right)^{\alpha-1}-1\right)/(\alpha-1)=\ln(\frac{p}{q})$ możemy obliczyć

$$\begin{aligned}
\left|\frac{dD_{KL}}{dt}\right| &\leq \lim_{\alpha\to 1}\left|\,|\alpha|\mathbb{E}_p\left[\frac{\dot{p}}{p}\left(\left(\frac{p}{q}\right)^{\alpha-1}-C_\alpha\right)/(\alpha-1)\right]\right| + \left|\mathbb{E}_q\left[\frac{\dot{q}}{q}\left(\left(\frac{p}{q}\right)^{\alpha}-C_\alpha\right)\right]\right| \\
&= \lim_{\alpha\to 1}\left|\,|\alpha|\mathbb{E}_p\left[\frac{\dot{p}}{p}\left(\left(\frac{p}{q}\right)^{\alpha-1}-1-(C_\alpha-1)\right)/(\alpha-1)\right]\right| + \left|\mathbb{E}_q\left[\frac{\dot{q}}{q}\left(\left(\frac{p}{q}\right)^{\alpha}-C_\alpha\right)\right]\right| \\
&\leq \mathbb{E}_p\left[\left|\frac{\dot{p}}{p}\right|\left|\ln\left(\frac{p}{q}\right)-(C_1-1)-D_{KL}\right|\right] + \mathbb{E}_q\left[\left|\frac{\dot{q}}{q}\left[\frac{p}{q}-C_1\right]\right|\right] \\
&= \mathbb{E}_p\left[\left|\frac{\dot{p}}{p}\right|\left|\ln\left(\frac{p}{q}\right)-D_{KL}\right|\right] + \mathbb{E}_q\left[\left|\frac{\dot{q}}{q}\left(\frac{p}{q}-1\right)\right|\right] \\
&= \mathbb{E}_p\left[\left|\frac{\dot{p}}{p}\right|\left|\ln\left(\frac{p}{q}\right)-D_{KL}\right|\right] + \mathbb{E}_q\left[\frac{d}{dt}\ln\left(\frac{p}{q}\right)\right]. \qquad (2.32)
\end{aligned}$$

W takim razie zmiana dywergencji KL jest ograniczona przez zmianę zarówno logarytmu ilorazu rozkładów jak i różnice tych logarytmów od samej dywergencji KL. W dalszej części przydatny będzie fakt, że

$$\frac{dC_\alpha}{dt}=(\alpha-1)\frac{dT_\alpha}{dt}. \qquad (2.33)$$

Skoro zachodzi $C_\alpha=\exp\left((\alpha-1)R_\alpha\right)$ to mamy, że

$$\begin{aligned}
\frac{dC_\alpha}{dt} &= \frac{d}{dt}\left(\exp((\alpha-1)R_\alpha)\right) \\
&= \exp((\alpha-1)R_\alpha)\cdot\frac{d}{dt}\left((\alpha-1)R_\alpha\right) \\
&= (\alpha-1)\exp((\alpha-1)R_\alpha)\cdot\frac{d}{dt}R_\alpha. \qquad (2.34)
\end{aligned}$$

W takim razie widzimy, że C_α zmienia się eskponencjalnie w porównaniu do R_α i mocno zależy nie tylko od zmiany samej dywergencji Renyiego, ale też tego jaką miała wartość do momentu t. Zdefiniujemy teraz informację Fishera od czasu, którą można interpretować jako kwadrat globalnej prędkości zmian systemu dynamicznego wyrażonego przez rozkład p [10]

Definicja 2.7. Niech $p(t)$ będzie pewnym rozkładem prawdopodobieństwa różniczkowalnym oraz zależnym od czasu. Wówczas *informacją czasową Fishera* nazywamy wyrażenie dane wzorem:

$$I_F(p)=\int_{\mathbb{R}}p(x)\left(\frac{\dot{p}}{p}\right)^2dx=\mathbb{E}_p\left[\left(\frac{\dot{p}}{p}\right)^2\right]. \qquad (2.35)$$

Informacja Fishera mierzy, ile średnio bitów informacji będziemy potrzebowali do zmiany stanu w czasie, lub bywa interpretowana jako po prostu szybkość danego układu podniesiona do kwadratu [18]. Formalnie, jest to wartość oczekiwana drugiej pochodnej logarytmu funkcji względem parametru czasu, co odzwierciedla czułość tego logarytmu na zmiany parametru.

Jako przykład obliczymy ją dla rozkładu normalnego ze średnią $\mu(t)$ zależną od czasu oraz różniczkowalną. Zakładamy, że wariancja σ^2 nie zmienia się w czasie. Wówczas gęstość tego

rozkładu jest równa $p_t(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu_t)^2}{2\sigma^2}\right)$. Obliczamy więc Informację Fishera:

$$\begin{aligned}
I_F(p) &= \mathbb{E}_p\left[\left(\frac{\dot{p}}{p}\right)^2\right] \\
&= \int_{\mathbb{R}} p(x)\left(\frac{\dot{p(x)}}{p(x)}\right)^2 dx \\
&= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x-\mu_t)^2}{2\sigma^2}\right)\left(\frac{\frac{d}{dt}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x-\mu_t)^2}{2\sigma^2}\right)}{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x-\mu_t)^2}{2\sigma^2}\right)}\right)^2 dx \\
&= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x-\mu_t)^2}{2\sigma^2}\right)\left(\frac{\frac{d}{dt}\exp\left(\frac{-(x-\mu_t)^2}{2\sigma^2}\right)}{\exp\left(\frac{-(x-\mu_t)^2}{2\sigma^2}\right)}\right)^2 dx \\
&= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x-\mu_t)^2}{2\sigma^2}\right)\left(\frac{d}{dt}\left(\frac{-(x-\mu_t)^2}{2\sigma^2}\right)\right)^2 dx \\
&= \frac{(\dot{\mu_t})^2}{\sigma^4}\int_{\mathbb{R}}(x-\mu_t)^2\frac{\exp(-(x-\mu_t)^2/2\sigma^2)}{\sqrt{2\pi\sigma^2}}dx \\
&= \frac{(\dot{\mu_t})^2}{\sigma^4}\cdot\sigma^2 \\
&= \left(\frac{\dot{\mu_t}}{\sigma}\right)^2. \qquad (2.36)
\end{aligned}$$

Z otrzymanej w we wzorze (2.36) równości zaobserwowano, że tempo zmian tego rozkładu ($\sqrt{I_F(p)}$) jest liniowe względem wartości bezwzględnej pochodnej średniej. Oznacza to, że im bardziej średnia naszego układu będzie się zwiększała, tym bardziej średnia liczba bitów potrzebnych do zakodowania będzie się zwiększać.

Korzystając z nierówności Höldera (2.2) dla dwóch procesów stochastycznych $X = \frac{\dot{p}}{p}$, $Y = \left(\frac{p}{q}\right)^{\alpha-1} - C_\alpha$ oraz z (2.30) otrzymujemy [10]:

$$\begin{aligned}
\left|\dot{C_\alpha}\right| \leq |\alpha|&\sqrt{\mathbb{E}_p\left[\left(\frac{\dot{p}}{p}\right)^2\right]}\sqrt{\mathbb{E}_p\left[\left(\left(\frac{p}{q}\right)^{\alpha-1} - C_\alpha\right)^2\right]} \\
&+ |\alpha-1|\sqrt{\mathbb{E}_q\left[\left(\frac{\dot{q}}{q}\right)^2\right]}\sqrt{\mathbb{E}_q\left[\left(\left(\frac{p}{q}\right)^{\alpha} - C_\alpha\right)^2\right]}. \qquad (2.37)
\end{aligned}$$

Następnie korzystając z (2.4) obserwujemy, że

$$\begin{aligned}
\mathbb{E}_p\left[\left(\left(\frac{p}{q}\right)^{\alpha-1} - C_\alpha\right)^2\right] &= \mathbb{E}_p\left[\left(\frac{p}{q}\right)^{2\alpha-2} - 2C_\alpha\left(\frac{p}{q}\right)^{\alpha-1} + C_\alpha^2\right] \\
&= \mathbb{E}_p\left[\left(\frac{p}{q}\right)^{2\alpha-2}\right] - 2C_\alpha\mathbb{E}_p\left[\left(\frac{p}{q}\right)^{\alpha-1}\right] + C_\alpha^2 \\
&= C_{2\alpha-1} - 2C_\alpha^2 + C_\alpha^2 \\
&= C_{2\alpha-1} - C_\alpha^2 \qquad (2.38)
\end{aligned}$$

oraz

$$\begin{aligned}
\mathbb{E}_q\left[\left(\left(\frac{p}{q}\right)^{\alpha}-C_\alpha\right)^2\right] &= \mathbb{E}_q\left[\left(\frac{p}{q}\right)^{2\alpha}-2C_\alpha\left(\frac{p}{q}\right)^{\alpha}+C_\alpha^2\right] \\
&= \int_{\mathbb{R}}\left(\frac{p(x)}{q(x)}\right)^{2\alpha}q(x)dx-2C_\alpha\int_{\mathbb{R}}\left(\frac{p(x)}{q(x)}\right)^{\alpha}q(x)dx+C_\alpha^2 \\
&= \int_{\mathbb{R}}\left(\frac{p(x)}{q(x)}\right)^{2\alpha-1}p(x)dx-2C_\alpha\int_{\mathbb{R}}\left(\frac{p(x)}{q(x)}\right)^{\alpha-1}p(x)dx+C_\alpha^2 \\
&= \mathbb{E}_p\left[\left(\frac{p}{q}\right)^{2\alpha-1}\right]-2C_\alpha\mathbb{E}_p\left[\left(\frac{p}{q}\right)^{\alpha-1}\right] \\
&= C_{2\alpha}-C_\alpha^2. \qquad (2.39)
\end{aligned}$$

Podstawiając (2.39) oraz informację Fishera (2.35) do (2.38) otrzymujemy

$$\left|\frac{d}{dt}C_\alpha\right| \le |\alpha|\sqrt{I_F(p)(C_{2\alpha-1}-C_\alpha^2)}+|\alpha-1|\sqrt{I_F(q)(C_{2\alpha}-C_\alpha^2)}. \qquad (2.40)$$

W takim razie, korzystając z definicji (2.5), możemy łatwo obliczyć, że

$$\begin{aligned}
\left|\dot{T}_\alpha\right| &= \frac{1}{|\alpha-1|}\left|\dot{C}_\alpha\right| \\
&\le \left|\frac{\alpha}{\alpha-1}\right|\sqrt{I_F(p)\,(C_{2\alpha-1}-C_\alpha^2)}+\sqrt{I_F(q)\,(C_{2\alpha}-C_\alpha^2)} \\
&= \left|\frac{\alpha}{\alpha-1}\right|\sqrt{I_F(p)\,((2\alpha-2)(T_{2\alpha-1}-T_\alpha)-(\alpha-1)^2T_\alpha^2)} \\
&\qquad +\sqrt{I_F(q)\,((2\alpha-1)T_{2\alpha}-(2\alpha-2)T_\alpha-(\alpha-1)^2T-\alpha^2)}. \qquad (2.41)
\end{aligned}$$

Analogicznie

$$\begin{aligned}
\left|\dot{R}_\alpha\right| \le \left|\frac{\alpha}{\alpha-1}\right|&\sqrt{I_F(p)\,(\exp\left(2(\alpha-1)(R_{2\alpha-1}-R_\alpha)\right)-1)} \\
&+\sqrt{I_F(q)\,(\exp\left((2\alpha-1)R_{2\alpha}-(\alpha-1)R_\alpha\right)-1)}. \qquad (2.42)
\end{aligned}$$

Stąd widzimy, że zmiany R_α oraz T_α zależą od prędkości zmian obydwu z rozkładów p i q poprzez $I_F(p)$ i $I_F(q)$. Dodatkowo aby zmiany tych dywergencji były jak najmniejsze to oba rozkłady muszą się nie zmienić (nie wystarczy aby oba się zmieniały w podobnym stopniu), albo musi zachodzić przybliżenie $(2\alpha-1)/(\alpha-1)\approx R_\alpha/R_{2\alpha}$ oraz $R_\alpha\approx R_{2\alpha-1}$.

