Korkeaulotteisten äärellisten joukkojen upotukset

Timi Turpeinen

Perustieteiden korkeakoulu

Kandidaatintyö
Espoo 23.8.2024

Vastuuopettaja

FT Pekka Alestalo

Työn ohjaaja

FT Pekka Alestalo

Aalto-yliopisto
Perustieteiden
korkeakoulu

Aalto-yliopisto
Perustieteiden korkeakoulu

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Tekniikan kandidaatintyön tiivistelmä

Tekijä Timi Turpeinen

Työn nimi Korkeaulotteisten äärellisten joukkojen upotukset

Koulutusohjelma Teknistieteellinen kandidaattiohjelma

Pääaine Matematiikka ja systeemitieteet **Pääaineen koodi** SCI3029

Vastuuopettaja FT Pekka Alestalo

Työn ohjaaja FT Pekka Alestalo

Päivämäärä 23.8.2024 **Sivumäärä** 31+2 **Kieli** Suomi

Tiivistelmä

Tässä työssä johdetaan Johnson-Lindenstrauss-teoreema sekä käydään läpi kyseisen teoreeman sovelluksia. Teoreema käsittelee korkeaulotteisten pisteiden upottamista pienempään ulottuvuuteen, niin että pisteiden väliset etäisyydet pysyvät likimääräisesti samana. Teoreeman käyttötarkoituksia löytyy monessa laskennallisessa menetelmässä.

Teoreeman johtamiseksi tarvitaan taustatietoa n-ulotteisten joukkojen tilavuuksien käsitteistä, mittateoriasta, metrisistä avaruuksista, topologiasta ja todennäköisyyslaskennasta. Päätuloksen johtaminen aloitetaan käymällä aluksi läpi Brunn-Minkowski-epäyhtälö, joka on voimassa kaikille kompakteille joukoille. Kyseistä epäyhtälöä voidaan hyödyntää yleisen isoperimetrisen epäyhtälön osoittamisessa. Isoperimetrinen epäyhtälö on puolestaan hyödyllinen tarkasteltaessa mitan keskittymistä pallon pinnalla. Työssä sovelletaan mitan keskittymistä todistettaessa Lévyn apulausetta, joka pätee yleisesti 1-Lipschitz-funktioille. Projisointifunktioiden ominaisuuksista johtuen Lévyn apulausetta hyödynnetään päälauseen todistuksessa. Lopuksi käsitellään teoreeman tarkoitusta ja sen hyödyllisiä sovelluksia laskennallisissa menetelmissä.

Avainsanat Metriset avaruudet, Upotukset, Mitan keskittyminen, Isoperimetrinen epäyhtälö, Analyysi

Aalto-yliopisto
Perustieteiden korkeakoulu

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the bachelor's thesis

Author Timi Turpeinen

Title Thesis template

Degree programme Bachelor's Programme in Science and Technology

Major Mathematics and Systems Sciences **Code of major** SCI3029

Teacher in charge PhD Pekka Alestalo

Advisor PhD Pekka Alestalo

Date 23.8.2024 **Number of pages** 31+2 **Language** Finnish

Abstract

In this thesis, we derive the Johnson-Lindenstrauss theorem and explore its applications. The theorem addresses the embedding of high-dimensional points into lower-dimensional spaces while preserving distances, a process known as high-dimensional reduction. The theorem also has numerous applications in various computational methods.

Deriving the theorem requires background knowledge in the concepts of volumes of n-dimensional sets, measure theory, metric spaces, topology, and probability theory. The derivation of the main result begins with the Brunn-Minkwoski inequality, which holds for all compact sets. This inequality can be used to prove the general isoperimetric inequality. The isoperimetric inequality is, in turn, useful for examining the measure concentration on the sphere. In this thesis, the measure concentration is used to prove Lévy's lemma which generally applies to every 1-Lipschitz function. As a result of the properties of projection functions, Lévy's lemma can be used in the proof of the main theorem. Finally, in the end we discuss the purpose and useful applications of the main theorem in numerical methods.

Keywords Metric spaces, Embeddings, Measure concentration, Isoperimetric inequality, Analysis

Sisällys

1 Johdanto

Tässä työssä on tarkoitus johtaa likimääräisiin upotuksiin liittyvä Johnson-Lindenstrauss-teoreema. Teoreeman johtaminen perustuu suurimmalta osin kirjaan *Lectures on Discrete Geometry* - Jiří Matoušek [9].

Johnson-Lindenstrauss-teoreema käsittelee korkeaulotteisessa euklidisessa avaruudessa $\mathbb{R}^n$ olevien äärellisen N:n pisteen joukon upotusta pienempään ulottuvuuteen $\mathbb{R}^k$, jossa $k < n$. Teoreema tulkitaan seuraavasti [9, s. 358]:

"*Olkoon X äärellinen N:n pisteen joukko euklidisessa avaruudessa $\mathbb{R}^n$ ja $\epsilon \in (0, 1]$ annettu. Tällöin joukolle X on olemassa $(1+\epsilon)$-upotus avaruuteen $\mathbb{R}^k$, jossa* $k = O(\epsilon^{-2} \ln(N))$."

Väittämän mukaisesti äärellinen määrä pisteitä voidaan upottaa korkeasta ulottuvuudesta pienempään ulottuvuuteen sallimalla ϵ suuruisen vääristymän pisteiden välisissä etäisyyksissä.

Teoreemaa voidaan hyödyntää tästä ominaisuudesta johtuen monissa käytännön sovelluksissa, jotka vievät laskennallisesti valtavasti aikaa suuren datamäärän takia. Teoreeman avulla voidaan pienentää tutkittavan datan määrää sallimalla vain pienen informaation menetyksen, mikä parantaa algoritmien tehokkuutta [9, 4, 12].

Työssä tarvitaan taustatietoa metrisistä avaruuksista, topologiasta, mittateoriasta sekä n-ulotteisten joukkojen tilavuuksista. Tämän jälkeen teoreeman johtaminen voidaan aloittaa Brunn-Minkowski-epäyhtälön todistamisesta [9, s. 297–301]. Epäyhtälö tarkoittaa, että kahden kompaktin joukon Minkowski-summan tilavuus on suurempi tai yhtäsuuri kuin joukkojen tilavuuksien summa kyseiseen ulottuvuuteen nähden. Epäyhtälön todistamisessa käytetään yleisen topologian ja mittateorian tunnettuja käsitteitä. Brunn-Minkowski-epäyhtälöä hyödynnetään yleisen isoperimetrisen epäyhtälön osoittamisessa [9, s. 333–334]. Isoperimetrinen epäyhtälö on hyödyllinen, kun arvioidaan mitan keskittymistä pallolla, missä saadaan määritetylle mitalle hyödyllinen epäyhtälö [9, s. 330–331, s. 334]. Määriteltyä mittaa voidaan pitää analogisena todennäköisyyteen. Todennäköisyysmitalle saatua epäyhtälöä voidaan hyödyntää vastaavasti todistettaessa Lévyn apulausetta, joka pätee kaikille 1-Lipschitz-funktiolle. Lévyn apulause väittää, että funktioiden arvot pallon pinnalla eroavat niiden keskiarvoista tietyllä todennäköisyydellä, joka on verrattain pieni [9, s. 337–338]. Lopuksi Lévyn apulausetta hyödynnetään Johnson-Lindenstrauss-teoreeman osoittamisessa, kun havaitaan, että yleisesti upotuksissa käytetyt projisointifunktiot ovat 1-Lipschitz-funktiota [9, s. 358–360].

