НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
Фізико-математичний факультет
Кафедра математичного аналізу та теорії ймовірностей

УДК 519.234

До захисту допущено
завідувач кафедри
_____ Клесов О. І.
«____» _____ 2025 р.

Магістерська робота

на здобуття ступеня магістр
за освітньо-науковою програмою «Страхова та фінансова математика»
спеціальності 111 «Математика»
на тему: «Глибина Тьюкі та алгоритмічні методи її обчислення»

Виконав:
студент II курсу, групи ОМ-31мн
Панаско Віталій Євгенович

Керівник:
доцент, к.ф-м.н. Ільєнко Андрій Борисович

Рецензент:
доцент, к.ф.-м.н. Подколзін Гліб Борисович

Засвідчую, що у цій дипломній роботі
немає запозичень з праць інших авторів
без відповідних посилань.
Студент: Панаско Віталій Євгенович

Київ – 2025 року

Зміст

1 Теоретичні основи глибини Тьюкі 4
1.1 Визначення та позначення . . . 4
1.2 Властивості . . . 5
1.3 Аналітичні вирази для деяких розподілів . . . 6
1.4 Висновки . . . 10

2 Обчислення глибини для ймовірнісних розподілів 12
2.1 Узагальнення глибини Тьюкі в рівномірно розподіленій одиничній кулі в $\mathbb{R}^d$ 12
2.1.1 Глибина Тьюкі для рівномірної міри на B^d . . . 12
2.1.2 Геометрія півпросторів у кулі . . . 12
2.1.3 Інтегральна форма для об'єму сегмента кулі . . . 13
2.1.4 Остаточна формула для глибини . . . 14
2.1.5 Приклади . . . 16
2.2 Глибина Тьюкі для α-стійких розподілів . . . 16
2.2.1 Постановка задачі для двовимірного випадку . . . 16
2.2.2 Аналіз порогу спрямованості . . . 17
2.2.3 Інтегральне представлення для формули глибини . . . 18
2.2.4 Візуалізація контурів глибини . . . 19
2.2.5 Узагальнення для R^d . . . 20
2.3 Глибина Тьюкі для двовимірного експоненційного розподілу . . . 21
2.3.1 Обчислення $I(k)$ . . . 21
2.3.2 Асимптотичний аналіз оптимального нахилу . . . 22
2.4 Глибина Тьюкі для двовимірного t-розподілу . . . 23

3 Рандомізована глибина Тьюкі 25
3.1 Основні визначення . . . 25
3.2 Збіжність за мірою Гаусса P . . . 26
3.2.1 Вибірка точок за глибиною . . . 26
3.2.2 Дизайн експерименту . . . 26
3.2.3 Адаптивний вибір випадкових напрямків . . . 28
3.2.4 Результати . . . 28
3.2.5 Короткий висновок . . . 30
3.3 Збіжність на гауссівських вибірках . . . 30

4 Висновки 34

Вступ

Глибина Тьюкі є фундаментальним поняттям у багатовимірній статистиці, що дозволяє вимірювати центральність точки в багатовимірному просторі даних. Завдяки своїй стійкості до викидів, вона є потужним інструментом у статистичному аналізі.

Мета та завдання. Мета роботи — поглибити розуміння глибини Тьюкі та розробити практичні інструменти для її обчислення в робастній статистиці й аналізі даних.

Об'єкт і предмет дослідження. Ця дипломна робота присвячена дослідженню теоретичних і обчислювальних аспектів глибини Тьюкі, зосереджуючись на її властивостях, явних формах для певних розподілів та алгоритмічних методах її обчислення.

Структура роботи. Робота складається з трьох розділів: перший аналізує теоретичні основи та явні форми глибини, другий розглядає аналітичні вирази та чисельні апроксимації, а третій присвячений рандомізованому методу апроксимації.

Розділ 1

Теоретичні основи глибини Тьюкі

Глибина напівпростору, запропонована Тьюкі в 1975 році, є потужним інструментом багатовимірної статистики (6), який визначає, наскільки центрально розташована точка в багатовимірному просторі. Глибина пропонує надійну альтернативу традиційним мірам центральності, таким як середнє або медіана. Ця секція надає всебічну теоретичну основу для глибини напівпростору, охоплюючи її визначення, ключові властивості та аналітичні вирази для конкретних розподілів популяції: двовимірний гаусівський, двовимірний розподіл Коші та рівномірний розподіл на квадраті. Ми розширюємо ці результати на вищі розмірності, де це можливо, закладаючи основу для вивчення розподілів, таких як гіперкуб у $\mathbb{R}^d$.

1.1 Визначення та позначення

Визначення 1 (Глибина напівпростору для емпіричного розподілу). Для скінченного набору даних $X_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ і точки $\boldsymbol{\theta} \in \mathbb{R}^d$ глибина напівпростору визначається як

$$\mathrm{HD}(\boldsymbol{\theta}, X_n) = \frac{1}{n} \min_{\|\mathbf{u}\|=1} \#\big(H_{\boldsymbol{\theta},\mathbf{u}} \cap X_n\big),$$

де $H_{\boldsymbol{\theta},\mathbf{u}} = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{u}^\mathsf{T}\mathbf{x} \geq \mathbf{u}^\mathsf{T}\boldsymbol{\theta}\}$ — замкнений напівпростір з одиничним нормальним вектором $\mathbf{u}$. Позначення $\#(A)$ означає потужність множини A.

Інакше кажучи, найменша міра півпросторів, що містять задану точку $\boldsymbol{\theta}$. Точки за межами опуклої оболонки X_n мають нульову глибину; чим центральнішою є точка всередині хмари даних, тим вона глибша.

Визначення 2 (Глибина напівпростору для теоретичного розподілу). Нехай P — ймовірнісний розподіл на $\mathbb{R}^d$ (тобто $P(\mathbb{R}^d) = 1$). Для точки $\boldsymbol{\theta} \in \mathbb{R}^d$ глибина напівпростору визначається як

$$\mathrm{HD}_P(\boldsymbol{\theta}) = \inf_{\|\mathbf{u}\|=1} P\big(H_{\boldsymbol{\theta},\mathbf{u}}\big).$$

Оскільки P може бути атомарною, inf не завжди досягається; проте для неперервних розподілів він збігається з мінімумом. Значення $\mathrm{HD}_P(\boldsymbol{\theta})$ належить інтервалу $[0, 1]$.

Визначення 3 (Множини рівня та медіана Тьюкі). Для $\alpha \geq 0$ визначимо множину рівня

$$D_\alpha = \big\{\boldsymbol{\theta} \in \mathbb{R}^d : \mathrm{HD}(\boldsymbol{\theta}) \geq \alpha\big\}.$$

Медіаною Тьюкі називається точка $\boldsymbol{\theta}^*$, що максимізує функцію глибини: $\mathrm{HD}(\boldsymbol{\theta}^*) = \sup_{\boldsymbol{\theta}} \mathrm{HD}(\boldsymbol{\theta})$. Якщо множина D_{α^*} при $\alpha^* = \sup_{\boldsymbol{\theta}} \mathrm{HD}(\boldsymbol{\theta})$ не є зв'язною, за медіану беруть центр ваги D_{α^*}.

Множини D_α є вкладеними ($\alpha' > \alpha \implies D_{\alpha'} \subseteq D_\alpha$) та допомагають візуалізувати контури розподілу. Медіана Тьюкі забезпечує робастну оцінку положення, покладаючись на геометрію напівпросторів, а не на моменти.

1.2 Властивості

Глибина напівпростору має низку математичних властивостей, що підсилюють її практичну цінність у статистичному аналізі. У цьому пункті властивості, отримані з першоджерела (8), подано з докладними поясненнями і, де доречно, зі стислими ескізами доведень, які прояснюють їхню сутність.

Твердження 1 (Афінна інваріантність). Глибина напівпростору є афінно інваріантною. Для афінного перетворення $g(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$, де $A \in \mathbb{R}^{d\times d}$ з $\det(A) \neq 0$ та $\mathbf{b} \in \mathbb{R}^d$, і міри μ, маємо

$$\mathrm{HD}(g(\boldsymbol{\theta}), \mu_g) = \mathrm{HD}(\boldsymbol{\theta}, \mu),$$

де $\mu_g(B) = \mu(g^{-1}(B))$ для будь-якої борелівської множини $B \subset \mathbb{R}^d$.

Короткий виклад доведення. Розглянемо напівпростір $H_{g(\boldsymbol{\theta}),\mathbf{u}} = \{\mathbf{y} : \mathbf{u}^\top \mathbf{y} \geq \mathbf{u}^\top (A\boldsymbol{\theta} + \mathbf{b})\}$. Під час перетворення g це відповідає напівпростору в початковому просторі з нормаллю $\mathbf{v} = A^\top \mathbf{u}/\|A^\top \mathbf{u}\|$. Міра $\mu_g(H_{g(\boldsymbol{\theta}),\mathbf{u}}) = \mu(g^{-1}(H_{g(\boldsymbol{\theta}),\mathbf{u}}))$ дорівнює мірі відповідного напівпростору в початковому просторі, зберігаючи інфімум. □

Афінна інваріантність гарантує, що глибина лишається незмінною за масштабування, обертання та паралельного перенесення, тож вона коректно працює з розподілами з довільним рівнем коваріації.

Твердження 2 (Квазіувігнутість). Для будь-якої позитивної міри μ, функція глибини HD_μ є квазіувігнутою, тобто для $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$ та $0 \leq \gamma \leq 1$,

$$\mathrm{HD}_\mu(\gamma\boldsymbol{\theta}_1 + (1-\gamma)\boldsymbol{\theta}_2) \geq \min\{\mathrm{HD}_\mu(\boldsymbol{\theta}_1), \mathrm{HD}_\mu(\boldsymbol{\theta}_2)\}.$$

Наслідок 1. Для будь-якого $\alpha \geq 0$, область глибини D_α є опуклою.

Квазіувігнутість означає, що функція глибини не зменшується занадто швидко вздовж відрізків прямих, що гарантує опуклість областей глибини та полегшує визначення центральних регіонів.

Твердження 3 (Верхня напівнеперервність). Для будь-якого ймовірнісного розподілу P, функція глибини $\mathrm{HD}(\boldsymbol{\theta})$ є верхньою напівнеперервною. Якщо P має щільність відносно міри Лебега, то $\mathrm{HD}(\boldsymbol{\theta})$ також є нижньою напівнеперервною, а отже, неперервною.

Верхня напівнеперервність гарантує, що малі зміни в $\boldsymbol{\theta}$ не викликають різких зростань HD, тоді як неперервність для розподілів із щільністю спрощує аналітичні та обчислювальні підходи.

Твердження 4 (Обмеженість). Для будь-якого ймовірнісного розподілу P та $\alpha > 0$, множина D_α є обмеженою та компактною.

Короткий виклад доведення. Оскільки $P(\mathbb{R}^d) = 1$, існує куля $B_m = \{\mathbf{x} : \|\mathbf{x}\| \leq m\}$ з $P(B_m) > 1 - \alpha$. Для $\boldsymbol{\theta} \notin B_m$, розділяюча гіперплощина визначає напівпростір $H_{\boldsymbol{\theta},\mathbf{u}}$ з $P(H_{\boldsymbol{\theta},\mathbf{u}}) < \alpha$, тому $\boldsymbol{\theta} \notin D_\alpha$. Отже, $D_\alpha \subset B_m$, а компактність випливає із замкненості (завдяки верхній напівнеперервності). □

Твердження 5 (Представлення множин рівня D_α). Для будь-якого ймовірнісного розподілу P та $\alpha > 0$,

$$D_\alpha = \bigcap \{H : H \text{ — замкнений напівпростір, для якого } P(H^c) < \alpha\}.$$

Еквівалентно,

$$D_\alpha = \bigcap \{H : H \text{ — замкнений напівпростір, для якого } P(H) > 1 - \alpha\}.$$

Це представлення підкреслює геометричну природу множин рівня глибини як перетинів напівпросторів, що узгоджується з інтуїтивним поняттям центральності.

Твердження 6 (Існування максимальної глибини). Для будь-якої ймовірнісної міри P на $\mathbb{R}^d$, існує принаймні одна точка $\boldsymbol{\theta}^*$, для якої

$$\mathrm{HD}(\boldsymbol{\theta}^*) = \sup_{\boldsymbol{\theta}} \mathrm{HD}(\boldsymbol{\theta}).$$

Короткий виклад доведення. Нехай $\alpha^* = \sup_{\boldsymbol{\theta}} \mathrm{HD}(\boldsymbol{\theta}) \leq 1$. Для $\alpha < \alpha^*$, D_α є непорожньою, замкненою та обмеженою. Перетин $D = \bigcap_{0<\alpha<\alpha^*} D_\alpha$ є непорожнім, і будь-яка $\boldsymbol{\theta} \in D$ має $\mathrm{HD}(\boldsymbol{\theta}) \geq \alpha^*$, отже, $\mathrm{HD}(\boldsymbol{\theta}) = \alpha^*$. □

Твердження 7 (Теорема про базис променів). Якщо існує точка $\boldsymbol{\theta}^*$ та набір одиничних векторів $J = \{\mathbf{u}_1, \ldots\}$, таких що

$$P(H_{\boldsymbol{\theta}^*, \mathbf{u}_j}) = \mathrm{HD}(\boldsymbol{\theta}^*) \quad \forall j, \quad \text{і} \quad \bigcup_{j \in J} H_{\boldsymbol{\theta}^*, \mathbf{u}_j} = \mathbb{R}^d,$$

то $\mathrm{HD}(\boldsymbol{\theta}^*) = \max_{\boldsymbol{\theta}} \mathrm{HD}(\boldsymbol{\theta})$. Набір J можна вибрати так, щоб він містив не більше $d + 1$ елементів.