2.4. Ograniczenia na szybkość zmian dywergencji KL

W niniejszym podrozdziale zostaną przedstawione kilka przydatnych nierówności, które zastosujemy w dalszej części pracy. Korzystając z (2.32) oraz nierówności Cauchy'ego–Schwarza

(2.3) obliczono moduł pochodnej $\left|\dot{D_{KL}}\right|$ że

$$\begin{aligned}\left|\dot{D_{KL}}\right| &\leq \mathbb{E}_p\left[\left|\frac{\dot{p}}{p}\right|\left|\ln\left(\frac{p}{q}\right)-D_{KL}\right|\right]+\mathbb{E}_q\left[\left|\frac{\dot{q}}{q}\left(\frac{p}{q}-1\right)\right|\right]\\ &\leq \sqrt{\mathbb{E}_p\left[\left|\frac{\dot{p}}{p}\right|^2\right]}\sqrt{\mathbb{E}_p\left[\ln^2\left(\frac{p}{q}\right)\right]-D_{KL}^2}+\sqrt{\mathbb{E}_q\left[\left|\frac{\dot{q}}{q}\right|^2\right]}\sqrt{\mathbb{E}_q\left[\left|\left(\frac{p}{q}-1\right)^2\right|\right]}\\ &= \sqrt{I_F(p)\left(\mathbb{E}_p\left[\ln^2\left(\frac{p}{q}\right)\right]-D_{KL}^2\right)}+\sqrt{I_F(q)T_2}. \end{aligned} \tag{2.43}$$

Nierówność (2.43) zawiera składnik $\mathbb{E}_p\left[\ln^2\left(\frac{p}{q}\right)\right]$, który może być problematyczny przy obliczeniach w wielu praktycznych sytuacjach. Z tego powodu dobrze jest mieć górne ograniczenie używając f-dywergencji [10]. Następnie użyto nierówności między średnią geometryczną i logarytmiczna [19] dla dwóch dodatnich liczb x, y zachodzi:

$$\sqrt{xy} \leq \frac{x-y}{\ln(x)-\ln(y)}. \tag{2.44}$$

Podstawiając $x=\sqrt{p/q}$ oraz $y=\sqrt{q/p}$ dostajemy

$$\begin{aligned} 1 &= \sqrt{\frac{p}{q}\cdot\frac{q}{p}}\\ &\leq \frac{\sqrt{p/q}-\sqrt{q/p}}{\ln(p/q)^{1/2}-\ln(q/p)^{1/2}}\\ &= \frac{\sqrt{p/q}-\sqrt{q/p}}{\frac{1}{2}\ln(p/q)+\frac{1}{2}\ln(p/q)} \end{aligned} \tag{2.45}$$

$$. \tag{2.46}$$

Podnosząc obie strony nierówności do kwadratu oraz mnożąc przez $\ln^2(p/q)$ mamy

$$\ln^2\left(\frac{p}{q}\right) \leq \left(\sqrt{p/q}-\sqrt{q/p}\right)^2. \tag{2.47}$$

Następnie obliczamy, że:

$$\begin{aligned}\mathbb{E}_p\left[\ln^2\left(\frac{p}{q}\right)\right] &\leq \mathbb{E}_p\left[\left(\sqrt{p/q}-\sqrt{q/p}\right)^2\right]\\ &= \mathbb{E}_p\left[\frac{p}{q}+\frac{q}{p}-2\right]\\ &= \mathbb{E}_p\left[\frac{p}{q}\right]+\int_{\mathbb{R}}\frac{q(x)}{p(x)}p(x)dx-2\\ &= \mathbb{E}_p\left[\frac{p}{q}\right]-1\\ &= \frac{C_2-1}{2-1}\\ &= T_2. \end{aligned} \tag{2.48}$$

Stąd otrzymujemy nierówność na szybkość zmian dywergencji KL, nad którą skupimy się w dalszej części pracy:

$$\left|\frac{d}{dt}D_{KL}\right| \leq \sqrt{I_F(p)}\sqrt{T_2 - D_{KL}^2} + \sqrt{I_F(q)T_2}, \tag{2.49}$$

która jest równoważna nierówności

$$1 \leq \frac{\sqrt{I_F(p)}\sqrt{T_2 - D_{KL}^2} + \sqrt{I_F(q)T_2}}{\left|\frac{d}{dt}D_{KL}\right|}. \tag{2.50}$$

Widzimy stąd, że jeśli zarówno p jak i q nie ulega zmianie, to pochodna dywergencji KL również będzie bliska zero. Jeśli jednak zarówno p jak i q będzie się zmieniało dynamicznie (nawet jeśli w tym samym tempie oraz tym samym "kierunku"), to górne ograniczenie będzie niskie tylko jeśli zarówno T_2 jak i D_{KL} również będą bliskie zeru. Nierówność (2.49) być wykorzystywana w modelach opartych na danych empirycznych (np. zmieniających się rozkładach statystycznych w czasie, jak w piłce nożnej), pozwala badać, jak szybko dwa rozkłady przestają być do siebie podobne.

Rozdział 3

Przykłady użycia dywergencji Kullbacka-Leiblera

3.1. Proste przykłady

3.1.1. Rozkład dwupunktowy z permutacją

Rozważmy dwa rozkłady dwupunktowe $P(A)$ oraz $Q(A)$, takie że A odpowiada pewnemu zdarzeniu przyjmującemu dwie wartości, gdzie:

$$P(A=0)=\frac{1}{3},\quad P(A=1)=\frac{2}{3} \tag{3.1}$$

oraz

$$Q(A=0)=\frac{2}{3},\quad Q(A=1)=\frac{1}{3}. \tag{3.2}$$

Możemy teraz obliczyć dywergencję KL między $P(A)$ oraz $Q(A)$:

$$D_{KL}(P||Q)=\frac{1}{3}\ln\left(\frac{1/3}{2/3}\right)+\frac{2}{3}\ln\left(\frac{2/3}{1/3}\right)=-\frac{1}{3}\ln(2)+\frac{2}{3}\ln(2)=\frac{1}{3}\ln(2). \tag{3.3}$$

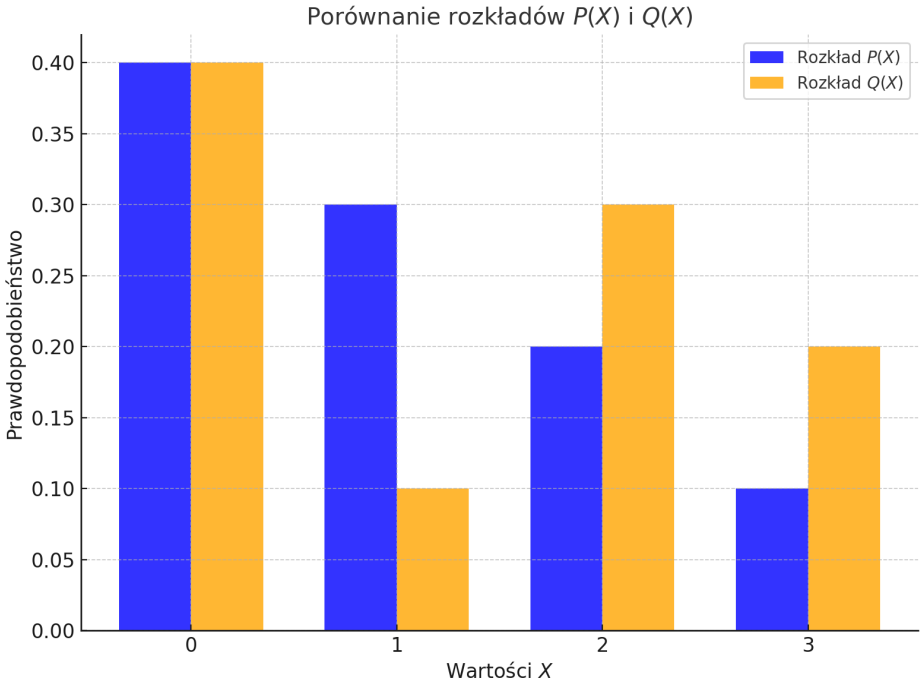
Przedstawiony przykład dobrze ilustruje sposób obliczenia dywergencji Kullbacka-Leiblera. To znaczy, że jeśli dla pewnego zdarzenia $(A=a)$ iloraz miara $P(A=a)/Q(A=a)=2$, a dla zdarzenia $(A=b)$ zachodzi $P(A=b)/Q(A=b)=\frac{1}{2}$, to w miarę skuteczne zostanie to skrócone przy obliczaniu D_{KL}. Łatwo można zauważyć, że $D_{KL}(P||Q)=D_{KL}(Q||P)$. W następnym przykładzie pokażemy brak symetryczności tej dywergencji.

3.1.2. Rozkład czteropunktowy z permutacją

Rozważmy dwa rozkłady czteropunktowe: $P(X)$ oraz $Q(X)$, takie że X odpowiada pewnemu zdarzeniu przyjmującemu cztery wartości, gdzie

$$P(X=0)=\frac{4}{10},\quad P(X=1)=\frac{3}{10},\quad P(X=2)=\frac{2}{10},\quad P(X=3)=\frac{1}{10}. \tag{3.4}$$

$$Q(X=0)=\frac{4}{10},\quad Q(X=1)=\frac{1}{10},\quad Q(X=2)=\frac{3}{10},\quad Q(X=3)=\frac{2}{10}, \tag{3.5}$$



Rysunek 3.1: Porównanie rozkładów $P(X)$ i $Q(X)$, które opisują różne prawdopodobieństwa wystąpienia zdarzeń dla czterech możliwych wartości zmiennej losowej X.

czyli jest "zamienione"prawdopodobieństwo wystąpienia 1, 2 oraz 3. Obliczając dywergencję KL z definicji (2.6) między P oraz Q wynosi

$$D_{KL}(P||Q) = \frac{4}{10}\ln\left(\frac{0.4}{0.4}\right) + \frac{3}{10}\ln\left(\frac{0.3}{0.1}\right) + \frac{2}{10}\ln\left(\frac{0.2}{0.3}\right) + \frac{1}{10}\ln\left(\frac{0.1}{0.2}\right) = 0.1\ln(6) \quad (3.6)$$

oraz symetrycznie

$$D_{KL}(Q||P) = \frac{4}{10}\ln\left(\frac{0.4}{0.4}\right) + \frac{1}{10}\ln\left(\frac{0.1}{0.3}\right) + \frac{3}{10}\ln\left(\frac{0.3}{0.2}\right) + \frac{2}{10}\ln\left(\frac{0.2}{0.1}\right) = 0.1\ln(4.5). \quad (3.7)$$

Stąd widzimy, że dywergencja KL nie jest symetryczna. Dywergencję KL dla rozkładów dyskretnych P i Q można interpretować jako sumę składników postaci: $d_n = P(A = n) \cdot \ln(\frac{P(A=n)}{Q(A=n)})$. Możemy zauważyć zaletę D_{KL}, czyli jeśli dla tego samego zdarzenia $(A = n)$ prawdopodobieństwa P oraz Q są takie same, to wówczas $d_n = 0$.

3.2. Wielowymiarowy rozkład normalny oraz lognormalny

W tym podrozdziale udowodnimy, iż f-dywergencja między dwoma rozkładami normalnymi d-wymiarowymi $X_p \sim \mathcal{N}(\mu_p, \Sigma_p)$ i $X_q \sim \mathcal{N}(\mu_q, \Sigma_q)$ jest identyczna jak f-dywergencja między odpowiadającymi im rozkładami lognormalnymi $Y_p \sim \Lambda(\mu_p, \Sigma_p)$ i $Y_q \sim \Lambda(\mu_q, \Sigma_q)$, gdzie $\Lambda(\mu, \Sigma)$ oznacza rozkład lognormalny o parametrach μ oraz Σ.

Oznacza to, że:

$$p_{Y_p}(y) = p_{X_p}(\log(y)) \prod_{i=1}^{d} \frac{1}{y_i}, \tag{3.8}$$

gdzie $p_{X_p}(x)$ jest gęstością rozkładu normalnego $\mathcal{N}(\mu_p, \Sigma_p)$ dla X_p, a $y = \exp(x)$.

Podobnie, dla Y_q mamy:

$$p_{Y_q}(y) = p_{X_q}(\log(y)) \prod_{i=1}^{d} \frac{1}{y_i}. \tag{3.9}$$

Jakobian tej transformacji dla $i, j = 1, ..., d$ będzie miał postać:

$$J = \frac{\partial y_i}{\partial x_j} = \delta_{ij} \exp(x_i), \tag{3.10}$$

gdzie $d_{ij} = \begin{cases} 1 & \text{dla } i = j, \\ 0 & \text{dla } i \neq j. \end{cases}$, czyli:

$$J = \operatorname{diag}(\exp(x_1), \exp(x_2), \ldots, \exp(x_d)). \tag{3.11}$$

Wyznacznik wynosi:

$$\det(J) = \exp\left(\sum_{i=1}^{d} x_i\right). \tag{3.12}$$

Zatem, stosując twierdzenie o zamianie zmiennych, gęstość rozkładu lognormalnego $p_Y(y)$ jest wyrażona przez gęstość rozkładu normalnego $p_X(x)$ i wyznacznik Jakobiana:

$$p_Y(y) = p_X(\log(y)) \prod_{i=1}^{d} \frac{1}{y_i}. \tag{3.13}$$

f-dywergencja między rozkładami Y_p i Y_q jest zdefiniowana jako:

$$D_f(Y_p \| Y_q) = \int_{\mathbb{R}_+^d} f\left(\frac{p_{Y_p}(y)}{p_{Y_q}(y)}\right) p_{Y_q}(y)\, dy. \tag{3.14}$$

Z transformacji zmiennych $y = \exp(x)$, mamy:

$$D_f(Y_p \| Y_q) = \int_{\mathbb{R}^d} f\left(\frac{p_{X_p}(x)}{p_{X_q}(x)}\right) p_{X_q}(x)\, dx. \tag{3.15}$$

Jest to istotnie f-dywergencja między rozkładami normalnymi $X_p \sim \mathcal{N}(\mu_p, \Sigma_p)$ oraz $X_q \sim \mathcal{N}(\mu_q, \Sigma_q)$.

Z powyższego wynika, że f-dywergencja między dwoma rozkładami normalnymi X_p i X_q jest identyczna jak f-dywergencja między odpowiadającymi im rozkładami lognormalnymi Y_p i Y_q:

$$D_f(Y_p \| Y_q) = \int f\left(\frac{p_X(x)}{q_X(x)}\right) q_X(x)\, dx = D_f(X_p \| X_q). \tag{3.16}$$

Stąd widzimy, że f-dywergencja między dwiema wielowymiarowymi lognormalnymi rozkładami jest taka sama jak f-dywergencja między odpowiadającymi im rozkładami normalnymi. Wynika to z faktu, że transformacja logarytmiczna zachowuje strukturę stosunku gęstości i odpowiednio przekształca całkowanie.