Tässä työssä esitellään luvussa 2 tarvittavat taustatiedot. Teoreeman todistus vastaavasti aloitetaan luvussa 3 käsittelemällä ensiksi Brunn-Minkowski-epäyhtälö luvussa 3.1. Tätä epäyhtälöä käytetään hyödyksi luvussa 3.2 isoperimetrisen epäyhtälön todistamisessa. Luvussa 3.3 vastaavasti osoitetaan epäyhtälö mitan keskittymiselle pallolla, mikä on avuksi Lévyn apulauseen todistamisessa luvussa 3.4. Lopulta pääteoreeman todistaminen todistetaan luvussa 3.5 Lévyn apulausetta hyödyntäen. Teoreeman seurauksiin ja sovelluksiin keskitytään luvussa 4, jossa käsitellään teoreeman tarkoitusta sekä käytännön ongelmia, joihin teoreemaa on käytetty.

2 Taustatietoa

2.1 Mittateoria

Mittateoria on laaja matematiikan ala, joka koostuu yleisistä määritelmistä mitoille ja niiden sovelluksille. Mittojen tärkeys korostuu, kun tarkastellaan korkeaulotteisia kappaleita, jolloin tilavuuden määritteleminen Riemannin mitan avulla tuottaa hankaluuksia. Haasteet ilmenevät etenkin, kun integroidaan korkeaulotteisten joukkojen yli. Kyseinen ongelma on ratkaistu määrittelemällä uudenlainen mitta, joka tunnetaan Lebesguen mittana [6, s. 3]. Lebesguen mitta joukolle $A \subset \mathbb{R}^n$ yleisesti määritellään

$$\mu(A) = \inf\left\{\sum_{i=1}^{\infty} \mathrm{vol}(I_i) : \ A \subset \bigcup_{i=1}^{\infty} I_i\right\}, \tag{1}$$

missä I_i on n-ulotteinen särmiö

$$I_i = \{x \in \mathbb{R}^n : \ a_j \leq x_j \leq b_j,\ j = 1,...,n\} = [a_1,\, b_1] \times ... \times [a_n,\, b_n] \tag{2}$$

kaikilla arvoilla $i = 1, 2, 3, ...$ Särmiöiden (2) tilavuudet vastaavasti määritellään arvoina

$$\mathrm{vol}(I_i) = (b_1 - a_1)(b_2 - a_2) \cdots (b_n - a_n).$$

Lebesguen mitan määritelmä (1) voidaan ajatella olevan kaksivaiheinen: ensiksi otetaan summa numeroituvasta määrästä särmiöiden tilavuuksia, jotka peittävät joukon A kokonaan, ja sen jälkeen otetaan pienin mahdollinen joukko näitä särmiöitä, jotka toteuttavat halutun ehdon. Mittateoriassa [9, s. 1] yleisen mitan määritelmä on

Määritelmä 2.1 (**Mitan määritelmä**).
Kuvaus $\mu\colon \{A\colon A \subset X\} \to [0, \infty)$ on mitta avaruudessa X, jos

1. $\mu(\varnothing) = 0$,
2. $\mu(A) \leq \mu(B)$, kun $A \subset B \subset X$ ja
3. $\mu(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu(A_i)$.

Määritelmässä (2.1) toinen ehto tunnetaan monotonisuutena ja viimeinen numeroituvana subadditiivisuutena. Nämä ehdot pätevät myös yleisesti tilavuuksille. Viimeisessä ehdossa pätee yhtäsuuruus, jos tarkasteltavat joukot ovat mitallisia ja erillisiä [6, s. 13]. Tässä työssä käsitellään erityisesti mitallisia ja erillisiä joukkoja, joten viimeinen ehto muuttuu numeroituvaksi additiivisuudeksi:

$$\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i),$$

missä $A_i \cap A_j = \varnothing$ kaikilla $i \neq j$ ja joukot A_i, $i = 1, 2, ...$ ovat mitallisia. Mittateoriassa voidaan myös määrittää raja-arvo jonolle mitallisia joukkoja, jotka ovat monotonisia

toisiinsa nähden [6, s. 17]. Yleisesti ottaen, jos $A_1 \subset A_2 \subset ...$, niin kyseessä on kasvava suppeneminen, jolle pätee raja-arvo

$$\lim_{i\to\infty}\mu(A_i) = \mu(\bigcup_{i=1}^{\infty} A_i), \tag{3}$$

eli mitan raja-arvo kyseisistä mitallisista joukoista on yhtäsuuri kuin joukkojen yhdistelmän mitta. Vastaavasti, jos $A_1 \supset A_2 \supset ...$ ja $\mu(A_j) < \infty$ jollain indeksillä j, niin

$$\lim_{i\to\infty}\mu(A_i) = \mu(\bigcap_{i=1}^{\infty} A_i), \tag{4}$$

jota kutsutaan laskevaksi suppenemiseksi.

Lopuksi mainitaan hyödyllisiä ominaisuuksia Lebesguen mitalle euklidisessa avaruudessa:

1. Jos $A \subset \mathbb{R}^n$ on Lebesgue-mitallinen joukko ja $x \in \mathbb{R}^n$, niin $\mu(A + x) = \mu(A)$.
2. Jos $A \subset B \subset \mathbb{R}^n$ ovat mitallisia joukkoja, niin $\mu(B \setminus A) = \mu(B) - \mu(A)$.
3. Jos $A \subset \mathbb{R}^n$ on mitallinen joukko ja $t \in \mathbb{R}, n \in \mathbb{N}$, niin $\mu(tA) = t^n \cdot \mu(A)$

[6, s. 51–52]. Työssä määritellään n-ulottuvuudessa olevan joukon tilavuus Lebesguen mitan kautta, eli joukon $A \subset \mathbb{R}^n$ tilavuutena voidaan pitää $\mathrm{vol}(A) = \mu(A)$. Tällöin Lebesguen mitan ominaisuudet voidaan pitää totena mitallisten joukkojen tilavuuksille.

2.2 n-ulotteinen tilavuus

Erityistä huomiota n-ulottuvuudessa saavat n-ulotteiset kuulat ja niiden säännönmukaisuudet. Lähteen [1] mukaisesti määritellään n-ulottuvuudessa olevan R-säteisen kuulan olevan

$$B_R^n = \{(x_1, ..., x_n) \in \mathbb{R}^n \mid \ x_1^2 + ... + x_n^2 \leq R^2\},$$

jonka tilavuus saadaan kaavalla

$$\mathrm{vol}(B_R^n) = \frac{2\pi^{\frac{n}{2}}}{n\Gamma\left(\frac{n}{2}\right)} R^n. \tag{5}$$

Nimittäjässä esiintyvä funktio Γ tunnetaan Eulerin gamma-funktiona:

$$\Gamma(x) = \int_0^\infty s^{x-1} e^{-s} \mathrm{d}s.$$

Eulerin gamma-funktiolle löytyy hyödylliset palautuskaavat

$$\Gamma(x+1) = x\Gamma(x) \text{ ja } \Gamma(n+1) = n!, \text{ kun } n \in \mathbb{N}.$$

Huomioon on otettava myös yksikkökuulan pinnan S^{n-1} pinta-ala $\mathrm{vol}(\partial B_1^n)$, joka saadaan määritettyä kaavalla

$$\mathrm{vol}(\partial B_1^n) = n \cdot \mathrm{vol}(B_1^n), \tag{6}$$

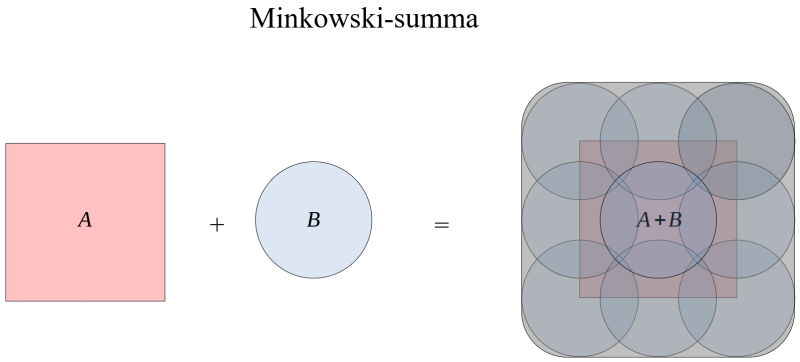
mikä voidaan johtaa hyödyntämällä polaarikoordinaatteja [1]. Kaavaa (6) tarkastelemalla saadaan määritettyä yleisesti R-säteisen kuulan pinnan pinta-ala derivoimalla säteen R suhteen kuulan tilavuutta, eli

$$\mathrm{vol}(\partial B_R^n) = \frac{d}{dR}(\mathrm{vol}(B_R^n)) = \lim_{t\to 0} \frac{\mathrm{vol}\,(B_R^n + tB_1^n) - \mathrm{vol}\,(B_R^n)}{t}. \tag{7}$$

Kahden joukon summa $A + B$ yhtälössä (7) määritellään asettamalla

$$A + B = \mathcal{A}(A \times B) = \{a + b \colon a \in A, b \in B\}, \tag{8}$$

mikä tunnetaan joukon A ja B Minkowski-summana [9, s. 297].



