Politechnika Wrocławska

Wydział Matematyki
Kierunek studiów: Matematyka Stosowana
Specjalność: –

Praca dyplomowa – inżynierska

WYZNACZANIE WYKŁADNIKA DYFUZJI ANOMALNEJ METODAMI UCZENIA MASZYNOWEGO.

Mateusz Gorczyca

słowa kluczowe:
dyfuzja anomalna, uczenie maszynowe, Python, śledzenie pojedynczych cząstek

krótkie streszczenie:
W pracy zajęto się szukaniem czynników wpływających na jakość wyznaczania wykładnika dyfuzji anomalnej metodami uczenia maszynowego dla trajektorii otrzymanych metodą śledzenia pojedynczych cząstek. Wyniki porównano również z tradycyjną metodą analizy ruchów cząstek, aby mieć odpowiedź na pytanie: „czy uczenie maszynowe to krok w dobrą stronę w analizie dyfuzji?”.

Opiekun pracy dyplomowej	dr hab. Janusz Szwabiński	.....	.....
	Tytuł/stopień naukowy/imię i nazwisko	ocena	podpis

*Do celów archiwalnych pracę dyplomową zakwalifikowano do:**

a) kategorii A (akta wieczyste)

b) kategorii BE 50 (po 50 latach podlegające ekspertyzie)

** niepotrzebne skreślić*

pieczątka wydziałowa

Wrocław, rok 2021

Wrocław University
of Science and Technology

Faculty of Pure and Applied Mathematics
Field of study: Applied Mathematics
Specialty: –

Engineering Thesis

DETERMINATION OF THE ANOMALOUS EXPONENT VIA MACHINE LEARNING METHODS.

Mateusz Gorczyca

keywords:
anomalous diffusion, machine learning, Python, single particle tracking

short summary:
The goal of this work is to estimate anomalous exponents for trajectories from SPT experiments. Different factors effecting the quality of the estimation are studied. The results are compared with the traditional method based on mean squared displacement in order to answear thequestion: "Is machine learning the right step towards an exhaustive SPT analysis?"

Supervisor	dr hab. Janusz Szwabiński	.....	.....
	Title/degree/name and surname	grade	signature

*For the purposes of archival thesis qualified to:**

a) category A (perpetual files)

b) category BE 50 (subject to expertise after 50 years)

** delete as appropriate*

stamp of the faculty

Wrocław, 2021

Spis treści

Wstęp

Dyfuzja to jeden z fundamentalnych procesów w otaczającym nas świecie. To dzięki niej zachodzą procesy odpowiedzialne między innymi za reakcje w żywych komórkach. Używaną zazwyczaj metodą badania dyfuzji jest śledzenie pojedynczych cząstek (ang. *single-particle tracking*, STP) [15]. Technika ta jest szeroko stosowana ze względu na to, że w wielu przypadkach nie wpływa na ruch cząstek w badanym obiekcie, co jest kluczowe dla otrzymania jak najdokładniejszych wyników oraz pozwala na dokonanie pomiarów in vivo (wewnątrz komórki/organellum). Jednym z często spotykanych w systemach biologicznych ruchów cząstek jest dyfuzja anomalna.

Celem niniejszej pracy jest wyznaczanie wykładnika dyfuzji anomalnej z użyciem uczenia maszynowego. W tym celu wykorzystano znane algorytmy uczenia maszynowego takie jak regresja liniowa [11], drzewo decyzyjne [2], las losowy [28] i wzmocnienie gradientowe [6].

Otrzymane wszystkimi metodami wyniki porównano między sobą oraz z wynikami otrzymanymi metodą porównania krzywych uśrednionego po czasie średniego przesunięcia kwadratowego (ang. *time-averaged mean squared displacement*, TAMSD) [20], które jest podstawową metodą używaną do analizy ruchu pojedynczych cząstek. TAMSD jednak jest mało skuteczne dla krótkich trajektorii oraz przy dużym ich zaszumieniu. Sposobem na poradzenie sobie z powyższymi problemami może być właśnie uczenie maszynowe (ang. *machine learning*, ML).

W rozdziale pierwszym zostaną omówione podstawowe twierdzenia, definicje i wzory wykorzystane w dalszej analizie, a także zostanie przedstawiony jeden ze sposobów wykonania SPT. Zostaną tam także zademonstrowane algorytmy uczenia maszynowego.

W kolejnym rozdziale różnymi metodami wyestymowany zostanie wykładnik dyfuzji anomalnej α dla ułamkowego ruchu Browna [4], w którym różne parametry ruchu, takie jak zaszumienie oraz współczynnik dyfuzji D, będą losowe. Badania będą prowadzone na trajektoriach o długości 20 kroków, 100 kroków oraz na takich o losowej długości kroku w zakresie od 20 do 100.

W trzecim rozdziale analiza zostanie rozszerzona na inne procesy dyfuzyjne: ciągłe błądzenie losowe (CTRW) [7], ułamkowy ruch Browna (FBM) [4], spacer Lévy'ego (LF) [22], model ruchu wyżarzonege (annealead transient time model, ATTM) [29] i skalowany ruch Browna (SBM) [27].

Następnie przebadany zostanie wpływ szumu. Do badań wygenerowane zostaną trajektorie o rożnym poziomie zaszumienia. W pierwszym etapie na każdym z poziomów szumu wytrenowano model i zbadano jego skuteczność na tym samym poziomie szumu co dane treningowe. W drugim etapie przebadano model, który był szkolony na danych o mieszanym zaszumieniu, przy wyznaczaniu wykładnika dla różnych poziomów zaszumień. W trzecim etapie przebadano, jak model szkolony bez szumu poradzi sobie z danymi testowymi o mieszanym poziomie szumu.

W przedostatnim rozdziale przebadany zostanie wpływ ilości danych treningowych na poprawność estymacji wykładnika dla trajektorii o mieszanym zaszumieniu. Zbadany zostanie także czas nauczania i szukania hiperparametrów modeli przy różnej liczbie danych treningowych.

Ostatni rozdział to podsumowanie wyników oraz płynące z nich wnioski. Przedstawione zostaną również perspektywy jakie niesie za sobą dalszy rozwój uczenia maszynowego dla analizy SPT.

Rozdział 1

Pojęcia podstawowe

1.1 Śledzenie pojedynczych cząstek

Definicja 1.1. Śledzenie pojedynczych cząstek (ang. *single particle tracking*, SPT) to proces rejestrowania położenia cząstek metodami mikroskopowymi w pewnych (najczęściej stałych) odstępach czasu Δt. Stała Δt jest zwana **rozdzielczością czasową** pomiaru. W wyniku pomiaru otrzymujemy szereg czasowy $\{(x_i, y_i)\}_{i=0,1,...,N}$, który podlega dalszej analizie [15].

Dokładność pomiarów oraz rozdzielczość czasowa Δt zależą od wykorzystanej do pomiarów aparatury. Najczęściej badane cząstki są barwione wskaźnikiem fluorescencyjnym. Podczas obserwacji są oświetlane światłem lasera, co stymuluje barwnik do emitowania światła. Właśnie to wyemitowane światło jest rejestrowane przez czujnik, a pozycja cząstek — zapisana [8].

Największymi problemami w analizie otrzymanych trajektorii są:

- stochastyczna natura procesu,
- niewielka długość trajektorii (procesy mogą być bardzo krótkie),
- brak zauważalnych asymptotycznych zachowań,
- szum środowiskowy.

W analizach często zamiast położenia cząstki będzie używane jej **przesunięcie** względem położenia początkowego (x_0, y_0) definiowane jako:

$$X_i = (x_i - x_0, y_i - y_0), \quad i = 0, 1, ..., N. \tag{1.1}$$

Jako **krok** w trajektorii rozumiemy odległość pomiędzy dwoma sąsiednimi pomiarami położenia cząstki:

$$S_i = (x_i - x_{i-1}, y_i - y_{i-1}), \quad i = 1, 2, ..., N. \tag{1.2}$$

1.2 Dyfuzja

Definicja 1.2. Dyfuzja to spontaniczny proces fizyczny polegający na przemieszczaniu się cząstek w ośrodku pod wpływem różnicy stężeń badanych cząstek w różnych punktach ośrodka. Cząstki poruszają się w kierunku ich niższego stężenia [23].
Jako cząstkę należy rozumieć dowolne ciało umieszczone w dowolnym ośrodku (np. cząstka cukru w wodzie, białko w komórce).

Definicja 1.3. Współczynnik dyfuzji D jest miarą tempa mieszania się substancji wyrażoną w $[\frac{\text{mm}}{\text{s}^2}]$ [19]. Jego wartość jest otrzymywana doświadczalnie.

Dyfuzja może występować w czterech formach [12]:

- **dyfuzja normalna** — ruch Browna $B_{n\Delta t}$.
- **ruch ukierunkowany** — ruch Browna z dryfem,
- **dyfuzja ograniczona** — ruch Browna na ograniczonej przestrzeni,
- **dyfuzja anomalna** — będzie omówiona w dalszej części pracy.

Definicja 1.4. Średnie przesunięcie (ang. *average displacement*) cząstki $\bar{X}$ do chwili $n\Delta t$ zdefiniowane jest jako:

$$\bar{X}(n\Delta t) = \frac{1}{n}\sum_{i=1}^{n} \|X_i\|. \tag{1.3}$$

gdzie:
n — liczba obserwacji,
X_i — przesunięcie cząstki liczone ze wzoru (1.1).
Jako $\bar{X}$ będziemy oznaczać $\bar{X}(N\Delta t)$, gdzie N to liczba zmierzonych kroków.

Definicja 1.5. Dyfuzja jest **dyfuzją anomalną** jeśli istnieje takie $\alpha \neq 1$, że zachodzi [7]:

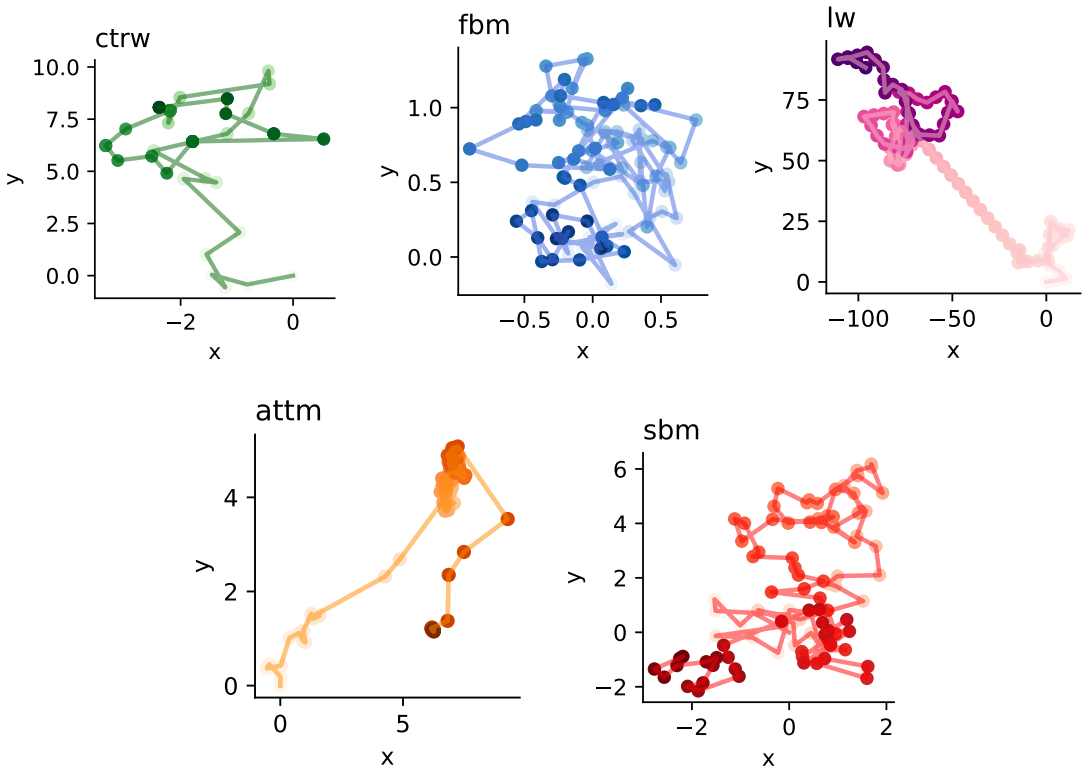
$$\left(\bar{X}(n\Delta t)\right)^2 \sim t^{\alpha},$$

gdzie:
$\bar{X}(n\Delta t)$ — średnie oddalenie cząstki do chwili $n\Delta t$ zgodne z równaniem (1.3).
Liczbę $\alpha \neq 1$ nazywamy **wykładnikiem dyfuzji**. Przypadek, gdy $\alpha = 1$, odpowiada dyfuzji normalnej.

Gdy wykładnik dyfuzji α jest większy od 1 to mamy do czynienia z **superdyfuzją**, natomiast gdy jest mniejszy — z **subdyfuzją**. Dyfuzję anomalną mogą opisywać następujące procesy stochastyczne [14]:

- **ciągłe błądzenie losowe** (ang. *continuous-time random walk*, CTRW),
- **ułamkowy ruch Browna** (ang. *fractional Brownian motion*, FBM),
- **spacer Lèvy'ego** (ang. *Lèvy Walk*, LW),
- **annealead transit time model** (ATTM),
- **skalowany ruch Browna** (ang. *scaled Brownian motion*, SBM).

Przykładowe trajektorie dla każdego ruchu zostały pokazane na rysunku 1.1.



Rysunek 1.1: Przykładowe trajektorie poszczególnych rodzajów ruchów. Im ciemniejszy kolor danego pomiaru, tym później on wystąpił. Zaprezentowane trajektorie mają czas trwania równy $T = 100$ i $n = 100$ kroków, a wykładnik anomalny wynosi $\alpha = 0,7$.

Definicja 1.6. Ciągłe błądzenie losowe to proces stochastyczny, w którym kroki czasu są niezależne od kroków w przestrzeni. Kroki w czasie muszą spełniać prawo potęgowe [7]:

$$f(t) \overset{t\to\infty}{=} \frac{1}{t^{a+1}}, \ a \in \mathcal{R}_+.$$

Kroki w czasie mogą pochodzić z innego rozkładu niż kroki w przestrzeni.

Ruch ten charakteryzuje się czasami oczekiwania pomiędzy kolejnymi skokami. Liczba skoków przeważnie jest mniejsza niż liczba kroków trajektorii.

Definicja 1.7. Ułamkowy ruch Browna to proces stochastyczny spełniający [4]:

- $\mathbb{E}B_H(t) = 0$
- $\mathrm{Var}B_H(t) = t^{2H}$
- $\mathrm{Cov}\left(B_H(t), B_H(s)\right) = R(t,s) = \frac{1}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right)$
- ma stacjonarne i niezależne przyrosty.

Stała H nazywana jest **wykładnikiem Hursta** i spełnia zależność:

$$\alpha = 2H$$

rodzaj procesu	**skrót**	**zakres α**
ciągłe błądzenie losowe	CTRW	$\alpha \in (0, 1]$
annealed transit time model	ATTM	
spacer Lévy'ego	LF	$\alpha \in (1, 2]$
ułamkowy ruch Browna	FBM	$\alpha \in (0, 2]$
skalowany ruch Browna	SBM	

Tabela 1.1: Używane przy analizach rodzaje procesów, ich skróty używane w dalszej części pracy oraz zakresy wykładnika α trajektorii, jakie mogą być z nich generowane [16].

Definicja 1.8. Spacer Lévy'ego to ruch, w którym wielkości skoków są generowane z ciężko-ogonowego rozkładu Lévy'ego. Skoki w czasie (jeśli prędkość ruchu jest stała) lub wariancja rozkładu, z którego są one losowane (jeśli prędkość ruchu nie jest stała), są proporcjonalne do długości skoków w przestrzeni [22].

Definicja 1.9. Annealead transit time model to model ruchu, który zbiega coraz mocniej do pewnego losowego punktu (celu). Co losowy czas cel jest wybierany ponownie. W założeniu ma to przypominać ruch atomów w podgrzanych metalach, które czasem przeskakują na inne miejsca [29]. W związku z tym, że ruch ma nieustannie zmieniający się przebieg. Jest to ruch nieergodyczny. Definicję nieergodyczności podano poniżej.

Definicja 1.10. Proces nieergodyczny to proces, którego podstawowe parametry statystyczne zmieniają się w sposób zależny nie tylko od czasu [3].

Definicja 1.11. Skalowany ruch Browna to proces o parametrach jak ruch Browna, ale ze zmieniającą się w czasie wariancją [27],

$$\sigma^2(t) = \sigma^2 t,$$

gdzie σ to współczynnik skali.