3.3. Szybkość zmian D_{KL} dla rozkładu normalnego

W tym podrozdziale przyjrzymy się wielowymiarowym rozkładom normalnym. Znajdują one szerokie zastosowanie w przewidywaniu kursów na giełdzie czy przy treningu modeli w głębokim uczeniu neuronowym, więc nasze obliczenia mogą być wykorzystane między innymi w tych dziedzinach. Nazwijmy zmienną losową $X = (X_1, ..., X_n)$ jako n-wymiarowy wektor losowy o rozkładzie Gaussa z wektorem średnich $\mu = \mathbb{E}[X] = (\mathbb{E}[X_1], ..., \mathbb{E}[X_n])^\intercal$ oraz $n \times n$ macierzą kowariancji Σ, gdzie $\Sigma_{ij} = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] = Cov[X_i, X_j]$. Wówczas gęstość X jest dana wzorem

$$p(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_p|}} \exp\left((-1/2)(x-\mu)^\intercal \Sigma^{-1}(x-\mu)\right). \tag{3.17}$$

Rozważmy funkcję $g = (\mathbf{x}-\mu)^\intercal \Sigma^{-1} (\mathbf{x}-\mu)$ jako formę kwadratową. Wówczas możemy łatwo policzyć, że

$$\dot{g} = \frac{d}{dt} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - \mu_i) s_{ij} (x_j - \mu_j), \tag{3.18}$$

gdzie $\Sigma^{-1} = [s_{ij}]$ oraz $\Sigma = [\sigma_{ij}]$. Następnie

$$\begin{aligned}
\dot{g} &= \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\dot{s_{ij}}(x_i - \mu_i)(x_j - \mu_j) + s_{ij} \frac{d}{dt}\left((x_i - \mu_i)(x_j - \mu_j)\right) \right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} (\dot{s_{ij}}(x_i - \mu_i)(x_j - \mu_j) + s_{ij}\left(-\dot{\mu_i} x_j - \dot{\mu_j} x_i + \dot{\mu_i}\mu_j + \mu_i \dot{\mu_j}\right)) \\
&= (x-\mu)^\intercal (\dot{\Sigma^{-1}})(x-\mu) - \dot{\mu}^\intercal \Sigma^{-1} x - x^\intercal \Sigma^{-1} \dot{\mu} + \dot{\mu}^\intercal \Sigma^{-1} \mu + \mu^\intercal \Sigma^{-1} \dot{\mu} \\
&= (x-\mu)^\intercal (\dot{\Sigma^{-1}})(x-\mu) - 2\dot{\mu}^\intercal \Sigma^{-1} x + 2\dot{\mu}^\intercal \Sigma^{-1} \mu \\
&= (x-\mu)^\intercal (\dot{\Sigma^{-1}})(x-\mu) - 2\dot{\mu}^\intercal \Sigma^{-1} (x-\mu),
\end{aligned} \tag{3.19}$$

gdzie $(\dot{\Sigma^{-1}}) = \frac{d}{dt}\left(\Sigma^{-1}\right)$. Używając wzoru Jacobiego [20] możemy obliczyć pochodną n- wymiarowej gęstości rozkładu normalnego:

$$\begin{aligned}
\dot{p} &= \frac{d}{dt} \frac{\exp(-g/2)}{\sqrt{(2\pi)^d |\Sigma|}} \\
&= \frac{1}{\sqrt{(2\pi)^d}} \left(\frac{\exp(-g/2) \cdot (-1/2) \cdot \dot{g} \cdot \sqrt{|\Sigma|} - \exp(-g/2) \frac{1}{2\sqrt{|\Sigma|}} |\Sigma| \operatorname{Tr}\left(\Sigma^{-1}\dot{\Sigma}\right)}{|\Sigma|} \right) \\
&= \frac{1}{\sqrt{(2\pi)^d}} \left(\frac{-\exp(-g/2) \cdot \dot{g} - \exp(-g/2) \operatorname{Tr}\left(\Sigma^{-1}\dot{\Sigma}\right)}{2\sqrt{|\Sigma|}} \right) \\
&= -p \cdot \left(\dot{g} + \operatorname{Tr}\left(\Sigma^{-1}\dot{\Sigma}\right) \right)/2,
\end{aligned} \tag{3.20}$$

gdzie $|\Sigma|$ oznacza wyznacznik macierzy Σ, a " $\cdot$ " oznacza mnożenie. Skoro (3.20), to zachodzi:

$$\begin{aligned} I_F(p) &= \mathbb{E}_p\left[\left(\frac{\dot{p}}{p}\right)^2\right] \\ &= \frac{1}{4}\mathbb{E}_p\left[\left(\dot{g}^2 + 2\dot{g}\,\mathrm{Tr}\left(\Sigma^{-1}\dot{\Sigma}\right) + \mathrm{Tr}^2\left(\Sigma^{-1}\dot{\Sigma}\right)\right)\right] \\ &= \frac{1}{4}\left(\mathrm{Tr}^2\left(\Sigma^{-1}\dot{\Sigma}\right) + 2\,\mathrm{Tr}\left(\Sigma^{-1}\dot{\Sigma}\right)\mathbb{E}_p\left[\dot{g}\right] + \mathbb{E}_p\left[\dot{g}^2\right]\right). \end{aligned} \tag{3.21}$$

Następnie, przeanalizujemy każdy ze składników z równania powyżej:

$$\begin{aligned} \mathbb{E}_p\left[\dot{g}\right] &= \mathbb{E}_p\left[(x-\mu)^\intercal(\dot{\Sigma^{-1}})(x-\mu) - 2\dot{\mu}^\intercal\Sigma^{-1}(x-\mu)\right] \\ &= \sum_{i,j}\mathbb{E}_p\left[(x_i-\mu_i)(x_j-\mu_j)\dot{s}_{ij} - 2\dot{\mu}_i s_{ij}(x_j-\mu_j)\right] \\ &= \sum_{i,j}\sigma_{ij}\dot{s}_{ij} \\ &= \mathrm{Tr}\left(\Sigma\frac{d}{dt}\left(\Sigma^{-1}\right)\right). \end{aligned} \tag{3.22}$$

Isserlis udowodnił wcześniej [21], że dla nieparzystej liczby zmiennych gaussowskich o średniej 0 wartość oczekiwana ich iloczynu również jest równa 0, więc możemy zapisać, że dla $i,j,k,l = 1,...n$ zachodzi:

$$\mathbb{E}_p\left[\sum_{i,j,k,l}(x_i-\mu_i)\dot{s_{ij}}(x_j-\mu_j)\dot{\mu_k}s_{k,l}(x_l-\mu_l)\right] = 0 \tag{3.23}$$

oraz dla parzystej liczby zmiennych losowych [22] w jednym składniku

$$\begin{aligned} \mathbb{E}_p\left[\left((x-\mu)^\intercal(\dot{\Sigma^{-1}})(x-\mu)\right)^2\right] &= \sum_{i,j,k,l}\mathbb{E}_p\left[(x_i-\mu_i)(x_j-\mu_j)(x_k-\mu_k)(x_l-\mu_l)\dot{s_{ij}}\dot{s_{kl}}\right] \\ &= \sum_{k,l}\sum_{i,j}(\sigma_{ij}\sigma_{kl}+\sigma_{ik}\sigma_{jl}+\sigma_{il}\sigma_{jk})\dot{s_{ij}}\dot{s_{kl}} \\ &= \sum_{k,l}\dot{s_{kl}}\left(\mathrm{Tr}\left(\Sigma\frac{d}{dt}\left(\Sigma^{-1}\right)\right)\sigma_{kl} + 2\bar{\sigma}_k\frac{d}{dt}\left(\Sigma^{-1}\right)\hat{\sigma}_l\right) \\ &= \mathrm{Tr}\left(\Sigma\frac{d}{dt}\left(\Sigma^{-1}\right)\right) + 2\,\mathrm{Tr}\left(\left(\frac{d}{dt}\left(\Sigma^{-1}\right)\Sigma\right)^2\right), \end{aligned} \tag{3.24}$$

gdzie $\hat{\sigma}_l$ to wektor stworzony z l-tej kolumny macierzy Σ oraz $\bar{\sigma}_k$ oznacza wektor stworzony z k-tego wiersza macierzy Σ, zaś $\mathrm{Tr}^2(\cdot) = [\mathrm{Tr}(\cdot)]^2$. Następnie używając własności rozkładu normalnego dostajemy

$$\begin{aligned} \mathbb{E}_p\left[\left(\dot{\mu}\Sigma^{-1}(x-\mu)\right)^2\right] &= \mathbb{E}_p\left[\mathcal{N}(0,\dot{\mu}\Sigma^{-1}\Sigma\Sigma^{-1})\dot{\mu}^\intercal\right] \\ &= \dot{\mu}\Sigma^{-1}\dot{\mu}^\intercal. \end{aligned} \tag{3.25}$$

Skoro dla macierzy $n \times n$ oznaczonej $\mathbf{0}$, która zawiera same zera zachodzi tożsamość:

$$\mathbf{0} = \frac{d}{dt}\left(\Sigma\Sigma^{-1}\right) = \dot{\Sigma}\Sigma^{-1} + \Sigma\frac{d}{dt}\left(\Sigma^{-1}\right), \tag{3.26}$$

to równoważnie zachodzi

$$\frac{d}{dt}\left(\Sigma^{-1}\right) = -\Sigma^{-1}\dot{\Sigma}\Sigma^{-1}. \tag{3.27}$$

W takim razie zachodzi:

$$\begin{aligned} I_F(p) &= \frac{1}{4}\mathbb{E}_p\left[\left(\dot{g}^2 + 2\dot{g}\,\mathrm{Tr}\left(\Sigma^{-1}\dot{\Sigma}\right) + \mathrm{Tr}^2\left(\Sigma^{-1}\dot{\Sigma}\right)\right)\right] \\ &= \frac{1}{4}\left(\mathrm{Tr}^2\left(\Sigma^{-1}\dot{\Sigma}\right) + 2\,\mathrm{Tr}\left(\Sigma^{-1}\dot{\Sigma}\right)\mathrm{Tr}\left(-\Sigma\Sigma^{-1}\dot{\Sigma}\Sigma^{-1}\right)\right. \\ &\quad \left. + \mathrm{Tr}^2\left(\Sigma^{-1}\dot{\Sigma}\right) + \mathrm{Tr}\left(\left(\dot{\Sigma^{-1}\Sigma}\right)^2\right) + 2^2\dot{\mu}\Sigma^{-1}\dot{\mu}^\intercal\right) \\ &= \frac{1}{2}\,\mathrm{Tr}\left(\left(\dot{\Sigma^{-1}\Sigma}\right)^2\right) + \dot{\mu}\Sigma^{-1}\dot{\mu}^\intercal. \end{aligned} \tag{3.28}$$

Możemy obliczyć dywergencję KL [23] między dwoma d-wymiarowymi rozkładami Gaussa $X \sim \mathcal{N}(\mu_p, \Sigma_p)$ oraz $Y \sim \mathcal{N}(\mu_q, \Sigma_q)$

$$D_{KL} = \frac{1}{2}\left(\ln\left(\frac{|\Sigma_q|}{|\Sigma_p|}\right) + \mathrm{Tr}(\Sigma_q^{-1}\Sigma_p) + (\mu_q - \mu_p)^\intercal\Sigma_q^{-1}(\mu_q - \mu_p) - d\right). \tag{3.29}$$

Stąd możemy obliczyć, że:

$$\dot{D_{KL}} = \frac{1}{2}\,\mathrm{Tr}\left(\Sigma_q^{-1}\dot{\Sigma_q} - \Sigma_p^{-1}\dot{\Sigma_p} + \frac{d}{dt}\left(\Sigma_q^{-1}\right)\Sigma_p + \Sigma_q^{-1}\dot{\Sigma_p}\right) + \frac{1}{2}\frac{d}{dt}\left((\mu_q - \mu_p)^\intercal\Sigma_q^{-1}(\mu_q - \mu_p)\right). \tag{3.30}$$

Korzystając z (3.19) dostajemy:

$$\begin{aligned} \dot{D_{KL}} = \frac{1}{2}\,\mathrm{Tr}[&\Sigma_q^{-1}\dot{\Sigma_q} - \Sigma_p^{-1}\dot{\Sigma_p} + \\ &+ \frac{d}{dt}\left(\Sigma_q^{-1}\right)\Sigma_p + \Sigma_q^{-1}\dot{\Sigma_p}] + \frac{1}{2}\left((\mu_p - \mu_q)^\intercal\frac{d}{dt}(\Sigma_q^{-1})(\mu_p - \mu_q) - 2\frac{d}{dt}(\mu_p - \mu_q)^\intercal\Sigma_q^{-1}(\mu_p - \mu_q)\right). \end{aligned} \tag{3.31}$$

Dodatkowo, możemy policzyć [24], że w tym przypadku T_2 jest równe:

$$T_2 = \exp\left((\mu_p - \mu_q)^\intercal[(2\Sigma_q - \Sigma_p)]^{-1}(\mu_p - \mu_q) - \frac{1}{2}\ln\frac{|2\Sigma_q - \Sigma_p|}{|\Sigma_p|^{-1}|\Sigma_q|^2}\right) - 1, \tag{3.32}$$

co zachodzi wtedy i tylko wtedy gdy $2\Sigma_q - \Sigma_p \in (0, \infty)$, w przeciwnym wypadku całka do wyliczania T_2 nie jest zbieżna.