Kuva 1: Mikonwski-summa joukolle A ja B

Yksi tapa tulkita kyseinen summa on kuvan 1 mukaisesti pitämällä joukko A kiinnitettynä ja valitsemalla piste $b \in B$ ja kääntämällä joukko B erilaisiin mahdollisiin asentoihin pisteen b suhteen [9, s. 297]. Vastaavasti kyseistä joukkojen summaa voidaan ajatella yhtälön (8) tapaisesti jatkuvana funktiona $\mathcal{A} : X \times X \to X$, joka ottaa joukkojen $A, B \subset X$ karteesisen tulon ja kuvaantuu projisoimalla samaan avaruuteen X [7, s. 55].

Lisäksi voidaan määrittää jokaisen joukon $A \subset \mathbb{R}^n$ pinnan pinta-ala yhtälön (8) avulla, käyttämällä joukon t-ympäristöä, joka euklidisessa avaruudessa on määritelty

$$A_t = \{x \in \mathbb{R}^n \colon \ \mathrm{dist}(A, x) < t\} = A + B_t^n = A + tB_1^n, \tag{9}$$

missä $\mathrm{dist}(A, x) = \inf\{|x - y| : y \in A\}$, ja viimeinen yhtäsuuruus pätee skaalaamalla yksikkökuulaa säteen t verran. Yhtälöä (9) hyödyntämällä saadaan joukon A pinnan ∂A pinta-ala $\mathrm{vol}(\partial A)$ määriteltyä raja-arvona

$$\mathrm{vol}(\partial A) = \lim_{t\to 0} \frac{\mathrm{vol}(A_t) - \mathrm{vol}(A)}{t} = \lim_{t\to 0} \frac{\mathrm{vol}(A + tB_1^n) - \mathrm{vol}(A)}{t}. \tag{10}$$

Määritelmää (10) hyödynnetään tutkittaessa isoperimetristä epäyhtälöä pallon pinnalla.

2.3 Äärelliset metriset avaruudet ja upotukset

Käydään vielä lyhyesti läpi tarvittavia määritelmiä äärellisistä metrisistä avaruuksista ja upotuksista Jiří Matoušekin kirjaan [9, s. 355–357] nojautuen.

Jos joukko $X \subset \mathbb{R}^n$ on äärellinen N:n pisteen joukko, joukon X metriikka d voidaan määrittää $N \times N$ matriisilla, jossa jokainen elementti kuvaa kahden pisteen välistä etäisyyttä:

$$\begin{pmatrix} d\,(1,1) & d\,(2,\,1) & \dots & d\,(N,\,1) \\ d\,(1,\,2) & d\,(2,2) & & \vdots \\ \vdots & & \ddots & \vdots \\ d\,(1,\,N) & \dots & \dots & d\,(N,\,N) \end{pmatrix} \in \mathbb{R}^{N\times N}.$$

Toisaalta symmetrisyyden perusteella kahden pisteen väliset etäisyydet toistuvat kahdesti kyseisessä matriisissa. Etäisyydet voidaan siis kuvata $\binom{N}{2}$ positiivisella luvulla. Kyseisiä taulukoita käytetään esimerkiksi mikrobiologiassa [9, s. 355]. Tärkeää on huomata, että jos N on suuri luku, niin on vaikeaa nähdä minkälainen rakenne metriikalla d on. Tästä johtuen on hyödyllistä määritellä helpompi tapa kuvata kyseinen metrinen avaruus.

Eräs tapa on määrittää funktio $f\colon X \to \mathbb{R}$ kaikille arvoille $x \in X$ siten, että etäisyys kahden pisteen välillä on sama kuin funktion arvojen erotus, eli $|f(x) - f(y)| = d(x, y)$ kaikilla $x, y \in X$. Tällöin funktio f kuvaa hyvin kyseisen metrisen avaruuden ominaisuuksia, jolloin nähdään esimerkiksi tiukat ryhmittymät, tiiviit joukot ja eristetyt pisteet. Toinen etu on, että kyseinen metriikka d voidaan tällöin kuvata vain $2N$ positiivisella luvulla. Hyödyllisyydestä huolimatta nämä funktiot ovat erittäin harvinaisia ja vaikeasti löydettävissä [9, s. 355–356]. Jos tällaisia funktioita on, niitä yleisesti kutsutaan upotuksiksi.

Upotukset voidaan mieltää monella tavalla. Yleisen topologian perusteella upotukset määritellään homeomorfismin kautta [7, s. 26]. Injektiota $f\colon X \to Y$ sanotaan homeomorfismiksi, jos f ja sen käänteisfunktio f^{-1} ovat jatkuvia. Tällöin sanotaan, että avaruudet X ja Y ovat homeomorfisia ($X \approx Y$), mikä tarkoittaa, että jokaista pistettä avaruudessa X vastaa yksikäsitteisesti jokin piste avaruudessa Y. Upotus on erikoistapaus homeomorfisuudesta. Tällöin ei välttämättä ole olemassa homeomorfismin ehtoa täyttävää funktiota f, mutta sen sijaan löytyy ehdon täyttävä jatkuva injektio $f'\colon X \to Z$, jossa $Z \subset Y$. Avaruus X on siis homeomorfinen aliavaruuden Z kanssa. Intuitiivisesti funktio f' siis upottaa joukon X joukkoon Z, joka peittää osan avaruudesta Y. Tämä määritelmä voidaan ottaa yleiseksi määritelmäksi myös metrisissä avaruuksissa.

Metrisissä avaruuksissa upotus voidaan määritellä kuitenkin tarkemmin. Kuvaus $f\colon X \to Y$, jossa (X, d_X) ja (Y, d_Y) ovat metrisiä avaruuksia, on isometrinen upotus, jos se säilyttää etäisyydet. Lyhyesti:

$$d_Y(f(x), f(y)) = d_X(x, y)\ \forall x, y \in X.$$

Isometriset upotukset ovat harvinaisia, ja monissa sovelluksissa täydellinen etäisyyksien säilyminen on tarpeetonta kuten Johnson-Lindenstrauss-teoreemassa. Tä-

män takia on hyödyllistä tarkastella likimääräisiä upotuksia, jotka voidaan kuvata seuraavasti:

Määritelmä 2.2 (**T-upotus**).
Olkoot (X, d_X) ja (Y, d_Y) metrisiä avaruuksia ja $T \geq 1$. Kuvausta $f\colon X \to Y$ kutsutaan T-upotukseksi, jos on olemassa luku $r > 0$ siten, että

$$r \cdot d_X(x, y) \leq d_Y(f(x), f(y)) \leq T \cdot r \cdot d_X(x, y) \tag{11}$$

kaikilla $x, y \in X$. Suurinta alarajaa luvulle T, jolla f on T-upotus, kutsutaan f:n vääristymäksi.