Ця теорема встановлює достатню умову для того, щоб точка була медіаною Тьюкі, використовуючи геометричне покриття $\mathbb{R}^d$ напівпросторами.

Твердження 8 (Нижня межа максимальної глибини). Для будь-якої ймовірнісної міри P на $\mathbb{R}^d$,

$$\max_{\boldsymbol{\theta}} \mathrm{HD}(\boldsymbol{\theta}) \geq \frac{1}{d+1}.$$

Короткий виклад доведення. За теоремою Хеллі (3), для будь-якого $\alpha^* = \max_{\boldsymbol{\theta}} \mathrm{HD}(\boldsymbol{\theta})$, існують $d + 1$ напівпростори H_j з $P(H_j^c) < \alpha^* + \epsilon$, такі що їхній перетин порожній. Отже, $1 = P(\mathbb{R}^d) \leq \sum_{j=1}^{d+1} P(H_j^c) < (d+1)(\alpha^* + \epsilon)$, що означає $\alpha^* \geq 1/(d+1)$ при $\epsilon \to 0$. □

В сукупності ці властивості виділяють глибину напівпростору як надійну, геометрично інтуїтивну та математично обґрунтовану міру багатовимірної центральності. Нижня оцінка максимальної глибини гарантує, зокрема, що навіть у високих розмірностях найглибша точка охоплює істотну частку маси розподілу.

1.3 Аналітичні вирази для деяких розподілів

Далі наведемо приклад функцій глибину напівпростору для трьох двовимірних розподілів: стандартного двовимірного гаусівського, двовимірного розподілу Коші та рівномірного розподілу на квадраті (4-кутнику) (8). Узагальнення на $\mathbb{R}^d$ наведені для полегшення вивчення аналогів у вищих розмірностях (наприклад гіперкуб).

Двовимірний гаусівський розподіл

Стандартний двовимірний гаусівський розподіл у $\mathbb{R}^2$ має щільність

$$f(x,y) = \frac{1}{2\pi}\exp\bigl(-\tfrac{x^2+y^2}{2}\bigr).$$

За властивістю радіальної симетрії глибина у точці (x,y) залежить лише від $r = \sqrt{x^2+y^2}$.

Твердження 9 (Глибина для стандартного двовимірного гаусівського розподілу).

$$\mathrm{HD}(x,y) \;=\; 1 - \Phi\bigl(\sqrt{x^2+y^2}\bigr),$$

де Φ — функція розподілу стандартного нормального розподілу. Область HD визначається як

$$D_\alpha \;=\; \{(x,y) : x^2+y^2 \le [\Phi^{-1}(1-\alpha)]^2\}.$$

Максимальний HD дорівнює $1/2$, а медіана Тьюкі — $(0,0)$.

Доведення. Для точки $(r,0)$ розглянемо напівплощину $H = \{(x,y) : x \ge r\}$. Тоді

$$P(H) \;=\; \int_r^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx \;=\; 1 - \Phi(r).$$

За симетрією це є мінімумом серед усіх напівплощин, що проходять через $(r,0)$. Для загального (x,y) покладемо $r = \sqrt{x^2+y^2}$. □

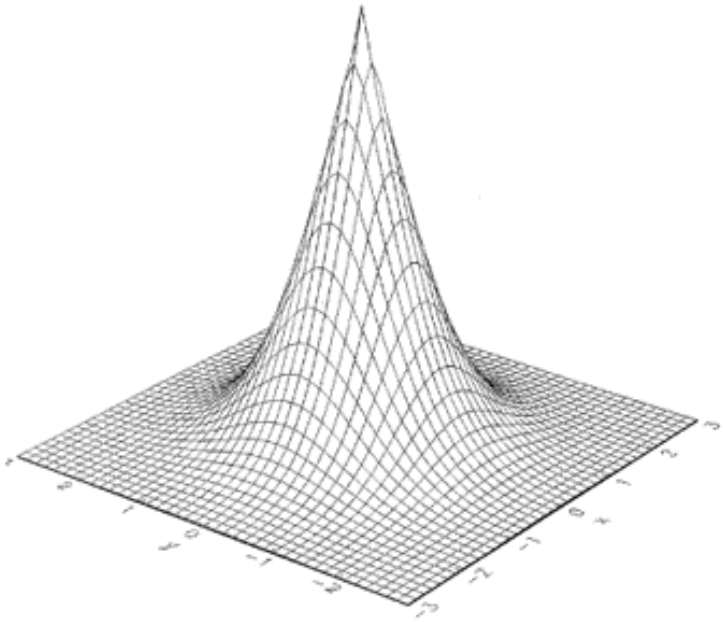

Рис. 1.1: Поверхня напівпросторової глибини для стандартного двовимірного гаусівського розподілу $\mathrm{HD}(x,y) = 1 - \Phi\bigl(\sqrt{x^2+y^2}\bigr)$. Центральний пік відповідає максимальній глибині $1/2$, а кругові контури відображають рівні глибини.

Твердження 10 (Глибина для двовимірного гаусівського розподілу з коваріацією Σ). Нехай $\mathbf{X} \sim \mathcal{N}_2(\boldsymbol{\mu}, \Sigma)$ з додатно-визначеною коваріаційною матрицею Σ. Позначимо радіус Махаланобіса

$$\delta(\mathbf{x}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}.$$

Тоді

$$\mathrm{HD}(\mathbf{x}) \;=\; 1 - \Phi\big(\delta(\mathbf{x})\big), \qquad D_\alpha \;=\; \big\{\mathbf{x} : \delta(\mathbf{x}) \le \Phi^{-1}(1-\alpha)\big\}.$$

Максимальна HD досягається в $\boldsymbol{\mu}$ та залишається рівною $1/2$.

Доведення. Застосуємо афінне відображення $\mathbf{z} = \Sigma^{-1/2}(\mathbf{x}-\boldsymbol{\mu})$, яке переводить $\mathcal{N}_2(\boldsymbol{\mu}, \Sigma)$ у стандартний двовимірний гаусівський розподіл. HD є афінно інваріантним, тому $\mathrm{HD}(\mathbf{x}) = \mathrm{HD}(\mathbf{z}) = 1 - \Phi(\|\mathbf{z}\|_2)$, а $\|\mathbf{z}\|_2 = \delta(\mathbf{x})$. □

Зауваження 1 (Узагальнення на $\mathbb{R}^d$). Для $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$ ($\Sigma \succ 0$), позначимо $\delta(\mathbf{x}) = \|\Sigma^{-1/2}(\mathbf{x}-\boldsymbol{\mu})\|_2$. Тоді

$$\mathrm{HD}(\mathbf{x}) = 1 - F_{\chi_p}\big(\delta(\mathbf{x})\big), \qquad D_\alpha = \big\{\mathbf{x} : \delta(\mathbf{x}) \le F^{-1}_{\chi_p}(1-\alpha)\big\}.$$

Максимальний HD становить $1/2$ і досягається в $\boldsymbol{\mu}$. Контури глибини є колами, коли $\Sigma = I_2$, і еліпсами, орієнтованими за власними векторами Σ у загальному випадку.

Зауваження 2. Зв'язок між відстанню Махаланобіса та напівпросторовою глибиною є ключовою особливістю еліптично-симетричних розподілів, таких як двовимірний нормальний. Ця властивість дозволяє ранжувати точки за центральністю для надійних наближень. Цей зв'язок використовується в обчислювальних методах, таких як (2).

Двовимірний розподіл Коші

Двовимірний розподіл Коші, утворений незалежними одновимірними змінними Коші X та Y, має щільність

$$f(x,y) = \frac{1}{\pi^2}\left(\frac{1}{1+x^2}\right)\left(\frac{1}{1+y^2}\right).$$

Його важкі хвости призводять до унікальних контурів HD порівняно з гаусівським розподілом.

Твердження 11 (Глибина для двовимірного розподілу Коші). Для двовимірного розподілу Коші функція напівпросторова глибина має вигляд

$$\mathrm{HD}(x,y) = \frac{1}{2} - \frac{1}{\pi}\arctan\left(\max\{|x|,|y|\}\right).$$

Контури HD є квадратами з центром у початку координат, тобто

$$D_\alpha = \{(x,y) : \max\{|x|,|y|\} \le \tan\left(\pi\left(\frac{1}{2}-\alpha\right)\right)\}.$$

Максимальна глибина становить $1/2$, а медіана Тьюкі — $(0,0)$.

Доведення. Розглянемо (u,v) з $0 \le u < v$ та напівплощину $H_{a,b} = \{(x,y) : ax+by \ge au+bv\}$, $a,b>0$. Лема $P(aX+bY) = P((|a|+|b|)X)$ означає

$$P(H_{a,b}) = P\left(X \ge \frac{au+bv}{a+b}\right) = P(X \ge \tilde{a}u + (1-\tilde{a})v),$$

де $\tilde{a} = a/(a+b)$. Оскільки $u<v$, вираз максимізується при $\tilde{a}=0$, що дає $P(X \ge v) = \frac{1}{2} - \frac{1}{\pi}\arctan(v)$. Симетрія розширює це на $\max\{|x|,|y|\}$. □

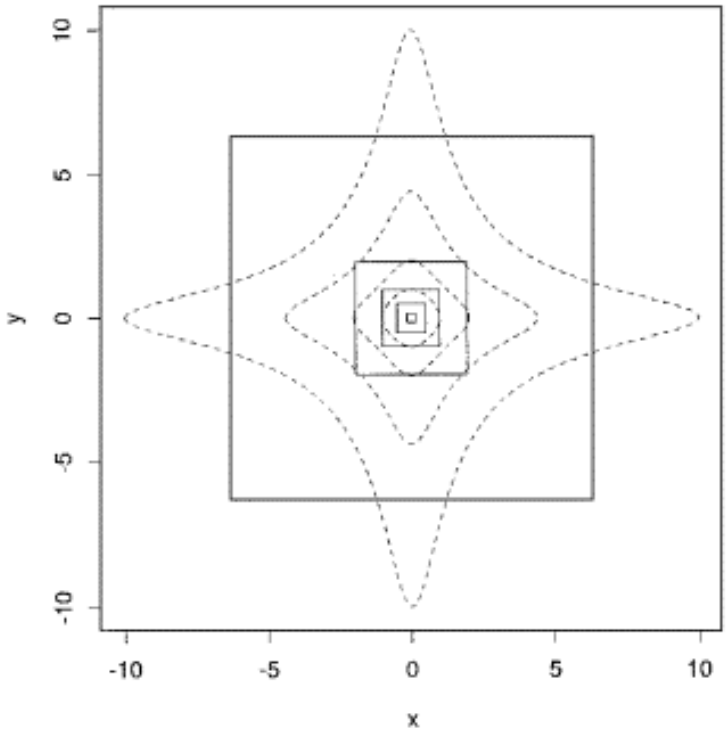

Рис. 1.2: Контури напівпросторової глибини для двовимірного розподілу Коші, що описуються формулою $\mathrm{HD}(x,y) = \frac{1}{2} - \frac{1}{\pi}\arctan\big(\max\{|x|,|y|\}\big)$. Рівні глибини утворюють вкладені квадрати (суцільні лінії)

Зауваження 3 (Узагальнення на $\mathbb{R}^d$). Для d-вимірного розподілу Коші з незалежними стандартними компонентами Коші функція HD має вигляд

$$\mathrm{HD}(\mathbf{x}) = \frac{1}{2} - \frac{1}{\pi}\arctan\left(\|\mathbf{x}\|_\infty\right),$$

де $\|\mathbf{x}\|_\infty = \max_{i=1,\dots,p}|x_i|$. Множини рівня визначаються як

$$D_\alpha = \{\mathbf{x} : \|\mathbf{x}\|_\infty \le \tan\left(\pi\left(\frac{1}{2} - \alpha\right)\right)\},$$

що є гіперкубом у $\mathbb{R}^d$. Максимальна глибина становить $1/2$ та досягається в точці $\mathbf{0}$.

Квадратні контури виникають через важкі хвости розподілу Коші, що контрастує з круговими контурами гаусівського розподілу та підкреслює вплив поведінки хвостів на HD.

Рівномірний розподіл на квадраті (4-кутнику)

Рівномірний розподіл на квадраті $Q = [0,1]\times[0,1] \subset \mathbb{R}^2$ має щільність

$$f(x,y) = I((x,y) \in Q).$$

Правильний чотирикутник є зручним базовим прикладом для аналізу рівневих контурів глибини, коли розподіл має обмежену (многокутну) опору.

Твердження 12 (HD для рівномірного розподілу на квадраті). Для рівномірного розподілу на $Q = [0,1] \times [0,1]$, функція HD має вигляд

$$\mathrm{HD}(x,y) = 2\min(x, 1-x)\min(y, 1-y) \quad \text{для } (x,y) \in Q,$$

і $\mathrm{HD}(x,y) = 0$ поза Q. Область HD визначається як

$$D_\alpha = \{(x,y) \in Q : \min(x, 1-x)\min(y, 1-y) \geq \frac{\alpha}{2}\}.$$

Максимальна HD становить $1/2$, та досягається в точці $(1/2, 1/2)$.