Zanim pokażemy przykład zastosowania otrzymanych wyników policzmy jeszcze informację Fishera dla analogicznych rozkładów lognormalnych.

3.4. Czasowa informacja Fishera dla wielowymiarowego rozkładu lognormalnego

Policzono informację Fishera między dwoma rozkładami lognormalnymi d-wymiarowymi $X \sim \Lambda(\mu_p, \Sigma_p)$ oraz $Y \sim \Lambda(\mu_q, \Sigma_q)$, co oznacza, że gęstość X jest dana wzorem

$$p(x) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_p|}\prod_i^d x_i}\exp\left((-1/2)(\ln(x) - \mu_p)^\intercal\Sigma_p^{-1}(\ln(x) - \mu_p)\right), \tag{3.33}$$

gdzie x jest d-wymiarowym wektorem i $x = (x_1, ..., x_d)$. Podobnie, dla Y i jej gęstości q.

Analogicznie jak w podrozdziale (3.3) możemy policzyć pochodną gęstości tego rozkładu i otrzymujemy

$$\dot{p} = -p \cdot \left(\dot{h} + \operatorname{Tr}\left(\Sigma^{-1} \dot{\Sigma} \right) \right), \tag{3.34}$$

gdzie $h = (\ln(x) - \mu)^\intercal \Sigma^{-1} (\ln(x) - \mu)$. Widzimy więc, że pochodna gęstości zarówno z wielowymiarowego normalnego rozkładu oraz z lognormalnego są bardzo podobne. W następnym kroku również analogicznie obliczymy informację Fishera:

$$\begin{aligned} I_F(p) &= \mathbb{E}_p \left[\left(\frac{\dot{p}}{p} \right)^2 \right] \\ &= \frac{1}{2} \operatorname{Tr}\left(\left(\dot{\Sigma^{-1}} \Sigma \right)^2 \right) + \dot{\mu} \Sigma^{-1} \dot{\mu}^\intercal. \end{aligned} \tag{3.35}$$

Ostatnia równość jest analogiczna do (3.3).

Stąd widzimy, że każda z przedstawionych wcześniej dywergencji oraz informacja Fishera są takie same między rozkładami normalnymi oraz lognormalnymi, o ile mają te same parametry średnich oraz macierzy kowariancji. W związku z tym, w dalszych rozważaniach badano jedynie rozkłady normalne. Pokazano następnie zastosowanie otrzymanych wyników.

Rozdział 4

Zastosowania

W tym rozdziale przyjerzano się dwyergencji Kullbacka-Leiblera (2.3) oraz nierówności (2.50).

4.1. Równanie dyfuzji dla rozkładu normalnego

W tym podrozdziale przyjrzano się zastosowaniu dywergencji Kullbacka-Leiblera (2.6) dla pewnego rozwiązania równania dyfuzji. Dyfuzją nazywamy "proces rozprzestrzeniania się cząsteczek lub energii w danym ośrodku (np. w gazie, cieczy lub ciele stałym), będący konsekwencją chaotycznych zderzeń cząsteczek dyfundującej substancji między sobą i/lub z cząsteczkami otaczającego ją ośrodka"[25]. Jest ona zjawiskiem występującym w różnych procesach fizycznych, chemicznych i biologicznych. Często jest związana z oddziaływaniami na złożonych i niejednorodnych strukturach. Można ją zaobserwować na przykład w procesach związanych z nagłymi zmianami temperatury. Tradycyjne jednowymiarowe równanie dyfuzji opisuje się za pomocą równania różniczkowego cząstkowego [26, 27, 28] danego wzorem:

$$\frac{du}{dt} = D\frac{d^2u}{dx^2}, \tag{4.1}$$

gdzie $u(x,t)$ jest gęstością prawdopodobieństwa zmiennej x w czasie t, a λ stałą dyfuzji [26]. Z prostych obliczeń wynika, że jeśli warunkiem początkowym będzie $u(x,0) = \frac{1}{\sqrt{2\pi}\sigma_0}\exp\left(\frac{-(x-\mu_0)^2}{2\sigma_0^2}\right)$, to rozwiązaniem równania (4.1) jest:

$$u(x,t) = \frac{1}{\sqrt{2\pi(\sigma_0^2+2\lambda t)}}\exp\left(-\frac{(x-\mu_0)^2}{2(\sigma_0^2+2\lambda t)}\right). \tag{4.2}$$

Wynik ten możemy interpretować jako "rozmywanie"się rozkładu początkowego względem czasu. Przebadamy teraz przebieg wartości dywergencji Kullbacka-Leiblera dla tego układu. Korzystając z (3.29) oraz (3.31) otrzymujemy w naszym przypadku:

$$D_{KL}\left(u(x,t)||u(x,0)\right) = \frac{1}{2}\left(\ln\left(\frac{\sigma_0^2}{\sigma_0^2+2\lambda t}\right) + \frac{\sigma_0^2+2\lambda t}{\sigma_0^2} - 1\right) \tag{4.3}$$

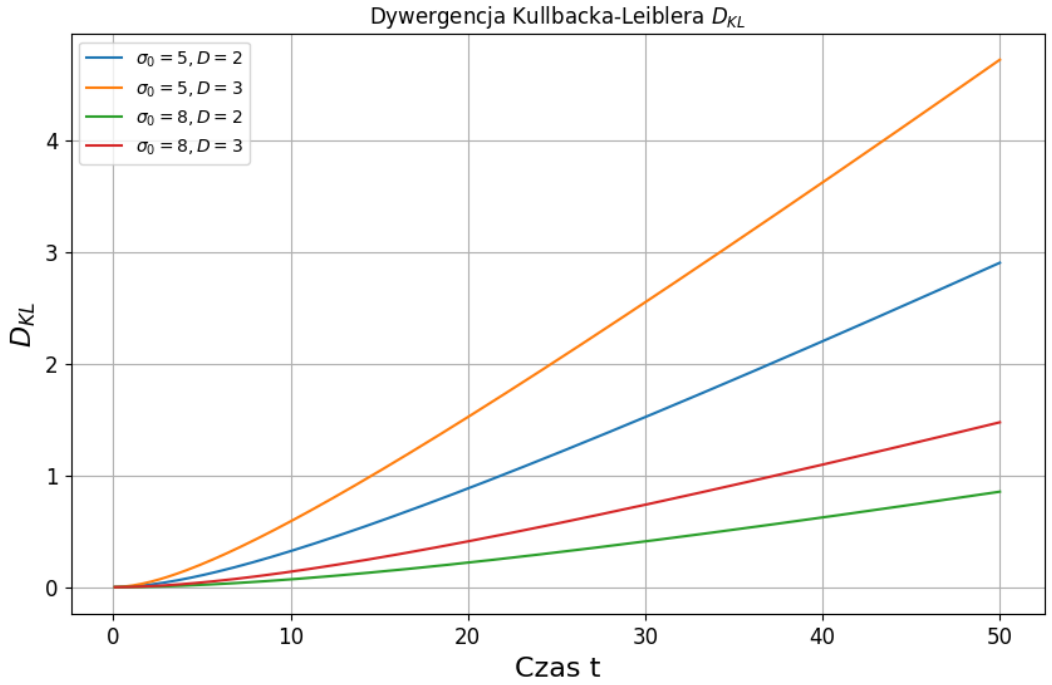
oraz

$$\frac{d}{dt}D_{KL} = -\frac{\lambda}{\sigma_0^2+2\lambda t} + \frac{\lambda}{\sigma_0^2} \tag{4.4}$$

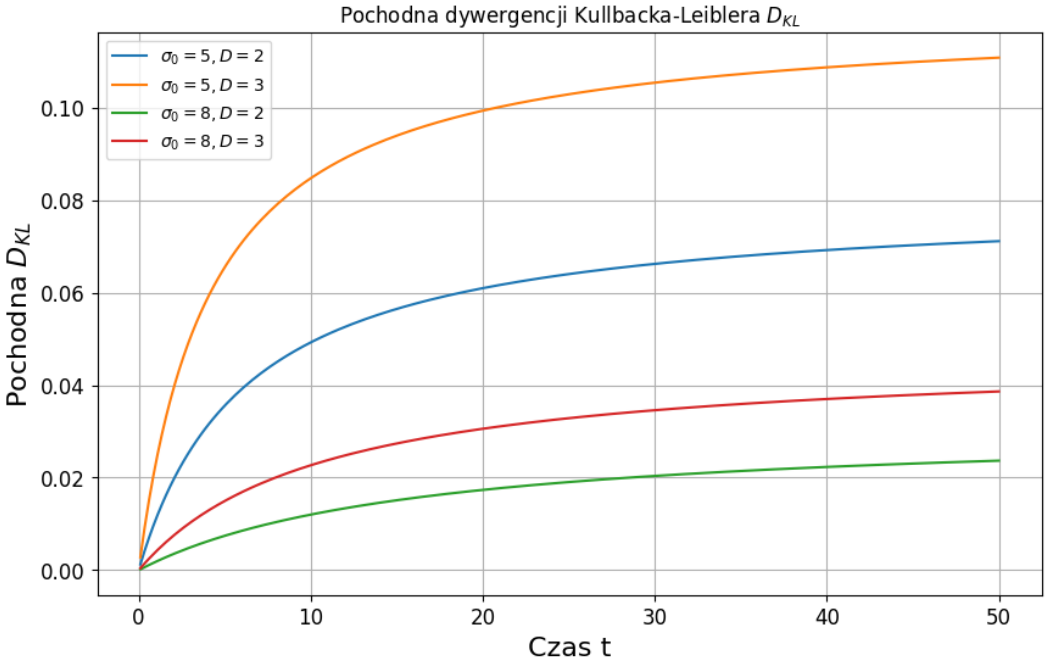
Wzrost dywergencji Kullbacka-Leiblera (KL) między układem obecnym a początkowym można zinterpretować następująco: ponieważ D_{KL} opiera się na logarytmie, a wariancja rośnie liniowo, dla dużych wartości czasu t dywergencja KL zaczyna zachowywać się liniowo.

Taka zależność jest zgodna z intuicją fizyczną: na początku procesy dyfuzji cząstek lub rozprzestrzeniania się ciepła zachodzą szybko, ponieważ gradienty są duże. W miarę wyrównywania się koncentracji (lub temperatury), zmiany stają się coraz mniej intensywne, choć równocześnie rozprzestrzeniają się na coraz dalsze obszary.

Warto jednak zauważyć, że analiza wzorów (co można również zaobserwować na załączonych wykresach pochodnej) wskazuje, iż pochodna dywergencji zbliża się do wartości stałej równej $\frac{\lambda}{\sigma_0^2}$.



Rysunek 4.1: Zmiany dywergencji Kullbacka-Leiblera ($D_{KL}(u(x,t)||u(x,0))$) w czasie t dla różnych wartości parametrów σ_0 (inicjalna wariancja) oraz λ (dyfuzji).



Rysunek 4.2: Zmiany pochodnej dywergencji $\frac{\mathrm{d}D_{KL}(u(x,t)\|u(x,0))}{\mathrm{d}t}$ względem czasu t dla różnych wartości parametrów σ_0^2 (inicjalna wariancja) oraz λ (współczynnik dyfuzji).

Możemy teraz obliczyć tez inne wielkości charakteryzujące między rozkładami o gęstościach $u(x,t)$ oraz $u(x,0)$, aby ostatecznie sprawdzić jak wygląda w tym przypadku nierówność (2.50). Zacznijmy od informacji Fishera (3.28):

$$I_F(u(x,t)) = \frac{2\lambda^2}{\left(\sigma_0^2 + 2\lambda t\right)^2} \tag{4.5}$$

oraz

$$I_F(u(x,0)) = 0. \tag{4.6}$$

Ostatnia równość wynika z tego, iż informacja Fishera bada szybkość zmian, a $u(x,0)$ nie zmienia się w czasie. Następnie musimy jeszcze policzyć dywergencję χ^2, czyli T_2 (3.32):

$$\begin{aligned} T_2(u(x,t)||u(x,0)) &= \left(\frac{2\sigma_0^2 - \sigma_0^2 - 2\lambda t}{\frac{\sigma_0^4}{\sigma_0^2 + 2\lambda t}}\right)^{-1/2} - 1 \\ &= \sqrt{\frac{\sigma_0^4}{\sigma_0^4 - 4\lambda^2 t^2}} - 1. \end{aligned} \tag{4.7}$$

Równość (4.7) staje się rozbieżna dla $t \geq \frac{\sigma_0^2}{2\lambda}$, ponieważ:

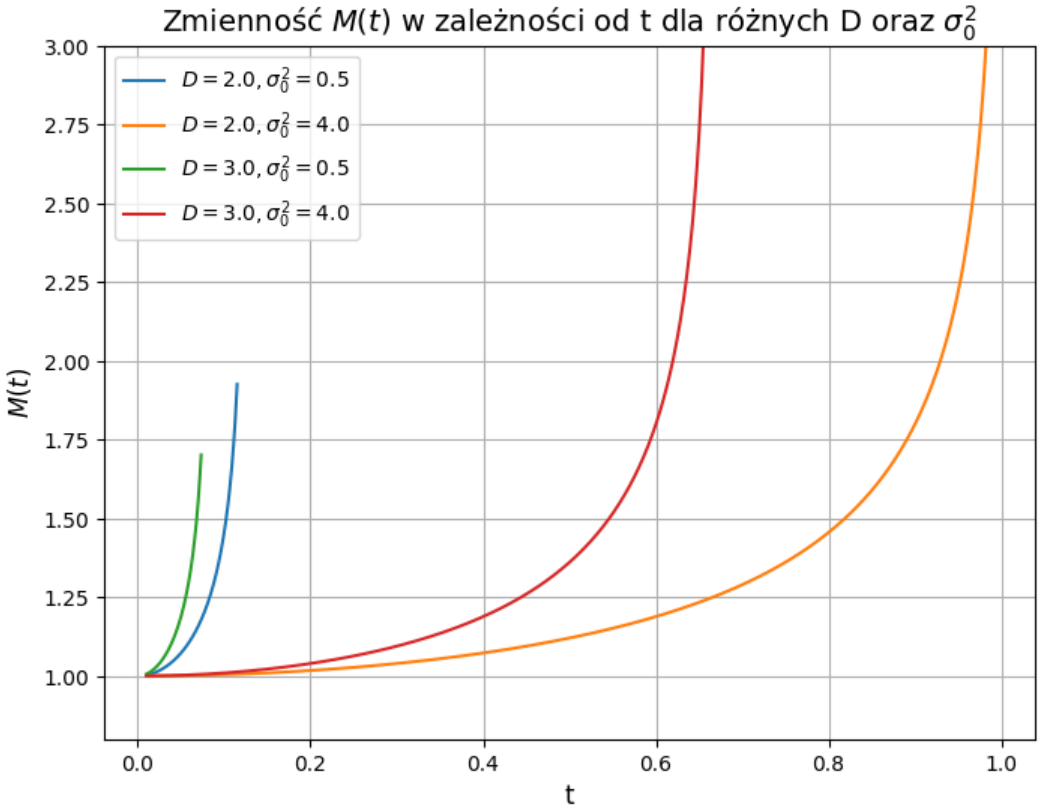
$$\begin{aligned} T_2(u(x,t)||u(x,0)) &= \int_{\mathbb{R}} \frac{u^2(x,t)}{u(x,0)} dx \\ &= \int_{\mathbb{R}} \frac{\sigma_0}{\sqrt{2\pi}(\sigma_0^2+2\lambda t)} \exp\left(-\frac{(x-\mu_0)^2}{\sigma_0^2+2\lambda t} + \frac{(x-\mu_0)^2}{2\sigma_0^2}\right) dx \\ &= \int_{\mathbb{R}} \frac{\sigma_0}{\sqrt{2\pi}(\sigma_0^2+2\lambda t)} \exp\left((x-\mu_0)^2 \cdot \frac{2\lambda t - \sigma_0^2}{\sigma_0^2(\sigma_0^2+2Dt)}\right) dx, \end{aligned} \tag{4.8}$$

co jest zbieżne wtedy i tylko wtedy, gdy $2\lambda t - \sigma_0^2 < 0$. W takim razie nierówność (2.49) w tym przypadku przyjmuje formę:

$$\begin{aligned} &\left|\frac{\lambda}{\sigma_0^2} - \frac{\lambda}{\sigma_0^2+2\lambda t}\right| \\ &\leq \sqrt{\frac{2\lambda^2}{\left(\sigma_0^2+2\lambda t\right)^2} \cdot \left(\sqrt{\frac{\sigma_0^4}{\sigma_0^4-4\lambda^2t^2}} - 1 - \left(\frac{1}{2}\left(\ln\left(\frac{\sigma_0^2}{\sigma_0^2+2\lambda t}\right) + \frac{\sigma_0^2+2\lambda t}{\sigma_0^2} - 1\right)\right)^2\right)}. \end{aligned} \tag{4.9}$$

Jest to jest równoważne formule:

$$\begin{aligned} 1 &\leq \frac{\sqrt{\frac{2}{\left(\sigma_0^2+2\lambda t\right)^2} \cdot \left(\sqrt{\frac{\sigma_0^4}{\sigma_0^4-4\lambda^2t^2}} - 1 - \frac{1}{2}\left(\ln\left(\frac{\sigma_0^2}{\sigma_0^2+2\lambda t}\right) + \frac{\sigma_0^2+2\lambda t}{\sigma_0^2} - 1\right)^2\right)}}{\left|\frac{2\lambda t}{\sigma_0^2(\sigma_0^2+2\lambda t)}\right|} \\ &= \frac{\sigma_0^2\sqrt{\left(\sqrt{\frac{\sigma_0^4}{\sigma_0^4-4\lambda^2t^2}} - 1 - \left(\frac{1}{2}\left(\ln\left(\frac{\sigma_0^2}{\sigma_0^2+2\lambda t}\right) + \frac{\sigma_0^2+2\lambda t}{\sigma_0^2} - 1\right)\right)^2\right)}}{\sqrt{2}\lambda t} := M(t). \end{aligned} \tag{4.10}$$



Rysunek 4.3: Wykres prawej strony nierówności (2.50), gdzie funkcja $M(t)$ jest zdefiniowana jako:

$$M(t) = \frac{\sigma_0^2 \sqrt{\left(\sqrt{\frac{\sigma_0^4}{\sigma_0^4 - 4\lambda^2 t^2}} - 1 - \left(\frac{1}{2}\left(\ln\left(\frac{\sigma_0^2}{\sigma_0^2 + 2\lambda t}\right) + \frac{\sigma_0^2 + 2\lambda t}{\sigma_0^2} - 1\right)\right)^2\right)}}{\sqrt{2}\lambda t}.$$

Skoro dla rozkładów lognormalnych oraz normalnych f-dywergencje oraz informacja Fishera zachowują się tak samo (co pokazaliśmy w rozdziale (3.2)), to możemy skorzystać z silniejszej nierówności (2.43) dla naszego przypadku o postaci:

$$1 \leq \frac{\sqrt{I_F(u(x,t)) \cdot \left[\mathbb{E}_p[\ln^2(u(x,t)/u(x,0))] - D_{KL}^2\right]}}{|\frac{d}{dt} D_{KL}|}. \tag{4.11}$$

Zostało udowodnione [10], że w nierówności (4.12) zachodzi równość dla rozkładów lognormalnych kiedy średnie obydwu rozkładów są takie same. W przypadku (4.2) średnie rzeczywiście są takie same, więc dla każdego $t < \frac{\sigma_0^2}{2D}$ zachodzi

$$1 = \frac{\sqrt{I_F(u(x,t)) \cdot \left[\mathbb{E}_p[\ln^2(u(x,t)/u(x,0))] - D_{KL}^2\right]}}{|\frac{d}{dt} D_{KL}|}. \tag{4.12}$$

4.2. Aproksymacja skorelowanego rozkładu nieskorelowanym

W tym podrozdziale rozpatrzono szczególny przypadek zastosowania dywergencji Kullbacka-Leiblera. Mając 2-wymiarową dystrybucję skorelowaną $X_t \sim \mathcal{N}(\mu_x(t), \Sigma_x(t))$ zaproponowano zasymulowanie jej jak najdokładniej dwoma niezależnymi zmiennymi losowymi normalnymi, czyli innymi słowy zmienną $Y_t \sim \mathcal{N}(\mu_y(t), \Sigma_y(t))$ nieskorelowaną. W związku z tym niech $\Sigma_x = \begin{bmatrix} \sigma_{11}(t) & \sigma_{12}(t) \\ \sigma_{12}(t) & \sigma_{22}(t) \end{bmatrix}$. Podobnie $\Sigma_y = \begin{bmatrix} \sigma_{11}(t)/f(t) & 0 \\ 0 & \sigma_{22}(t)/g(t) \end{bmatrix}$, gdzie f oraz g są różniczkowalnymi funkcjami rzeczywistymi nieujemnymi, oraz Σ_x i Σ_y są macierzami kowariancji. Dodatkowo niech $\mu_x(t) = [\mu_{x1}(t), \mu_{x2}(t)]^\intercal$ i $\mu_y = [a(t) + \mu_{x1}(t), b(t) + \mu_{x2}(t)]^\intercal$, gdzie a oraz b są funkcjami różniczkowalnymi (nie muszą być dodatnie). Korzystając z (3.29) otrzymujemy, że:

$$\begin{aligned}
D_{KL}(Y_t||X_t) &= \frac{1}{2}([\ln\left(f(t)g(t)\frac{\sigma_{11}(t)\sigma_{22}(t) - \sigma_{12}^2(t)}{\sigma_{11}(t)\sigma_{22}(t)}\right) + \frac{\sigma_{22}(t)\sigma_{11}}{f(t)(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} + \\
&\frac{\sigma_{11}(t)\sigma_{22}(t)}{g(t)(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} + \frac{\sigma_{22}(t)a^2(t) + \sigma_{11}(t)b^2(t) - 2\sigma_{12}(t)a(t)b(t)}{\sigma_{11}(t)\sigma_{22}(t) - \sigma_{12}^2(t)} - 2]) \\
&\geq \frac{1}{2}\left(\ln(C) + \ln(f) + \ln(g) + \frac{1}{Cf} + \frac{1}{Cg}\right) - 1 \\
&\geq \frac{1}{2}\left(\ln(C) + \ln(Cf) - \ln(C) + \ln(Cg) - \ln(C) + \frac{1}{Cf} + \frac{1}{Cg}\right) - 1 \\
&= \frac{1}{2}\left(\ln\left(\frac{1}{C}\right) + \ln(Cf) + \frac{1}{Cf} + \ln(Cg) + \frac{1}{Cg}\right) - 1 \\
&\geq \frac{1}{2}\left(\ln\left(\frac{\sigma_{11}\sigma_{22}}{(\sigma_{11}\sigma_{22} - \sigma_{12}^2)}\right)\right) \\
&= -\frac{1}{2}\ln\left(1 - \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}}\right),
\end{aligned} \tag{4.13}$$

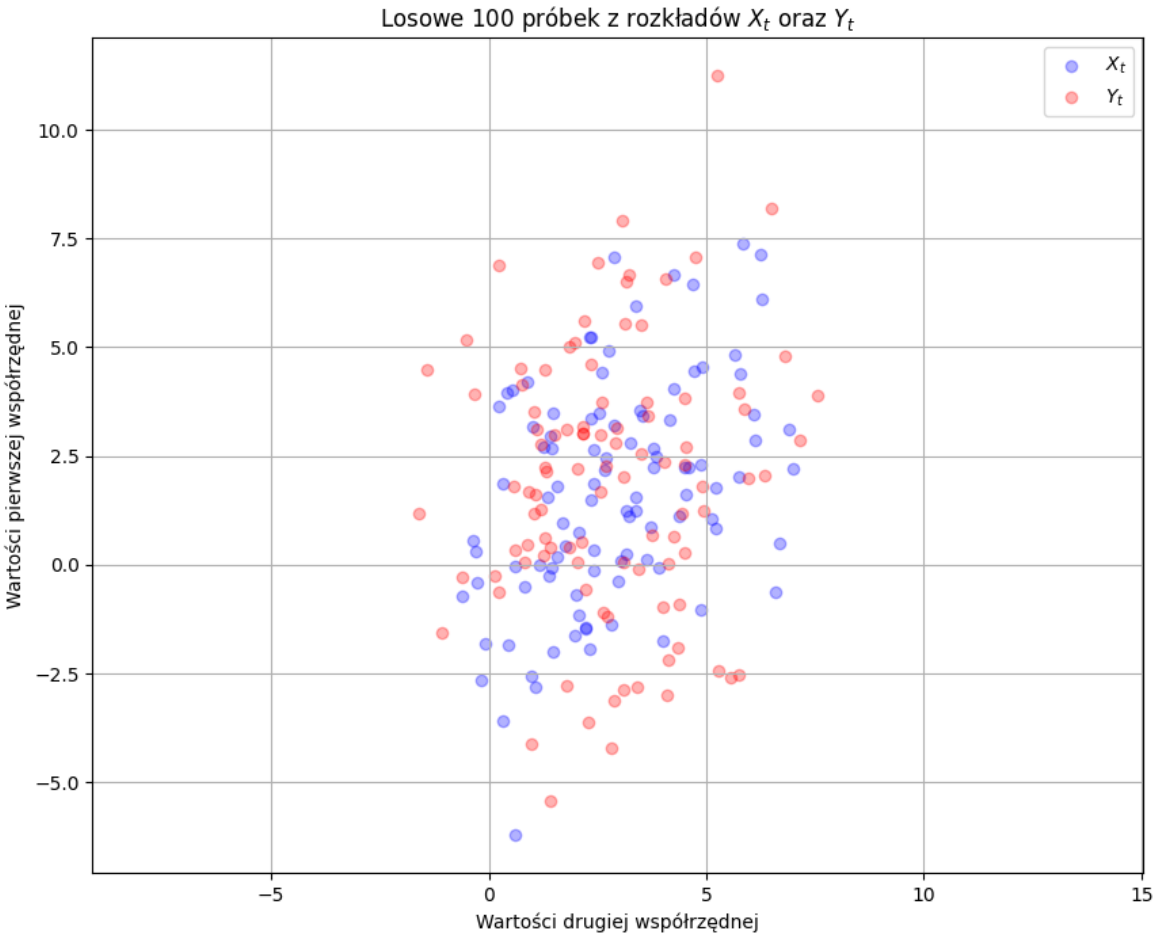
gdzie $C = \frac{\sigma_{11}(t)\sigma_{22}(t) - \sigma_{12}^2(t)}{\sigma_{11}(t)\sigma_{22}(t)}$. Przedostatnie przejście wynika z faktu, iż przekształcenie $x \mapsto \frac{1}{x} + \ln(x)$ osiąga minimum w punkcie $x = 1$. Nierówność (4.13) zostaje nasycona gdy $f = g = 1/C$ oraz $a = b = 0$. Możemy stąd wywnioskować, że jeśli zachodzi $\sigma_{12}^2 \ll \sigma_{11}\sigma_{22}$ (czyli że σ_{12}^2 jest znacznie mniejsze niż $\sigma_{11}\sigma_{22}$), czyli dla słabych korelacji, to nasza aproksymacja staje się wysoce dokładna. Co ciekawe, wynika stąd, że najlepsze przybliżenie skorelowanego rozkładu gaussowskiego dwuwymiarowego przez nieskorelowany jest osiągane dla dokładnie tych samych parametrów średnich oraz macierzy kowariancji poza wartościami σ_{12}, które muszą być równe zero (aby był to rozkład nieskorelowany). Na rysunku (4.4) przedstawiono po 100 losowych próbek dla rozkładu X_t oraz Y_t o parametrach kolejno

$$\mu_x = [3, 2], \quad \Sigma_x = \begin{bmatrix} 4 & 3 \\ 3 & 9, \end{bmatrix} \tag{4.14}$$

oraz

$$\mu_y = [3, 2], \quad \Sigma_y = \begin{bmatrix} 4C & 0 \\ 0 & 9C, \end{bmatrix} \tag{4.15}$$

gdzie $C = \frac{4 \cdot 9}{4 \cdot 9 - 3^2} = \frac{4}{3}$.