Niektóre z wymienionych rodzajów ruchów mogą produkować trajektorie dla ograniczonych wykładników α. Właściwości tych ruchów oraz używane w pracy skróty zostały przedstawione w tabeli 1.1.

1.3 Uśrednione po czasie średnie przesunięcie kwadratowe (TAMSD)

Jedną z podstawowych metod analizy trajektorii jest eksperymentalne wyznaczenie krzywej średniego przesunięcia kwadratowego (MSD), a następnie dopasowanie do niej modelu teoretycznego [7].

Definicja 1.12. MSD $\hat{\rho}(t)$ jest zdefiniowane jako [20]:

$$\hat{\rho}(n\Delta t) = \hat{\rho}(t) = \frac{1}{M}\sum_{i=1}^{M} \|X_t^i\|^2,$$

gdzie:
M — liczba badanych cząstek,
X_t^i — przesunięcie cząstki i do chwili t.

W związku z tym, że w doświadczeniach SPT dysponujemy tylko jedną trajektorią dla każdej cząstki, użyjemy uśrednionego po czasie średniego przesunięcia kwadratowego (TAMSD).

Definicja 1.13. TAMSD wyraża się wzorem [20]:

$$\rho(n\Delta t) = \rho(t) = \frac{1}{N-n}\sum_{i=0}^{N-n} \|X_{i+n} - X_i\|^2, \tag{1.4}$$

gdzie:
N — liczba kroków w trajektorii,
X_i — przesunięcie cząstki do chwili $i\Delta t,\ i = 1, 2, ..., N$.

TAMSD dla dyfuzji anomalnej wynosi [5]:

$$\rho(t) = 4Dt^{\alpha}, \tag{1.5}$$

gdzie:
D — współczynnik dyfuzji,
t — czas, w którym liczony jest wyznacznik TAMSD,
α — wykładnik dyfuzji anomalnej.

Współczynnik dyfuzji D otrzymano zgodnie z równaniem (1.5) dla czasu $t = 1$:

$$D = \frac{\rho(1)}{4}, \tag{1.6}$$

gdzie $\rho(1)$ wyznaczono ze wzoru (1.5).

Do wyznaczenia wykładnika dyfuzji w metodzie TAMSD często używa się metody najmniejszych kwadratów. Z równania (1.5) otrzymujemy:

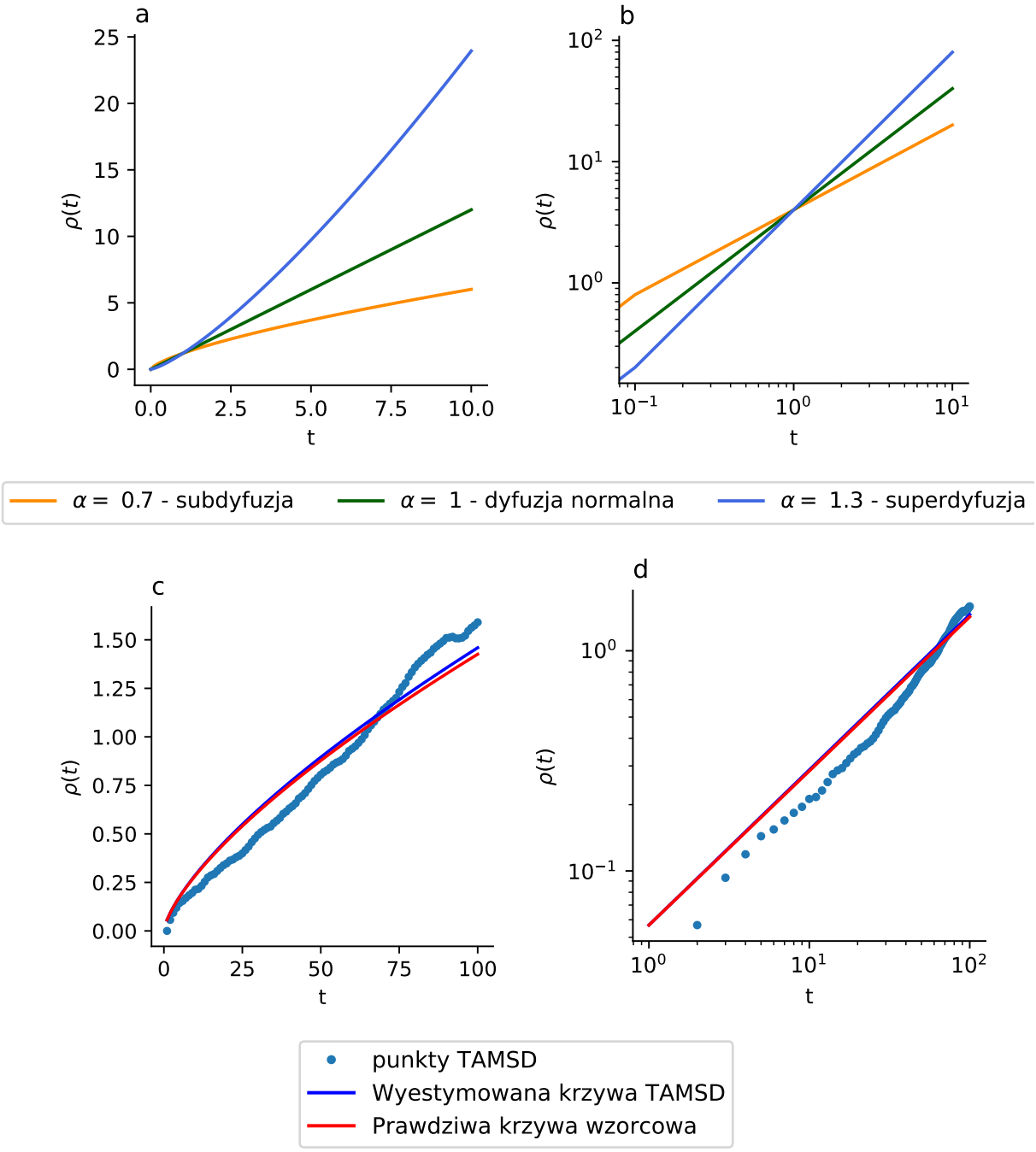
$$\ln(\rho(t)) = \ln(4D) + \alpha \ln(t), \tag{1.7}$$

gdzie $\rho(t)$ wyznaczono zgodnie z równaniem (1.5). Zatem estymator parametru $\widehat{\alpha}$ otrzymany metodą najmniejszych kwadratów wynosi[1]:

$$\widehat{\alpha} = \frac{\sum_{i=1}^{n} \ln(i\Delta t) \cdot \ln\Big(\rho(i\Delta t)\Big) - \ln(4D)\sum_{i=1}^{n} \ln(i\Delta t)}{\sum_{i=1}^{n} \Big(\ln(i\Delta t)\Big)^2}. \tag{1.8}$$

Na rysunku 1.2 przedstawiono krzywe TAMSD dla superdyfuzji, subdyfuzji i dyfuzji normalnej oraz krzywe TAMSD dla przykładowej trajektorii.

[1]Dowód znajduje się w sekcji *Dodatki*-B.1.



Rysunek 1.2: Wzorcowe krzywe TAMSD dla $\alpha \in \{0,7;\ 1;\ 1,3\}$ i $D = 1$ w skali liniowej (a) i logarytmicznej (b). Przykład porównania wyznaczonych TAMSD z trajektorii i krzywa prawdziwej wartości α w skali liniowej (c) i logarytmicznej (d) dla trajektorii o wykładniku $\alpha = 0,7$.

1.4 Algorytmy uczenia maszynowego

1.4.1 Właściwości trajektorii

Do analizy trajektorii metodą uczenia maszynowego trajektorie wymagają uprzedniego przetworzenia, a dokładniej wyznaczenia z nich pewnych parametrów [12, 21]:

- **Właściwości policzone z długości kroków**:
  współczynnik dyfuzji D, wydajność E, współczynnik spowolnienia s, iloraz MSD κ, współczynnik anty-gaussowości G, liniowość S, autokorelacja R.
- **Właściwości policzone z położenia cząstki**:
  maksymalna odległość między punktami $d_{\max}$, prawdopodobieństwo uwięzienia P, wymiar fraktalny D_f.

Współczynnik dyfuzji D otrzymano zgodnie z równaniem (1.6).

Jako **wydajność** E trajektorii oznaczymy jej kwadrat całkowitego przesunięcia w stosunku do sumy kwadratów kroków [12]:

$$E = \frac{N \cdot \|X_N\|^2}{\sum\limits_{i=1}^{N} \|S_i\|^2},$$

gdzie:
N — liczba kroków trajektorii,
X_i — przesunięcie cząstki od położenia początkowego dla chwili i,
S_i — krok policzony zgodnie ze wzorem (1.2).

Jako **współczynnik spowolnienia** s_n trajektorii będziemy uznawali najprostszą estymację wykładnika α, czyli stosunek średniej długości kroku dla n ostatnich kroków do średniej długości kroku dla n pierwszych kroków:

$$s_n = \frac{\sum\limits_{i=N-n+1}^{N} \|S_i\|}{\sum\limits_{i=1}^{n} \|S_i\|}.$$

Iloraz MSD jest zdefiniowany jako [12]:

$$\kappa(n) = \frac{\rho(n)}{\rho(n+1)} - \frac{n}{n+1},$$

gdzie $\rho(n)$ policzono ze wzoru (1.4).

Współczynnik anty-gaussowości G można wyliczyć ze wzoru [21]:

$$G(\Delta) = \frac{d}{d+2} \cdot \frac{\bar{\delta}^4(\Delta)}{\bar{\delta}^2(\Delta)} \overset{d=2}{=} \frac{\bar{\delta}^4(\Delta)}{2\bar{\delta}^2(\Delta)},$$

gdzie:
d — liczba wymiarów (w rozważanym przypadku $d = 2$), a $\bar{\delta}^a(\Delta)$ zostało wyliczone jako:

$$\bar{\delta}^a(\Delta) = \frac{1}{N-\Delta} \sum_{i=0}^{N-\Delta} \|X_{i+\Delta} - X_i\|^a.$$

W przypadku ruchu Browna współczynnik anty-gaussowości wynosi $G(\Delta) = 0$.

Liniowość mówi o tym, jak prosta jest trajektoria. Wyliczana jest jako stosunek przemieszczenia cząstki do przebytej przez nią drogi [12]:

$$S = \frac{\|X_N\|}{\sum_{i=1}^{N} \|S_i\|}.$$

Do zbadania zależności pomiędzy następującymi po sobie krokami zostanie policzona **autokorelacja**. Autokorelacja dwóch kroków oddalonych o Δ pomiarów od siebie wynosi:

$$\gamma(\Delta) = \frac{1}{\sigma^2 N} \sum_{i=1}^{N-\Delta} \left(\|S_{i+\Delta}\| - \bar{S}\right)\left(\|S_i\| - \bar{S}\right),$$

gdzie:
σ^2 — wariancja długości kroków trajektorii,
$\bar{S}$ — średnia długość kroku w trajektorii.

Jako **maksymalną odległość** $d_{\max}$ przyjęto maksymalną odległość między dwoma dowolnymi położeniami cząstki w trajektorii:

$$d_{\max} = \max_{\substack{i,j=1,\ldots,N \\ i>j}} \{\|X_i - X_j\|\}.$$

Prawdopodobieństwo uwięzienia P^T liczone jest jako szansa, że cząstka jest ograniczona przez otoczenie o promieniu r_0 i wyrażone jest wzorem [26]:

$$P^T = P(D, t, r_0) = 1 - \exp\left(0.2045 - 0.25117\left(\frac{Dt}{r_0^2}\right)\right).$$

Jako t przyjęto łączny czas trwania całej trajektorii czyli

$$t = T \overset{\Delta t=1}{=} N.$$

Jako r_0 przyjęto połowę maksymalnej odległości między dwoma dowolnymi położeniami cząstki:

$$r_0 = \frac{1}{2} d_{\max}.$$

Wymiar fraktalny D_f można wyznaczyć jako [10]:

$$D_f = \frac{\log N}{\log\left(N \frac{d_{\max}}{L}\right)},$$

gdzie L to długość trajektorii wyliczona jako:

$$L = \sum_{i=1}^{N} \|S_i\|.$$

1.4.2 Podstawowe informacje o uczeniu maszynowym

Definicja 1.14. Metody **uczenia maszynowego** to algorytmy, które mają zdolność do automatycznego poprawiania się poprzez naukę (algorytmy samouczące).
Najczęściej budują one modele matematyczne na podstawie pewnych danych zwanych **danymi treningowymi**.

Uczenie często polega na dążeniu do zmniejszenia tzw. **funkcji błędu** [9]. W przypadku analiz przeprowadzanych w niniejszej pracy, jako funkcję błędu będziemy rozumieć błąd średnio-kwadratowy (ang. *mean squared error*, MSE) wyrażony wzorem:

$$\mathrm{MSE}(\hat{\alpha}) = \frac{1}{M}\sum_{i=1}^{M}\left(\hat{\alpha}_i - \alpha_i\right)^2,$$

gdzie:
M — liczba trajektorii,
$\hat{\alpha}_i$ — wyestymowana przez model wartość parametru α,
α — prawdziwa wartość parametru α.

W danych treningowych musi być zawarta informacja na temat prawdziwej wartości estymowanego parametru, aby umożliwić modelowi sprawdzanie poprawności wytrenowanego modelu.

Gdy model jest zbyt skomplikowany następuje tak zwane **przetrenowanie**.

Definicja 1.15. Przetrenowanie (ang. *overfitting*) występuje, gdy model nauczy się danych tak mocno, że zaczyna traktować szum jako ważną cechę, przez co podejmowane przez niego decyzje dla danych, których wcześniej nie widział, są błędne lub nie tak dokładne [31].

Przetrenowanie jest częstym problemem w uczeniu maszynowym.

1.4.3 Wielowymiarowa regresja liniowa

Jednym z algorytmów uczenia maszynowego jest wielowymiarowa regresja liniowa.

Definicja 1.16. Wielowymiarowa regresja liniowa polega na optymalnym dobraniu parametrów (wag) $a, b_1, b_2, ..., b_w$ [11], gdzie w to liczba parametrów w równaniu:

$$y_i = a + b_1 x_1 + \cdots + b_w x_w,$$

gdzie x_i to wartość i-tej własności ruchu dla danej trajektorii.

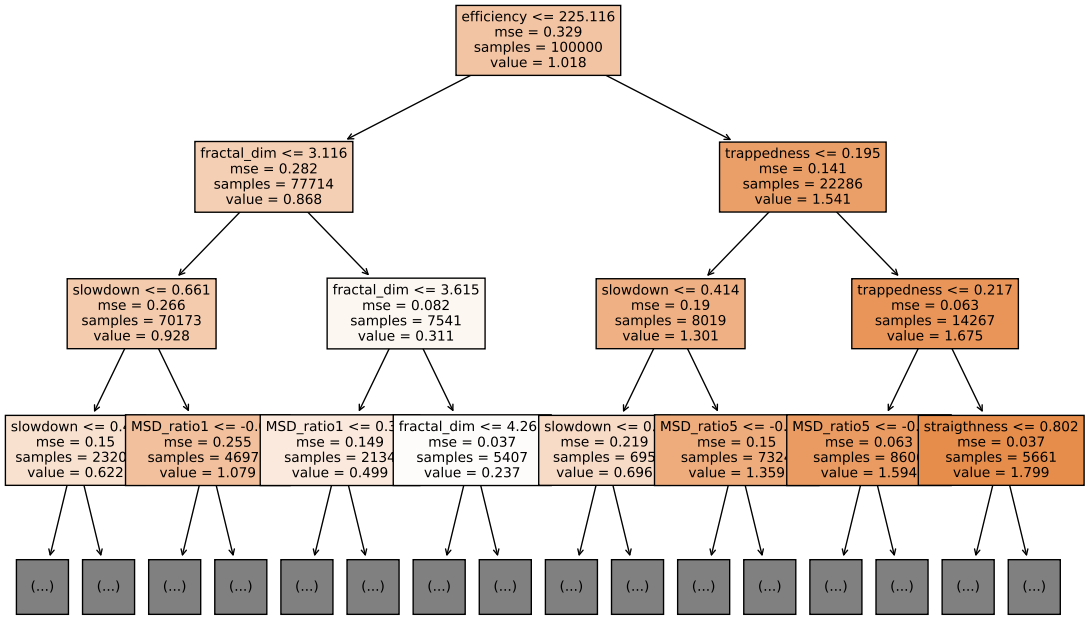
1.4.4 Drzewo decyzyjne

Definicja 1.17. Metoda **drzewa decyzyjnego** polega na sekwencyjnym dzieleniu danych na podzbiory. Podział następuje na podstawie jednego z parametrów wejściowych w tzw. **węzłach**. Węzeł dzieli dane na dwa podzbiory, jeśli są spełnione określone w modelu warunki. Węzeł, który nie ulega podziałowi nazywany jest **liściem**. Podczas estymacji model kwalifikuje dane do jednego liścia, który określa ostateczną decyzję drzewa [2].