Короткий виклад доведення. Квадрат можна розглядати як перетин чотирьох напівпросторів. Для $(x,y) \in Q$, мінімальна ймовірнісна маса в напівплощині досягається при розрізі вздовж координатних осей, що дає добуток відстаней до меж. Поза Q, принаймні одна напівплощина має нульову масу. □

Функція HD формує гіперболічний параболоїд у кожному квадранті квадрата, з гребенями вздовж $x = 1/2$ та $y = 1/2$. Контури згладжують кути квадрата, створюючи характерний візерунок.

Зауваження 4 (Узагальнення на $\mathbb{R}^d$). Для рівномірного розподілу на гіперкубі $Q = [0,1]^p \subset \mathbb{R}^d$ з щільністю

$$f(\mathbf{x}) = I(\mathbf{x} \in Q),$$

функція глибини має вигляд

$$\mathrm{HD}(\mathbf{x}) = 2^p \prod_{i=1}^{p} \min(x_i, 1-x_i) \quad \text{для } \mathbf{x} \in Q,$$

і $\mathrm{HD}(\mathbf{x}) = 0$ поза Q. Область HD визначається як

$$D_\alpha = \left\{ \mathbf{x} \in Q : \prod_{i=1}^{p} \min(x_i, 1-x_i) \geq \frac{\alpha}{2^p} \right\}.$$

Максимальний глибина становить $1/2$ і досягається в $(1/2, \ldots, 1/2)$.

1.4 Висновки

1. Визначення глибини Тьюкі. Найменша міра півпросторів, що містять задану точку x. Глибина Тьюкі (або напівпросторова глибина HD) показує, наскільки точка є типовою як значення випадкового вектора з розподілом μ.

2. Ключові властивості. Глибина є афінно інваріантною, квазіувігнутою, верхньо-(а для неперервних мір — й нижньо-) напівнеперервною; множини D_α опуклі й компактні. Доведено існування медіани Тьюкі та універсальну нижню межу $\max_\theta \mathrm{HD}(\theta) \geq \frac{1}{d+1}$, що гарантує змістовність поняття «найглибшої» точки навіть у високих розмірностях.

3. Аналітичні формули для репрезентативних розподілів.

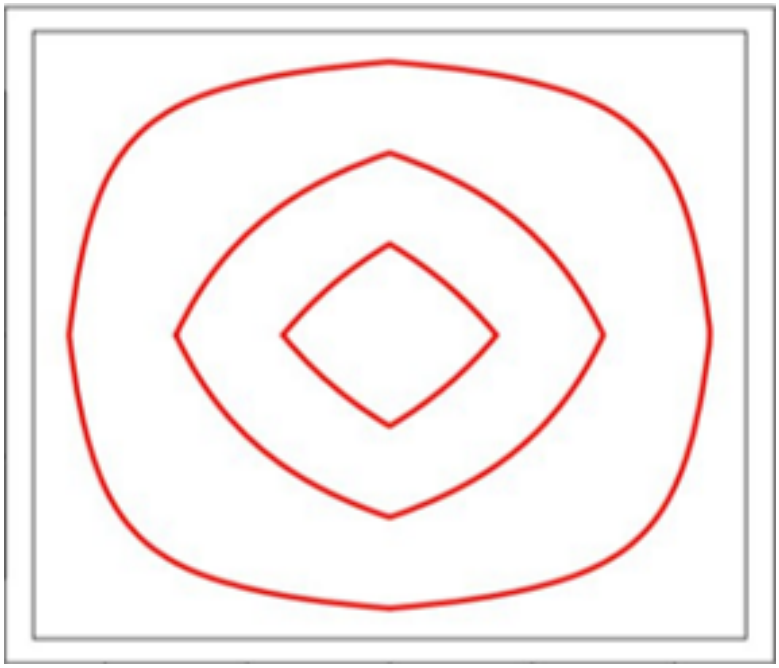

Рис. 1.3: Контури напівпросторової глибини для рівномірного розподілу на квадраті, що описуються формулою $\mathrm{HD}(x, y) = 2\min(x, 1-x)\min(y, 1-y)$, утворюють форму вкладених згладжених квадратів.

- Гаусівський. $\mathrm{HD}(x, y) = 1 - \Phi(r)$ (або $1 - \Phi(\delta)$ для довільної Σ); контури — концентричні кола / еліпси.
- Коші. $\mathrm{HD}(x, y) = \frac{1}{2} - \frac{1}{\pi}\arctan\big(\max\{|x|, |y|\}\big)$; контури — квадрати (гіперкуби у $\mathbb{R}^d$).
- Рівномірний на квадраті. $\mathrm{HD}(x, y) = 2\min(x, 1-x)\min(y, 1-y)$; контури згладжують вершини й вирівнюються вздовж $x = \frac{1}{2}, y = \frac{1}{2}$.

В усіх трьох випадках максимальна глибина дорівнює $1/2$ і досягається в центрі симетрії.

Розділ 2

Обчислення глибини для ймовірнісних розподілів

2.1 Узагальнення глибини Тьюкі в рівномірно розподіленій одиничній кулі в $\mathbb{R}^d$

2.1.1 Глибина Тьюкі для рівномірної міри на B^d

Нехай $B^d = \{x \in \mathbb{R}^d \mid \|x\| \leq 1\}$ — замкнена одинична куля у $\mathbb{R}^d$. Розглядаємо рівномірну (ізотропну) ймовірнісну міру: для будь-якої вимірної множини $A \subset B^d$

$$F(A) \;=\; \frac{\mathrm{Vol}_d(A)}{\mathrm{Vol}_d(B^d)},$$

де $\mathrm{Vol}_d(\cdot)$ - d-вимірна міра Лебега.

.

Для точки $\theta \in \mathbb{R}^d$ та міри F визначимо

$$\mathrm{HD}_F(\theta) \;=\; \inf_{u \in S^{d-1}} F\big(\{\, x \in \mathbb{R}^d : \langle x - \theta,\, u\rangle \geq 0\}\big),$$

тобто мінімальну ймовірність півпростору, відмежованого гіперплощиною $\langle x - \theta,\, u\rangle = 0$.

2.1.2 Геометрія півпросторів у кулі

Щоб проілюструвати ідею, розглянемо одиничний диск B^2 із центром O і точкою θ всередині нього (з $\|\theta\| < 1$). Міра замкненого півпростору $H(u, \theta) \cap B^2$ залежить від перпендикулярної відстані від O до гіперплощини $\langle x - \theta, u\rangle = 0$. Для заданого напрямку $u \in S^1$ (одиничне коло) ця відстань дорівнює

$$l = |\langle \theta, u\rangle|.$$

Серед усіх напрямків u, які забезпечують $\theta \in H(u, \theta)$, мінімальний об'єм $H(u, \theta) \cap B^2$ досягається, коли l максимізується. Геометрично ця оптимальна ситуація виникає, коли u вибирається паралельним до вектора від початку координат до θ, тобто

$$u = \frac{\theta}{\|\theta\|}.$$

У цьому випадку гранична гіперплощина має вигляд

$$\langle x, \theta \rangle = \|\theta\|^2,$$

а відповідний півпростір задається як

$$H\left(\frac{\theta}{\|\theta\|}, \theta\right) = \left\{x \in \mathbb{R}^2 : \langle x, \theta \rangle \geq \|\theta\|^2\right\}.$$

Глибина Тьюкі точки θ визначається як відношення площі шапки $H\left(\frac{\theta}{\|\theta\|}, \theta\right) \cap B^2$ до загальної площі B^2:

$$\mathrm{HD_F}(\theta) = \frac{\mathrm{Vol}\left(H\left(\frac{\theta}{\|\theta\|}, \theta\right) \cap B^2\right)}{\mathrm{Vol}(B^2)}.$$

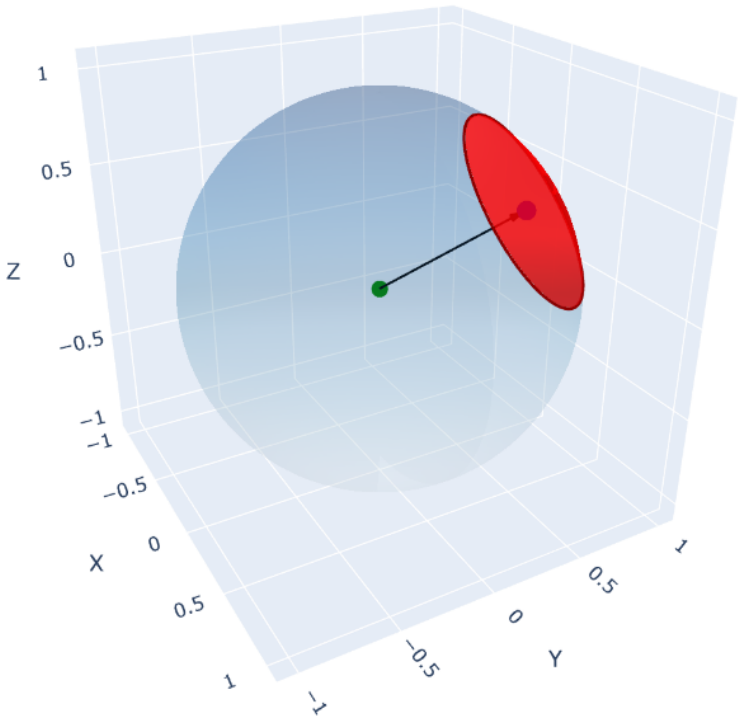

Рис. 2.1: Одинична куля з центром у точці O та внутрішня точка, позначимо її як θ. Серед усіх напівпросторів, чия гранична площина проходить через θ, оптимальним є той, чия площина є ортогональною до вектора $O\theta$; вона розташована якнайдалі від початку координат і, отже, відсікає найменшу сферичну шапку від кулі. Відносний об'єм цієї шапки (щодо всієї кулі) і дорівнює напівпросторовій глибині точки θ.

2.1.3 Інтегральна форма для об'єму сегмента кулі

Без втрати загальності, скориставшись інваріантністю відносно обертання, ми можемо сумістити θ з першою координатною віссю. Тоді умова $\langle x, \theta \rangle \geq \|\theta\|^2$ спрощується до

$$x_1 \geq \|\theta\|,$$

де x_1 — перша координата вектора x. Відтак

$$\{x \in B^d : \langle x, \theta \rangle \geq \|\theta\|^2\} \;=\; \{x : \|x\| \leq 1,\ x_1 \geq \|\theta\|\}.$$

Щоб обчислити його об'єм, розіб'ємо одиничну кулю $\|x\| \leq 1$ на гіперплощинні перерізи за кожного фіксованого $r \in [\|\theta\|, 1]$. Кожен такий переріз є $(d-1)$-вимірною кулею радіуса $\sqrt{1-r^2}$. Отже

$$\mathrm{Vol}\big(H(\theta) \cap B^d\big) = \int_{\|\theta\|}^{1} \mathrm{vol}\big(B^{d-1}\big)\,(1-r^2)^{\frac{d-1}{2}}\,dr.$$

2.1.4 Остаточна формула для глибини

Ми починаємо з виразу

$$\mathrm{HD}(\theta; F) \;=\; \frac{1}{\mathrm{Vol}(B^d)} \int_{\|\theta\|}^{1} \mathrm{vol}\big(B^{d-1}\big)\,\big(1-r^2\big)^{\frac{d-1}{2}}\,dr,$$

де загальний об'єм d-вимірної кулі та $(d-1)$-вимірної кулі задано як

$$\mathrm{Vol}(B^d) \;=\; \frac{\pi^{\frac{d}{2}}}{\Gamma\Big(\frac{d}{2}+1\Big)} \quad \text{і} \quad \mathrm{vol}\big(B^{d-1}\big) \;=\; \frac{\pi^{\frac{d-1}{2}}}{\Gamma\Big(\frac{d-1}{2}+1\Big)}.$$

Підставляючи ці вирази у формулу HD, отримуємо

$$\mathrm{HD}(\theta; F) \;=\; \frac{\Gamma\Big(\frac{d}{2}+1\Big)}{\pi^{\frac{d}{2}}}\,\frac{\pi^{\frac{d-1}{2}}}{\Gamma\Big(\frac{d-1}{2}+1\Big)} \int_{\|\theta\|}^{1} \big(1-r^2\big)^{\frac{d-1}{2}}\,dr.$$

Цей множник спрощується до

$$\frac{\Gamma\Big(\frac{d}{2}+1\Big)}{\Gamma\Big(\frac{d-1}{2}+1\Big)}\,\frac{1}{\pi^{1/2}}.$$

Щоб обчислити інтеграл, ми підставляємо $u = r^2$, так що $du = 2r\,dr$, а отже

$$dr = \frac{du}{2\sqrt{u}},$$

з межами інтегрування, що змінюються від $r = \|\theta\|$ до $u = \|\theta\|^2$ і від $r = 1$ до $u = 1$. Таким чином, інтеграл перетворюється на

$$\int_{\|\theta\|}^{1} \big(1-r^2\big)^{\frac{d-1}{2}}\,dr \;=\; \frac{1}{2}\int_{\|\theta\|^2}^{1} u^{-\frac{1}{2}}\,\big(1-u\big)^{\frac{d-1}{2}}\,du.$$