Rysunek 4.4: Losowe 100 próbek z rozkładów X_t oraz Y_t

4.3. Przewidywanie skuteczności drużyny piłkarskiej

W ostatnich latach coraz popularniejsze stały się modele do przewidywania padania gola po jednym strzale [29] zwane golami oczekiwanymi (xG). Są one używane zarówno przez analityków sportowych, jak i telewizje do ukazania przebiegu gry oraz oceny stwarzanych sytuacji, oraz gry zespołu w meczu. Stworzymy własne takie modele, bardzo uproszczone, ale pokazujące w jaki sposób możemy wykorzystywać D_{KL} oraz nierówność (2.50) do analizy oraz porównywania skuteczności tych modeli. Wykorzystamy w tym celu rzeczywiste dane z sezonu 2018/2019 angielskiej Premier League. Pochodzą one ze strony internetowej [30], skąd można pobrać plik CSV zawierający statystyki meczowe. Plik CSV zawiera dane dotyczące meczów piłkarskich z English Premier League z sezonu 2018/2019, w tym informacje o drużynach, które wzięły udział w każdym spotkaniu, numerze kolejki oraz liczbie strzelonych goli przez drużyny gospodarzy i gości. W pliku znajdują się następujące kolumny: home_team_name (nazwa drużyny gospodarzy), away_team_name (nazwa drużyny gości), Game Week (numer kolejki ligowej), home_team_goal_count (liczba goli strzelonych przez drużynę gospodarzy), away_team_goal_count (liczba goli strzelonych przez drużynę gości), home_team_shots (liczba strzałów oddanych przez drużynę gospodarzy), away_team_shots (liczba strzałów oddanych przez drużynę gości), home_team_shots_on_target (liczba strzałów celnych drużyny gospodarzy), away_team_shots_on_target (liczba strzałów celnych drużyny gości). Pierwsze 5 rzędów oraz 5 kolumn danych przedstawiono w tabeli (4.1).

home_team_name	away_team_name	Game Week	home_team_goal_count	away_team_goal_count
Manchester United	Leicester City	1	2	1
Newcastle United	Tottenham Hotspur	1	1	2
AFC Bournemouth	Cardiff City	1	2	0
Fulham	Crystal Palace	1	0	2
Huddersfield Town	Chelsea	1	0	3
Watford	Brighton Hove Albion	1	2	0
Wolverhampton Wanderers	Everton	1	2	2

Tabela 4.1: Tabela przedstawiająca przykładowe dane dotyczące meczów piłkarskich, w tym drużyny gospodarzy, drużyny gości, numery kolejek oraz liczbę goli strzelonych przez obie drużyny.

Analizowane dane zostały zebrane z meczów piłkarskich, w których brał udział Liverpool, a analiza skupiła się jedynie na wynikach skuteczności tej drużyny. Z pliku zostały wybrane tylko te informacje, które dotyczą liczby strzałów oraz goli zdobytych przez Liverpool w poszczególnych spotkaniach. Plik stanowi podstawę analizy efektywności ataku drużyny w kontekście liczby strzałów oddanych na bramkę oraz zdobytych goli w różnych meczach, co pozwala na ocenę wydajności tej drużyny w danej części rozgrywek. W tej pracy policzono pewien model xG dla danej Liverpoolu F.C.

W naszym podejściu uznano, iż rzeczywistą skutecznością zespołu w danej kolejce będzie liczba goli w niej strzelonych przez liczbę strzałów (to znaczy założono, że prawdopodobieństwo strzelenia gola z każdej okazji (w danym meczu) zakończonej strzałem jest takie samo). Rozkład ten oznaczymy jako Q. Niech A_t oznacza zdarzenie czy w t-tej kolejce po dowolnym strzale padł gol($Q(A_t = 1) = q_t$), czy nie ($Q(A_t = 0) = 1 - q_t$). Ustalmy, że $G(t)$ = liczba goli w t- tej kolejce oraz $S(t)$ = liczba strzałów w t- tej kolejce. Wówczas $q_t = G(t)/S(t)$. Ponieważ nasz model jest bardzo prosty, to czasami może się zdarzać sytuacja, że $q_t = 0$. Jeżeli tak się

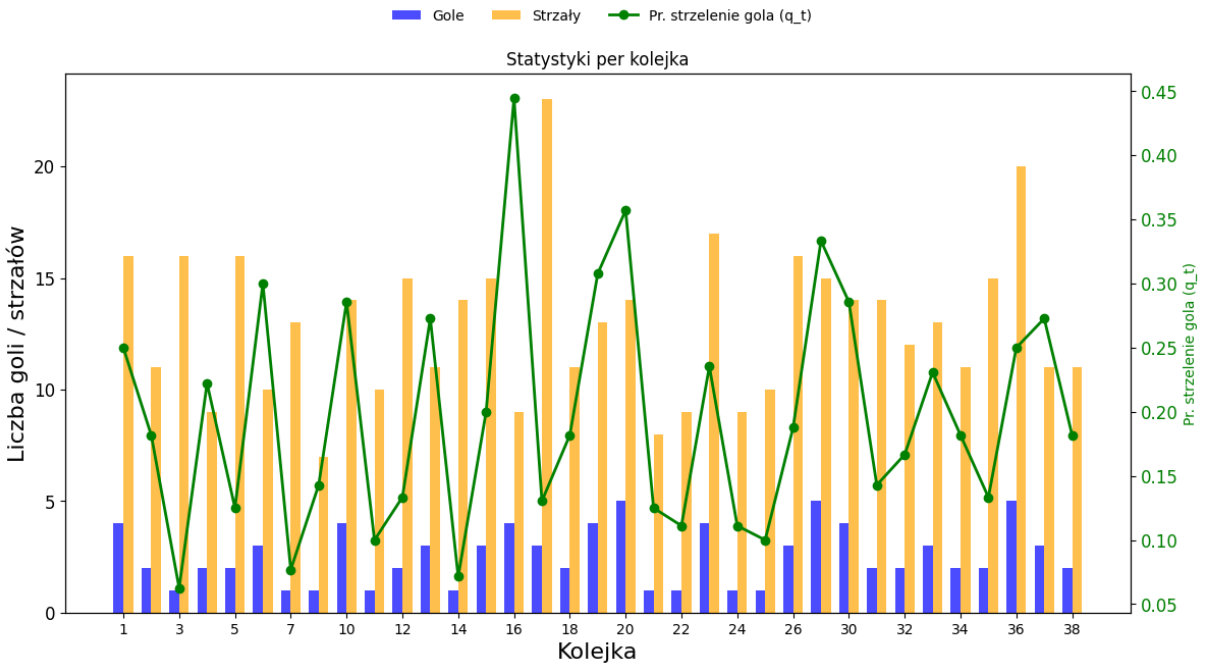
wydarzy, wówczas po prostu wyrzucimy dane dla tej kolejki z rozpatrywanych.

Dane po przetworzeniu zaprezentowano w tabeli (4.2).

kolejka	gole	strzały	q
1	4	16	0.250000
2	2	11	0.181818
3	1	16	0.062500
4	2	9	0.222222
5	2	16	0.125000
6	3	10	0.300000
7	1	13	0.076923
9	1	7	0.142857

Tabela 4.2: Tabela przedstawiająca dane dotyczące meczów Liverpoolu, w tym numer kolejki, liczbę goli, strzałów oraz wartość zmiennej Q.

Dynamika prawdopodobieństw q_t w czasie przedstawia poniższy rysunek (4.5):



Rysunek 4.5: Żółte kolumny oznaczają liczbę strzałów, niebieskie liczbę goli, zaś zielona linia to iloraz wartości tych kolumn.

W następnych podrozdziałach przedstawiono kilka metod wyliczania xG oraz pokazano wizualizacje otrzymanych wyników.

4.3.1. Iloraz wszystkich goli oraz strzałów do tej pory

W tym podejściu będziemy estymowali procent skuteczności w t-tej kolejce za pomocą prawdopodobieństwa ze skumulowanej liczby goli oraz strzałów do $(t-1)$ kolejki włącznie. Oznaczymy

ten rozkład jako J_t. Spodziewamy się, że wówczas będzie on estymował wokół średniej i będzie dla nas bardziej odnośnikiem dla innych metod. Formalnie możemy zapisać, że:

$$J_t(A_t = 1) = \left(\sum_{i=1}^{t-1} G(i)\right) / \left(\sum_{i=1}^{t-1} S(i)\right) \tag{4.16}$$

oraz

$$J_t(A_t = 0) = 1 - J_t(A_t = 1). \tag{4.17}$$

4.3.2. Średnia krocząca sum

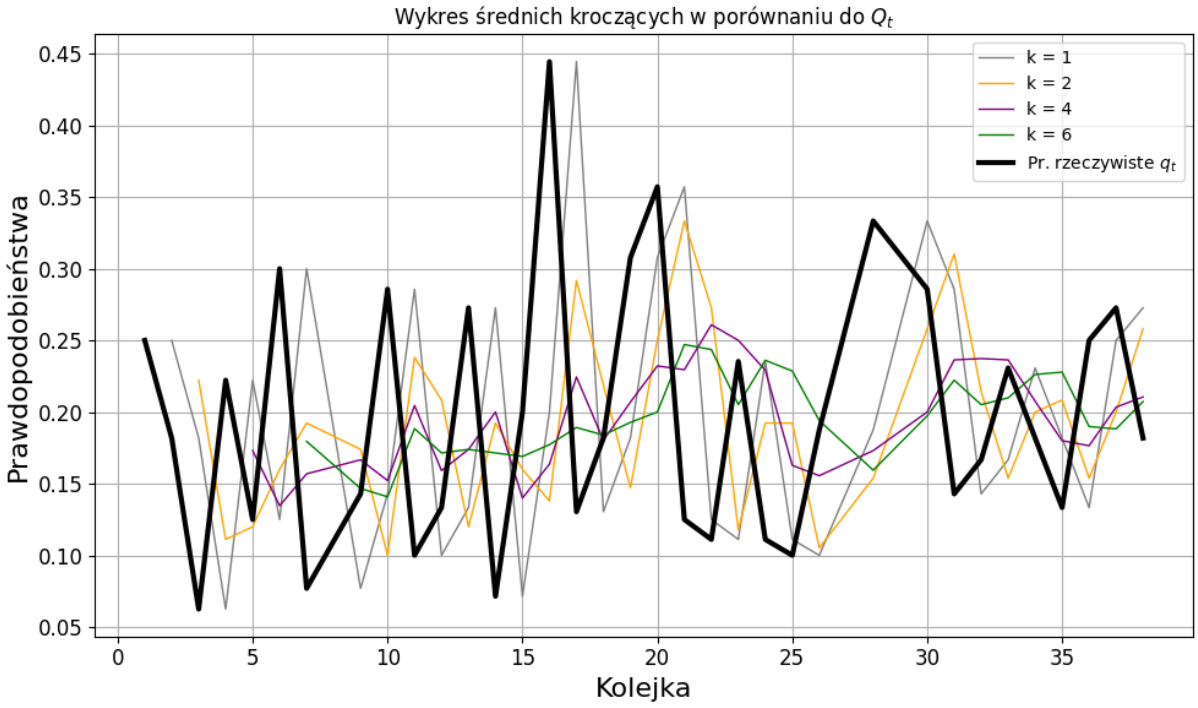
Tworząc ten model postarano się bardziej brać pod uwagę "formę"zespołu, czyli będziemy brali średnią arytmetyczną z poprzednich k meczów, czyli tak zwaną średnią kroczącą. Oznaczmy tę zmienną jako K_t, która przyjmie formułę

$$K_t(A_t = 1) = \left(\sum_{i=1}^{k} G(t-i)\right) / \left(\sum_{i=1}^{k} S(t-i)\right) \tag{4.18}$$