Przykładowe drzewo decyzyjne zostało przedstawione na rysunkach 1.3 i *Dodatki*-A.1.

Drzewo decyzyjne posiada tak zwane **hiperparametry**, czyli parametry, których algorytm nie jest w stanie sam określić i które regulują strukturę drzewa. Hiperaparametrami drzewa decyzyjnego są [18]:

- maksymalna głębokość $m_{\max}$ — określa, ile maksymalnie decyzji może podjąć drzewo decyzyjne,
- minimalna liczba próbek potrzebna do podziału gałęzi $n_{\mathrm{min,parent}}$,



Rysunek 1.3: Początek przykładowego drzewa decyzyjnego dla 100 000 trajektorii. Kolor poszczególnych komórek decyzyjnych odpowiada przewidywanej wartości parametru α. Pełne drzewo o głębokości maksymalnej $m_{\max} = 10$ zostało zawarte w sekcji *Dodatki*-A.1.

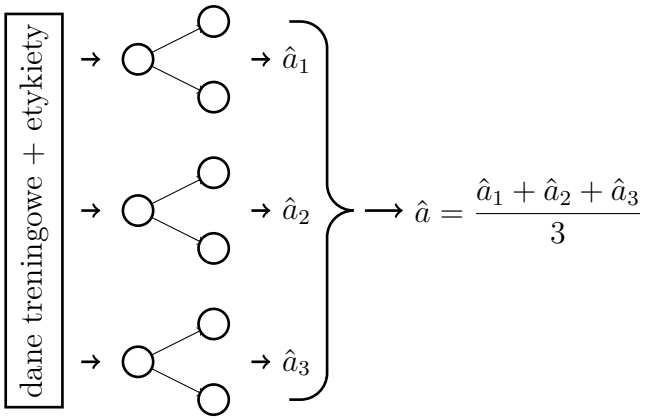
- minimalna liczba próbek na pochodnej gałęzi, która pozwala na podział $n_{\text{min,child}}$,
- liczba właściwości na które ma patrzeć węzeł, aby podjąć najlepszą decyzję w,
- maksymalna liczba węzłów $l_{\max}$.

Aby dany węzeł podzielił się na dwa kolejne muszą zachodzić następujące warunki:

- aktualna liczba podjętych decyzji (głębokość drzewa) prowadzących do tego węzła musi być mniejsza niż $m_{\max}$,
- liczba próbek zawartych w tym węźle musi być co najmniej $n_{\text{min,parent}}$,
- żaden z liści powstałych z tego węzła nie może mieć mniej próbek niż $n_{\text{min,child}}$,
- aktualna liczba węzłów w drzewie jest mniejsza niż $l_{\max}$.

Algorytmy używane do podziałów w regresyjnych drzewach decyzyjnych to [18]:

- minimalizacja wariancji (MSE),
- minimalizacja wariancji z ulepszeniem Friedmana (Friedman MSE),
- minimalizacja błędu (MAE),
- minimalizacja odchylenia Poissona.



Rysunek 1.4: Schemat działania regresyjnego lasu losowego.

1.4.5 Las losowy

Bardziej skomplikowaną metodą od pojedynczego drzewa decyzyjnego jest zespół drzew, który wspólnie podejmuje decyzję na temat wyniku. Najprostszym rozwiązaniem tego typu jest las losowy (ang. *random forest*).

Definicja 1.18. Las losowy to metoda uczenia maszynowego na podstawie $N \in \mathbb{N}$ drzew decyzyjnych. Każde z drzew jest uczone niezależnie na innym zbiorze danych testowych. Przy estymacji każde drzewo podejmuje równoważną decyzję [28]. Zbiory danych dla poszczególnych drzew mogą być wybierane albo poprzez podział danych (trajektorii) albo podział ze względu na parametry wejściowe [18]. Schemat działania lasu losowego został przedstawiony na rysunku 1.4.

Ostateczna wartość wyliczona przez model ma postać:

$$\hat{a} = \frac{1}{N} \sum_{i=1}^{N} \hat{a}_i,$$

gdzie:
N — liczba drzew w lesie losowym,
$\hat{a}_i$ — wartość parametru przewidziana przez i-te drzewo w lesie losowym.

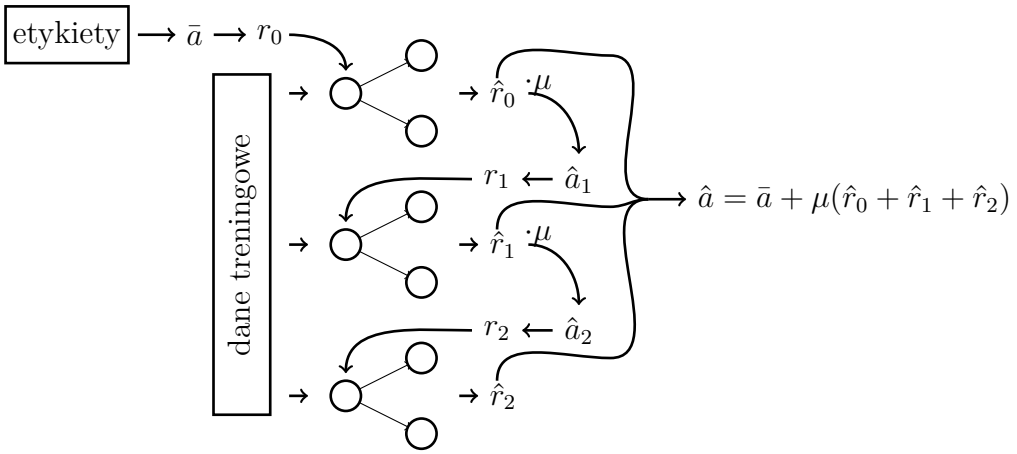
Model lasu losowego posiada te hiperparametry, co drzewo decyzyjne, a dodatkowo [18]:

- liczbę drzew decyzyjnych N,
- część danych lub parametrów, na podstawie której generowane jest pojedyncze drzewo p.

1.4.6 Wzmocnienie gradientowe

Kolejnym rozszerzeniem, jakie można wprowadzić do zespołu drzew, jest umożliwienie drzewom nauki na błędach poprzedniego drzewa. Jednym z algorytmów tego rodzaju jest wzmocnienie gradientowe.

Definicja 1.19. Wzmocnienie gradientowe to metoda uczenia maszynowego na podstawie $N \in \mathbb{N}$ drzew decyzyjnych. W odróżnieniu od lasu losowego (definicja 1.18) każde kolejne drzewo $i + 1$ jest nauczane do przewidywania residuów poprzedniego drzewa r_i. Ostateczny wynik jest sumą średniej wartości szukanego parametru z danych treningowych $\bar{r}$ i przeskalowanej sumy przewidzianych przez wszystkie drzewa residuów [2, 18, 30].



Rysunek 1.5: Schemat działania wzmocnienia gradientowego.

Ze względu na to, że model jest rekurencyjny i wynik każdego drzewa opiera się na wynikach poprzednika, potrzebna jest początkowa wartość residuów r_0^j dla każdej trajektorii $j = 1, 2, ..., M$. Jest ona wyliczana jako kwadrat różnicy między prawdziwą wartością parametru a^j a średnią wartością wszystkich parametrów $\bar{a}$ [18, 30]:

$$r_0^j = \left(a^j - \sum_{k=1}^{M} a^k\right)^2 = \left(a^j - \bar{a}\right)^2,$$

gdzie M to liczba wszystkich trajektorii.
Czasem jako residuum jest brany po prostu błąd całkowity estymacji parametru:

$$r_0^j = a^j - \bar{a}.$$

Model wzmocnienia gradientowego posiada te hiperparametry, co drzewo losowe z wyjątkiem p, a do tego [18]:

- funkcję błędu — funkcja z której liczone są residua,
- współczynnik nauczania μ — wartość, przez jaką zostaje przemnożona wartość przewidzianych residuów przed wyznaczeniem estymatora parametru a.

Poniżej zaprezentowano **algorytm** obliczania błędu r_i popełnionego przez aktualne drzewo na podstawie estymacji błędu poprzednich drzew $r_0, ..., r_{i-1}$ [2, 18].

1. Wyliczana jest przewidywana wartość szukanego parametru parametru

$$\hat{a}_i = \bar{a} \ + \mu \left(\sum_{k=1}^{i-1} \hat{r}_k\right).$$

2. Wartość ta porównywana jest z prawdziwą wartością a tego parametru. Błąd przybliżenia wyliczany jest z funkcji błędu, która przeważnie ma postać

$$r_i = (a - \hat{a}_i)^2. \tag{1.9}$$

W niektórych przypadkach przyjmuje się

$$r_i = a - \hat{a}_i.$$

Ostateczna wynik działania modelu ma postać:

$$\hat{a} = \bar{a} + \mu \left(\sum_{i=1}^{N} \hat{r}_{i-1} \right),$$

gdzie:
$\bar{a}$ — wartość średnia szukanego parametru wyliczona z danych testowych,
μ — **współczynnik nauczania**,
N — liczba drzew w modelu,
$\hat{r}_i$ — wartość residuum przewidziana przez i-te drzewo wzmocnienia gradientowego.

Schemat modelu przedstawiono na rysunku 1.5.

1.5 Wskaźniki skuteczności modeli

W celu oceny jakości wytrenowanych modeli wprowadza się 3 statystyki testowe:

1. współczynnik determinacji,
2. średni błąd całkowity,
3. błąd średnio-kwadratowy.

Definicja 1.20. Współczynnik determinacji (R2, R^2) opisuje w jakim stopniu model pasuje do próbki danych. Liczony jest on jako kwadrat współczynnika korelacji [13]:

$$R^2 = \left(\frac{\mathrm{Cov}(\alpha, \hat{\alpha})}{\sigma_\alpha \cdot \sigma_{\hat{\alpha}}} \right)^2$$

gdzie:
$\mathrm{Cov}(\alpha, \hat{\alpha})$ — współczynnik kowariancji pomiędzy wartościami prawdziwymi α i tymi estymowanymi przez model $\hat{\alpha}$,
σ_α — odchylenie standardowe dla estymowanych wartości $\hat{\alpha}$, analogicznie $\sigma_{\hat{\alpha}}$.

Współczynnik determinacji przyjmuje wartości od 0 do 1. Im wynik jest bliżej 1 tym model jest lepiej dopasowany. W tabeli 1.2 przedstawiono słowne oceny dopasowania trajektorii na podstawie wartości współczynnika R2 [1].

Definicja 1.21. Średni błąd całkowity (ang. *mean absolute error*, MAE) to średnia wartość różnicy pomiędzy estymowaną wartością parametru $\hat{\alpha}$, a prawdziwą wartością parametru α i wyrażony jest wzorem [24]:

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{\alpha} - \alpha|,$$

gdzie N to rozmiar badanej próbki.

Im mniejsza wartość MAE tym lepsze dopasowanie.

Definicja 1.22. Błąd średnio-kwadratowy (ang. *mean-squared error*, MSE) to średnia wartość kwadratu różnicy pomiędzy estymowaną wartością parametru $\hat{\alpha}$, a prawdziwą wartością parametru α i wyrażony jest wzorem [25]:

$$\mathrm{MSE} = \frac{1}{N} \sum_{i=1}^{N} (\hat{\alpha} - \alpha)^2,$$

gdzie N to rozmiar badanej próbki. Im mniejsza wartość MSE tym lepsze dopasowanie. MSE jest bardziej podatny na wartości odstające niż MAE.

R2	poziom dopasowania
$0.0 - 0.5$	dopasowanie błędne
$0.5 - 0.6$	dopasowanie słabe
$0.6 - 0.8$	dopasowanie zadowalające
$0.8 - 0.9$	dopasowanie dobre
$0.9 - 1.0$	dopasowanie bardzo dobre

Tabela 1.2: Słowna ocena dopasowania modelu na podstawie wartości współczynnika determinacji [1].

Rozdział 2

Ułamkowy ruch Browna

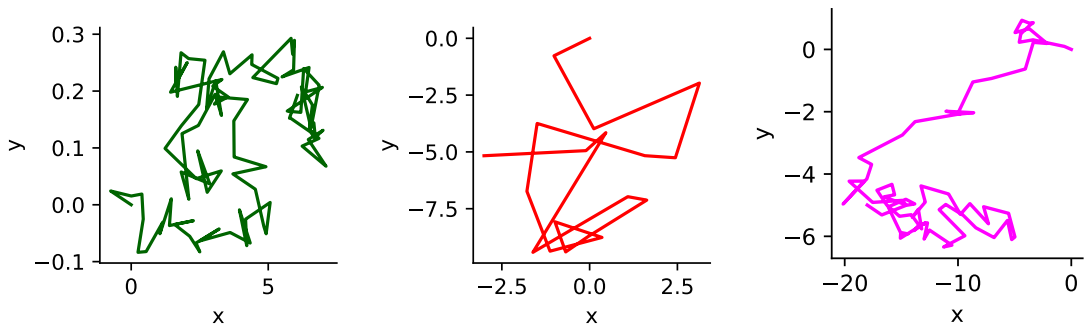
Przed rozpoczęciem nauki modeli na wszystkich wymienionych pod definicją 1.18 rodzajach ruchu, warto sprawdzić, jak modele będą sobie radziły z estymacją wykładnika dyfuzji. W tym celu warto wybrać model, który nie będzie komplikował algorytmom pracy i umożliwi generowanie trajektorii będących subdyfuzjami lub superdyfuzjiami. Takim modelem jest ułamkowy ruch Browna [16].

Do wygenerowania danych używanych w kolejnych analizach posłużono się biblioteką „andi-datasets" [16] w języku Python przygotowaną na potrzeby „Anomalous Diffusion (AnDi) Challenge" [17]. Wszystkie wygenerowane trajektorie zaczynają się w punkcie $(0, 0)$. Posiadają one losowy współczynnik dyfuzji D, losowy wykładnik dyfuzji $\alpha \in \{0,05; 0,1; ...; 1,95\}$ oraz nałożony gaussowski szum o losowej intensywności spośród $\{0.1, 0.3, 1\}$.

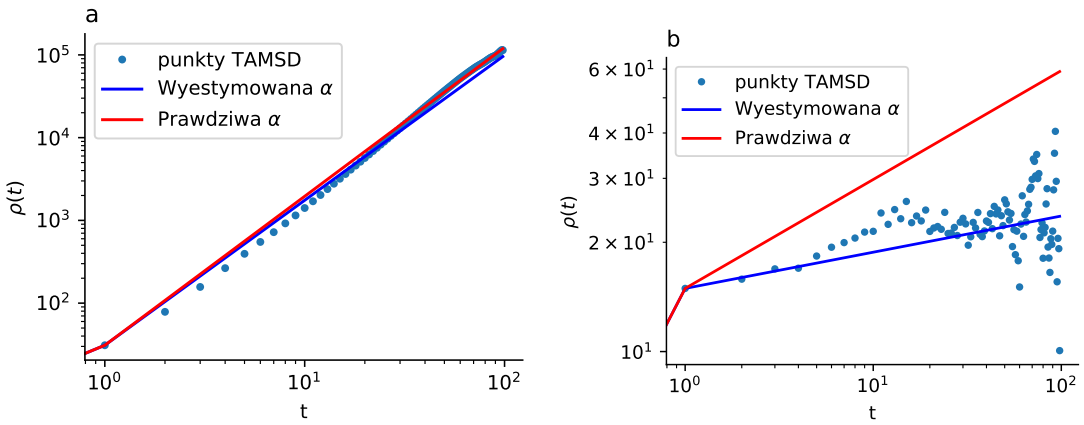
W pierwszej części eksperymentu wygenerowano 3 grupy danych:

1. pierwsza grupa A zawiera trajektorie o długości $T = 100$ kroków.
2. druga grupa B zawiera trajektorie o długości $T = 20$ kroków.
3. trzecia grupa C zawiera trajektorie o losowej długości od $T = 20$ do $T = 100$ kroków.

Dla każdej z grup wysymulowano 110 000 trajektorii. Przy czym 100 000 trajektorii to dane treningowe, a 10 000 trajektorii to dane testowe.



Rysunek 2.1: Przykłady trajektorii wygenerowanych za pomocą AnDi-datasets. Przedstawione trajektorie posiadają losowe modele, zaszumienie, współczynnik dyfuzji i wykładnik α. Od lewej kolejno trajektorie z grup A, B i C.