За визначенням неповної бета-функції,

$$B_z(a,b) \;=\; \int_0^z t^{\,a-1}\,(1-t)^{\,b-1}\,dt,$$

з $a = \frac{1}{2}$ та $b = \frac{d+1}{2}$, наведений інтеграл можна записати як

$$\int_{\|\theta\|^2}^{1} u^{-\frac{1}{2}}\,\big(1-u\big)^{\frac{d-1}{2}}\,du \;=\; B\Big(\tfrac{1}{2}, \tfrac{d+1}{2}\Big) - B_{\|\theta\|^2}\Big(\tfrac{1}{2}, \tfrac{d+1}{2}\Big).$$

Отже,

$$\int_{\|\theta\|}^{1} \left(1-r^2\right)^{\frac{d-1}{2}} dr \;=\; \frac{1}{2}\left[B\Big(\tfrac{1}{2},\tfrac{d+1}{2}\Big) - B_{\|\theta\|^2}\Big(\tfrac{1}{2},\tfrac{d+1}{2}\Big)\right].$$

Далі, використаємо тотожність, яка пов'язує бета-функцію з гамма-функціями:

$$B(a,b) \;=\; \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}.$$

Отже,

$$B\Big(\tfrac{1}{2},\tfrac{d+1}{2}\Big) \;=\; \frac{\Gamma\Big(\frac{1}{2}\Big)\,\Gamma\Big(\frac{d+1}{2}\Big)}{\Gamma\Big(\frac{d}{2}+1\Big)}.$$

Регуляризована неповна бета-функція визначається як

$$I_z(a,b) \;=\; \frac{B_z(a,b)}{B(a,b)}.$$

Поєднуючи ці співвідношення, ми виражаємо інтеграл як

$$\int_{\|\theta\|}^{1} \left(1-r^2\right)^{\frac{d-1}{2}} dr \;=\; \frac{1}{2}\,B\Big(\tfrac{1}{2},\tfrac{d+1}{2}\Big)\Big[1 - I_{\|\theta\|^2}\Big(\tfrac{1}{2},\tfrac{d+1}{2}\Big)\Big].$$

Підставляючи назад у вираз для $\mathrm{HD}(\theta;F)$, отримуємо

$$\mathrm{HD}(\theta;F) \;=\; \frac{\Gamma\Big(\frac{d}{2}+1\Big)}{\Gamma\Big(\frac{d-1}{2}+1\Big)}\,\frac{1}{\pi^{1/2}} \cdot \frac{1}{2}\,B\Big(\tfrac{1}{2},\tfrac{d+1}{2}\Big)\Big[1 - I_{\|\theta\|^2}\Big(\tfrac{1}{2},\tfrac{d+1}{2}\Big)\Big].$$

Підставимо бета-функцію через гамма-функції:

$$\frac{1}{2}\,B\Big(\tfrac{1}{2},\tfrac{d+1}{2}\Big) \;=\; \frac{1}{2}\,\frac{\Gamma\Big(\frac{1}{2}\Big)\,\Gamma\Big(\frac{d+1}{2}\Big)}{\Gamma\Big(\frac{d}{2}+1\Big)}.$$

Після скорочення $\Gamma\Big(\frac{d}{2}+1\Big)$:

$$\mathrm{HD}(\theta;F) \;=\; \frac{1}{2}\,\frac{\Gamma\Big(\frac{1}{2}\Big)\,\Gamma\Big(\frac{d+1}{2}\Big)}{\Gamma\Big(\frac{d-1}{2}+1\Big)\,\pi^{1/2}}\,\Big[1 - I_{\|\theta\|^2}\Big(\tfrac{1}{2},\tfrac{d+1}{2}\Big)\Big].$$

Оскільки $\Gamma\Big(\frac{1}{2}\Big) = \sqrt{\pi}$, отримуємо остаточний вираз у замкненій формі:

$$\boxed{\mathrm{HD}(\theta;F) \;=\; \frac{1}{2}\Big[1 - I_{\|\theta\|^2}\Big(\tfrac{1}{2},\tfrac{d+1}{2}\Big)\Big], \quad \text{для } \|\theta\| < 1.}$$

Аналітична формула явно демонструє, як об'єми $(d-1)$- та d-вимірних куль (що включають π та Г-функції) поєднуються з інтегральним виразом, щоб у підсумковій формулі отримати регуляризовану неповну бета-функцію.

2.1.5 Приклади

Випадок $d = 1$. Одинична «куля» тут — інтервал $B^1 = [-1, 1]$. Для точки $\theta \in [-1, 1]$ напівпростір мінімальної ймовірності — промінь $H = \{x \geq \theta\}$ (якщо $\theta > 0$), довжина якого дорівнює $1 - \theta$. Імовірність (відношення довжин)

$$\mathrm{HD}(\theta) = \frac{1 - |\theta|}{2}, \qquad \theta \in [-1, 1].$$

Випадок $d = 2$. Нехай $\theta \in B^2$ з радіусом $r = \|\theta\| \in [0, 1)$. Завдяки обертовій інваріантності спрямуємо θ уздовж осі x. Оптимальний напівпростір обмежує пряма $x = r$; вона відтинає від одиничного диска круговий сегмент площею

$$A_{\mathrm{cap}}(r) = 2 \int_r^1 \sqrt{1 - t^2}\, dt = \arccos r - r\sqrt{1 - r^2}.$$

Оскільки $\mathrm{Vol}(B^2) = \pi$, маємо

$$\mathrm{HD}(\theta) = \frac{A_{\mathrm{cap}}(r)}{\pi} = \frac{\arccos r - r\sqrt{1 - r^2}}{\pi}, \qquad 0 \leq r \leq 1.$$

Максимальна глибина дорівнює $1/2$ (досягається у центрі).

Випадок $d = 3$. Для $\theta \in B^3$ позначимо $r = \|\theta\| \in [0, 1)$. Площина $x = r$ (ортогональна до $O\theta$) відсікає від кулі сферичний сегмент висоти $h = 1 - r$ і об'єму

$$V_{\mathrm{cap}}(r) = \frac{\pi h^2 (3 - h)}{3} = \frac{\pi (1 - r)^2 (2 + r)}{3}.$$

Оскільки $\mathrm{Vol}(B^3) = 4\pi/3$,

$$\mathrm{HD}(\theta) = \frac{V_{\mathrm{cap}}(r)}{4\pi/3} = \frac{(1 - r)^2 (2 + r)}{4}, \qquad 0 \leq r \leq 1.$$

Як і раніше, $\max \mathrm{HD} = 1/2$ у точці O.

2.2 Глибина Тьюкі для α-стійких розподілів

2.2.1 Постановка задачі для двовимірного випадку

Симетрична α-стійка випадкова величина X має характеристичну функцію

$$\varphi_X(t) = \exp\bigl(-|t|^\alpha\bigr), \qquad 0 < \alpha \leq 2.$$

$$X \sim S(\alpha, 0, 1, 0)$$

з характеристичною функцією

$$\varphi_X(t) = \exp(-|t|^\alpha).$$

Нехай незалежні $X, Y \sim S(\alpha, 0, 1, 0)$. Тоді для будь-яких констант a, b лінійна комбінація

$$Z = aX + bY$$

матиме характеристичну функцію

$$\varphi_Z(t) = \exp\Big[-\Big(|a|^\alpha + |b|^\alpha\Big)|t|^\alpha\Big],$$

що вказує на те, що

$$Z \sim S\Big(\alpha, 0,\ c',\ 0\Big), \quad \text{де} \quad c' = \Big(|a|^\alpha + |b|^\alpha\Big)^{1/\alpha}.$$

Ця властивість масштабування дозволяє стандартизувати задачу так, що

$$P\Big(aX + bY \geq ax + by\Big) = P\left(\frac{aX + bY}{c'} \geq \frac{ax + by}{c'}\right).$$

Для стислості ми позначимо

$$c = \frac{ax + by}{c'},$$

та назвемо його є порогом спрямованості.

2.2.2 Аналіз порогу спрямованості

Ми досліджуємо функцію, визначену як

$$c(a, b) = \frac{a\,x + b\,y}{\big(|a|^\alpha + |b|^\alpha\big)^{1/\alpha}},$$

яка є однорідною функцією першого степеня:

$$|a|^\alpha + |b|^\alpha = 1.$$

Введемо наступну параметризацію:

$$a = t^{1/\alpha}, \quad b = (1 - t)^{1/\alpha}, \quad t \in [0, 1],$$

функція набуває вигляду:

$$c(t) = x\,t^{1/\alpha} + y\,(1 - t)^{1/\alpha}.$$

Увігнутість проти опуклості. Характер відображення $t \mapsto t^{1/\alpha}$ є ключовим:

- Для $\alpha > 1$: Оскільки $1/\alpha < 1$, функція $t^{1/\alpha}$ є увігнутою. Таким чином, $c(t)$ є увігнутою функцією на $(1, 2]$ і її максимум досягається в єдиній внутрішній точці.
- Для $\alpha \leq 1$: Тут $1/\alpha \geq 1$, тому $t^{1/\alpha}$ є опуклою; отже, $c(t)$ є опуклою на $(0, 1]$, і максимум досягається в одній із граничних точок.

Випадок 1: $1 < \alpha \leq 2$

Диференціюючи $c(t)$ і прирівнюючи $c'(t) = 0$, швидко отримуємо

$$\left(\frac{1 - t}{t}\right)^{\frac{\alpha-1}{\alpha}} = \frac{y}{x},$$

що дає оптимальне значення

$$t^* = \frac{1}{1 + \left(\frac{y}{x}\right)^{\frac{\alpha}{\alpha-1}}}.$$

Підставляючи t^* у $f(t)$, отримуємо

$$c_{\max} = x \left[1 + \left(\frac{y}{x}\right)^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}.$$

Визначаючи

$$\beta = \frac{\alpha}{\alpha - 1},$$

це спрощується до вигляду

$$\boxed{c_{\max} = \left(x^\beta + y^\beta\right)^{1/\beta}.}$$

Випадок 2: $0 < \alpha \leq 1$

Коли $0 < \alpha \leq 1$, функція $c(t)$ є опуклою, і тому її максимум досягається в одній із граничних точок інтервалу. Отже, у цьому випадку маємо

$$\boxed{c_{\max} = \max\{x, y\}.}$$

2.2.3 Інтегральне представлення для формули глибини

Після визначення $c_{\max}$, ми обчислюємо ймовірність для стандартної симетричної стійкої змінної наступним чином. Для

$$Z \sim S(\alpha, 0, 1, 0),$$

за формулою інверсії Гіля–Пелаеза (5) маємо функцію розподілу:

$$F(z) = \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \frac{\sin(tz)}{t} \exp(-t^\alpha)\, dt.$$

Отримуємо:

$$P(Z \geq z) = 1 - F(z) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{\sin(tz)}{t} \exp(-t^\alpha)\, dt.$$

Підставивши у формулу поріг спрямованості, ймовірність, що нас цікавить, дорівнює:

$$P\Big(aX + bY \geq ax + by\Big) = P\Big(Z \geq c\Big),$$

де

$$c = \frac{ax + by}{\Big(|a|^\alpha + |b|^\alpha\Big)^{1/\alpha}}.$$

Підставляючи оптимальний поріг $c_{\max}$ (який тепер залежить від співвідношення значень x і y та від α), отримуємо остаточну формулу для HD:

$$\boxed{\mathrm{HD}(x, y) = P\Big(aX + bY \geq ax + by\Big) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{\sin\Big(t\, c_{\max}\Big)}{t} \exp(-t^\alpha)\, dt.}$$

Тут зауважимо, що:

- Для $1 < \alpha \leq 2$: $c_{\max} = \left(x^\beta + y^\beta\right)^{1/\beta}$, де $\beta = \frac{\alpha}{\alpha-1}$.
- Для $0 < \alpha \leq 1$: $c_{\max} = \max\{x, y\}$.

2.2.4 Візуалізація контурів глибини

Розглянемо два сценарії:

- $0 < \alpha \leq 1$ (наприклад, $\alpha = 0.8$), функція $f(t) = x\,t^{1/\alpha} + y\,(1-t)^{1/\alpha}$ є опуклою, тому максимум досягається в одній із граничних точок. Отже, оптимальний поріг становить
$$c_{\max} = \max\{x, y\}.$$
- $1 < \alpha \leq 2$ (наприклад, $\alpha = 1.3$), оптимум визначається як
$$c_{\max} = \left(x^{\beta} + y^{\beta}\right)^{1/\beta}, \quad \beta = \frac{1.3}{1.3-1} \approx 4.33.$$

На рисунку 2.2 показано контури глибини для $\alpha = 0.8$ та $\alpha = 1.3$. Для $1 < \alpha \leq 2$ контури мають більш округлу форму, тоді як для $0 < \alpha \leq 1$ - залежать лише від більшої координати, що призводить до виродженого випадку - завжди квадратних форм контурів. Ці приклади узгоджується з відомими розподілами: для $\alpha = 1$ (розподіл Коші) отримуємо $c_{\max} = \max\{x, y\}$ з ймовірністю
$$P(Z \geq z) = \frac{1}{2} - \frac{1}{\pi}\arctan(z),$$
а для $\alpha = 2$ (гауссівський розподіл) з $\beta = 2$, поріг стає
$$c_{\max} = \sqrt{x^2 + y^2},$$
що є знайомою евклідовою нормою та частиною формули глибини для стандартного гауссівського розподілу. Ці особливі випадки разом із контурними графіками яскраво ілюструють геометричну поведінку функції глибини при зміні α.