Määritelmässä 2.2 luku r tunnetaan skaalautumisvakiona, josta voidaan päästä eroon euklidisessa avaruudessa kertomalla puolittain luvun käänteisluvulla $\frac{1}{r}$. Määritelmä 2.2 voidaan myös kuvata toisessa muodossa, jota käytetään tässä työssä. Puhutaan $(1+\epsilon)$-upotuksesta, jossa $\epsilon > 0$. Tällöin (11) kuvataan muodossa

$$(1-\epsilon) \cdot d_X(x, y) \leq d_Y(f(x), f(y)) \leq (1+\epsilon) \cdot d_X(x, y). \tag{12}$$

Voidaan havaita, että ehto (12) näyttää samalta kuin bilipschitz-funktion ehto. Seuraavassa osiossa määritellään kyseinen ehto ja muita työssä tarvittavia määritelmiä.

2.4 Määritelmiä

Määritelmissä käytetään metrisiä avaruuksia (X, d_X) ja (Y, d_Y), missä d_X on joukon X metriikka ja vastaavasti d_Y joukon Y metriikka.

Määritelmä 2.3 (**Lipschitz-funktiot**).
Funktio $f\colon X \to Y$ on Lipschitz-funktio, jos on olemassa $L > 0$ siten, että

$$d_Y(f(x), f(y)) \leq L \cdot d_X(x, y) \tag{13}$$

kaikilla $x \in X$ ja $y \in Y$.

Määritelmä 2.4 (**bilipschitz-funktiot**).
Funktio $f\colon X \to Y$ on bilipschitz-funktio, jos on olemassa $M \geq 1$ siten, että

$$\frac{1}{M} \cdot d_X(x, y) \leq d_Y(f(x), f(y)) \leq M \cdot d_X(x, y) \tag{14}$$

kaikilla $x \in X$ ja $y \in Y$.

Koska $1 \geq \frac{1}{M} > 0$ ja $M \geq 1$, voidaan havaita, että jokainen bilipshcitz-funktio on likimääräinen upotus. Lisäksi funktion f Lipschitz-ominaisuuden 2.3 avulla voidaan määrittää funktion f vääristymä 2.2 käyttämällä Lipschitz-vakiota ja käänteiskuvausta f^{-1}. Jos funktion f Lipschitz-normi määritellään olevan

$$||f||_{\text{Lip}} = \sup\left\{ \frac{d_Y(f(x),\, f(y))}{d_X(x,\, y)} :\ \ x,\, y \in X,\, x \neq y \right\}, \tag{15}$$

niin tällöin funktion f vääristymä on $||f||_{\text{Lip}} \cdot ||f^{-1}||_{\text{Lip}}$.

Määritelmä 2.5 (**funktion odotusarvo ja mediaani**).
Mitallinen reaalifunktio $f\colon S^{n-1} \to \mathbb{R}$ pallon pinnalla S^{n-1} voidaan tulkita satunnaismuuttujana, ja sen odotusarvo on

$$\mathbb{E}[f] = \int_{S^{n-1}} f(x)dP(x), \tag{16}$$

jossa P on normalisoitu mitta pallon pinnalla. Tämä määritellään myöhemmin luvussa 3.3. Vastaavasti funktion f mediaani on

$$\operatorname{med}(f) = \sup\{t \in \mathbb{R}\colon P[f \leq t] \leq \frac{1}{2}\}, \tag{17}$$

jossa $P[f \leq t] = P[x \in S^{n-1}\colon f(x) \leq t]$.

Työssä osoitetaan luvussa 3.4 Lipschitz-funktioiden olevan keskittyneitä mediaanin $\operatorname{med}(f)$ lähettyville. Vastaavasti samat tulokset voidaan osoittaa odotusarvolle $\mathbb{E}[f]$ [11, s. 109]. Tässä työssä mediaanille saatuja tuloksia voidaan siis pitää yhtäpitävinä odotusarvolle.

Määritelmä 2.6 (**tiilijoukko**).
Olkoon A joukko, joka koostuu eri mittaisista koordinaattiakselien suuntaisista n-ulotteisista särmiöistä $A_i, i = \{1, 2, ...\}$ siten, että särmiöiden sisukset ovat erillisiä, eli

$$\operatorname{int}(A_i) \cap \operatorname{int}(A_j) = \varnothing \tag{18}$$

kaikilla $i \neq j$. Kyseistä joukkoa kutsutaan tiilijoukoksi.

3 Menetelmät ja päätulokset

3.1 Brunn-Minkowski-epäyhtälö

Seuraavaksi käsitellään Brunn-Minkowski-epäyhtälö [9, s. 297–301]. Brunn-Minkowski-epäyhtälö muotoillaan seuraavasti:

Teoreema 3.1 (Brunn-Minkowski-epäyhtälö).
Olkoot A ja B epätyhjiä kompakteja joukkoja euklidisessa avaruudessa $\mathbb{R}^n$. Tällöin

$$\mathrm{vol}(A+B)^{1/n} \geq \mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n}.$$

Sanallisesti tulkittuna siis joukon $A+B$ tilavuus on vähintään joukon A ja B tilavuuksien summa suhteessa ulottuvuuteen n. Teoreema 3.1 voidaan todistaa helposti osoittamalla ensiksi, että kyseinen epäyhtälö riittää todistaa tiilijoukoille A' ja B'.

Apulause 3.2.
Jos Brunn-Minkowski-epäyhtälö pätee kaikille epätyhjille tiilijoukoille A', $B' \subset \mathbb{R}^n$, niin tällöin se pätee myös kaikille epätyhjille kompakteille joukoille A, $B \subset \mathbb{R}^n$.

Todistus (Apulause 3.2).
Käytetään mittateorian yleistä käsitettä laskevista suppenevista mitallisista joukoista (4). Olkoot siis A, $B \subset \mathbb{R}^n$ epätyhjiä ja kompakteja. Seuraavaksi otetaan indekseillä $k = 1, 2, \ldots$ suljetut akselien suuntaiset särmiöt C_k, joiden sivujen pituudet ovat 2^{-k} ja joiden keskipisteet ovat kohdissa $2^{-k}\mathbb{Z}^n$. Nämä särmiöt peittävät koko avaruuden $\mathbb{R}^n$ ja niillä on erilliset sisustat. Määritellään $A_k = \bigcup\{C_k : C_k \cap A \neq \varnothing\}$ ja $B_k = \bigcup\{C_k : C_k \cap A \neq \varnothing\}$, eli A_k ja B_k ovat yhdistelmiä kaikista kuutiosta C_k, jotka leikkaavat joukkoja A ja B. Joukot A_k ja B_k ovat siis tiilijoukkoja kaikilla $k \in \mathbb{Z}_+$.

Voidaan myös havaita, että jonot (A_k) ja (B_k) ovat molemmat laskevia suppenevia mitallisia joukkoja, joten $A = \bigcap_{k=1}^{\infty} A_k$ ja $B = \bigcap_{k=1}^{\infty} B_k$. Näin ollen yhtälön (4) perusteella saadaan

$$\mathrm{vol}(A) = \lim_{k\to\infty}\mathrm{vol}(A_k) \text{ ja } \mathrm{vol}(B) = \lim_{k\to\infty}\mathrm{vol}(B_k).$$

Oletetaan seuraavaksi, että $x \in A_k + B_k$ kaikilla k arvoilla. Valitaan pisteet $y_k \in A_k$ ja $z_k \in B_k$ siten, että $x = y_k + z_k$. Koska joukot A ja B ovat kompakteja[1], voidaan suppenevien osajonojen nojalla olettaa, että $\lim_{k\to\infty} y_k = y \in A$ ja $\lim_{k\to\infty} z_k = z \in B$. Tästä seuraa, että $x = y + z \in A + B$, joka osoittaa inkluusion $A + B \subset \bigcap_{k=1}^{\infty}(A_k + B_k)$.