Rysunek 2.2: Przykładowe wykresy zawierające wyliczone współczynniki TAMSD oraz wykresy krzywych TAMSD zgodnych z równaniem (1.5) dla wyestymowanego i prawdziwego wykładnika dyfuzji anomalnej.

Na danych treningowych każdej grupy wyszkolono osobne modele:

1. model **A** na trajektoriach z grupy A,
2. model **B** na trajektoriach z grupy B,
3. model **C** na trajektoriach z grupy C.

Przykłady wygenerowanych trajektorii przedstawiono na rysunku 2.1. Każdy z wyszkolonych modeli sprawdzono dla danych testowych grup A, B i C.

Wykładnik dyfuzji α metodą TAMSD wyznaczono zgodnie ze wzorem (1.8). Dla wygenerowanych trajektorii estymacja TAMSD często przynosiła wyniki α bliskie prawdziwemu jak zaprezentowano na rysunku 2.2(a), a czasem dawała wyniki dość odległe od prawdziwych (rysunek 2.2(b)). Jak można zauważyć, na rysunku 2.2(b) współczynniki TAMSD mają niewykładniczy przyrost. Może to oznaczać, że trajektoria jest silnie zaszumiona, co mocno utrudniło wyznaczenie wykładnika.

W celu przeprowadzenia estymacji wygładnika dyfuzji anomalnej α metodami uczenia maszynowego wykorzystano bibliotekę „scikit-learn” [18] w języku Python. W tabeli 2.1 przedstawiono modele z tej biblioteki, których użyto do analizy trajektorii.

Do algorytmów uczenia maszynowego z trajektorii treningowych zostały wyliczone parametry zawarte w sekcji 1.4.1. Do wyznaczenia hiperparametrów dla drzewa decyzyjnego, lasu losowego i wzmocnienia gradientowego użyto wbudowanego w scikit-learn obiektu `RandomizedSearchCV`. Obiekt operował na 10% danych treningowych (10 000 trajektorii). Otrzymane hiperparametry dla każdego z modeli przedstawiono w tabeli 2.2.

model	obiekt
regresja liniowa	`linear_model.LinearRegression`
drzewo decyzyjne	`tree.DecisionTreeRegressor`
las losowy	`ensemble.RandomForestRegressor`
wzmocnienie gradientowe	`ensemble.GradientBoostingRegressor`
szukanie hiperparametrów	`model_selection.RandomizedSearchCV`

Tabela 2.1: Obiekty biblioteki scikit-learn do poszczególnych algorytmów uczenia maszynowego.

modele **A**	$m_{\max}$	$n_{\min,\text{parent}}$	$n_{\min,\text{child}}$	w	N	p	μ
drzewo decyzyjne	7	9	4	10	—	—	—
las losowy	18	3	2	10	900	0,30	—
wzmocnienie gradientowe	5	8	5	4	700	—	0,014
modele **B**	$m_{\max}$	$n_{\min,\text{parent}}$	$n_{\min,\text{child}}$	w	N	p	μ
drzewo decyzyjne	5	3	3	4	—	—	—
las losowy	16	10	2	4	1000	0,60	—
wzmocnienie gradientowe	5	8	5	4	700	—	0,014
modele **C**	$m_{\max}$	$n_{\min,\text{parent}}$	$n_{\min,\text{child}}$	w	N	p	μ
drzewo decyzyjne	7	9	4	10	—	—	—
las losowy	14	3	3	4	500	0,60	—
wzmocnienie gradientowe	5	8	5	4	700	—	0,014

Tabela 2.2: Hiperparametry otrzymane w wyniku działania „RandomizedSearchCV” dla modeli z kategorii **A**, **B** i **C**.

Modele wyszkolono na danych treningowych danej grupy. Aby nie dopuścić do przetrenowania pomimo dużej liczby drzew w modelu w przypadku lasu losowego i wzmocnienia gradientowego, ich drzewa mają niewielką głębokość $m_{\max}$. Dla lasu losowego jest ona na poziomie $m_{\max} \in [14; 18]$, a dla wzmocnienia gradientowego — zawsze równa $m_{\max} = 5$.

Wszystkie modele przetestowano na trajektoriach testowych każdej grupy. Przykładową dokładność przewidywań dla poszczególnych wartości wykładnika $\hat{\alpha}$ przedstawiono na wykresach w sekcji *Dodatki*-C.1,C.2.

Współczynniki determinacji przy estymacjach wykładnika dyfuzji α dla każdego z modeli przedstawiono w tabeli 2.3. W tabeli 2.4 zaprezentowano dodatkowo średnie błędy absolutne dla każdego z przypadków, a w tabeli 2.5 błędy średnio-kwadratowe.

dane testowe z grupy:	dane A			dane B		
dane treningowe z grupy:	**A**	**B**	**C**	**A**	**B**	**C**
TAMSD	0.861	0.861	0.861	0.680	0.680	0.680
regresja liniowa	0.876	0.010	0.868	0.646	0.679	0.660
drzewo decyzyjne	0.886	0.774	0.882	0.637	0.687	0.681
las losowy	**0.898**	0.862	0.896	0.653	0.714	0.707
wzmocnienie gradientowe	**0.898**	0.861	0.896	0.674	**0.715**	0.708

dane testowe z grupy:	dane C		
dane treningowe z grupy:	**A**	**B**	**C**
TAMSD	0.810	0.810	0.810
regresja liniowa	0.802	0.234	0.812
drzewo decyzyjne	0.816	0.753	0.826
las losowy	0.832	0.820	**0.844**
wzmocnienie gradientowe	0.839	0.822	**0.844**

Tabela 2.3: Współczynniki determinacji **R2** modeli przy estymacji wykładnika dyfuzji dla danych z poszczególnych grup. Kolorem zielonym oznaczono najlepszą estymację w danej grupie danych testowych. Pogrubieniem natomiast najlepszy model dla każdej z zielonych kolumn.

dane testowe z grupy:	dane A			dane B		
dane treningowe z grupy:	**A**	**B**	**C**	**A**	**B**	**C**
TAMSD	0.085	0.085	0.085	0.232	0.232	0.232
regresja liniowa	0.039	1.108	0.043	0.196	0.100	0.117
drzewo decyzyjne	0.036	0.097	0.037	0.120	0.098	0.101
las losowy	**0.032**	0.044	0.033	0.115	**0.089**	0.092
wzmocnienie gradientowe	**0.032**	0.044	0.033	0.105	**0.089**	0.091
dane testowe z grupy:	dane C					
dane treningowe z grupy:	**A**	**B**	**C**			
TAMSD	0.122	0.122	0.122			
regresja liniowa	0.074	0.346	0.059			
drzewo decyzyjne	0.058	0.091	0.055			
las losowy	0.054	0.056	**0.049**			
wzmocnienie gradientowe	0.051	0.056	**0.049**			

Tabela 2.4: Średnie błędy absolutne **MAE** popełniane przez modele przy estymacji wykładnika dyfuzji dla danych z poszczególnych grup. Kolorem zielonym oznaczono najlepszą estymację w danej grupie danych testowych. Pogrubieniem natomiast najlepszy model dla każdej z zielonych kolumn.

Dla każdego rodzaju danych testowych, najlepsze wyniki otrzymano dla modeli, które były uczone na danych o tej samej długości trajektorii co dane treningowe. Dla danych testowych A i B, drugą najlepszą grupą modeli jest grupa **C**, która miała trajektorie mieszanej długości. Najgorzej wyniki dają modele, które były szkolone na zupełnie innych danych treningowych niż dane testowe.

Najlepszymi modelami okazały się wzmocnienie gradientowe i las losowy. Nie jest to nic dziwnego, gdyż są to najbardziej zaawansowane wśród wybranych modeli. Wybór najlepszych modeli patrząc na każdą ze statystyk (R2, MAE, MSE) pokrywa się.

dane testowe z grupy:	dane A			dane B		
dane treningowe z grupy:	**A**	**B**	**C**	**A**	**B**	**C**
TAMSD	0.225	0.225	0.225	0.387	0.387	0.387
regresja liniowa	0.156	0.577	0.164	0.360	0.250	0.270
drzewo decyzyjne	0.147	0.242	0.150	0.274	0.244	0.248
las losowy	**0.139**	0.169	0.141	0.270	0.233	0.237
wzmocnienie gradientowe	0.140	0.168	0.141	0.257	**0.232**	0.235
dane testowe z grupy:	dane C					
dane treningowe z grupy:	**A**	**B**	**C**			
TAMSD	0.270	0.270	0.270			
regresja liniowa	0.215	0.354	0.191			
drzewo decyzyjne	0.188	0.233	0.180			
las losowy	0.180	0.188	0.171			
wzmocnienie gradientowe	0.175	0.186	**0.170**			

Tabela 2.5: Błędy średnio-kwadratowe **MSE** popełniane przez modele przy estymacji wykładnika dyfuzji dla danych z poszczególnych grup. Kolorem zielonym oznaczono najlepszą estymację w danej grupie danych testowych. Pogrubieniem natomiast najlepszy model dla każdej z zielonych kolumn.

Klasa modelu	Model	szukanie hiperparametrów mm:ss	nauka modelu mm:ss	estymowanie parametru ss.ms
A	TAMSD	—	—	13.875
	regresja liniowa	—	00:00	00.005
	drzewo decyzyjne	00:06	00:01	00.004
	las losowy	07:57	08:09	02.459
	wzmocnienie gradientowe	19:49	15:09	00.360
B	TAMSD	—	—	03.480
	regresja liniowa	—	00:00	00.006
	drzewo decyzyjne	00:04	00:01	00.004
	las losowy	07:50	03:59	02.328
	wzmocnienie gradientowe	19:38	16:09	00.288
C	TAMSD	—	—	08.544
	regresja liniowa	—	00:00	00.007
	drzewo decyzyjne	00:04	00:01	00.004
	las losowy	07:31	01:52	01.065
	wzmocnienie gradientowe	19:35	13:38	00.275

Tabela 2.6: Czasy trwania różnych etapów pracy modeli z kategorii A, B i C. Szukanie hiperparametrów odbyło się na 90 próbkach z 10 000 trajektorii. Nauka modelu odbyła się na 100 000 trajektorii. Estymowanie parametru odbyło się na 10 000 trajektorii. Dane techniczne komputera zostały zawarte w sekcji *Dodatki*-D.1.

Najlepsze wyniki wśród badanych danych treningowych uzyskano dla danych testowych z grupy A (długość $T = 100$), natomiast najgorsze dla danych z grupy B (długość $T = 10$).

Dane w grupie C zostały przygotowane, aby sprawdzić jak modele **A** i **B** radzą sobie z nieznanymi dla siebie danymi. Biorąc pod uwagę te 2 modele — lepiej poradził sobie model **A**, który był uczony na dłuższych trajektoriach niż model **B**.

Z powyższego wynika, że aby zwiększyć skuteczność modelu (R2) w estymowaniu wykładnika dyfuzji α, należy używać modelu nauczonego na tej samej długości trajektorii, co badana próbka, a gdy nie posiada się wystarczającej liczby próbek — na modelu uczonym na różnych długościach trajektorii.

Mimo, że dokładność jest najważniejsza, to nie można zapomnieć o koszcie czasowym użycia modelu. W tabeli 2.6 przedstawiono czas szacowania hiperparametrów, nauki i estymacji wykładnika. W tabeli zawarto jedynie przypadki, gdy trajektorie testowe pochodziły z tej samej grupy co trajektorie treningowe danego modelu.

Z tabeli wynika, że najwięcej czasu zajmuje wzmocnienie gradientowe. Jednak jest to prawda tylko wtedy, gdy patrzymy na cały proces, w którym zawarta jest również nauka modelu. Jednak model, który raz został nauczony, może być wykorzystany wiele razy do estymacji parametrów, więc najistotniejszym czasem jest czas estymacji.

Jak można zauważyć, szukanie hiperparametrów oraz nauka modelu zajmuje nieporównywalnie więcej czasu niż estymowanie wykładnika dyfuzji. W takim razie, jeśli posiadamy wyszkolony model, to czas samej estymacji wykładnika dyfuzji α jest zaniedbywalnie mały (w tabeli dla 10 000 trajektorii). Biorąc pod uwagę samą estymację parametru α najwolniejszym modelem jest las losowy. Wynika to z dużo większej głębokości drzew $m_{\max}$, niż w przypadku wzmocnienia gradientowego, przy porównywalnej ich liczbie N.

Rozdział 3

Inne rodzaje ruchu

Po zbadaniu trajektorii dla jednego rodzaju ruchu przyszedł czas na wprowadzenie pozostałych modeli.

Procedury generowania danych oraz nauczania i testowania modeli przeprowadzono tak, jak w rozdziale 2. z tą różnicą, że trajektorie wygenerowano z wszystkich ruchów zawartych pod definicją 1.18.

Hiperparametry znalezione przez `RandomizedSearchCV` zawarto w tabeli 3.1.

Aby nie dopuścić do przetrenowania, drzewa w modelach mają niewielką głębokość $m_{\max}$. W przypadku lasu losowego jest ona na poziomie $m_{\max} \in [14; 15]$, a w przypadku wzmocnienia gradientowego — z przedziału $m_{\max} \in [5; 8]$.

Wszystkie modele przetestowano na trajektoriach testowych każdej z grup. Przykładową dokładność przewidywań dla poszczególnych wartości wykładnika $\hat{\alpha}$ przedstawiono na wykresach w sekcji *Dodatki*-E.1-E.2.

Współczynniki determinacji przy estymacjach wykładnika dyfuzji α dla każdego z modeli przedstawiono w tabeli 3.2. W tabeli 3.3 przedstawiono dodatkowo średnie błędy absolutne dla każdego z przypadków, a w tabeli 3.4 błędy średnio-kwadratowe.

Tak jak w przypadku ułamkowego ruchu Browna, najlepsze wyniki dają trajektorie uczone na tej samej długości trajektorii co dane testowe. Po ich wykluczeniu nadal lepiej radzi sobie model szkolony na zmiennej długości trajektorii.

model **A**	$m_{\max}$	$n_{\min,\text{parent}}$	$n_{\min,\text{child}}$	w	N	p	μ
drzewo decyzyjne	7	9	4	10^1	—	—	—
las losowy	15	6	3	4	800	0,80	—
wzmocnienie gradientowe	8	5	4	4	600	—	0,007
model **B**	$m_{\max}$	$n_{\min,\text{parent}}$	$n_{\min,\text{child}}$	w	N	p	μ
drzewo decyzyjne	5	3	3	10	—	—	—
las losowy	14	5	4	4	600	0,50	—
wzmocnienie gradientowe	5	8	5	4	700	—	0,014
model **C**	$m_{\max}$	$n_{\min,\text{parent}}$	$n_{\min,\text{child}}$	w	N	p	μ
drzewo decyzyjne	7	10	3	4	—	—	—
las losowy	15	6	3	4	800	0,80	—
wzmocnienie gradientowe	8	5	4	4	600	—	0,007

Tabela 3.1: Hiperparametry otrzymane w wyniku działania „RandomizedSearchCV” dla kategorii modeli A, B i C.

dane testowe z grupy:	dane A			dane B		
dane treningowe z grupy:	**A**	**B**	**C**	**A**	**B**	**C**
TAMSD	0.000	0.000	0.000	0.000	0.000	0.000
regresja liniowa	0.455	0.109	0.102	0.212	0.315	0.260
drzewo decyzyjne	0.552	0.443	0.519	0.078	0.351	0.308
las losowy	0.637	0.501	0.618	0.327	0.420	0.389
wzmocnienie gradientowe	**0.641**	0.513	0.624	0.354	**0.422**	0.391
dane testowe z grupy:	dane C					
dane treningowe z grupy:	**A**	**B**	**C**			
TAMSD	0.000	0.000	0.000			
regresja liniowa	0.346	0.124	0.369			
drzewo decyzyjne	0.381	0.397	0.482			
las losowy	0.516	0.453	0.554			
wzmocnienie gradientowe	0.526	0.464	**0.557**			

Tabela 3.2: Współczynniki determinacji **R2** modeli przy estymacji wykładnika dyfuzji dla danych z poszczególnych grup. Kolorem zielonym oznaczono najlepszą estymację w danej grupie danych testowych. Pogrubieniem natomiast najlepszy model dla każdej z zielonych kolumn.

Również w tym przypadku najlepszymi modelami okazały się wzmocnienie gradientowe oraz las losowy. Występuje natomiast przewaga wzmocnienia gradientowego nad lasem losowym. Może to wynikać z większej różnorodności trajektorii.