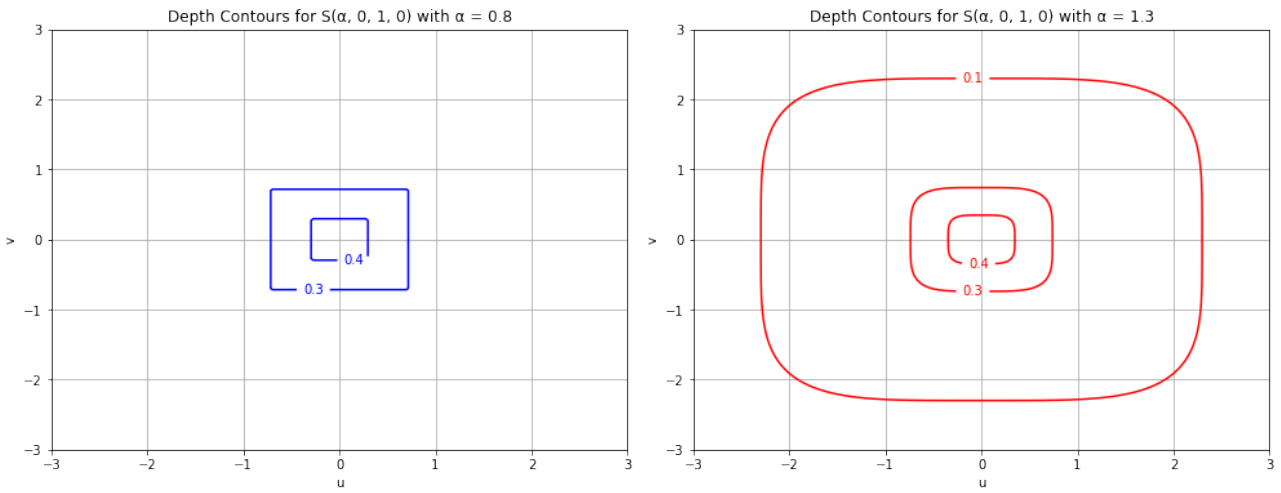

Рис. 2.2: Контури глибини для різних значень параметра α.

2.2.5 Узагальнення для R^d

Нехай $\mathbf{x} = (x_1, \ldots, x_d)$ — випадковий вектор із незалежними однаково розподіленими компонентами $x_i \sim S_\alpha(0, 1, 0)$. Для будь-якого вектора $\mathbf{a} = (a_1, \ldots, a_d) \neq \mathbf{0}$ визначимо поріг прямованості як:

$$c(\mathbf{a}) = \frac{\sum_{i=1}^{d} a_i x_i}{\left(\sum_{i=1}^{d} |a_i|^\alpha\right)^{1/\alpha}}.$$

Оскільки і чисельник, і знаменник є однорідними функціями першого степеня, ми можемо нормалізувати, накладаючи умову

$$\sum_{i=1}^{d} |a_i|^\alpha = 1.$$

Таким чином, задача оптимізації набуває вигляду

$$\max_{\mathbf{a}} \sum_{i=1}^{d} a_i x_i \quad \text{за умови} \quad \sum_{i=1}^{d} |a_i|^\alpha = 1.$$

Нерівність Гельдера ($1 < \alpha \leq 2$)

Нехай $\beta = \dfrac{\alpha}{\alpha - 1}$, так що $\frac{1}{\alpha} + \frac{1}{\beta} = 1$. За нерівністю Гельдера,

$$\sum_{i=1}^{d} a_i x_i \;\leq\; \Big(\sum_{i=1}^{d} |a_i|^\alpha\Big)^{1/\alpha} \Big(\sum_{i=1}^{d} |x_i|^\beta\Big)^{1/\beta} = \|\mathbf{x}\|_\beta,$$

і рівність досягається тоді і тільки тоді, коли

$$|a_i|^\alpha = \frac{|x_i|^\beta}{\sum\limits_{j=1}^{d} |x_j|^\beta}.$$

Враховуючи знак x_i, такі значення $a_i^\star$ будуть максимізувати лінійну комбінацію:

$$a_i^\star = \frac{|x_i|^{\beta-1} \operatorname{sgn}(x_i)}{\left(\sum_{j=1}^{d} |x_j|^\beta\right)^{1/\alpha}}, \qquad i = 1, \ldots, d,$$

Оптимальне значення $c_{\max}$ є L^β-нормою

$$\boxed{c_{\max} = \|\mathbf{x}\|_\beta = \Big(\sum_{i=1}^{d} |x_i|^\beta\Big)^{1/\beta}}.$$

Рішення, з використанням граничного значення $0 < \alpha \leq 1$

Коли $0 < \alpha \leq 1$, відображення $t \mapsto |t|^\alpha$ є опуклим, тому множина обмежень є опуклим тілом, а цільова функція $\sum a_i x_i$ — лінійна. За опуклістю можна стверджувати, що максимум досягається в екстремумі, тобто коли всі $|a_i|$, крім одного, дорівнюють нулю. Отже,

$$\boxed{c_{\max} = \|\mathbf{x}|_\infty = \max_{1 \leq i \leq d} |x_i|}.$$

2.3 Глибина Тьюкі для двовимірного експоненційного розподілу

2.3.1 Обчислення $I(k)$

Розглянемо дві незалежні випадкові величини $X \sim \text{Exp}(\lambda_1)$ та $Y \sim \text{Exp}(\lambda_2)$, із щільністю, заданою як

$$f_{X,Y}(x,y) = \lambda_1\lambda_2 e^{-\lambda_1 x - \lambda_2 y}, \quad x, y \geq 0.$$

У двовимірному просторі через точку (a, b) проходить пряма з нахилом $k = \tan\theta$, виділяючи дві комплементарні області:

- Нижче прямої: $I(k) = P(Y \leq k(X - a) + b)$.
- Вище прямої: $1 - I(k) = P(Y \geq k(X - a) + b)$.

Таким чином, глибину Тьюкі можна переписати як:

$$D(a,b) = \inf_k \min\{I(k), 1 - I(k)\}.$$

Ми розрізняємо три випадки для $I(k)$:

- Випадок I (Трикутник): Точки перетину з осями x та y додатні, тобто $x_0 = a - \frac{b}{k} > 0$ і $y_0 = b - ka > 0$. Область утворює трикутник, обмежений прямою та осями.
- Випадок II: Від'ємний перетин з віссю x і додатний перетин з віссю y, тобто $x_0 < 0$ і $y_0 > 0$.
- Випадок III: Додатний перетин з віссю x і від'ємний перетин з віссю y, тобто $x_0 > 0$ і $y_0 < 0$.

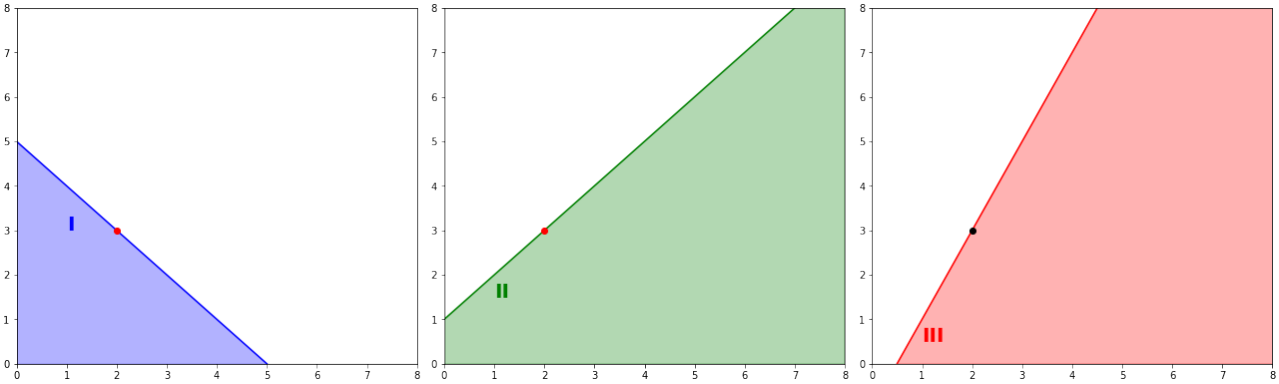

Рис. 2.3: Візуалізація областей, визначених прямими через (a, b): Випадок I показує трикутну область (обидва перетини додатні), Випадок II — область, коли перетин з віссю x від'ємний, і Випадок III — коли перетин з віссю y від'ємний.

Випадки II і III можна об'єднати, оскільки область інтегрування в першому квадранті подібна. Для уніфікованого підходу ми визначаємо

$$A = \max\left\{0, a - \frac{b}{k}\right\}.$$

Таким чином, ймовірнісна маса нижче прямої задається як

$$I(k) = \begin{cases} I_1(k) & \text{якщо } x_0 > 0 \text{ і } y_0 > 0 \quad \text{(Випадок I)}, \\ I_2(k) & \text{в іншому випадку (Випадки II і III)}, \end{cases}$$

де:

$$I_1(k) = \int_0^{x_0} \int_0^{k(x-a)+b} \lambda_1 \lambda_2 e^{-\lambda_1 x - \lambda_2 y} \, dy \, dx,$$

$$I_2(k) = \int_A^{\infty} \int_0^{k(x-a)+b} \lambda_1 \lambda_2 e^{-\lambda_1 x - \lambda_2 y} \, dy \, dx.$$

Обчислюючи внутрішній інтеграл за y,

$$\int_0^{k(x-a)+b} \lambda_2 e^{-\lambda_2 y} \, dy = 1 - e^{-\lambda_2 [k(x-a)+b]},$$

ми отримуємо для Випадку I:

$$I_1(k) = \int_0^{x_0} \lambda_1 e^{-\lambda_1 x} \left[1 - e^{-\lambda_2 [k(x-a)+b]}\right] \, dx,$$

і для Випадків II та III:

$$I_2(k) = \int_A^{\infty} \lambda_1 e^{-\lambda_1 x} \left[1 - e^{-\lambda_2 [k(x-a)+b]}\right] \, dx.$$

Ці інтеграли не можна обчислити аналітично. Для Випадку I найближчий вираз закритого вигляду має вигляд

$$I_1(k) = \left[1 - e^{-\lambda_1 \left(a - \frac{b}{k}\right)}\right] - e^{-\lambda_2 (ka - b)} \frac{\lambda_1}{\lambda_1 + \lambda_2 k} \left[1 - e^{-(\lambda_1 + \lambda_2 k)\left(a - \frac{b}{k}\right)}\right],$$

а для Випадків II та III:

$$I_2(k) = e^{-\lambda_1 A} \left[1 - \frac{\lambda_1}{\lambda_1 + \lambda_2 k} e^{\lambda_2 k (a - A) - \lambda_2 b}\right].$$

Для визначення оптимального нахилу k^*, який мінімізує обидві форми $I(k)$, ми могли б диференціювати по k. Межі інтегрування залежать від k, тому необхідно застосовувати правило Лейбніца для диференціювання під знаком інтеграла. Через складність виразів ми покладаємося лише на чисельні методи (рис. 2.4), які будуть детально розглянуті в наступному розділі, або асимптотичні наближення для знаходження k^*.

2.3.2 Асимптотичний аналіз оптимального нахилу

Ми досліджуємо оптимальне k^* у двох режимах на основі функції $g(x) = k(x - a) + b$.

Режим малих значень Коли $g(x)$ є достатньо малим у межах області інтегрування, ми застосовуємо наближення:

$$1 - e^{-\lambda_2 g(x)} \approx \lambda_2 g(x) = \lambda_2 [k(x - a) + b].$$

Підставляючи в $I(k)$ і оптимізуючи, отримуємо:

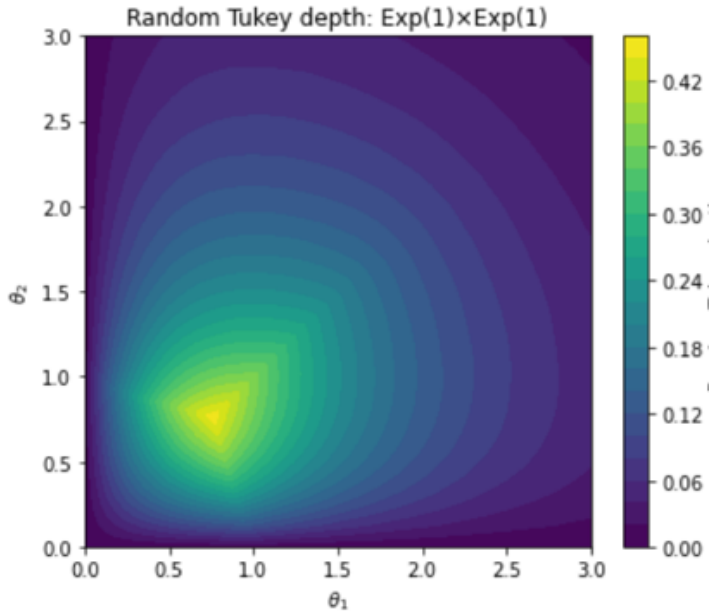

Рис. 2.4: Чисельна апроксимація для двовимірного експоненціального розподілу з незалежними координатами ($\lambda_1 = \lambda_2 = 1$).

$$k^* \approx \frac{\lambda_1}{\lambda_2}.$$

За умови $a - \frac{b\lambda_2}{\lambda_1} > 0$, об'єднана область (Випадки II і III) дає:

$$I_2(k^*) \approx \frac{1}{2} e^{-\lambda_1\left(a - \frac{b\lambda_2}{\lambda_1}\right)}.$$

Це наближення вказує на нижчу ймовірнісну масу в цьому режимі, хоча точна мінімізація вимагає оцінки всіх випадків по k.