[1] Työssä on tärkeätä myös muistaa määritelmä joukon kompaktiudelle. Lyhyesti kuvattuna joukko on peitekompakti, jos jokaiselle joukon avoimelle peitteelle löytyy äärellinen osapeite. Vastaavasti joukko on jonokompakti, jos jokaisella joukkoon sisältyvällä jonolla on suppeneva osajono. Metrisissä avaruuksissa peite- ja jonokompaktius ovat yhtäpitäviä, mutta yleisessä topologiassa kumpikaan ei seuraa toisesta. [8, 7]

Täten Lebesguen mitan monotonisuuden perusteella pätee $\lim_{k\to\infty}\mathrm{vol}(A_k+B_k) \leq \mathrm{vol}(A+B)$, jota hyödyntämällä saadaan:

$$\begin{aligned}\mathrm{vol}(A+B)^{1/n} &\geq \lim_{k\to\infty}\mathrm{vol}(A_k+B_k)^{1/n} \geq \lim_{k\to\infty}(\mathrm{vol}(A_k)^{1/n}+\mathrm{vol}(B_k)^{1/n}) \\ &= \mathrm{vol}(A)^{1/n}+\mathrm{vol}(B)^{1/n},\end{aligned}$$

missä toinen kohta seuraa Brunn-Minkowski-epäyhtälöstä tiilijoukoille. ■

Apulauseen 3.2 avulla voidaan todistaa Brunn-Minkowski-epäyhtälö käyttämällä tiilijoukkoja $A', B' \subset \mathbb{R}^n$ joukkojen A ja B sijaan. Hyödynnetään induktiotodistusta Brunn-Minkowski-epäyhtälön osoittamiseksi.

Todistus (Teoreema 3.1).
Olkoot A ja B tiilijoukkoja, jotka koostuvat yhteensä k määrästä n-ulotteisia särmiöitä.
Alkuaskel:
Jos $k = 2$, tällöin A, B ja $A+B$ ovat myös särmiöitä, missä joukon A särmät olkoot $x_1, x_2, ..., x_n$ ja joukon B olkoot $y_1, y_2, ..., y_n$. Särmiöiden (2) tilavuuksien määritelmien perusteella täytyy siis osoittaa

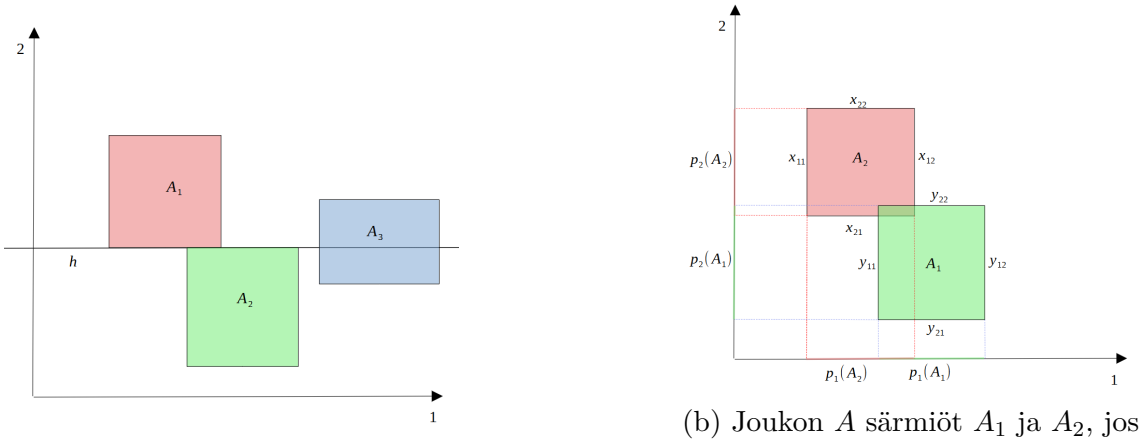
$$\Big(\prod_{i=1}^{n} x_i\Big)^{1/n} + \Big(\prod_{i=1}^{n} y_i\Big)^{1/n} \leq \Big(\prod_{i=1}^{n}(x_i+y_i)\Big)^{1/n} \tag{19}$$

kaikille $n \in \mathbb{Z}_+$. Epäyhtälö (19) seuraa aritmeettis-geometrisesta keskiarvojen epäyhtälöstä

$$\Big(\prod_{i=1}^{n} c_i\Big)^{1/n} \leq \frac{1}{n}\sum_{i=1}^{n} c_i, \tag{20}$$

mikä pätee kaikilla $c_i \in \mathbb{R}$ ja $n \in \mathbb{Z}_+$. Sijoittamalla arvot $\dfrac{x_i}{x_i+y_i}$ ja $\dfrac{y_i}{x_i+y_i}$ arvon c_i paikalle yhtälössä (20) ja ottamalla molempien epäyhtälöiden summan sekä kertomalla puolittain luvulla $\Big(\prod_{i=1}^{n}(x_i+y_i)\Big)^{1/n}$ saadaan yhtälö (19). Täten alkuaskel toteutuu kyseisessä tapauksessa.
Induktioaskel:
Olkoon nyt $k > 2$, ja oletetaan, että Brunn-Minkowski-epäyhtälö toteutuu kaikille A ja B -tiilijoukoille, joissa on yhteensä alle k särmiöitä. Olkoot joukossa A ja B yhteensä k särmiöitä, jotka valitaan siten, että joukossa A on vähintään kaksi särmiötä.

Etsitään seuraavaksi hypertaso h, joka jakaa joukon A siten, että hypertason molemmilla puolilla on ainakin yksi A:n kokonainen särmiö. Kuvataan tiilijoukon A särmiötä joukkoina A_i, jossa $i = \{1, 2, ..., k\}$.



(a) Joukon A särmiöitä.

(b) Joukon A särmiöt A_1 ja A_2, jos hypertasoa h ei löydy.

Kuva 2: Hypertason h poikkileikkaus joukon A särmiöihin nähden.

Kuvan 2a geometrian perusteella kyseisen hypertason h löytämiseksi täytyy vain osoittaa, että kahden särmiön välille löytyy niitä separoiva hypertaso. Osoitetaan väite hyödyntämällä ristiriitatodistusta. Täten olkoot kaksi särmiötä $A_1, A_2 \subset A$. Voidaan olettaa, että särmiö A_1 on särmiön A_2 vasemmalla puolella symmetrian perusteella. Merkitään särmiön A_1 särmiä arvoilla x_{ji}, jossa indeksi $i = 1, 2, ..., k$ kuvaa särmien akselin suuntaista ulottuvuutta ja $j = 1, 2$ särmää kyseisellä i:n arvolla siten, että $x_{1i} < x_{2i}$. Vastaavasti olkoon särmiön A_2 särmien arvot y_{ji}. Oletetaan, ettei kyseistä hypertasoa h löydy näiden kuutioiden välille. Kuvan 2b mukaisesti tästä seuraa, että särmiöiden välisillä särmillä pätee epäyhtälöt

$$x_{1i} < y_{1i} < x_{2i} \text{ kaikilla } i = \{1, ..., n\}.$$

Tästä seuraa, että jokaisen akselin suhteen tarkasteluna löytyy särmiöillä A_1 ja A_2 yhteinen jana, eli

$$p_i(\text{int}(A_1) \cap \text{int}(A_2)) \neq \varnothing \text{ kaikilla } i = \{1, ..., n\}, \tag{21}$$

missä $p_i : \mathbb{R}^n \to \mathbb{R}$ on projisointifunktio siten, että $p_i(z) = z_i$ jokaiselle $z \in \mathbb{R}^n$.