Niezmiennie, najlepsze wyniki są uzyskane dla danych testowych z grupy A (długość $T = 100$), natomiast najgorsze dla danych z grupy B (długość $T = 10$). Tożsame z wynikami z rozdziału 2. są również wyniki dla danych testowych grupy C, gdzie modele grupy **A** dają lepsze wyniki niż modele grupy **B**.

dane testowe z grupy:	dane A			dane B		
dane treningowe z grupy:	**A**	**B**	**C**	**A**	**B**	**C**
TAMSD	0.543	0.543	0.543	0.649	0.649	0.649
regresja liniowa	0.179	41.148	0.679	0.281	0.223	0.247
drzewo decyzyjne	0.147	0.193	0.159	0.349	0.211	0.227
las losowy	0.120	0.181	0.128	0.252	0.189	0.199
wzmocnienie gradientowe	**0.118**	0.178	0.125	0.247	**0.188**	0.199
dane testowe z grupy:	dane C					
dane treningowe z grupy:	**A**	**B**	**C**			
TAMSD	0.585	0.585	0.585			
regresja liniowa	0.216	7.069	0.207			
drzewo decyzyjne	0.210	0.200	0.170			
las losowy	0.168	0.183	0.147			
wzmocnienie gradientowe	0.164	0.180	**0.146**			

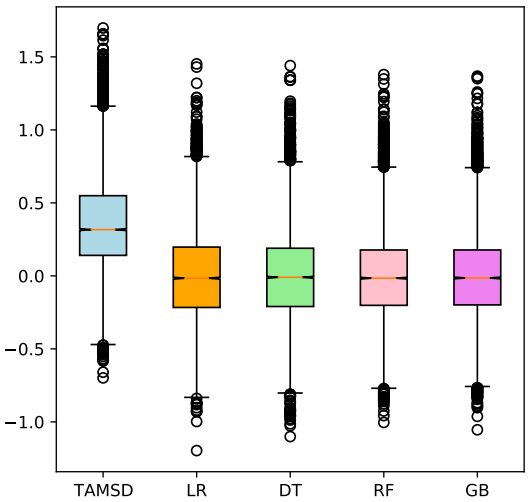
Tabela 3.3: Średnie błędy absolutne **MAE** popełniane przez modele przy estymacji wykładnika dyfuzji dla danych z poszczególnych grup. Kolorem zielonym oznaczono najlepszą estymację w danej grupie danych testowych. Pogrubieniem natomiast najlepszy model dla każdej z zielonych kolumn.

dane testowe z grupy:	dane A			dane B		
dane treningowe z grupy:	**A**	**B**	**C**	**A**	**B**	**C**
TAMSD	0.607	0.607	0.607	0.657	0.657	0.657
regresja liniowa	0.332	1.934	0.346	0.428	0.384	0.399
drzewo decyzyjne	0.294	0.344	0.310	0.469	0.368	0.382
las losowy	0.263	0.340	0.277	0.412	0.345	0.359
wzmocnienie gradientowe	**0.262**	0.334	0.273	0.406	**0.344**	0.358
dane testowe z grupy:	dane C					
dane treningowe z grupy:	**A**	**B**	**C**			
TAMSD	0.627	0.627	0.627			
regresja liniowa	0.371	0.959	0.359			
drzewo decyzyjne	0.350	0.349	0.320			
las losowy	0.318	0.336	0.296			
wzmocnienie gradientowe	0.313	0.331	**0.295**			

Tabela 3.4: Błędy średnio-kwadratowe **MSE** popełniane przez modele przy estymacji wykładnika dyfuzji dla danych z poszczególnych grup. Kolorem zielonym oznaczono najlepszą estymację w danej grupie danych testowych. Pogrubieniem natomiast najlepszy model dla każdej z zielonych kolumn.

Aby zwiększyć dokładność wyznaczania wykładnika dyfuzji, należy więc używać do badań modelu nauczonego na tej samej długości trajektorii, co badana próbka, a gdy nie posiada się wystarczającej liczby próbek — tego szkolonego na różnych długościach trajektorii.

Metoda TAMSD estymowania wykładnika dyfuzji anomalnej α dla ułamkowego ruchu Browna dawała dużo lepsze wyniki niż w tym przypadku. Na rysunku 3.1 przedstawiono wykresy pudełkowe residuów dla każdego z modeli z kategorii **B**, przy testowaniu na danych testowych B. Jak można zauważyć, średni błąd jest daleki od zera, co świadczy o tym, że wynik estymacji jest zawyżany. Przyczyną może być fakt, że niektóre ze zastosowanych



Rysunek 3.1: Residua estymacji wykładnika dyfuzji anomalnej α dla każdego z modeli kategorii **B** dla danych testowych B. Oznaczenia: LR — regresja liniowa, DT — drzewo decyzyjne, RF — las losowy, GB — wzmocnienie gradientowe.

Klasa modelu	Model	szukanie hiperparametrów mm:ss	nauka modelu mm:ss	estymowanie parametru ss.ms
A	TAMSD	—	—	18.527
	regresja liniowa	—	00:00	00.022
	drzewo decyzyjne	00:07	00:01	00.004
	las losowy	08:34	04:32	01.573
	wzmocnienie gradientowe	22:22	01:18	00.095
B	TAMSD	—	—	05.509
	regresja liniowa	—	00:00	00.010
	drzewo decyzyjne	00:06	00:01	00.005
	las losowy	11:18	02:50	01.161
	wzmocnienie gradientowe	25:59	19:46	00.271
C	TAMSD	—	—	09.992
	regresja liniowa	—	00:00	00.006
	drzewo decyzyjne	00:10	00:01	00.004
	las losowy	08:34	04:14	01.607
	wzmocnienie gradientowe	20:08	14:00	00.261

Tabela 3.5: Czasy trwania różnych etapów pracy modeli z kategorii A, B i C. Szukanie hiperparametrów odbyło się na 90 próbkach z 10 000 trajektorii. Nauka modelu odbyła się na 100 000 trajektorii. Estymowanie parametru odbyło się na 10 000 trajektorii. Dane techniczne komputera zostały zawarte w sekcji *Dodatki*-D.1.

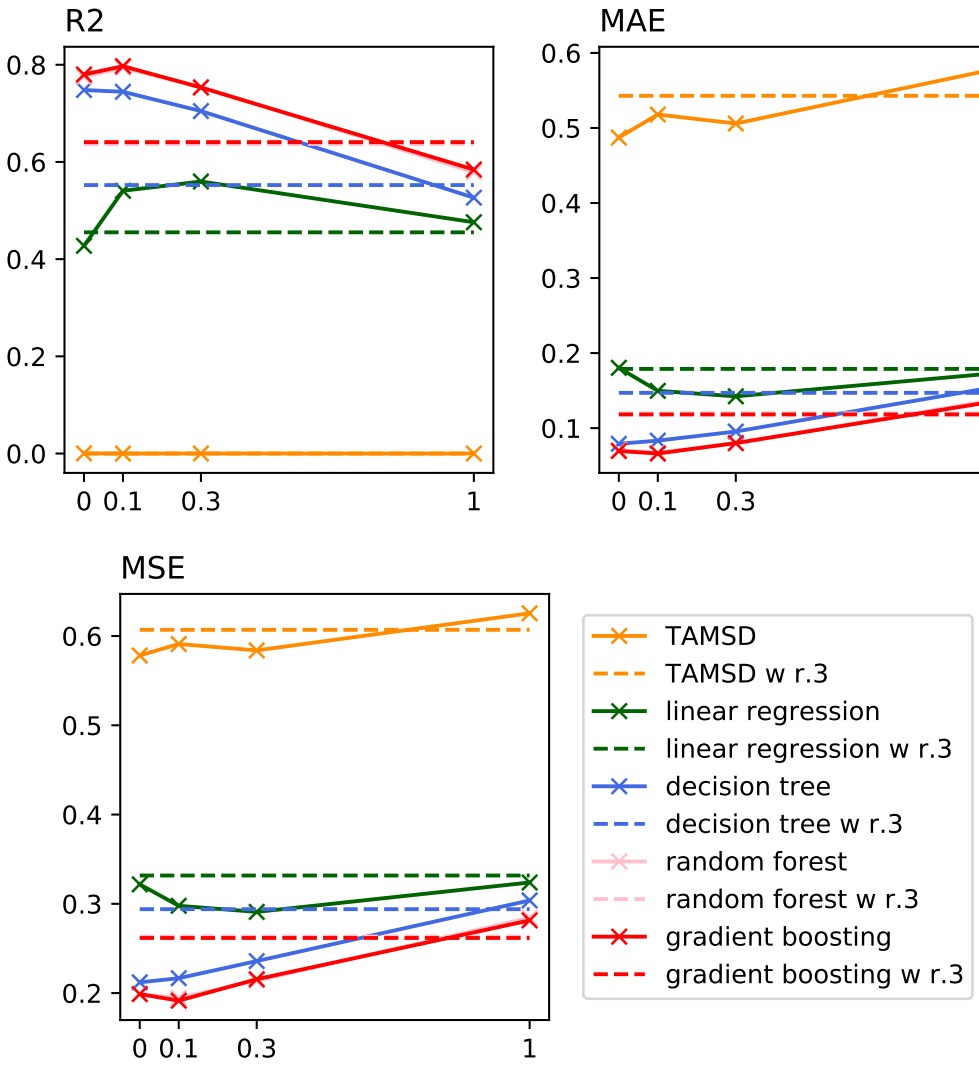
do generowania trajektorii procesów są procesami nieergodycznymi (definicja 1.10), a metoda TAMSD daje błędne wyniki, gdy proces jest nieergodyczny [16].

Wyniki pomiarów czasu pracy różnych modeli są podobne do tych z rozdziału 2, a zawarto je w tabeli 3.5. Wzmocnienie gradientowe wymaga najdłuższego czasu nauki dla każdej klasy modelu, za to ma krótszy czas przewidywania niż las losowy.

Rozdział 4

Wpływ szumu

W kolejnej części pracy zbadano wpływu szumu na dokładność przewidywania wykładnika dyfuzji anomalnej α. Aby skupić się na szumie — wygenerowano tylko trajektorie o długości

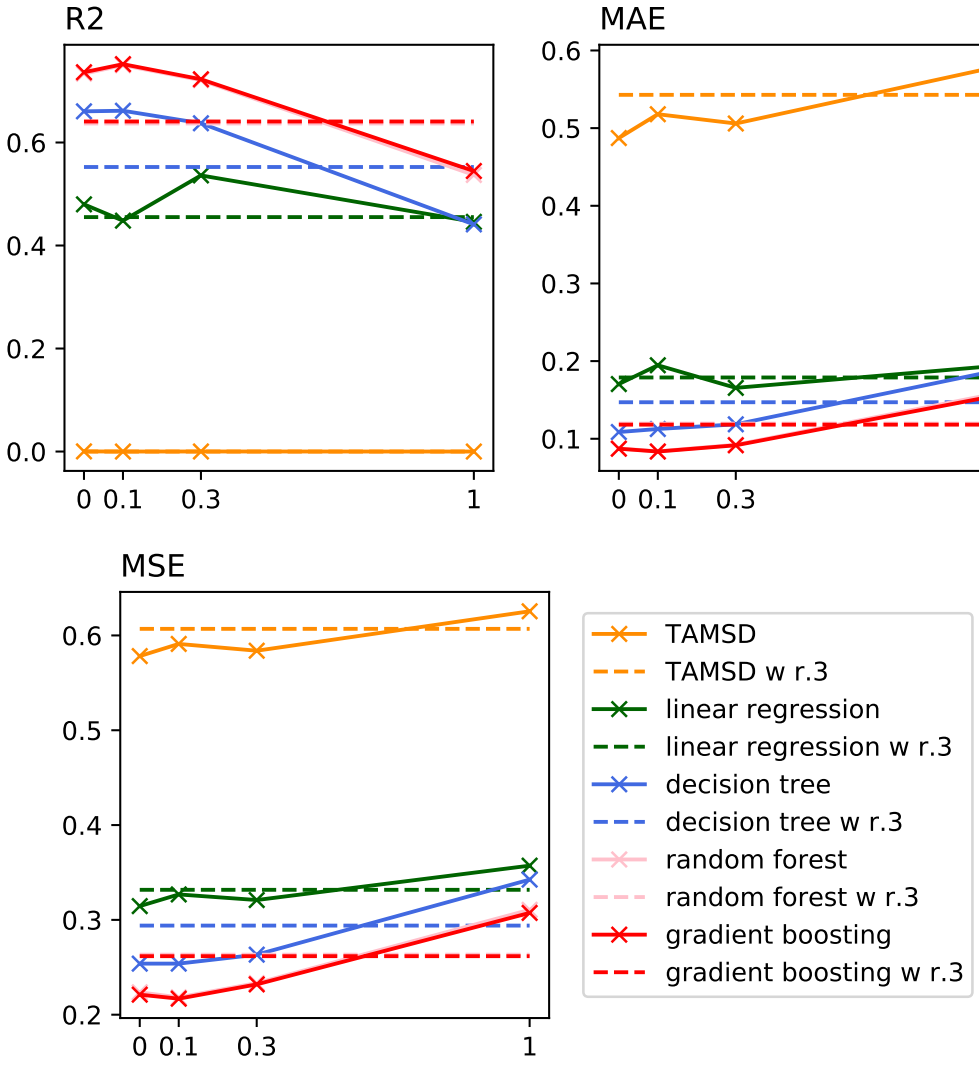


Rysunek 4.1: Wyniki pomiarów skuteczności estymacji parametru α oraz popełniane błędy przy różnym poziomie zaszumienia trajektorii dla modeli szkolonych na trajektoriach o danym poziomie zaszumienia. Dodatkowo zaznaczono poziomy jakie otrzymywały modele kategorii **A** przy danych testowych grupy A w rozdziale 3.

$T = 100$ kroków (grupa A). Utworzono w ten sposób 4 grupy trajektorii o stosunkach szumu do czystej trajektorii równych $0, 0.1, 0.3$ i 1. Każda grupa zawiera 100 000 trajektorii treningowych oraz 10 000 trajektorii testowych.

4.1 Zaszumienie testowe równe zaszumieniu treningowemu

Najpierw wytrenowano po jednym modelu na każdej grupie trajektorii, a następnie modele przetestowano na trajektoriach testowych o tym samym zaszumieniu, co trajektorie treningowe dla danego modelu. Wyniki przedstawiono na rysunku 4.1. Jak można zauważyć, modele trenowane na danych radzą sobie z estymacją wykładnika dyfuzji anomalnej α tym lepiej, im mniejsze jest zaszumienie. Dodatkowo, jedynie gdy szum ma taką samą intensywność jak ruch, modele dają gorsze wyniki niż w rozdziale 3. Gdy szum jest mniejszy, to widać znaczną poprawę jakości przewidywań wykładnika. Wzrost jakości



Rysunek 4.2: Wyniki pomiarów skuteczności estymacji parametru α oraz popełniane błędy przy różnym poziomie zaszumienia trajektorii dla modeli kategorii **A** wytrenowanych w rozdziale 3. Dodatkowo zaznaczono poziomy jakie otrzymywały modele kategorii **A** przy danych testowych grupy A w rozdziale 3.

przewidywań widać nawet w TAMSD, które w rozdziale 3. dawało niepoprawne wyniki. Jego wskazania nadal są niepoprawne, jednak widać zmniejszenie błędów MAE i MSE dla małego zaszumienia. Wyjątkiem jest regresja liniowa, która przy średnim zaszumieniu (0.3) ma delikatnie dokładniejsze predykcje niż przy niskim, a przy braku szumu — daje najgorsze wyniki.

4.2 Model z rozdziału 3

Mmodele grupy **A** z rozdziału 3. przetestowano na danych testowych z bieżącego rozdziału. Wyniki przedstawiono na rysunku 4.2. Wyniki są podobne do wyników z sekcji 4.1 (rysunek 4.1). Mocniej odbiegają tylko wyniki dla regresji liniowej przy braku szumu, które są lepsze niż w przypadku małego i dużego zaszumienia. Najlepsze wyniki osiągane są nadal przy średnim zaszumieniu.

4.3 Skuteczność modelu uczonego na danych bez szumu

model	R2	MAE	MSE
TAMSD	0.000	0.543	0.607
regresja liniowa	0.010	5.107	0.417
drzewo decyzyjne	0.446	0.213	0.353
las losowy	0.484	0.203	0.346
wzmocnienie gradientowe	0.537	0.178	0.326

Tabela 4.1: Wyniki pomiarów skuteczności estymacji parametru α oraz popełniane błędy przez modele trenowane na trajektoriach bez szumu na trajektoriach o mieszanym stopniu zaszumienia.

Model wyszkolony na trajektoriach bez szumu przetestowano na danych testowych o mieszanym szumie (kategoria *A* w rozdziale 3). Wyniki pomiaru skuteczności estymacji przedstawiono w tabeli 4.1. Wydajność estymacji jest znacznie niższa niż w przypadku wszystkich poprzednich testów. Największy spadek widać w przypadku regresji liniowej, gdzie prawie żadne wyniki estymacji nie zgadzają się z prawdziwymi wartościami wykładnika α.