Режим великих значень Коли $g(x)$ є великим, окрім як поблизу межі, інтеграл визначається вузькою областю, що дає:

$$k^* \approx \frac{\lambda_1}{\lambda_2}\left(1 + \frac{1}{c}\ln\frac{\lambda_1}{\lambda_2}\right),$$

де c відображає масштаб поблизу межі, враховуючи швидкий експоненційний спад.

Примітка: Ці асимптотичні формули можуть бути корисними, але здебільшого - дуже наближеними і непрактичними. Повний аналіз вимагає оцінки як Випадку I, так і Випадків II/III для всіх k, щоб точно визначити мінімальну ймовірнісну масу.

2.4 Глибина Тьюкі для двовимірного t-розподілу

У цьому розділі ми коротко відмітимо аналітичні та обчислювальні властивості глибини Тьюкі для двовимірного t-розподілу, з якими ми стикнулися в ході дослідження.

По-перше, аналітична форма для проекції існує в інтегральному вигляді (9); підінтегральна функція включає в себе бета-функцію та тригонометричну степеневу функцію.

Хоча точна формула є складною і тут не наводиться, її існування забезпечує теоретичну основу для вивчення функції глибини.

По-друге, еліптична симетрія t-розподілу забезпечує значну перевагу. Ця властивість дозволяє визначити оптимальний напрямок проекції для точки x як $-\frac{x}{\|x\|}$. Це контрастує з розподілами, такими як двовимірний експоненційний, які не мають еліптичної симетрії, що робить неможливим аналітично визначити оптимальний напрям. Завдяки цьому обчислення глибини Тьюкі можна звести до моделювання Монте-Карло, використовуючи інтегральне представлення функції щільності в точці.

Проте в цьому дослідженні контури глибини були обчислені за допомогою чисельних методів, а не симуляцій Монте-Карло. Точність цих контурів буде перевірена в наступному розділі. Для випадку, коли $v = 1$, що відповідає двовимірному розподілу Коші, обчислені контури глибини збігаються з теоретичними контурами, очікуваними для цього розподілу.

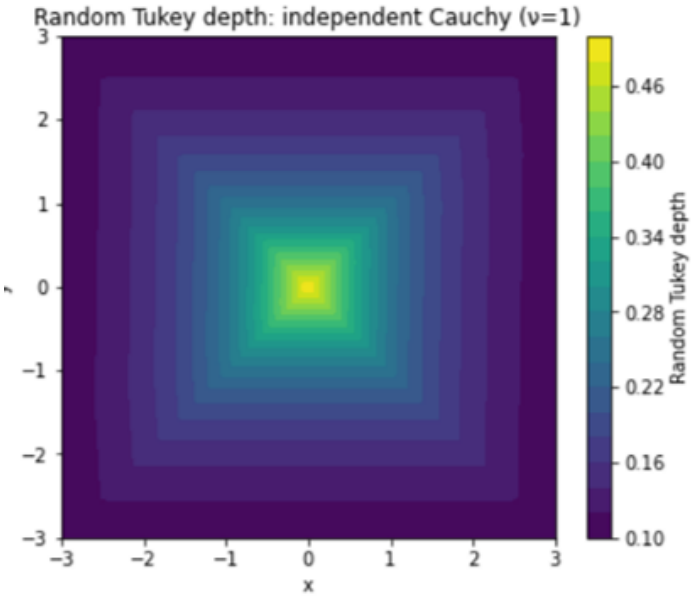

Рис. 2.5: Контурні графіки глибини Тьюкі для біваріатного t-розподілу з $v = 1$ (розподіл Коші). Чисельно обчислені контури відповідають теоретичним очікуванням для випадку Коші.

Розділ 3

Рандомізована глибина Тьюкі

3.1 Основні визначення

Метою цього розділу є оцінка ефективності рандомізованої оцінки для глибини Тьюкі

$$HD_k(x) = \min_{1\le j\le k} P\big(\langle u_j, X\rangle \le \langle u_j, x\rangle\big)$$

як наближення до точної глибини Тьюкі

$$HD(x) = \inf_{u\in S^{d-1}} P\big(\langle u, X\rangle \le \langle u, x\rangle\big),$$

де $\{u_j\}_{j=1}^k$ — це незалежні однаково розподілені рівномірні напрямки на одиничній сфері.

Ми розглядаємо регіон проміжної глибини

$$I_\varepsilon = \big\{\, x : \varepsilon_1 < HD(x) < \tfrac{1}{2} - \varepsilon_2 \big\}, \quad \varepsilon_1,\, \varepsilon_2 \to 0.$$

Ця множина виключає як екстремальні викиди, так і точки, близькі до медіани, забезпечуючи збалансовану основу для висновків на основі глибини.

Попередні теоретичні роботи встановлюють, що на всій області

$$\sup_{x\in\mathbb{R}^d} \big(HD_k(x) - HD(x)\big) = O\Big(\big(\tfrac{\ln k}{k}\big)^{\frac{2}{d-1}}\Big) \quad (7).$$

і що досягнення фіксованої точності на I_ε вимагає експоненційних витрат за d:

$$k \gtrsim \exp(c\,d) \quad [1.$$

У наших експериментах ми використовуємо максимум, відбираючи скінченну множину з N_{test} точок у I_ε та вимірюючи

$$\max_{i=1,\dots,N_{test}} \big(HD_k(x_i) - HD(x_i)\big).$$

Цей максимум над тестовою вибіркою апроксимує теоретичний супремум і забезпечує прості емпіричні порівняння.

3.2 Збіжність за мірою Гаусса P

3.2.1 Вибірка точок за глибиною

Пригадаємо формули глибини Тьюкі для стандартного двовимірного нормального розподілу. Якщо

$$X \sim \mathcal{N}(0, I_2), \quad x = (x_1, x_2) \in \mathbb{R}^2,$$

то (див. Rousseeuw & Ruts (8))

$$HD(x) = \inf_{u \in S^1} P\big(\langle u, X\rangle \le \langle u, x\rangle\big) = 1 - \Phi\big(\|x\|_2\big).$$

У загальному випадку $X \sim \mathcal{N}(0, \Sigma)$ евклідову норму замінюють на норму Махаланобіса

$$\|x\|_\Sigma = \sqrt{x^\top \Sigma^{-1} x},$$

тому

$$HD(x) = 1 - \Phi\big(\|x\|_\Sigma\big).$$

Для генерації точок, чия глибина Тьюкі належить до $\psi \in (\varepsilon_1,\, 0.5 - \varepsilon_2)$, виконуємо:

1. Вибираємо $\psi \sim \mathrm{Unif}(\varepsilon_1,\, 0.5 - \varepsilon_2)$.
2. Обчислюємо $r = \Phi^{-1}(1 - \psi)$.
3. Вибираємо $Z \sim \mathcal{N}(0, I_d)$, задаємо $u = Z/\|Z\|_2$.
4. Формуємо $z = r\,u$, а потім $x = L\,z$, де $LL^\top = \Sigma$.

Ця схема вибірки на основі глибини відрізняється від підходу з істинною розмірністю в (7), оскільки вона навмисно націлена на велику кількість «проблемних» точок проміжної глибини (1). Була отримана густина радіусу Махаланобіса $R = \|X\|_\Sigma$

$$f_R(r) = \frac{\phi(r)}{\Phi(r_{\max}) - \Phi(r_{\min})}\, \mathbf{1}\{r_{\min} \le r \le r_{\max}\},$$

де $\phi(r)$ — щільність стандартного нормального розподілу, а $r_{\min} = \Phi^{-1}(1 - \varepsilon_1)$, $r_{\max} = \Phi^{-1}(1 - 0.5 + \varepsilon_2)$. Отже, вибірка за глибиною концентрує точки ближче до центру (менші $\|x\|_\Sigma$) і недостатньо представляє справжні гауссові хвости. Але по значенню глибини ψ точки будуть представлені рівномірно.

3.2.2 Дизайн експерименту

В ході кожної з $N_{\text{iter}} = 100$ симуляцій ми вводимо параметризацію, подібно до (7), за допомогою

$$\{N_{\text{iter}},\ N_{\text{test}},\ k,\ \varepsilon_1,\ \varepsilon_2,\ d,\ \Sigma,\ \text{seed}\}.$$

1. Згенерувати N_{test} точок з I_ε.
2. Вибрати k напрямків на S^{d-1}.
3. Обчислити точну глибину $HD(x_i)$ та рандомізовану глибину $HD_k(x_i)$ для кожної точки.

4. Виміряти $\max_i |HD_k(x_i) - HD(x_i)|$.

Наші емпіричні результати (Таблиця 3.1, Таблиця 3.2, Рисунок 3.1) підтверджують, що помилки для [0, 0.5] та для проміжної глибини відрізняються менш ніж на 3%. Для $k \in \{10^2, 10^3, 10^4\}$ логарифмічно-логарифмічний графік медіанної помилки дає два локальні нахили:

$$\hat{\beta}_{100\to1000} \approx -0.93, \quad \hat{\beta}_{1000\to10000} \approx -0.60,$$

що дають середній емпіричний показник $\beta_{emp} \approx -0.76$. Таким чином ми фіксуємо значно швидше спадання похибки, ніж асимптотична оцінка $\beta = -\frac{2}{d-1} = -0.222$ для $d = 10$, отримане з обмеження

$$\sup_i (HD_k(x_i) - HD(x_i)) = O\big((\ln k/k)^{2/(d-1)}\big).$$

Таким чином, наші дані перебувають у преасимптотичному режимі: помилка все ще зменшується швидше, ніж передбачає остаточний закон $(\ln k/k)^{2/(d-1)}$, і буде співмірною лише тоді, коли k стане експоненціально великим відносно d.

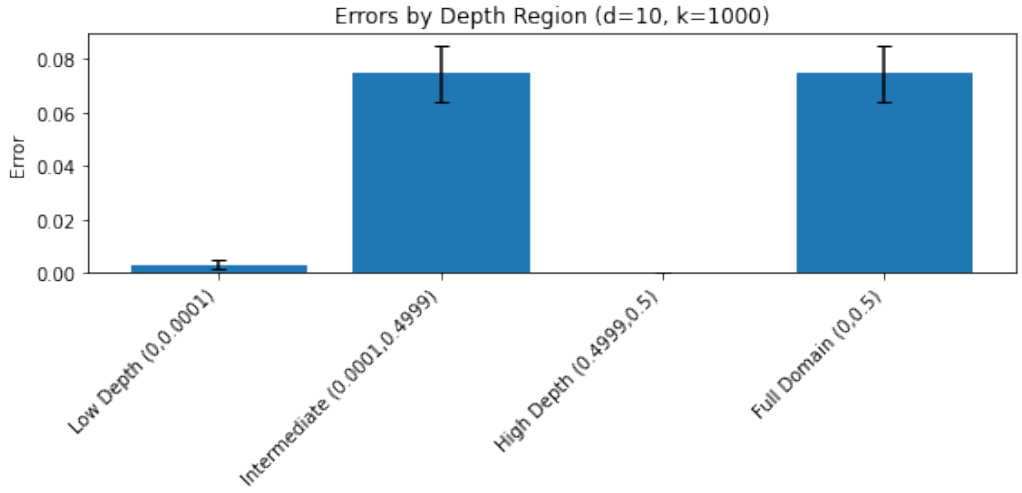

Рис. 3.1: Середнє значення та стандартна похибка для введеної похибки-максимума $max_i\big(HD_k(x_i) - HD(x_i)\big)$ для $N_{\text{iter}} = 100$ повторень, розподілені за регіонами глибини (Низька, Проміжна, Висока та увесь проміжок [0, 0.5]) для $d = 10$, $k_{\text{dirs}} = 1000$, $N_{\text{test}} = 500$.

Табл. 3.1: Медіанне значення $max_i\big(HD_k(x_i) - HD(x_i)\big)$ залежно від розмірності простору $d \in \{2, 5, 10, 20\}$, порівняння двох стратегій вибірки — увесь проміжок ($\psi \in (0, 0.5)$) та Проміжна Глибина ($\psi \in (10^{-4}, 0.5 - 10^{-4})$) — обчислено за $N_{\text{iter}} = 100$ повторень, кожне з $N_{\text{test}} = 500$ точками та $k_{\text{dirs}} = 1000$ випадковими напрямками, використовуючи аналітичні гауссові формули для HD і HD_k.

d	Full Median	Int. Median	Theory	Δ Full (%)	Δ Int (%)
2	0.000	0.000	0.000	-76.136	-76.128
5	0.021	0.021	0.040	-46.619	-46.614
10	0.074	0.074	0.122	-38.993	-38.996
20	0.139	0.139	0.219	-36.392	-36.392

Табл. 3.2: Порівняння медіани та 90-го перцентиля $max_i\big(HD_k(x_i) - HD(x_i)\big)$ для різної кількості випадкових напрямків $k \in \{100, 1000, 10000\}$ у Проміжній Глибині ($\psi \in (10^{-4},\, 0.5 - 10^{-4})$), оцінено при розмірності простору $d = 10$ з $N_{\text{test}} = 500$ та $N_{\text{iter}} = 100$. Стовпці містять змодельовану медіанну похибку, змодельовану 90-перцентильну похибку, теоретичну межу похибки та відносну різницю між змодельованими та теоретичними значеннями.

k	Median Error	90th Percentile	Theory Error	Diff (%)
100	0.034	0.065	0.270	-87.310
1000	0.004	0.005	0.122	-96.936
10000	0.001	0.001	0.065	-98.719

3.2.3 Адаптивний вибір випадкових напрямків

У цьому дослідженні ми провели два експерименти, щоб поглибити наше розуміння того, як необхідна кількість випадкових напрямків k^* залежить як від розмірності простору, від бажаної точності та для заданого рівня глибини.