Ottamalla karteesisen tulon jokaisen akselin suuntaisen janan (21) suhteen saadaan särmiöiden sisuksille yhteinen alue $\text{int}(A_1) \cap \text{int}(A_2) \neq \varnothing$, mikä on ristiriita tiilijoukon A määritelmän 2.6 perusteella. Täten on olemassa hypertaso h, joka jakaa joukon A siten, että hypertason molemmilla puolilla on ainakin yksi A:n kokonainen särmiö.

Koordinaatiston valinnalla voidaan olettaa, että kyseinen hypertaso h on akseli $h = \{x_1 = 0\}$. Olkoon nyt A' joukko, joka on hypertason h positiivisella puolella $h^{\oplus} = \{x_1 > 0\}$, eli $A' = \overline{A \cap h^{\oplus}}$. Vastaavasti $A'' = \overline{A \cap h^{\ominus}} = \overline{A \cap \{x_1 < 0\}}$. Seuraavaksi, käännetään tiilijoukko B akselin x_1 suuntaiseksi siten, että hypertaso h jakaa joukon B samassa suhteessa kuin joukon A. Tämä on mahdollista, sillä kääntäminen ei vaikuta Brunn-Minkowski-epäyhtälön väittämään. Merkitään joukkoja B' ja B'' vastaavalla tavalla kuin joukon A osajoukkoja. Tilavuuksien suhteelle ρ saadaan nyt

yhtälö

$$\rho = \frac{\mathrm{vol}(A')}{\mathrm{vol}(A)} = \frac{\mathrm{vol}(B')}{\mathrm{vol}(B)}. \tag{22}$$

Koska $A'' = A \setminus A'$ ja $B'' = B \setminus B'$ saadaan yhtälön (22) perusteella myös

$$1 - \rho = \frac{\mathrm{vol}(A'')}{\mathrm{vol}(A)} = \frac{\mathrm{vol}(B'')}{\mathrm{vol}(B)} \tag{23}$$

käyttämällä mittateorian ominaisuuksia. Voidaan olettaa, että $\mathrm{vol}(A) \neq 0 \neq \mathrm{vol}(B)$, sillä muuten epäyhtälön pitävyys on selvä.

Koska osajoukoilla $A', A'' \subset A$ ja $B', B'' \subset B$ on vähemmän kuin k kuutiota, voidaan käyttää induktio-oletusta, ja koska $A' + B' \subset A + B \supset A'' + B''$ saadaan mitan monotonisuuden sekä yhtälöiden (22) ja (23) perusteella:

$$\begin{aligned}
\mathrm{vol}(A+B) &\geq \mathrm{vol}(A'+B') + \mathrm{vol}(A''+B'') \\
\text{(Induktio-oletus)} &\geq [\mathrm{vol}(A')^{1/n} + \mathrm{vol}(B')^{1/n}]^n + [\mathrm{vol}(A'')^{1/n} + \mathrm{vol}(B'')^{1/n}]^n \\
&= \rho \cdot [\mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n}]^n + (1-\rho) \cdot [\mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n}]^n \\
&= [\mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n}]^n.
\end{aligned}$$

Lopulta ottamalla n-juuren molemmin puolin saadaan haluttu tulos. Täten Brunn-Minkowski-epäyhtälö pätee kaikille epätyhjille kompakteille joukoille A ja B. ■

3.2 Isoperimetriset epäyhtälöt

Tässä kohdassa käsitellään lyhyesti, mitä isoperimetriset epäyhtälöt ovat [9, s. 333–334]. Yksinkertaisin isoperimetrinen epäyhtälö on tasossa $\mathbb{R}^2$, joka muotoillaan seuraavasti: *"Olkoon $C \subset \mathbb{R}^2$ epätyhjä joukko ja olkoon A sen pinta-ala sekä P sen piirin suuruus. Tällöin*

$$4\pi A \leq P^2, \tag{24}$$

missä yhtäsuuruus pätee, jos ja vain jos C on kuula." [5, s. 6–7]

Yleisesti epäyhtälölle (24) ei löydy yksinkertaista todistusta. Yksi mahdollinen tapa on todistaa kyseinen epäyhtälö käyttämällä konveksi- ja differentiaaligeometriaa [5, s. 6–7].

Suurissa ulottuvuuksissa ei ole mielekästä puhua isoperimetrisessä epäyhtälössä joukon "piiristä", vaan yleisesti isoperimetriset epäyhtälöt kuvataan käyttämällä joukon ympäristöä. Jokainen isoperimetrinen epäyhtälö perustuu väittämään, että metrisissä avaruuksissa kaikkien joukkojen tilavuuksiin verrattuna tilavuudeltaan samankokoisella kuulalla on pienin t-ympäristön tilavuus. Täsmällisesti yleinen isoperimetrinen epäyhtälö on seuraavanlainen:

Propositio 3.3 (**Isoperimetrinen epäyhtälö**).
Jokaiselle kompaktille joukolle $A \subset \mathbb{R}^n$ ja jokaisella arvolla $t \geq 0$

$$\mathrm{vol}(A_t) \geq \mathrm{vol}(B_t), \tag{25}$$

missä B on kuula, jolla on sama tilavuus kuin joukolla A, eli $\mathrm{vol}(A) = \mathrm{vol}(B)$.

Yleinen isoperimetrinen epäyhtälö (25) on helppo johtaa Brunn-Minkowski-epäyhtälöstä.

Todistus (Isoperimetrinen epäyhtälö).
Skaalauksen perusteella voidaan olettaa, että kuulan B säde on 1. Oletetaan, että epätyhjällä kompaktilla joukolla $A \subset \mathbb{R}^n$ on sama tilavuus kuin yksikkökuulalla B, eli $\mathrm{vol}(A) = \mathrm{vol}(B)$.

Joukkojen t-ympäristön määritelmän (9) perusteella voidaan t-ympäristö joukolle A kuvata yksikkökuulan B avulla muodossa

$$A_t = A + t \cdot B. \tag{26}$$

Yksikkökuula B skaalattuna arvolla t on suljettu ja rajattu avaruudessa $\mathbb{R}^n$, joten Heine-Borel-teoreeman[2] perusteella $t \cdot B$ on myös kompakti joukko. Koska joukot A ja $t \cdot B$ ovat kompakteja, yhtälön (26) perusteella joukko A_t on epätyhjä kompakti joukko. Täten Brunn-Minkowski-epäyhtälön 3.1 nojalla:

$$\begin{aligned}\mathrm{vol}(A_t) = \mathrm{vol}(A + tB) &\geq [\mathrm{vol}(A)^{1/n} + \mathrm{vol}(tB)^{1/n}]^n = [\mathrm{vol}(A)^{1/n} + t \cdot \mathrm{vol}(B)^{1/n}]^n \\ &= [\mathrm{vol}(B)^{1/n} + t \cdot \mathrm{vol}(B)^{1/n}]^n = [(1+t)\mathrm{vol}(B)^{\frac{1}{n}}]^n = (1+t)^n \mathrm{vol}(B) = \mathrm{vol}(B_t),\end{aligned}$$

missä hyödynnettiin mittateorian käsitteitä ja viimeinen yhtäsuuruus seuraa B yksikkökuulan t-ympäristöstä ja mitan määritelmistä. Koska joukko A oli valittu mielivaltaisesti, pätee epäyhtälö (25) kaikille kompakteille joukoille $A \subset \mathbb{R}^n$. ■

Epäyhtälö (24) voidaan johtaa myös epäyhtälöstä (25) ottamalla raja-arvo $t \to 0$ yhtälöä (10) hyödyntämällä. Isoperimetrisiä epäyhtälöitä löytyy monia muitakin kuin euklidisessa avaruudessa $\mathbb{R}^n$ olevia. Tärkeä isoperimetrinen epäyhtälö on myös pallon pinnalla oleva isoperimetrinen epäyhtälö

$$P[A_t] \geq P[C_t], \tag{27}$$

jossa P on pallon pinnalla S^{n-1} oleva mitta ja C on sellainen kalotti, että $P[A] = P[C]$. Tämä voidaan johtaa hyödyntämällä yleistä isoperimetristä epäyhtälöä 3.3, mutta yksinkertaista todistusta tälle kuitenkaan ei löydy[3]. Eräs tapa osoittaa epäyhtälö (27) on käyttämällä kuulan symmetriaa hyödyksi [10, s. 7–8]. Oletetaan epäyhtälö (27) tunnetuksi, ja käytetään kyseistä epäyhtälöä, kun tarkastellaan mitan keskittymistä pallon pinnalla seuraavassa osiossa.