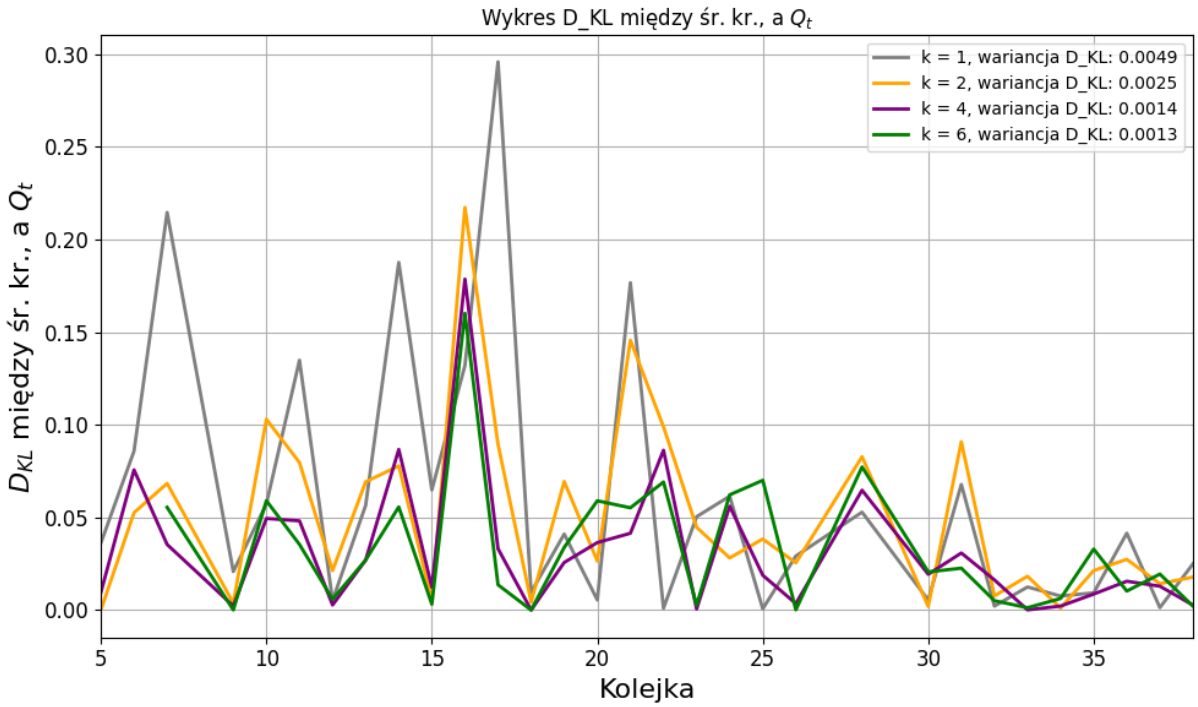
oraz

$$K_t(A_t = 0) = 1 - K_t(A_t = 1). \tag{4.19}$$

Ogonem średniej kroczącej nazywamy liczbę składników branych do liczenia średniej arytmetycznej, czyli k. Zanim porównano metodologie spróbowano znaleźć najdokładniejsze k dla średniej kroczącej. Przyjęto, że $k < 7$, tak żeby można było zastosować to rozumowanie dla jak największej liczby kolejek. Poniższe wykresy (4.6) i (4.7) odpowiadają na to pytanie, które k jest najlepsze w tym podejściu. Wywnioskowano z nich, że najlepszą metodologią jest wzięcie pod uwagę k z przedziału od 4., do 6. kolejkami wstecz, jednak trudno stwierdzić w jakim wariancie, która liczba będzie najlepsza.



Rysunek 4.6: Prawdopodobieństwa $K_t(A_t = 1)$ dla różnych długości ogonów rozkładu K_t



Rysunek 4.7: Zmiany dywergencji $D_{KL}(K_t||Q_t)$ dla różnych długości ogona średnich kroczących.

4.3.3. Średnia krocząca prawdopodobieństw

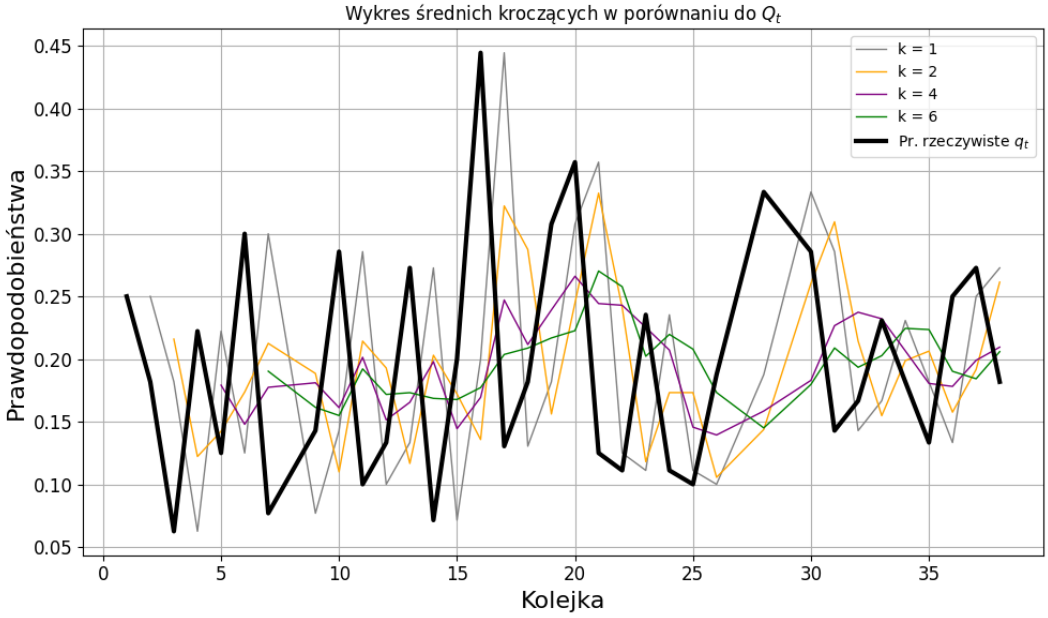
W tym podrozdziale przedstawimy podejście do tworzenia modelu bardzo podobne do (4.3.2), z tym że będziemy liczyli średnią kroczącą wartości q_t. Oznaczmy ten rozkład przez W_t, wówczas możemy zapisać, że:

$$W_t(A_t = 1) = \frac{1}{k} \sum_{i=1}^{k} q_{t-i} \tag{4.20}$$

oraz

$$W_t(A_t = 0) = 1 - W_t(A_t = 1). \tag{4.21}$$

Na rysunku (4.8) pokazano zmiany prawdopodobieństwa $W_t(A_t = 1)$ w czasie. Czarna, pogrubiona kreska oznacza rzeczywiste prawdopodobieństwo strzelenia gola z jednego strzału $Q_t(A_t = 1)$. Podobnie jak dla rozkładu K_t zauważono, że W_t najlepiej przybliża rozkład Q_t dla $4 \leq k \leq 6$. W tej części pracy nie przeprowadzono jednak szczegółowych analiz, ponieważ uznano, że wystarczy jedynie pewna estymacja k.



Rysunek 4.8: Prawdopodobieństwa strzelenia gola $Q_t(A_t = 1)$ dla różnych długości ogonów.

4.3.4. Porównanie metod predykcji przy użyciu dywergencji KL

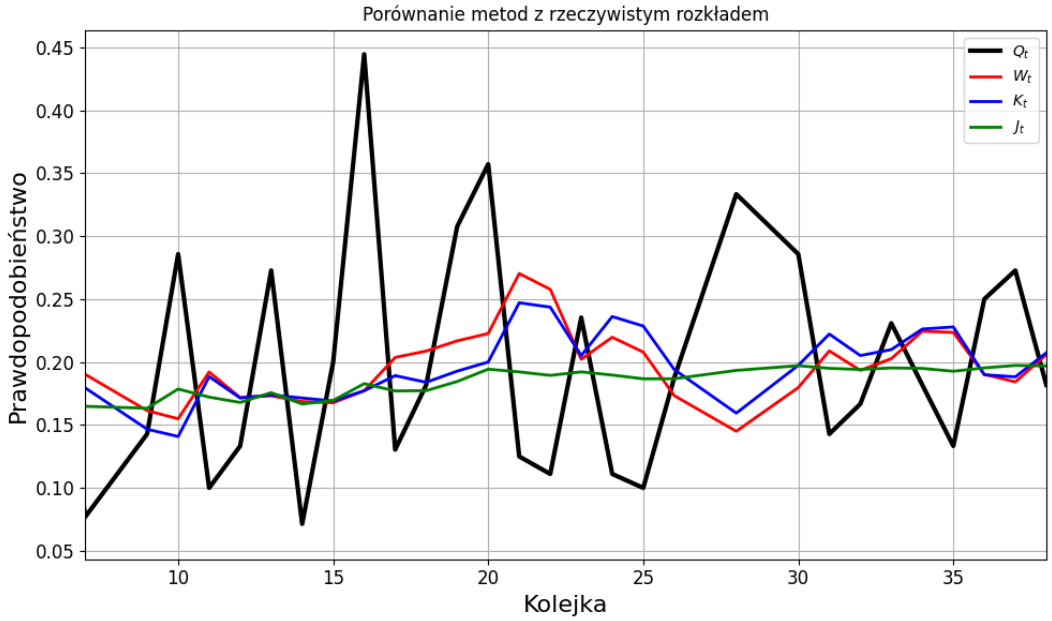
Porównamy teraz wyżej wymienione metody, z tym że dla (4.3.2) weźmiemy $k = 5$, zaś dla (4.3.3) $k = 6$. Wtedy możemy zapisać, że

$$D_{KL}(J_t||Q_t) = J_t(A_t = 1) \cdot \ln\left(\frac{J_t(A_t = 1)}{Q_t(A_t = 1)}\right) + J_t(A_t = 0) \cdot \ln\left(\frac{J_t(A_t = 0)}{Q_t(A_t = 0)}\right), \tag{4.22}$$

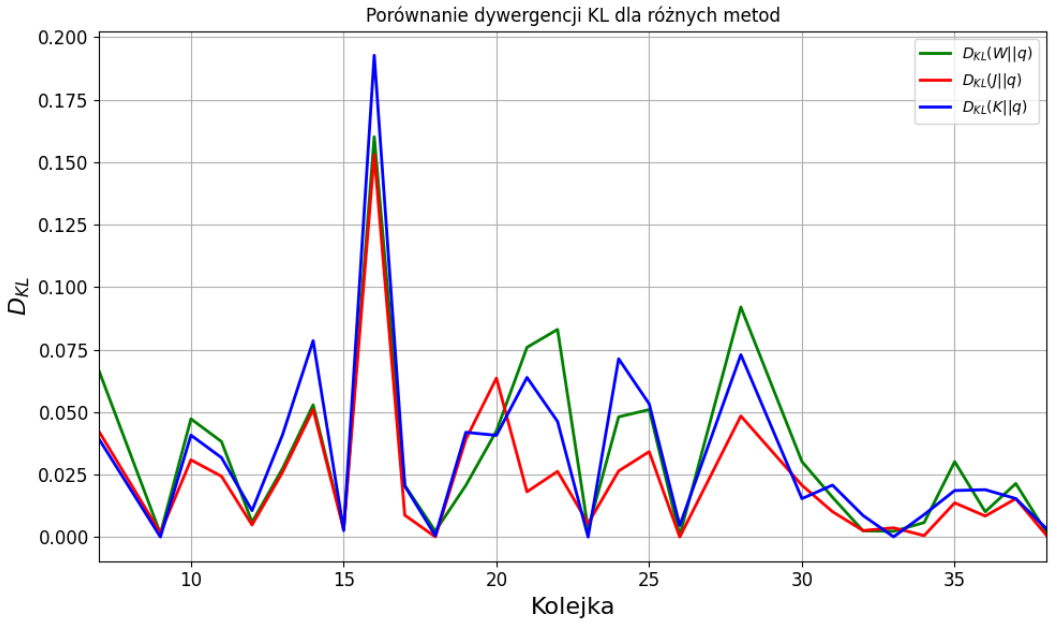
$$D_{KL}(K_t||Q_t) = K_t(A_t = 1) \cdot \ln\left(\frac{K_t(A_t = 1)}{Q_t(A_t = 1)}\right) + K_t(A_t = 0) \cdot \ln\left(\frac{K_t(A_t = 0)}{Q_t(A_t = 0)}\right), \tag{4.23}$$

$$D_{KL}(W_t||Q_t) = W_t(A_t = 1) \cdot \ln\left(\frac{W_t(A_t = 1)}{Q_t(A_t = 1)}\right) + W_t(A_t = 0) \cdot \ln\left(\frac{W_t(A_t = 0)}{Q_t(A_t = 0)}\right). \tag{4.24}$$

Na rysunkach (4.9) oraz (4.10) pokazano porównanie prawdopodobieństw zajścia zdarzenia ($A_t = 1$) dla rozkładów prawdopodobieństwa wymienionych w (4.22). Widać na nim, że rozkłady K_t oraz W_t posiadają dużo wyższą wariację niż J_t. Na rysunku (4.9) przedstawiono natomiast zmiany dywergencji KL K_t, W_t oraz J_t względem Q_t.