Опис алгоритму Алгоритм залишається таким же, як і раніше: для кожної точки x з істинною глибиною Тьюкі ψ, починаємо з $k = k_0$, вибираємо $u_i \sim \text{Unif}(S^{d-1})$, обчислюємо $HD_k(x)$ і подвоюємо k, доки

$$HD_k(x) - \psi \leq \varepsilon \quad \text{або} \quad k = k_{\max}.$$

Експеримент 1 (Варіювання розмірності) Ми фіксуємо $\varepsilon = 10^{-2}$, $k_0 = 32$, $k_{\max} = 2^{18}$, і вибираємо $N_{\text{per}} = 1000$ точок у кожному з п'яти інтервалів глибини:

$$(0, 10^{-4}),\ [10^{-4}, 0.1),\ [0.1, 0.3),\ [0.3, 0.5 - 10^{-4}),\ (0.5 - 10^{-4}, 0.5).$$

Для кожного $d \in \{2, 4, 6, 8, 10\}$, ми рівномірно вибираємо ψ в інтервалі, перетворюємо на радіуси через $\Phi^{-1}(1 - \psi)$, генеруємо $x \sim \mathcal{N}(0, I_d)$ і обчислюємо медіанне значення $k^*(\psi, d)$.

Експеримент 2 (Варіювання допустимого рівня похибки) Ми фіксуємо $d = 5$, $k_0 = 32$, $k_{\max} = 2^{18}$, $N_{\text{per}} = 1000$, і виконуємо те саме групування на інтервали, змінюючи лише

$$\varepsilon \in \{5 \times 10^{-3}, 3.775 \times 10^{-3}, 2.55 \times 10^{-3}, 1.325 \times 10^{-3}, 10^{-4}\}.$$

Як і в попередньому експерименті ми записуємо лише медіанне значення k^* у кожному інтервалі.

3.2.4 Результати

Малюнок 3.2 показує, що середні затрати на вибір k^* залишаються нижчими для крайніх інтервалів $(0, 10^{-4})$ та $(0.5 - 10^{-4}, 0.5)$ для всіх d, порівняно з інтервалами проміжної глибини. Примітно, що навіть «легкі для обчислення» інтервали демонструють експоненціальну тенденцію в d, хоча й зі зменшою швидкістю.

Таблиця 3.3 кількісно описує цю поведінку:

- Інтервали низької та високої глибини зберігають медіанне k^* на рівні 10^1–10^2, навіть коли d зростає.

- Інтервали проміжної глибини демонструють чітке експоненціальне зростання медіанного k^* з d, що підтверджує, що ці точки створюватимуть найбільшу похибку при обчисленнях.

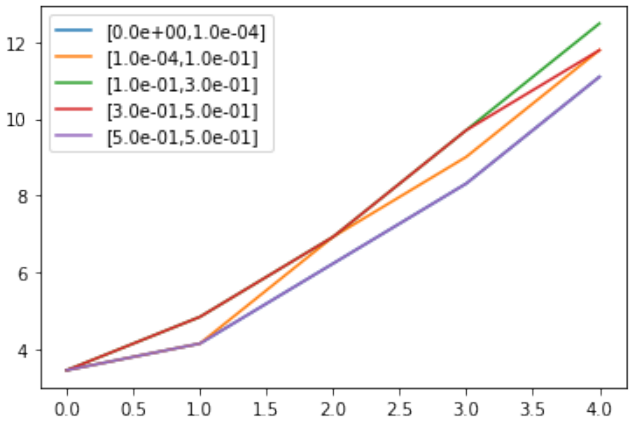

Рис. 3.2: Експеримент 1: медіанне $k^*(\psi, d)$ на логарифмічній шкалі залежно від інтервалу глибини для $d \in \{2, 4, 6, 8, 10\}$ при $\varepsilon = 10^{-2}$, $N_{\text{per}} = 1000$.

Табл. 3.3: Експеримент 1: медіанне $k^*(\psi, d)$ за інтервалами глибини та розмірністю простору d.

d	$(0, 10^{-4})$	$(10^{-4}, 0.1)$	$(0.1, 0.3)$	$(0.3, 0.5 - 10^{-4})$	$(0.5 - 10^{-4}, 0.5)$
2	32	32	32	32	32
4	64	64	128	128	64
6	512	1 024	1 024	1 024	512
8	4 096	8 192	16 384	16 384	4 096
10	65 536	131 072	262 144	131 072	65 536

Результати експерименту 2 показано у таблиці 3.4:

- Зменшення ε теж експоненційно підвищує медіанне k^* у всіх інтервалах.
- Розрив між «легкими» (низька/висока глибина) та «складними» (проміжна глибина) інтервалами зменшується, коли ε стає більш жорстким.
- У граничному випадку $\varepsilon \to 0$, відмінності, специфічні для глибини, зникають, і всі інтервали сходяться до однаково великих значень k^*, що вказує на те, що надзвичайно висока точність часто нівелює обчислювальні переваги для інтервалів низької/високої глибини.

Ці розширені результати підтверджують наш головний висновок: точки проміжної глибини спричиняють найшвидше зростання обчислювальних витрат із d, але за більш жорстких допусків навіть найбільш «швидкі» точки вимагають великої кількості напрямків, що нівелює перевагу, засновану на глибині.

Табл. 3.4: Експеримент 2: медіанне $k^*(\psi, d)$ за інтервалами глибини та допустимим рівнем похибки ϵ.

$(0, 1e-4)$	(1e-4, 0.1)	(0.1, 0.3)	(0.3, 0.5-1e-4)	(0.5-1e-4, 0.5)
512.000	1024.000	2048.000	1024.000	512.000
1024.000	2048.000	2048.000	2048.000	1024.000
2048.000	4096.000	8192.000	4096.000	2048.000
8192.000	8192.000	16384.000	16384.000	8192.000
262144.000	262144.000	262144.000	262144.000	262144.000

3.2.5 Короткий висновок

Адаптивна схема допомагає виділити витрати на апроксимацію глибини Тьюкі в різних інтервалах глибини. Наші експерименти показують:

- Хвостові та глибокі точки: припускалося, що вони потребують лише поліноміально малу кількість напрямків, але в ході експерименту це не вдалося підтвердити — було виявлено експоненціальне зростання k.
- Точки проміжної глибини вимагають k^*, яке теж зростає експоненціально з розмірністю d.
- Вибір наступної гібридної реалізації — запуск невеликого рівномірного k, а потім адаптивне підсилення лише для проміжних точок буде забезпечувати майже оптимальну помилку.

3.3 Збіжність на гауссівських вибірках

Налаштування. Нехай $X_{1:n} \overset{\text{iid}}{\sim} \mathcal{N}_d(0, I_d)$. Для будь-якої точки $x \in \mathbb{R}^d$ ми вибираємо k випадкових напрямків $u_1, \dots, u_k \overset{\text{iid}}{\sim} (\mathbb{S}^{d-1})$ і обчислюємо випадкову глибину Тьюкі точки x відносно вибірки:

$$HD_k^{(n)}(x) = \min_{1 \le j \le k} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\big\{\langle u_j, X_i \rangle \le \langle u_j, x \rangle\big\}.$$

Тут $\mathbf{1}(\cdot)$ — це індикаторна функція.

Глибина обчислюється точно на вибірці за допомогою алгоритму без напрямків, описаного в (4), який розв'язує задачу глибини напівпростору з машинною точністю. Ми позначаємо це як $HD^{(n)}(x)$.

Паралельно ми також підраховуємо теоретичну гауссівську глибину

$$HD(x) = 1 - \Phi\big(\|x\|_2\big),$$

щоб ми могли порівняти:

$$\Delta_{\text{emp}}(x) = \big|HD_k^{(n)}(x) - HD^{(n)}(x)\big|, \qquad \Delta_{\text{th}}(x) = \big|HD_k^{(n)}(x) - HD(x)\big|.$$

Для кожної конфігурації (d, n, k) ми проводимо $N_{\text{iter}=10}$ незалежних випробувань.

Етап 1: вплив n та k для $d = 2, 3, 4$

Ми провели два варіювання параметрів:

- Варіювання напрямків: змінюємо $k \in \{50, 100, 200, 400\}$, залишаючи $n = 1000$ фіксованим (Рис. 3.3);
- Варіювання розміру вибірки: змінюємо $n \in \{250, 500, 1000, 2000\}$, залишаючи $k = 250$ фіксованим (Рис. 3.4).

Усі результати представлені на логарифмічно-логарифмічних осях; пунктирна блакитна лінія — теоретична похибка, пунктирна червона — емпірична.

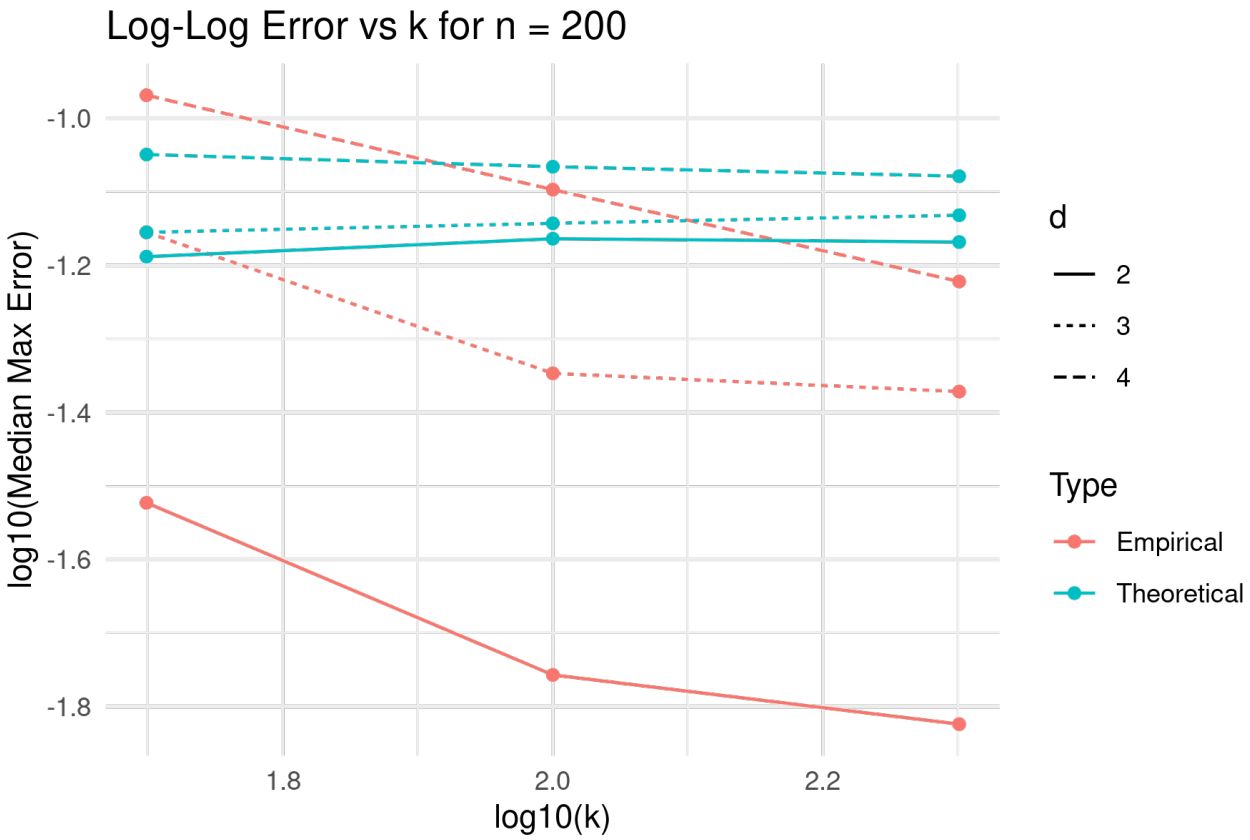

Рис. 3.3: Етап 1a: похибка залежно від кількості напрямків k (логарифмічно-логарифмічна шкала) при $n = 1000$

Варіювання напрямків. Для всіх d емпірична похибка стабільно зменшується з ростом k, але теоретична похибка залишається на одному і тому ж рівні (Малюнок 3.3). При $d = 4$ червона крива при $k = 400$ становить приблизно третину від блакитної. Лінійні наближення на логарифмічно-логарифмічній шкалі для $d = 4$ дають локальні нахили

$$\hat{\beta}_{50\to100} \approx -0.56, \quad \hat{\beta}_{100\to200} \approx -0.66, \quad \hat{\beta}_{200\to400} \approx -0.60,$$

тобто $\beta_{\text{emp}} \approx -0.61$ (близько до $-2/(d-1) = -2/3 \approx -0.667$), що підтверджує преасимптотичну узгодженість теоретичної межі.