[2] **Heine-Borel-teoreema:** Joukko $A \subset \mathbb{R}^n$ on kompakti, jos ja vain jos se on suljettu ja rajattu [7, s. 37].

[3] Kirjallisuuden perusteella alkuperäisen tavan epäyhtälön (27) osoittamiseksi löysi P. Lévy tai V. Milman [9, 10]. P. Lévy hyödynsi juuri kuulan symmetriaa epäyhtälön osoittamisessa. Tästä johtuen kyseinen epäyhtälö tunnetaan myös Lévyn isoperimetrisenä epäyhtälönä.

3.3 Mitan keskittyminen pallolla

Seuraavaksi käydään läpi mitan keskittyminen pallon pinnalla [9, s. 330–331, s. 334]. Määritellään ensiksi kuulan pinnalla oleva mitta. Olkoon mitta P yksikkökuulan S^{n-1} pinnalla euklidisessa avaruudessa siten, että koko pinnan S^{n-1} mitan arvo on 1. Pinnan mitta voidaan määritellä n-ulotteisen origokeskeisen kuulan B_R^n ja joukon A pintaa vastaavan pallosektorin $\tilde{A} = \{\alpha x \colon \ x \in A, \alpha \in [0, R]\}$ tilavuuksien välisenä suhteena

$$P[A] = \frac{\mathrm{vol}(\tilde{A})}{\mathrm{vol}(B_R^n)}. \tag{28}$$

Yhtälön (28) mukaisesti voidaan siis tulkita mitta P normalisoituna Lebesguen mittana, joka saa arvoja väliltä $P \in [0, 1]$. Kyseinen mitta voidaan myös tulkita analogisena todennäköisyyteen, eli $P[A] = \mathrm{Prob}[A]$, missä Prob kuvaa todennäköisyysteoriassa aluetta A vastaavaa arvoa. Skaalautumisen perusteella voidaan olettaa tässä kohdassa kuulan olevan yksikkökuula B_1^n. Tavoitteena on nyt tutkia, miten todennäköisyysmitta P keskittyy pallon pinnalla S^{n-1} suurissa ulottuvuuksissa n. Tarkasteltaessa voidaan havaita, että todennäköisyysmitta keskittyy juuri suurilla n arvoilla lähelle "päiväntasaajaa", eli mitä tahansa isoympyrää pallon S^{n-1} pinnalla [9, s. 330–331]. Tarkemmin tarkasteltuna on havaittavissa, että todennäköisyysmitta on keskittynyt lähelle mitä tahansa joukkoa $A \subset S^{n-1}$, joka peittää ainakin puolet pallon pinnasta $P[A] \geq \frac{1}{2}$. Kyseinen väittämä voidaan muotoilla seuraavasti:

Teoreema 3.4 (**Mitan keskittyminen pallolla**).
Olkoon $A \subset S^{n-1}$ sellainen mitallinen joukko, että $P[A] \geq \frac{1}{2}$ ja olkoon A_t joukon A t-ympäristö. Tällöin

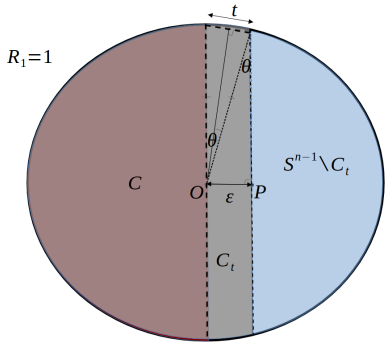
$$1 - P[A_t] \leq e^{-t^2 \cdot n/2}. \tag{29}$$

Etäisyyksistä puhuttaessa pallon pinnalla S^{n-1} on otettava huomioon, että puhutaan euklidisesta etäisyydestä $\|\cdot\|_2$. Toisin sanoen kahden pisteen välinen etäisyys pallon pinnalla S^{n-1} mitataan kuulan B_1^n sisäosan kautta. Seuraavaksi todistetaan kyseinen teoreema käyttämällä hyväksi isoperimetristä epäyhtälöä (27) pallon pinnalla [9, s. 334], ja arvioimalla tarvittavan joukon mittaa [2, s. 12].

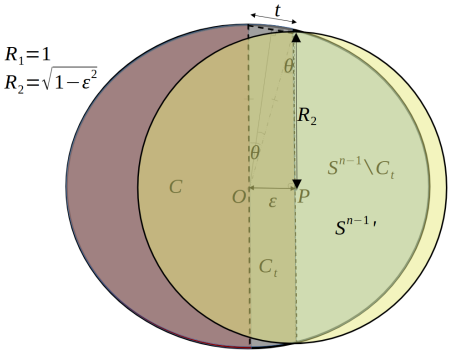
Todistus (Mitan keskittyminen pallolla).
Skaalauksen perusteella voidaan olettaa tutkittavan pallon pinnan S^{n-1} olevan yksikkökuulan B_1^n pinta. Olkoon joukko $A \subset S^{n-1}$ sellainen, että se peittää puolet kyseisen pallon pinnasta, eli $P[A] \geq \frac{1}{2}$. Isoperimetrisen epäyhtälön (27) mukaan vastaavan kokoisen kalotin C ($P[C] = P[A]$) t-ympäristö on pienempi kuin joukon A ympäristö. Tämän perusteella voidaan estimoida teoreeman 3.4 epäyhtälön (29) vasenta puolta $P[S^{n-1} \setminus A] = 1 - P[A_t]$ arvioimalla ylärajaa alueelle $P[S^{n-1} \setminus C_t] = 1 - P[C_t]$, sillä

$$P[A_t] \geq P[C_t] \Leftrightarrow 1 - P[A_t] \leq 1 - P[C_t] = P[S^{n-1} \setminus C_t], \tag{30}$$

missä viimeinen yhtäsuuruus johtuu mitallisten joukkojen ominaisuuksista. Voidaan olettaa, että kalotti C on pallonpuolisko, sillä tällöin joukon $S^{n-1} \setminus C_t$ alue on suurimmillaan.



(a) n-ulotteisen kuulan B_1^n pallon pinnan S^{n-1} poikkileikkaus hypertasoon nähden, mikä jakaa pallon pinnan joukkoon C_t ja sen komplementtiin.



(b) n-ulotteisen kuulan B_1^n pallon pinnan S^{n-1} poikkileikkaus ja kuulan B' pallon pinta $S^{n-1\prime}$.

Kuva 3: Poikkileikkaus mitan keskittymisestä pallon pinnalla S^{n-1}.

Täytyy osoittaa, että joukon $S^{n-1} \setminus C_t$ mitta on korkeintaan yhtälön (29) oikeanpuolinen arvo, eli $P[S^{n-1} \setminus C_t] \leq e^{-t^2 \cdot n/2}$. Olkoon $\epsilon > 0$ etäisyys keskipisteestä O pisteeseen P ja t ympäristön etäisyys kuvan 3a mukaisesti. Trigonometrian perusteella saadaan, että $\epsilon = \sin(\theta)$ ja $t = 2 \cdot \sin\left(\frac{\theta}{2}\right)$. Käyttämällä sinin kaksinkertaisen kulman kaavaa tästä seuraa, että $\epsilon = t \cdot \cos\left(\frac{\theta}{2}\right) \geq \frac{t}{\sqrt{2}}$, sillä poikkileikkauksen kulma saa arvoja väliltä $\theta \in [0, \frac{\pi}{2}]$.