4.4 Wpływ szumu

Współczynniki determinacji z części 4.1-4.3 oraz wyniki z testowania modeli **A** na danych testowych *A* w rozdziale 3. zestawiono w tabeli 4.2.

W prawie każdym przypadku, najlepszym modelem jest wzmocnienie gradientowe. Wyjątkiem jest model trenowany na trajektoriach o średnim zaszumieniu (0.3). Wtedy las losowy dał delikatnie lepszy wynik, niż wzmocnienie gradientowe. Niepoprawne wyniki dla każdej próby dawało TAMSD.

Zarówno dla modelu trenowanego na trajektoriach o zaszumieniu mieszanym, jak i dla tych trenowanych na zaszumieniu stałym, wyniki były tym lepsze, im mniejsze było zaszumienie trajektorii.

$$(\mathrm{m}/\mathrm{s}_1) \geq (\mathrm{m}/\mathrm{s}_2),\ (\mathrm{s}_1/\mathrm{s}_1) \geq (\mathrm{s}_2/\mathrm{s}_2),\ \mathrm{s}_1 \leq \mathrm{s}_2,\ \mathrm{s}_1, \mathrm{s}_2 \in \{0, 0.1, 0.3, 1\}, \tag{4.1}$$

gdzie zapis (a/b) oznacza poprawność estymacji wykładnika dyfuzji anomalnej α przez model trenowany na danych o zaszumieniu „a” i testowany na danych o zaszumieniu „b”, a „m” oznacza zaszumienie mieszane.

	zaszumienie: dane treningowe / testowe				
model	m/m	0/0	0.1/0.1	0.3/0.3	1/1
TAMSD	0.000	0.000	0.000	0.000	0.000
regresja liniowa	0.455	0.427	0.547	0.560	0.476
drzewo decyzyjne	0.552	0.748	0.744	0.705	0.527
las losowy	0.637	0.775	0.794	0.755	0.579
wzmocnienie gradientowe	0.641	0.780	0.797	0.754	0.584
	m/0	m/0.1	m/0.3	m/1	0/m
TAMSD	0.000	0.000	0.000	0.000	0.000
regresja liniowa	0.480	0.448	0.536	0.446	0.010
drzewo decyzyjne	0.661	0.662	0.637	0.441	0.446
las losowy	0.734	0.750	0.721	0.537	0.484
wzmocnienie gradientowe	0.736	0.752	0.723	0.545	0.537

Tabela 4.2: Współczynniki determinacji dla modeli trenowanych i testowanych na danych o różnych zaszumieniach, gdzie „m" oznacza, że zaszumienie było losowo wybrane z $\{0, 0.1, 0.3, 1\}$, a liczba oznacza stosunek szumu do czystego ruchu.

Wyjątkiem jest regresja liniowa, której estymacja była najlepsza dla średniego zaszumienia.

Dla każdego testu na stałych zaszumieniach, modele uczone na stałych zaszumieniach dały lepsze wyniki niż modele uczone na mieszanych zaszumieniach. Więc modele radzą sobie najlepiej z trajektoriami przypominającymi te, na których zostały wytrenowane.

$$(\mathrm{m/s}) \leq (\mathrm{s/s}),\ \mathrm{s} \in \{0, 0.1, 0.3, 1\},$$

w zgodzie z zapisem z równania (4.1).

Modele, które trenowano na danych bez szumu, dają gorsze wyniki w rozpoznawaniu danych z szumem (0/m) niż modele trenowane na danych z szumem (m/m).

$$(0/\mathrm{m}) \leq (\mathrm{m/m}), \tag{4.2}$$

w zgodzie z zapisem z równania (4.1).

Dla zaszumień słabych i średnich modele, które trenowano na danych z szumem mieszanym, dają lepsze wyniki przy testowaniu na danych z szumem stałym niż szumem mieszanym, natomiast gorsze wyniki — dla silnego zaszumienia.

$$(\mathrm{m}/1) \leq (\mathrm{m/m}) \leq (\mathrm{m/s}),\ \mathrm{s} \in \{0, 0.1, 0.3\}, \tag{4.3}$$

w zgodzie z zapisem z równania (4.1).

Ostatecznie, zależności pomiędzy skutecznościami testowanych modeli można zapisać jako:

$$\begin{cases} (0/\mathrm{m}) \leq (\mathrm{m/m}) \leq (\mathrm{m/s}) \leq (\mathrm{s/s}), & \text{gdy } s \in \{0, 0.1, 0.3\}, \\ (\mathrm{m/s}) \leq (\mathrm{s/s}) \leq (\mathrm{m/m}), & \text{gdy } s = 1, \end{cases}$$

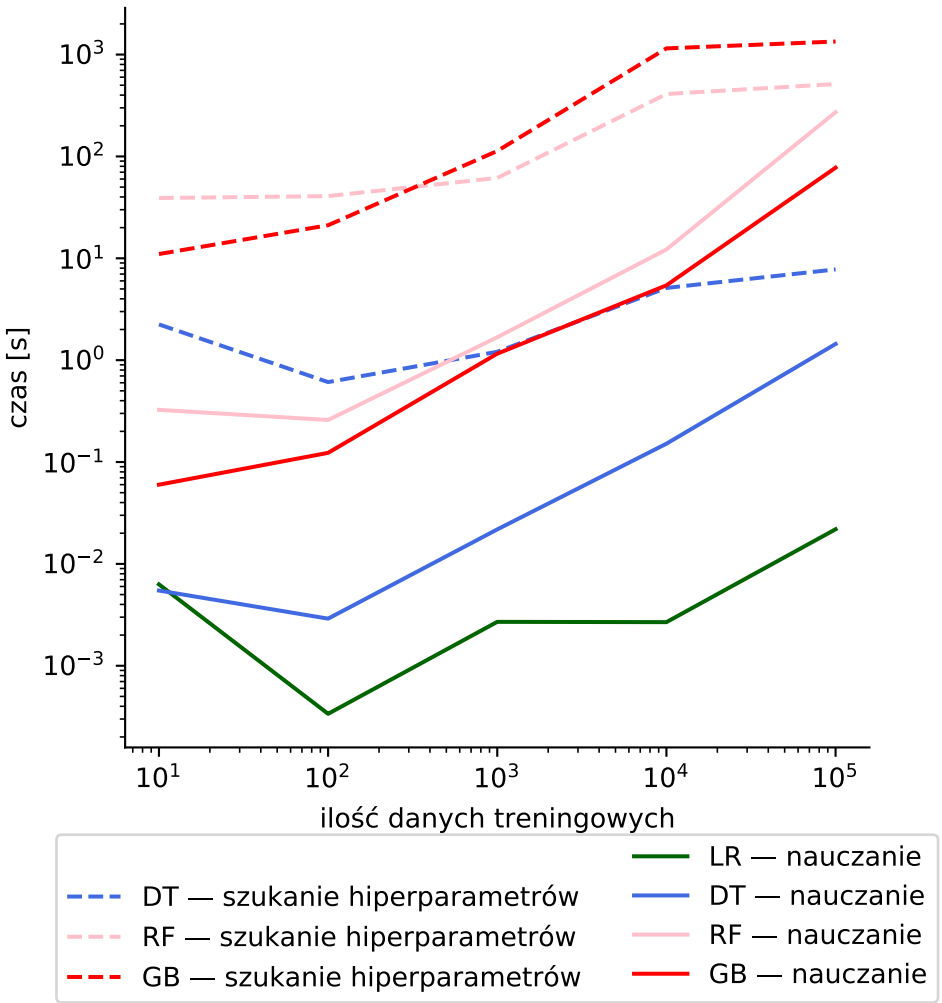
w zgodzie z zapisem z równania (4.1).

Rozdział 5

Wpływ liczby trajektorii

Jako ostatni zbadano wpływ liczby danych treningowych na szybkość nauczania i szukania hiperparametrów oraz skuteczność estymacji różnych metod uczenia maszynowego.

Dla każdego z algorytmów uczenia maszynowego wygenerowano dane treningowe o długości $T = 100$ tą samą metodą, którą generowano dane w rozdziale 3 (dane A). Następnie na różnej liczbie trajektorii treningowych (10^i, $i = 1, ..., 5$) wyszkolono niezależne modele.



Rysunek 5.1: Czasy szukania hiperparametrów i trenowania modeli dla różnej liczby danych treningowych. Oznaczenia: LR — regresja liniowa, DT — drzewo decyzyjne, RF — las losowy, GB — wzmocnienie gradientowe.

Wyniki pomiarów czasu szukania hiperparametrów i nauki modeli dla różnej liczby danych treningowych przedstawiono na wykresie 5.1.

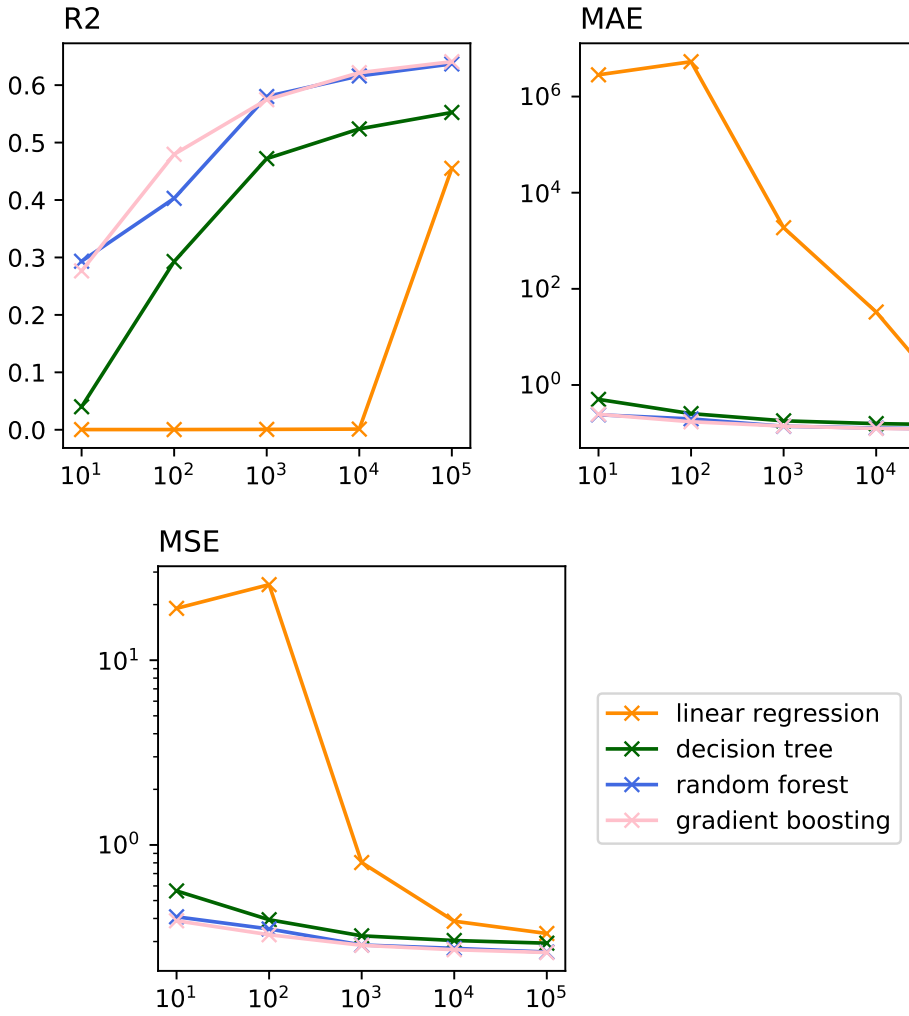
W każdym przypadku wraz ze wzrostem liczby badanych trajektorii, wzrasta czas szukania hiperparametrów. Wyjątkiem od tej reguły są czasy dla 10 trajektorii. Spowodowane jest to czynnościami, które wykonuje program przed rozpoczęciem właściwej pracy.

Dla każdego z algorytmów czas nauczania (linia ciągła na rysunku 5.1) jest proporcjonalny do liczby trajektorii treningowych. Czas szukania hiperparametrów (linia przerywana na rysunku 5.1) natomiast, ma charakter krzywej logistycznej.

Dla małej liczby danych treningowych wzmocnienie gradientowe jest szybsze niż las losowy, co zmienia się dla większej liczby danych (około 1000).

Wytrenowane modele przetestowano na danych testowych z grupy A z rozdziału 3. Współczynnik determinacji oraz błąd średni całkowity i średnio-kwadratowy przedstawiono na rysunku 5.2.

Dla każdego z przetestowanych modeli skuteczność estymacji wykładnika dyfuzji rośnie wraz ze zwiększaniem się liczby badanych trajektorii. Wzrost dokładności najwyraźniej widać w przypadku regresji liniowej.



Rysunek 5.2: Wyniki pomiarów skuteczności estymacji parametru α oraz popełniane błędy przy różnej liczbie danych treningowych typu A z rozdziału 3. Współczynnik determinacji R2 narysowano w skali liniowej, natomiast średni błąd całkowity i błąd średnio-kwadratowy — w sklali logarytmicznej.

Rozdział 6

Podsumowanie i wnioski.

Porównując do siebie wyniki z rozdziałów 2 i 3 łatwo zauważyć, że wprowadzenie dodatkowych rodzajów trajektorii do modelu mocno utrudniło estymację wykładnika dyfuzji α. Współczynniki determinacji dla każdego z testowanych modeli były niższe w przypadku danych pochodzących z wielu procesów. Analogicznie, w tym samym przypadku, błędy MAE i MSE były wyższe. Najlepszy współczynnik determinacji uzyskany w przypadku samego ułamkowego ruchu Browna był równy 0,898 (bardzo dobre dopasowanie), natomiast w przypadku wszystkich ruchów — 0,641 (zadowalające dopasowanie). Pogorszenie to świadczy o tym, że **wraz ze wzrostem skomplikowania danych, zmniejsza się skuteczność estymacji**. Największe pogorszenie estymacji nastąpiło w przypadku TAMSD. Przy samym FBM najwyższy współczynnik determinacji wynosił 0,861 (dobre dopasowanie), a po wprowadzeniu dodatkowych ruchów, był niemal równy 0 (błędne dopasowanie).

Zarówno w rozdziale 2, jak i 3 wszystkie modele radziły sobie lepiej z estymacją trajektorii o długości $T = 100$ kroków (A), niż tych o długości $T = 20$ kroków (B). Nawet modele, które były trenowane na krótszych trajektoriach (**B**), były bardziej skuteczne przy estymacji wykładnika α dla trajektorii długich, niż krótkich. Wynika z tego, że **wydajność modeli rośnie wraz ze wzrostem długości trajektorii testowych**.

W przypadkach zawartych w obu rozdziałach, jakość estymacji dla każdej grupy danych testowych (A, B, C), jest największa dla modeli, które były uczone na danych treningowych tej samej grupy. Oznacza to, że do estymacji **najlepiej używać modelu, który był trenowany na danych testowych podobnych do danych wejściowych**.

Dla danych testowych o losowej długości (C), zarówno dla FBM, jak i wszystkich rodzajów ruchu, modele trenowane na trajektoriach o długości $T = 100$ kroków (**A**), lepiej estymowały wykładnik dyfuzji α, niż modele uczone na trajektoriach uczonych na trajektoriach o długości $T = 20$ kroków. Stąd **wydajność modeli rośnie wraz ze wzrostem długości trajektorii treningowych**.

Z równania (4.1) wynika, że **wraz ze wzrostem zaszumienia trajektorii, spada skuteczność estymacji** wykładnika dyfuzji α. Ponadto z równania (4.3) mamy, że modele szkolone na danych treningowych o mieszanym zaszumieniu, dla danych testowych o niskim i średnim zaszumieniu (0, 0.1, 0.3), wykazują wyższy współczynnik determinacji, niż dla danych treningowym o zaszumieniu mieszanym. Odwrotny wynik otrzymano, gdy zaszumienie było wysokie (równe 1).

Dodatkowo, przy stałym zaszumieniu danych testowych modele, które były uczone na danych o takim samym zaszumieniu co dane testowe, wykazywały lepszą jakość estymacji wykładnika, niż modele, które były uczone na trajektoriach o zmiennym zaszumieniu.

Zgodnie z równaniem 4.2 **do wyznaczania wykładnika dla trajektorii posiadających szum, model powinien być uczony na danych zaszumionych**.

Zgodnie z rysunkiem 5.2, **wydajność estymacji wykładnika wzrasta, gdy rośnie liczba trajektorii treningowych**. Niestety, **rośnie wtedy również czas szukania hiperparametrów i nauczania** w modelach.