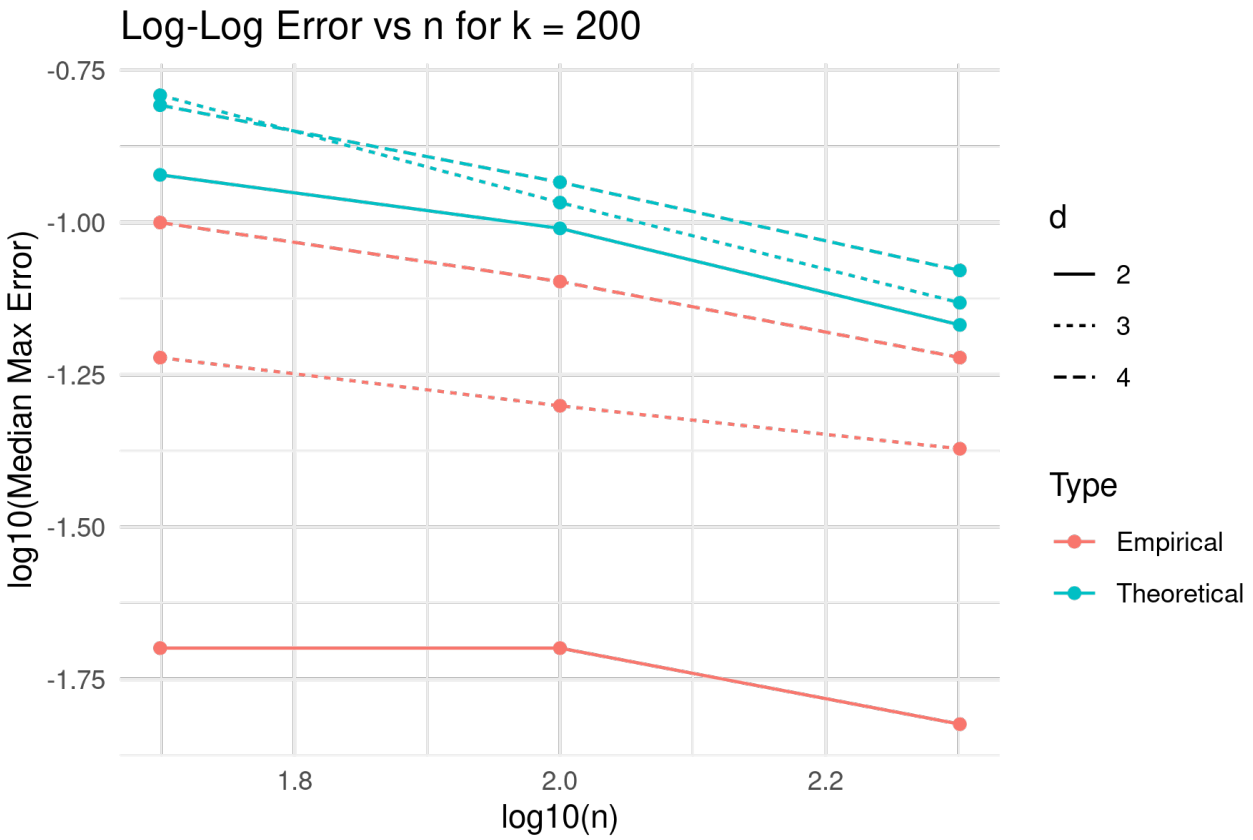

Рис. 3.4: Етап 1b: похибка залежно від розміру вибірки n (логарифмічно-логарифмічна шкала) при $k = 250$.

Варіювання розміру вибірки. При фіксованому $k = 250$ і $n \in \{250, 500, 1000, 2000\}$ (див. Мал. 3.4). Апроксимація логарифмічно-логарифмічних ліній дає $\beta_{\text{emp}}(d) \approx (-0.48, -0.45, -0.50)$ для $d = 2, 3, 4$. Ці результати підтверджують, що збільшення n ефективно зменшує шум вибірки, але при фіксованому k залишкова помилка напрямків зрештою обмежує загальну точність.

Етап 2: похибки, розподілені за глибиною, при $d = 4$

Фіксуємо $(d, k, n) = (4, 250, 1000)$ і використовуємо ті самі $m = 100$ тестових точок. Для кожної точки ми обчислюємо емпіричну та теоретичну точкові похибки, як визначено вище. Малюнок 3.5 показує ці похибки залежно від теоретичної глибини $HD(x)$.

Спостерігаємо дві типові поведінки для точок:

1. Стабільність емпіричної похибки. Сині точки залишаються в межах $[0.05, 0.10]$; мілкі ($HD \approx 10^{-4}$) та центральні ($HD \approx 0.5 - 10^{-4}$) відрізняються лише приблизно на 40%. Оскільки обидві глибини використовують ту саму вибірку, спільний шум зберігає різницю малою.

2. Зростання теоретичної похибки. Червоні точки зростають від 0.05 при низькій глибині до 0.15 поблизу $HD = 0.5$. З лише $k = 250$ напрямками в $d = 4$, багато критичних гіперплощин для центральних точок пропускаються, що спричиняє систематичне недооцінювання $HD(x)$.

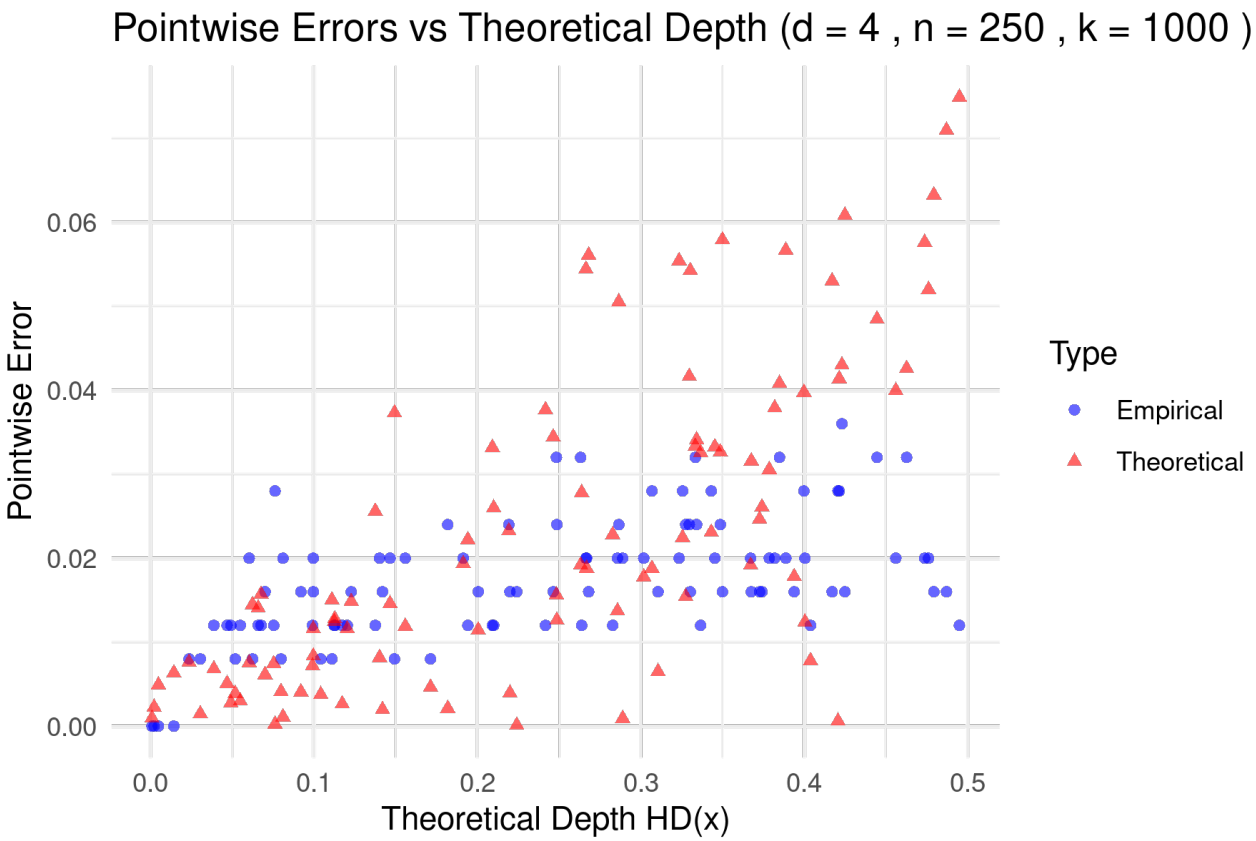

Рис. 3.5: Етап 2: точкова похибка залежно від теоретичної глибини для $(4, 250, 1000)$. Сині точки - емпірична похибка; червоні трикутники - теоретична

Висновки для дослідження емпіричного Гаусса.

1. Емпірична оцінка завжди перевершує теоретичну у наших експериментах.
2. Збільшення вибірки зменшує похибку зі швидкістю $n^{-1/2}$, але не може повністю компенсувати зростання розмірності при $d \geq 4$.
3. Регіони проміжної глибини залишаються найскладнішими для оцінки: теоретична похибка зростає втричі, тоді як емпірична залишається стабільною.

Розділ 4

Висновки

Ця дипломна робота дослідила глибина Тьюкі з теоретичної та обчислювальної перспектив, розкриваючи її властивості, явні форми та методи апроксимації через три основні розділи.

Перший розділ заклав теоретичну основу, надаючи формальне визначення глибини Тьюкі та аналізуючи її ключові властивості, описані в літературі [8] . Ескізи доведень сприяли кращому розумінню цих властивостей. Для трьох ймовірнісних розподілів — гауссівського розподілу, розподілу Коші та рівномірного розподілу на квадраті — було виведено явні вирази функції глибини у заданій точці. Контури глибини для кожного з цих розподілів було описано та візуалізовано.

Другий розділ розширив ці результати; ми обчислили аналітичні вирази функції глибини для рівномірного розподілу на кулі та α-стійких розподілів. Ці вирази супроводжувалися доведеннями та відповідними візуалізаціями для кращого розуміння. У випадках, коли аналітичні рішення були неможливі, зокрема для двовимірного експоненційного розподілу з незалежними координатами та двовимірного t-розподілу, застосовувалися чисельні апроксимації для оцінки контурів глибини. Було детально описано послідовність дій, надано рекомендації щодо вибору оптимальних напрямків і аналізу асимптотичної поведінки, а також представлено візуалізації апроксимованих контурів.

Третій розділ зосередився на рандомізованому методі апроксимації глибини Тьюкі, порівнюючи його з точним обчисленням із практичної точки зору. Особливу увагу приділено інтервалу проміжної глибини. Дослідження гауссівського розподілу з теоретичною мірою P та емпіричним гаусівським розподілом показало співмірні коефіцієнти збіжності похибок (-0.76 проти -0.61), попри різні розмірності ($d = 10$ та $d = 4$). Емпіричний аналіз виявив парадокс: більшість точок мають низьку глибину, що може бути пов'язано з формою опуклої оболонки точок, на яких будується глибина. Адаптивний метод підбору кількості напрямків k для гауссівського розподілу з теоретичною мірою P показав, що для всіх інтервалів глибини (малої, високої, проміжної) потрібна експоненційна кількість напрямків для досягнення якісної апроксимації. При заданому малому рівні похибки $\epsilon \to 0$ різниця між інтервалами зникає, і кількість напрямків стає однаковою для всіх випадків.

Ця робота внесла вклад у робастну статистику, запропонувавши нові інструменти для аналізу багатовимірних даних. Подальші дослідження можуть бути спрямовані на вирішення обчислювальних викликів у високих розмірностях та вдосконалення алгоритмічних методів.

Бібліографія

[1] Simon Briend, Gábor Lugosi, and Roberto I. Oliveira. On the quality of randomized approximations of tukey's depth, 2023. URL: https://arxiv.org/abs/2309.05657, arXiv:2309.05657.

[2] J. A. Cuesta-Albertos and A. Nieto-Reyes. The random tukey depth, 2007. URL: https://arxiv.org/abs/0707.0167, arXiv:0707.0167.

[3] Ludwig Danzer, Branko Grünbaum, and Victor Klee. Helly's theorem and its relatives, 2013. URL: https://www3.nd.edu/~andyp/teaching/2013SpringMath366/DanzerGruenbaumKlee.pdf.

[4] Rainer Dyckerhoff and Pavlo Mozharovskyi. Exact computation of the halfspace depth, 2016. URL: https://arxiv.org/abs/1411.6927, arXiv:1411.6927.

[5] J. Gil-Pelaez. Note on the inversion theorem. 1951. doi:10.1093/biomet/38.3-4.481.

[6] Stanislav Nagy. Halfspace Depth: Theory and Computation. PhD thesis, 2022. URL: https://www.mff.cuni.cz/cs/vedecka-rada/nagy-thesisshort.pdf.

[7] Stanislav Nagy, Rainer Dyckerhoff, and Pavlo Mozharovskyi. Uniform convergence rates for the approximated halfspace and projection depth. 2020. arXiv:1910.05956, doi:10.1214/20-EJS1759.

[8] Peter J. Rousseeuw and Ida Ruts. The depth function of a population distribution. 1999. doi:10.1007/PL00020903.

[9] Harold Ruben. On the distribution of the weighted difference of two independent student variables. 1960. doi:10.1111/j.2517-6161.1960.tb00365.x.