Rysunek 4.9: Zmiana prawdopodobieństw zajścia zdarzenia ($A_t = 1$) dla różnych rozkładów W_t, K_t, J_t oraz W_t



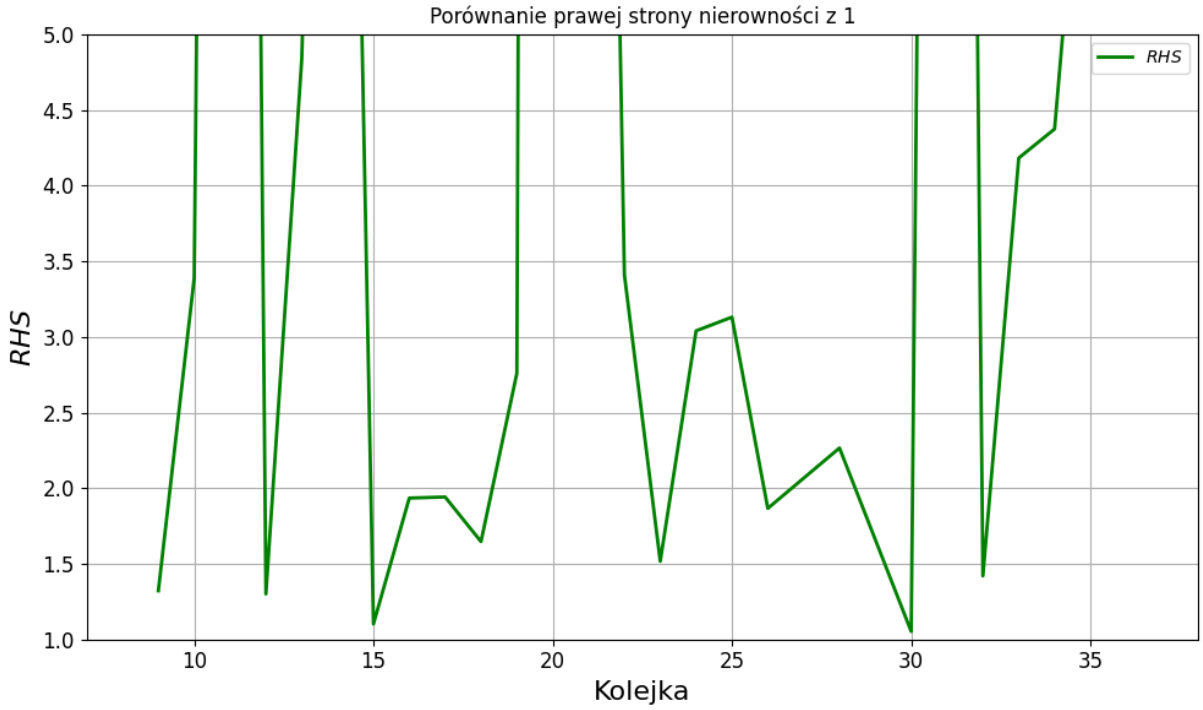
Rysunek 4.10: Zmiana dywergencji KL w czasie względem dla różnych rozkładów W_t, K_t, J_t oraz W_t

W związku z tym, dodatkowo porównano średnie oraz wariancje w czasie dywergencji Kullbacka-Leiblera dla wszystkich kolejek między rzeczywistym rozkładem Q oraz J, W i K. Z tabeli (4.3) widać, że średnie dywergencje w czasie $D_{KL}(J||Q)$ są znacznie niższe niż średnia dywergencja $D_{KL}(K||Q)$ oraz $D_{KL}(W||Q)$. Jednak wariancje są stosunkowo niskie, więc nie odbiegają one bardzo od średniej, co może sugerować, że istnieje możliwość poprawienia rezultatów.

D_{KL}	**Średnia**	**Wariancja**
$D_{KL}(W\|\|Q)$	0.033218	0.001313
$D_{KL}(K\|\|Q)$	0.033807	0.001474
$D_{KL}(J\|\|Q)$	0.022875	0.000852

Tabela 4.3: Tabela przedstawiająca dane dotyczące średniej i wariancji w czasie dla różnych metod predykcji.

Sprawdzono również czy rozważane dane spełniają nierówność (2.50) dla rozkładu K_t, porównując prawą stronę nierówności z liczbą 1.



Rysunek 4.11: $RHS = \frac{\sqrt{I_F(K_t)}\sqrt{T_2(K_t||q) - D_{KL}^2(K_t||q)} + \sqrt{I_F(q)T_2}}{\left|\frac{d}{dt} D_{KL}(K_t||q)\right|}$.

Jak widzimy nasza nierówność (2.50) jest spełniona dla prawdziwych danych i często jest bliska równości. Niestety przez prostotę naszego modelu często odbiega od liczby 1, co widać na powyższym rysunku.

Widzimy stąd, że wszystkie te modele są podobne do siebie, jednak celem tego przykładu miało być pokazanie w jaki sposób można analizować dane oraz przewidywania przy pomocy dywergencji KL oraz w szczególności nierówności (2.50). Powyższe rozumowanie można

uwzględnić również w innych eksperymentach, na przykład oceniając restauracje na serwisach internetowych z opiniami. Jeśli restauracja zmieniała poziom swoich usług w przeciągu ostatnich miesięcy, to średnia ze wszystkich ocen nie będzie dobrym predyktorem wrażeń kulinarnych jeśli przyjedziemy do niej jutro.

Rozdział 5

Podsumowanie

W niniejszej pracy przedstawiono analizę f-dywergencji między rozkładami prawdopodobieństwa. Głównie skupiono się na dywergencjach Tsallisa i Renyiego oraz ich szczególny przypadek dywergencji Kullbacka-Leiblera (KL). Pokazano, jak różne rodzaje dywergencji mogą być używane do opisywania relacji między rozkładami probabilistycznymi, analizując ich właściwości oraz zastosowania. Szczególną uwagę poświęcono dywergencji Kullbacka-Leiblera, która, mimo swojej asymetryczności i braku cech charakterystycznych dla metryk, pozostaje jednym z najważniejszych narzędzi analizy danych probabilistycznych w matematyce, statystyce oraz uczeniu maszynowym.

Zaprezentowane wyniki pracy pokazują, iż f-dywergencje, w tym dywergencja KL, mogą być użyteczne w analizie zmian zachodzących w systemach probabilistycznych. Wyprowadzenie nierówności ograniczających tempo zmian tych dywergencji dostarcza nie tylko nowych narzędzi analitycznych, ale także otwiera możliwości dalszych badań w zakresie modelowania procesów stochastycznych. Analiza ta jest szczególnie istotna w kontekście rozkładów wielowymiarowych, takich jak rozkłady normalne i lognormalne, co pozwala na lepsze zrozumienie dynamiki tych procesów.

Jednym z pierwszych elementów pracy było pokazanie relacji między dywergencją Kullbacka-Leiblera, Tsallisa oraz Rényiego, co pozwala na głębsze zrozumienie struktury tych narzędzi analitycznych. Ponadto, badanie nierówności ograniczających tempo zmian dywergencji KL dostarcza nowych metod analitycznych, które mogą być stosowane w modelowaniu dynamicznych systemów probabilistycznych. Praktyczne zastosowanie tych wyników w analizie danych piłkarskich pokazuje potencjał dywergencji KL w rzeczywistych problemach analizy danych. W szczególności, użycie tych narzędzi w prostych modelach predykcyjnych ilustruje, jak można je zastosować do analizy danych stochastycznych w różnych dziedzinach nauki, takich jak statystyka sportowa czy modelowanie procesów fizycznych.

Jednym z kluczowych elementów tej pracy było zaprezentowanie analizy górnych ograniczeń wartości bezwzględnych dla pochodnej po czasie dywergencji KL. Wyniki te zostały zastosowane w praktyce, w tym w analizie dynamicznych systemów probabilistycznych, takich jak dane sportowe, oraz w zastosowaniach do równań różniczkowych, takich jak równanie ciepła. W szczególności badano, kiedy nierówności te są nasycone, czyli zachodzi w nich równość, co może mieć praktyczne znaczenie w kontekście modelowania i analizy sygnałów stochastycznych. W pracy podkreślono uniwersalność dywergencji KL jako narzędzia analitycznego, co zostało zilustrowane na przykładzie danych piłkarskich z angielskiej Premier League z se-

zonu 2018/2019. Analiza ta ukazuje, jak proste modele predykcyjne mogą być wzbogacone o bardziej zaawansowane techniki statystyczne, aby lepiej opisywać rzeczywiste procesy zachodzące w danych. Otrzymane wyniki mają potencjał zarówno teoretyczny, jak i praktyczny, dostarczając narzędzi do eksploracji danych w różnych dziedzinach, od analizy sygnałów po modelowanie zmian w czasie.

Przedstawione w pracy wyniki otwierają wiele możliwości dla dalszych badań. Jednym z potencjalnych kierunków jest rozwijanie bardziej zaawansowanych metod analitycznych opartych na wyprowadzonych nierównościach, szczególnie w kontekście dynamicznych systemów probabilistycznych i ich zastosowań w statystyce. Dalsza eksploracja zastosowań dywergencji KL w modelach dynamicznych, szczególnie w kontekście danych wielowymiarowych, takich jak dane finansowe czy dane biologiczne, mogłaby dostarczyć nowych spostrzeżeń.

Podsumowując, niniejsza praca dostarcza wiele solidnych podstaw teoretycznych i praktycznych dla dalszych badań w zakresie teorii informacji oraz analizy statystycznej. Otrzymane wyniki mają potencjał, aby znacząco wpłynąć na sposób, w jaki badamy sygnały stochastyczne, co czyni je cennym wkładem zarówno w teorię, jak i praktykę.

Bibliografia

[1] A. Renyi. On measures of entropy and information. In *Proc. 4th Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561, Berkeley, CA, 1960. Univ. of California Press.

[2] K. R. Chernyshov. Applying tsallis divergence to proteins organization prediction problems. *IFAC-PapersOnLine*, **55**:513–519, 2022.

[3] C. E. Shannon and W. Weaver. A mathematical theory of communication. *The Bell System Technical Journal*, **27**:379–423, 623–656, 1948.

[4] J. Shlens. Notes on kullback-leibler divergence and likelihood, https://doi.org/10.48550/arxiv.1404.2000.

[5] D. Dimitrov A. Bulinski. Statistical estimation of the kullback-leibler divergence, https://doi.org/10.48550/arxiv.1907.00196.

[6] A. Clim, R. D. Zota, and G. Tinică. The kullback-leibler divergence used in machine learning algorithms for health care applications and hypertension prediction: A literature review. *Procedia Computer Science*, **141**:448–453, 2018.

[7] T. Mora A. M. Walczak F. Camaglia, I. Nemenman. Bayesian estimation of the kullback-leibler divergence for categorical sytems using mixtures of dirichlet priors.

[8] Fábio Mendonça, Sheikh Shanawaz Mostafa, Fernando Morgado-Dias, and Antonio G. Ravelo-García. On the use of kullback–leibler divergence for kernel selection and interpretation in variational autoencoders for feature creation. *Information*, 14(10), 2023.

[9] L. Xie, J. Zeng, U. Kruger, X. Wang, and J. Geluk. Fault detection in dynamic systems using the kullback–leibler divergence. *Control Engineering Practice*, **43**:39–48, 2015.

[10] Jan Karbowski. Bounds on the rates of statistical divergences and mutual information via stochastic thermodynamics. *Phys. Rev. E*, **109**:054126, May 2024.

[11] Z. Goldfeld and K. Khezeli. Ece 5630: Information theory for data transmission, security and machine learning, 2020.

[12] R. Czyż. Teoria miary i całki, http://www2.im.uj.edu.pl/leszekpieniazek/du/mic/test-20.html.

[13] P. Strzelecki. Skrypt analiza matematyczna ii, mimuw.edu.pl/-pawelst/am2/analiza-matematyczna-2/notatki-files/skrypt-amii-05-01-2016.pdf.

[14] A. Cichocki and S-I. Amari. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy*, **12**:1532–1568, 2010.

[15] I. Sason and S. Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, **62**(11):5973–6006, 2016.

[16] Kullback- leibler divergence explained, https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained.

[17] S. Simic. On a new moments inequality. *Statistics and Probability Letters*, **78**(16):2671–2678, 2008.

[18] B.R. Frieden. *Science from Fisher Information: A Unification, 2nd ed.* Cambridge Univ. Press, 2004. Cambridge, UK.

[19] B. C. Carlson. Some inequalities for hypergeometric functions. *Proceedings of the American Mathematical Society*, **17**:32–39, 1966.

[20] Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Wiley, revised edition, 1999.

[21] L. Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, **12**:134–139, 1918.

[22] J. Michalowicz, J. Nichols, Frank Bucholtz, and Colin Olson. An isserlis' theorem for mixed gaussian variables: Application to the auto-bispectral density. *Journal of Statistical Physics*, **136**:89–102, 07 2009.

[23] Y. Zhang, J. Pan, L. K. Li, W. Liu, Z. Chen, X. Liu, and J. Wang. On the properties of kullback-leibler divergence between multivariate gaussian distributions. In *Advances in Neural Information Processing Systems*, volume **36**, pages 58152–58165, 2023.

[24] M. Gil, F. Alajaji, and T. Linder. Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, **249**:124–131, 11 2013.

[25] Z. Grzesik. Podstawy dyfuzji, https://home.agh.edu.pl/ grzesik/fizykochemia_ns/fizykochemia_w5.pdf.

[26] T. Chwiej. Montecarlo, http://galaxy.agh.edu.pl/c̃hwiej/mc/lab/diffusion/dyfuzja_absorpcja.pdf.

[27] J. Leszczyński M. Ciesielski. Schemat mrs dla równania anormalnej dyfuzji z pochodną niecałkowitego rzędu po czasie. *Scientific Research of the Institute of Mathematics and Computer Science*, **1**(1):39–46, 2002.

[28] L. C. Evans. *Partial Differential Equations*, volume 19. American Mathematical Society, 2nd edition, 2010.

[29] James H. Hewitt and Oktay Karakuş. A machine learning approach for player and position adjusted expected goals in football (soccer). *Franklin Open*, **4**:100034, 2023.

[30] https://footystats.org/download-stats-csv FootyStats.org.