Biorąc pod uwagę dane z całej pracy, łatwo stwierdzić, że modele pod względem precyzji estymacji wykładnika dyfuzji, układają się w następującej kolejności (od najlepszego do najgorszego):

1. wzmocnienie gradientowe (najwyższy współczynnik determinacji oraz najniższe błędy MAE i MSE w prawie każdym przypadku),

2. las losowy (skuteczniejszy od wzmocnienia gradientowego przy bardzo małej liczbie próbek [10, 100]),

3. drzewo decyzyjne,

4. regresja liniowa (słabe dopasowanie przy małej liczbie trajektorii)

5. uśrednione po czasie średnie przesunięcie kwadratowe (słabe dopasowanie przy danych zawierających wszystkie modele, najniższe wyniki w każdym rozpatrywanym przypadku).

Najlepszym modelem, pod względem jakości estymowanych danych, jest wzmocnienie gradientowe. Natomiast najgorsze — średnie przesunięcie kwadratowe. Stąd **metody uczenia maszynowego są krokiem w dobrą stronę w analizie trajektorii ze śledzenia pojedynczych cząstek**.

Im bardziej skomplikowany model, tym jego estymacje są dokładniejsze, ale również jego czasy nauki i szukania hiperparametrów są dłuższe. Czas nauki rośnie również, gdy rośnie liczba trajektorii. Wzrost jest liniowy, ale znacznie zwiększa dokładność estymacji.

Aby zmaksymalizować poprawność estymacji wykładnika dyfuzji anomalnej α należy zadbać o jak najmniejsze zaszumienie trajektorii. Jeśli mamy sposób mierzenia zaszumienia, to lepiej jest użyć modelu wytrenowanego na trajektoriach o podobnym zaszumieniu co badane trajektorie, niż takiego, który był uczony na trajektoriach o różnym szumie.

Warto też starać się uzyskać jak najwięcej zmierzonych kroków ruchu cząstki na przykład poprawiając oświetlenie próbki (w rezultacie mniej czasu będzie wymagał jeden odczyt), używając lepszej aparatury.

Jeśli czas nie gra roli lub posiadamy niewiele danych treningowych, to najlepszym modelem jest wzmocnienie gradientowe. Jeśli danych jest bardzo dużo, a zależy nam na minimalizacji czasu nauki modelu, to należy ocenić, czy lepiej zmniejszyć liczbę danych treningowych, a może lepiej użyć mniej skomplikowanego modelu.

Czy uczenie maszynowe to krok w dobrą stronę w analizie trajektorii otrzymanych metodą SPT? Każdy z zaprezentowanych algorytmów uczenia maszynowego przyniósł lepsze efekty niż TAMSD, więc z pewnością uczenie maszynowe jest krokiem w dobrą stronę w analizie SPT. Radzi sobie ono dużo lepiej zarówno ze silnie zaszumionymi trajektoriami, jak i z niewielką długością trajektorii czy niewielką ich liczbą.

Dalszy rozwój techniki pozwoli uzyskiwać trajektorie o większej rozdzielczości czasowej, co mocno poprawi przewidywanie wykładnika dyfuzji anomalnej. a to z kolei pomoże lepiej zrozumieć procesy zachodzące między innymi w komórkach. Możliwe stanie się badanie metodą SPT procesów, które trwają bardzo krótko.

Rozwój informatyki prowadzi do powstawania coraz szybszych komputerów, które są w stanie szybciej analizować większe ilości danych, dzięki czemu czasy nauki modeli stają się nieistotnie małe.

Powstanie też jeszcze więcej skuteczniejszych programów samouczących, które będą w stanie jeszcze dokładniej przewidywać wykładnik dyfuzji. Aktualnie nauka zmierza ku wykorzystaniu w analizie SPT głębokich sieci neuronowych [12].

Bibliografia

[1] Aczel−Sounderpandian. *Complete Business Statistics: How Good Is the Regression?* The McGrawHill Companies, Inc., 2009, pp. 438–443.

[2] Barros, R. C., Winck, A. T., Machado, K. S., Basgalupp, M. P., de Carvalho, A. C., Ruiz, D. D., de Souza, O. N. Automatic design of decision-tree induction algorithms tailored to flexible-receptor docking data. *BMC bioinformatics 13* (2012), 310.

[3] Boltzmann, L. *Vorlesungen über Gastheorie: Abeschmitt: Erdogen.* 1898, pp. 89–92.

[4] Brémaud, P. *Probability Theory and Stochastic Processes: Fractal Brownian Motion.* Springer, Cham, Switzerland, pp. 461–463.

[5] Feder, T., Brust-Mascher, I., Slattery, J., Baird, B., Webb, W. Constrained diffusion or immobile fraction on cell surfaces: a new interpretation. *Biophysical Journal 70*, 6 (1996), 2767 – 2773.

[6] Friedman, J. H. Stochastic gradient boosting.

[7] Granik, N., Weiss, L. E., Nehme, E., Levin, M., Chein, M., Perlson, E., Roichman, Y., Shechtman, Y. Single-particle diffusion characterization by deep learning. *Biophysical Journal 117* (2019), 185–192.

[8] Hansen, A. S., Woringer, M., Grimm, J. B., Lavis, L. D., Tjian, R., Darzacq, X. Robust model-based analysis of single-particle tracking experiments with spot-on. *eLife 7* (jan 2018), e33125.

[9] James, A. P. *Deep Learning Classifiers with Memristive Networks - Theory and Applications*, vol. 14. Springer US, Boston, MA, 2020.

[10] Katz, M. J., George, E. B. Fractals and the analysis of growth paths. bltn mathcal biology. *Bulletin of Mathematical Biology 47*, 273–286.

[11] Kelleher, J. D., Namee, B. M., D'Arcy, A. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.* The MIT Press, Cambridge, Massachusetts, London, England, 2015, pp. 323–396.

[12] Kowalek, P., Loch-Olszewska, H., Szwabiński, J. Classification of diffusion modes in single-particle tracking data: Feature-based versus deep-learning approach. *Phys. Rev. E 100* (Sep 2019), 032410.

[13] Mcbratney, A. Everitt, b.s., 2002. the cambridge dictionary of statistics. 2nd edition. *Geoderma 121* (07 2004).

[14] Metzler, R., Jae-Hyung, J., Cherstvy, A., Barkai, E. *Anomalous diffusion models and their properties: non-stationarity, non-ergodicity, and ageing at the centenary of single particle tracking*, vol. 16. The Royal Society of Chemistry, 2014, pp. 24128–24164.

[15] Metzler, R., Jeon, J.-H., Cherstvy, A. G., Barkai, E. Anomalous diffusion models and their properties: non-stationarity, non-ergodicity, and ageing at the centenary of single particle tracking. *Phys. Chem. Chem. Phys. 16* (2014), 24128–24164.

[16] Munoz-Gil, G., Maciej Lewenstein, Carlo Manzo, G. V. M. A. G.-M. R. M. The anomalous diffusion (AnDi) challenge datasets documentation. *Cond-Mat.Stat-Mech* (2020), [github.com].

[17] Munoz-Gil, G., Maciej Lewenstein, Carlo Manzo, G. V. M. A. G.-M. R. M. The anomalous diffusion (AnDi) challenge overview. *Cond-Mat.Stat-Mech* (2020), [andi–challenge.org].

[18] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[19] Pollard, M., Thomas, D., William, D., Earnshaw, C. *Cell Biology: Biophysical Principles*. Saunders Elsevier, Philadelphia, PA, 2007, pp. 57–68.

[20] Qian, H., Sheetz, M., Elson, E. Single particle tracking. analysis of diffusion and flow in two-dimensional systems. *Biophysical Journal 60* (1991), 910–921.

[21] Ralf Metzler, Jae-Hyung Jeon, A. G. C., Barkai, E. Tanomalous diffusion models and their properties: non-stationarity, non-ergodicity, and ageing at the centenary of single particle tracking. *Phys. Chem. Chem. Phys.* (2014), 24128–24164.

[22] Randall, G. Random walks and diffusion: Levy flights ($\sigma = \infty$). [MIT OpenCourseWare].

[23] Rastogi, V. B. *Modern Biology: Reproduction in Flowering Plant*, vol. 2. Pitambar Publishing Company PVT. LTD., New Delhi, India, 1997, pp. III–3–45.

[24] Sammut, C., Webb, G. I., Eds. *Encyclopedia of Machine Learning: Mean Absolute Error*. Springer US, Boston, MA, 2010, pp. 652–652.

[25] Sammut, C., Webb, G. I., Eds. *Encyclopedia of Machine Learning: Mean Squared Error*. Springer US, Boston, MA, 2010, pp. 653–653.

[26] Saxton, M. J. Lateral diffusion in an archipelago effects of impermeable patches on diffusion in a cell membrane. *Biophysical Journal 39*, 165–173.

[27] Schurma, G. Brownian motion: The scaled random walk. [appliedbusinesseconomics.com].

[28] Senthilnathan, K., Shanmugam, B., Goyal, D., Annapoorani, I., Samikannu, R. *Deep Learning Applications and Intelligent Decision Making in Engineering: Liver Disease Detection Using Grey Wolf Optimization and Random Forest Classification.* IGI Global, PA, USA, 2020, pp. 130–160.

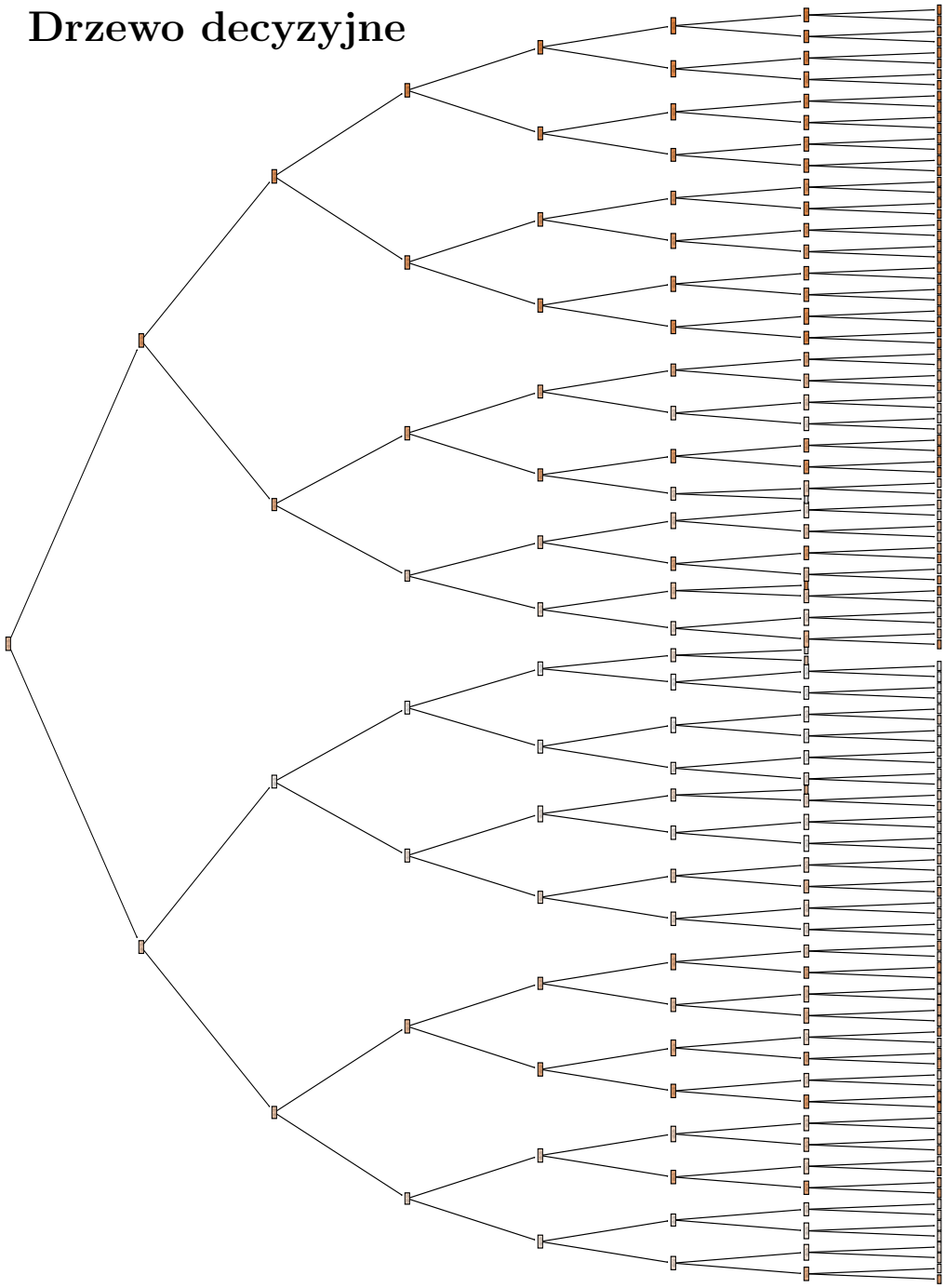
[29] Shraddha. Annealing of metals: 3 main stages | metallurgy. [engineeringenotes.com].

[30] StatQuest with Josh Starmer. Gradient boost part 1: Regression main ideas. [video], [youtube.com].

[31] Webb, G. I. *Encyclopedia of Machine Learning - Overfitting.* Springer US, Boston, MA, 2010, pp. 744–744.

Dodatki

A Drzewo decyzyjne



Rysunek A.1: Pełne drzewo decyzyjne o głębokości maksymalnej $m_{\max} = 10$. Widać, że niektóre węzły nie zostały podzielone.

B Dowody dla TAMSD

Dowód B.1.
Twierdzenie
Estymator parametru $\widehat{\alpha}$ otrzymany metodą najmniejszych kwadratów wynosi:

$$\widehat{\alpha} = \frac{\sum\limits_{i=1}^{n} \ln(i\Delta t) \cdot \ln\Big(\rho(i\Delta t)\Big) - \ln(4D) \sum\limits_{i=1}^{n} \ln(i\Delta t)}{\sum\limits_{i=1}^{n} \Big(\ln(i\Delta t)\Big)^2}.$$

Dowód
Metoda najmniejszych kwadratów dla równania $y(x) = ax + b$ polega na znalezieniu parametrów, które spełniają równanie:

$$\sum_{i=1}^{n}(y_i - ax - b)^2 = \min_a \sum_{i=1}^{n}(y_i - ax_i - b)^2 =$$
$$= \min_a \Bigg(\underbrace{\sum_{i=1}^{n} y_i^2 + a^2 \sum_{i=1}^{n} x_i^2 + nb^2 + 2ab \sum_{i=1}^{n} x_i - 2a \sum_{i=1}^{n} x_i y_i - 2b \sum_{i=1}^{n} y_i}_{F(a)} \Bigg).$$

Aby znaleźć minimum funkcji $F(a)$ należy sprawdzić warunek istnienia ekstremum:

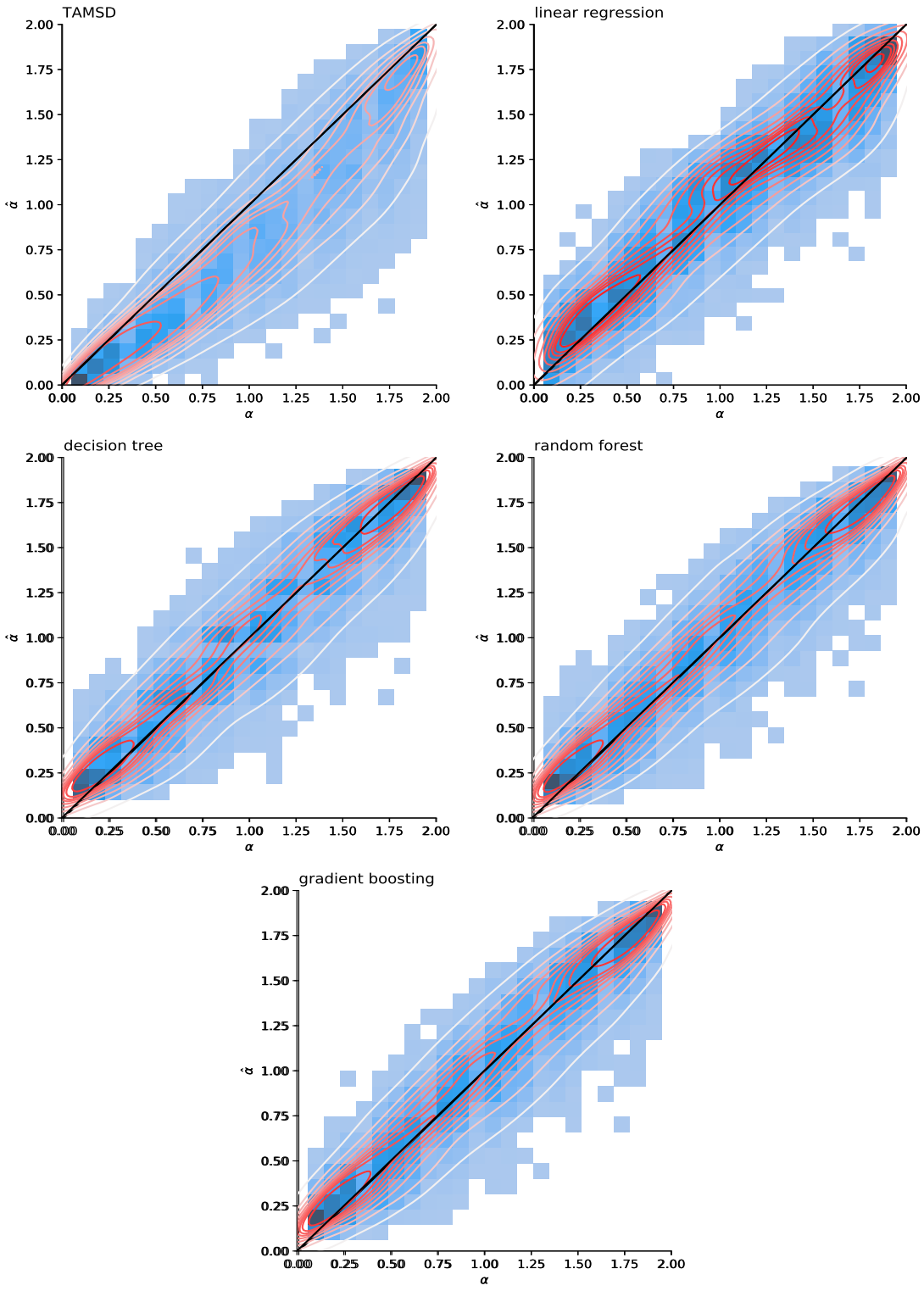
$$\frac{\partial F}{\partial a} = 2a \sum_{i=1}^{n} x_i^2 + 2b \sum_{i=1}^{n} x_i - 2 \sum_{i=1}^{n} x_i y_i = 0.$$

Rozwiązaniem powyższego równania jest:

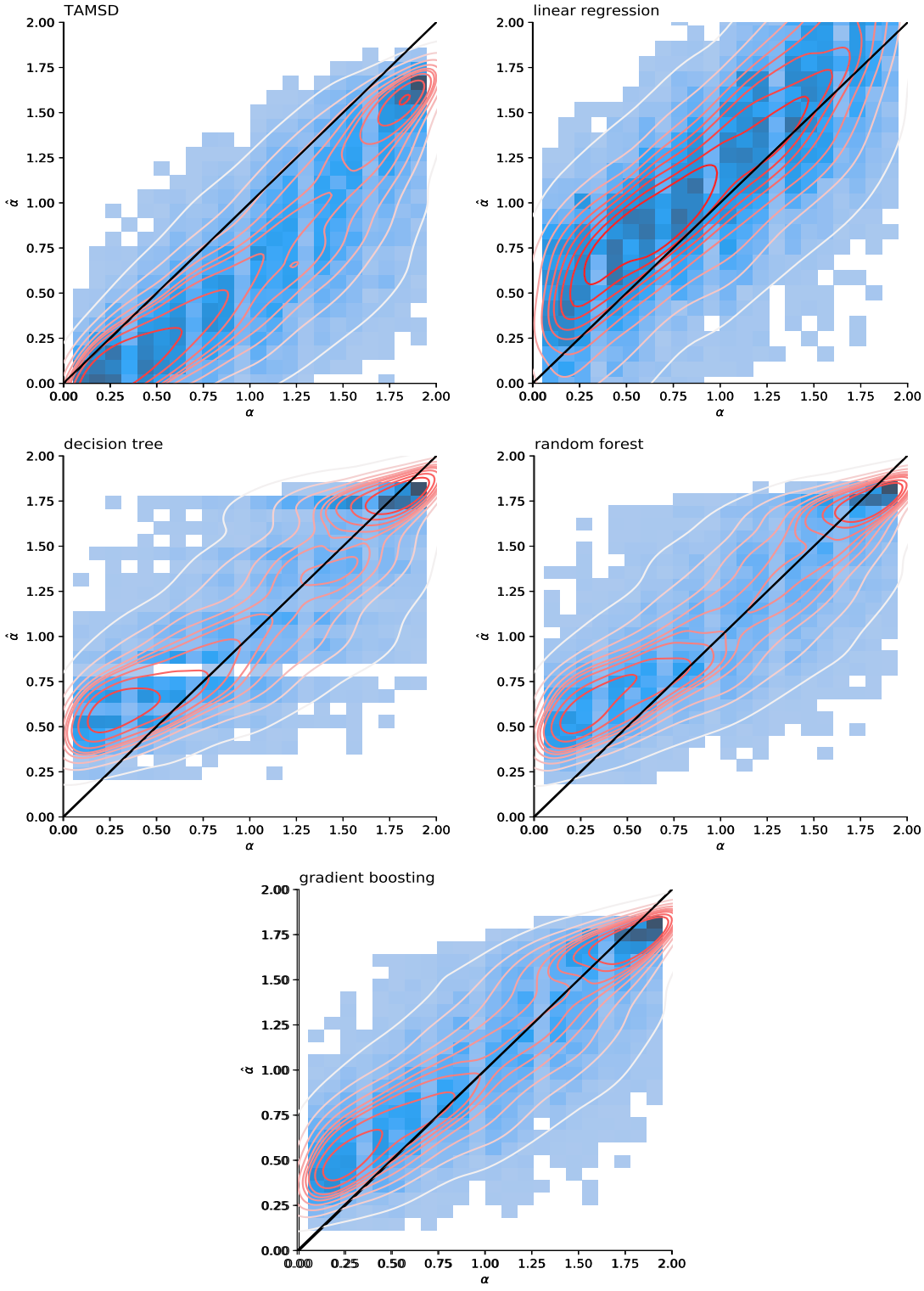
$$a = \frac{\sum\limits_{i=1}^{n} x_i y_i - b \sum\limits_{i=1}^{n} x_i}{\sum\limits_{i=1}^{n} x_i^2}.$$

Po podstawieniu zgodnie z równaniem (1.7) *— $y_i = \ln\Big(\rho(i\Delta t)\Big)$, $x_i = \ln(i\Delta t)$, $a = \widehat{\alpha}$, $b = \ln(4D)$ otrzymujemy równanie* (1.8)*. cnd.*

C Wyniki przewidywań dla ułamkowego ruchu Browna.



Rysunek C.1: Wyniki estymacji wykładnika dyfuzji anomalnej modelu wyszkolonego na trajektoriach z kategorii **A** różnymi modelami. do testowania posłużyły trajektorie testowe z grupy **A**. na osi x zaznaczono prawdziwą wartość parametru α, a na osi y — estymator $\hat{\alpha}$. Im ciemniejszy jest kolor, tym więcej wyników zawarto w jego przedziale. na czerwono zaznaczono poziomice ilości poprawnie przewidzianych wyników.



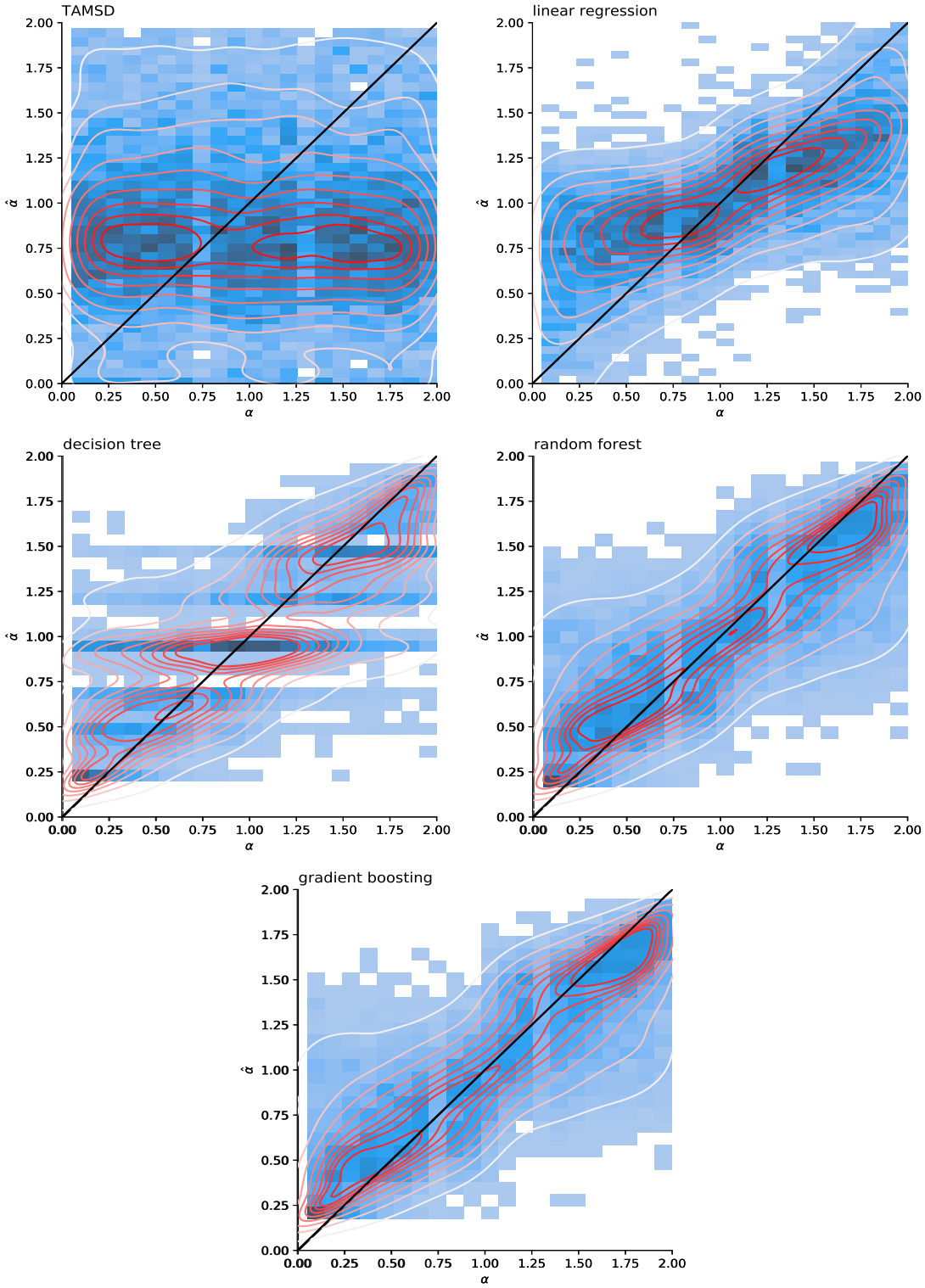
Rysunek C.2: Wyniki estymacji wykładnika dyfuzji anomalnej modelu wyszkolonego na trajektoriach z kategorii **A** różnymi modelami. do testowania posłużyły trajektorie testowe z grupy **B**. na osi x zaznaczono prawdziwą wartość parametru α, a na osi y — estymator $\hat{\alpha}$. Im ciemniejszy jest kolor, tym więcej wyników zawarto w jego przedziale. na czerwono zaznaczono poziomice ilości poprawnie przewidzianych wyników.

D Dane techniczne komputera.

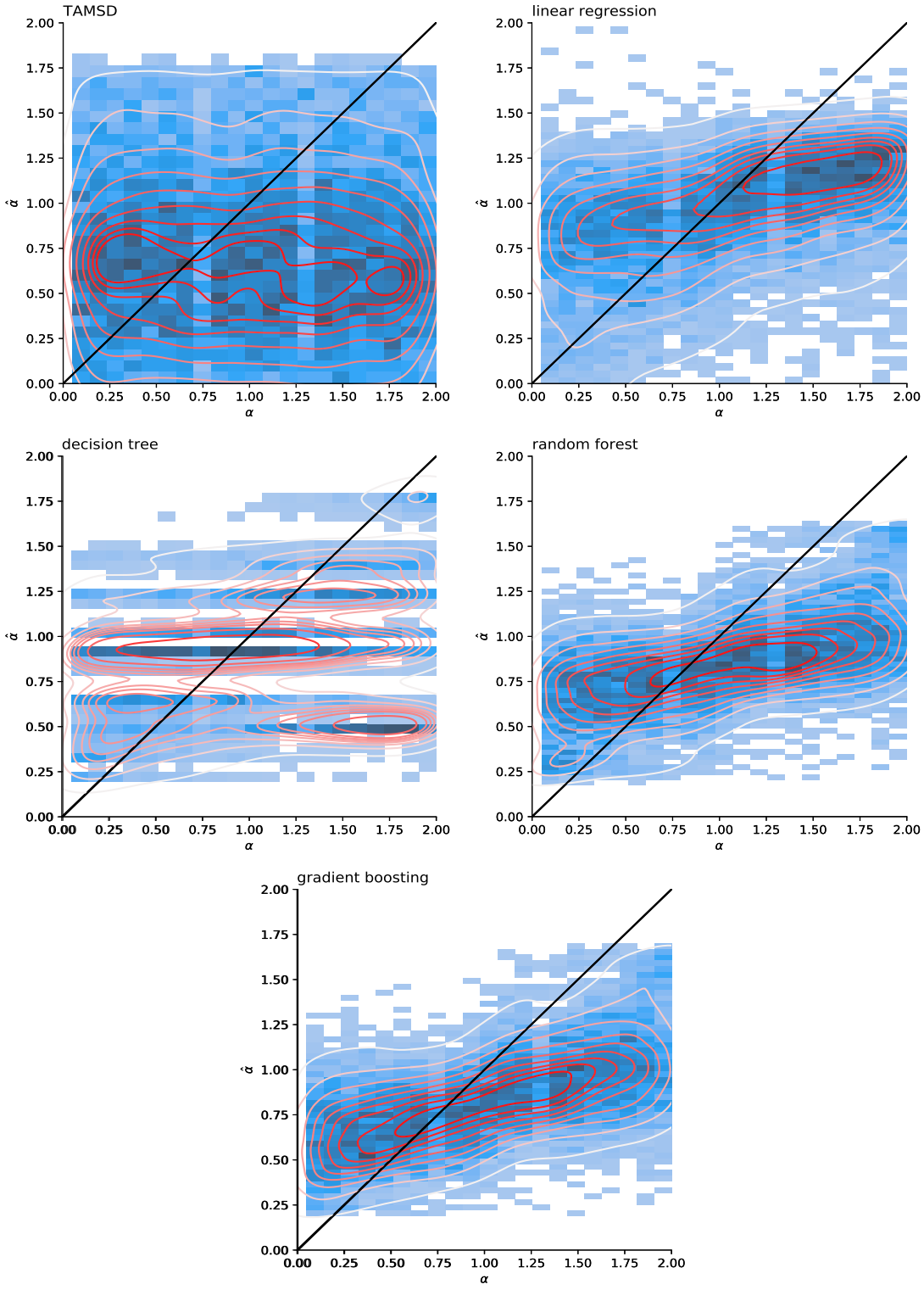
Model	
Rodzaj	laptop
Model	Lenovo Ideapad Z510
System operacyjny	Linux Mint 20.0.4
Procesor	
Model	Intel Core i3-4000M
Taktowanie	2400 MHz
Liczba rdzeni	2
Liczba wątków	4
Karta graficzna	
Model	NVIDIA GeForce GT 740M
Pamięć	2 GB
Pamięć RAM	
Pojemność	8 GB

Tabela D.1: Dane techniczne komputera, na którym przeprowadzano badania. Przy każdych obliczeniach procesor pracował na 3 z 4 wątków.

E Wyniki przewidywań dla wszystkich ruchów.



Rysunek E.1: Wyniki estymacji wykładnika dyfuzji anomalnej modelu wyszkolonego na trajektoriach z kategorii **A** różnymi modelami. do testowania posłużyły trajektorie testowe z grupy **A**. na osi x zaznaczono prawdziwą wartość parametru α, a na osi y — estymator $\hat{\alpha}$. Im ciemniejszy jest kolor, tym więcej wyników zawarto w jego przedziale. na czerwono zaznaczono poziomice ilości poprawnie przewidzianych wyników.



Rysunek E.2: Wyniki estymacji wykładnika dyfuzji anomalnej modelu wyszkolonego na trajektoriach z kategorii **A** różnymi modelami. do testowania posłużyły trajektorie testowe z grupy **B**. na osi x zaznaczono prawdziwą wartość parametru α, a na osi y — estymator $\hat{\alpha}$. Im ciemniejszy jest kolor, tym więcej wyników zawarto w jego przedziale. na czerwono zaznaczono poziomice ilości poprawnie przewidzianych wyników.