Joukon C_t komplementti voidaan mitan P määritelmän perusteella esittää muodossa

$$1 - P[C_t] = 1 - \frac{\text{vol}(\tilde{C}_t)}{\text{vol}(B_1^n)}, \tag{31}$$

missä $\tilde{C}_t \subset B_1^n \subset \mathbb{R}^n$ on joukkoa C_t vastaava pallosektorin alue. Käyttämällä yhtälöä (31) saadaan yläarvio mitalle $P[S^{n-1} \setminus C_t]$ ottamalla toinen kuula B' kuvan 3b mukaisesti. Kuvassa 3b keltaisella on kuulaa B' vastaava kuulan pinnan $S^{n-1\prime}$ poikkileikkaus. Kuvan 3b perusteella kuulan B' säde on $R_2 = \sqrt{1 - \epsilon^2}$, joka voidaan määrittää Pythagoraan lauseen avulla. Kuvasta 3b havaitaan kuulan B' peittävän kokonaan joukkoa $S^{n-1} \setminus C_t$ vastaavan pallosektorin alueen, joten yhtälön (31) perusteella saadaan yläarvioksi

$$1 - P[C_t] \leq \frac{\text{vol}(B')}{\text{vol}(B_1^n)} = \frac{R_2^n}{1^n} = (1 - \epsilon^2)^{n/2}. \tag{32}$$

Lopulta muokkaamalla yhtälön (32) oikeaa puolta saadaan

$$(1 - \epsilon^2)^{n/2} \leq e^{-\epsilon^2 \cdot n/2} \leq e^{-t^2 n} \leq e^{-t^2 \cdot n/2}, \tag{33}$$

missä viimeinen epäyhtälö seuraa eksponenttifunktion ominaisuuksista ja toinen epäsuuruus saadaan aikaisemmin todettua epäyhtälöä $\epsilon \geq \frac{t}{\sqrt{2}}$ käyttämällä. Ensimmäinen epäyhtälö

$$(1-\epsilon^2)^{n/2} \leq e^{-\epsilon^2 \cdot n/2} \tag{34}$$

voidaan todistaa seuraavasti:

Olkoon $n = 1$. Tällöin (34) on muotoa $(1-\epsilon^2)^{1/2} \leq e^{-\epsilon^2 \cdot 1/2}$. Ottamalla epäyhtälössä kaikki vasemmalle puolelle sekä yhteiseksi tekijäksi arvon $e^{-\epsilon^2/2} > 0$ saadaan lauseke muotoon

$$e^{\frac{\epsilon^2}{2}}(1-\epsilon^2)^{\frac{1}{2}} \leq 1 \Leftrightarrow \frac{1}{2}(\epsilon^2 + \ln\Big(1-\epsilon^2\Big)) \leq \ln(1) = 0 \tag{35}$$

käyttämällä luonnollisen logaritmin laskusääntöjä. Sijoittamalla $x = \epsilon^2 \in [0,1]$ saadaan kaavan (35) vasemmalle puolelle funktio $f(x) = \frac{1}{2}(x + \ln(1-x))$, joka derivoimalla havaitaan olevan laskeva funktio. Täten funktio saa suurimman arvonsa kyseisellä välillä, kun $x = \epsilon = 0$, eli $\max_{x\in[0,1]} f(x) = f(0) = 0 \leq 0$.

Täten saadaan $(1-\epsilon^2)^{1/2} \leq e^{-\epsilon^2 \cdot 1/2}$, missä molemmat puolet ovat positiivisia kaikilla $\epsilon \in [0,1]$ arvoilla. Korottamalla epäyhtälössä molemmat puolet arvolla n saadaan haluttu lopputulos (34).

Yhdistämällä yhtälön (31) yläarvioon (32) ja viimeiseksi epäyhtälöön (30) saadaan haluttu tulos:

$$1 - P[A_t] \leq 1 - P[C_t] \leq e^{-t^2 \cdot n/2} \tag{36}$$

■

Teoreema 3.4 tulee hyödylliseksi arvioitaessa satunnaismuuttujan $f(X)$ keskittymistä odotusarvon $\mathbb{E}[f(X)]$ ympärille, kun $X \sim N(0, I_n)$. Yleisesti vain ei-oskilloivat funktiot noudattavat kyseistä ominaisuutta [11, s. 104–105]. Tällaisia funktiota ovat Lipschitz-funktiot, joita tarkastellaan seuraavassa osiossa.

3.4 Lévyn apulause

Tässä kohdassa osoitetaan jokaisen Lipschitz-funktion olevan tiukasti keskittynyt funktion odotusarvon lähettyville pallon pinnalla S^{n-1}. Toisin sanoen Lipschitz-funktioiden $f\colon S^{n-1} \to \mathbb{R}$ arvot ovat keskittyneet odotusarvon $\mathbb{E}[f]$ lähettyville.

Osoitetaan tulos hyödyntämällä mitan keskittymistä pallon pinnalla [9, s. 337–338]. Ensiksi osoitetaan yksinkertainen apulause, joka on hyödyllinen Lévyn apulauseen osoittamisessa.

Apulause 3.5.
Olkoon $f\colon \Omega \to \mathbb{R}$ mitallinen funktio avaruudessa Ω todennäköisyysmitalla P. Tällöin

$$P[f < \operatorname{med}(f)] \leq \frac{1}{2} \text{ ja } P[f > \operatorname{med}(f)] \leq \frac{1}{2}$$

Apulauseen 3.5 osoittamiseksi käytetään perustietoja mittateoriasta.

Todistus (Apulause 3.5).
Olkoon $f\colon \Omega \to \mathbb{R}$ mitallinen funktio avaruudessa Ω todennäköisyysmitalla P. Ensiksi, jos $f < \text{med}(f) \Leftrightarrow \text{med}(f) - f > 0$, mikä tarkoittaa, että on olemassa pienin luku $k \in \mathbb{Z}_+$ siten, että $\text{med}(f) - f \geq \frac{1}{k}$ ja $\frac{1}{k-1} > \text{med}(f) - f$. Yhdistettynä ehdot saadaan muotoon

$$\frac{1}{k-1} > \text{med}(f) - f \geq \frac{1}{k} \Leftrightarrow \text{med}(f) - \frac{1}{k-1} < f \leq \text{med}(f) - \frac{1}{k}. \tag{37}$$

Ottamalla seuraavaksi erillisten joukkojen $A_k = \left\{\text{med}(f) - \frac{1}{k-1} < f \leq \text{med}(f) - \frac{1}{k}\right\}$ yhdiste kaikilla arvoilla $k = 1, 2, ...$, saadaan mitallisten joukkojen numeroituvan additiivisuuden 2.1 perusteella:

$$\begin{aligned} P[f < \text{med}(f)] = P[\bigcup_{k=1}^{\infty} A_k] &= \sum_{k=1}^{\infty} P[\text{med}(f) - \frac{1}{k-1} < f \leq \text{med}(f) - \frac{1}{k}] \\ &\leq \sup_{k\geq 1} P[f \leq \text{med}(f) - \frac{1}{k}] \leq \frac{1}{2}, \end{aligned} \tag{38}$$

missä toiseksi viimeinen epäyhtälö seuraa ominaisuudesta $A_k \subset \sup_{k\geq 1}\{f \leq \text{med}(f) - \frac{1}{k}\}$ kaikilla $k \in \mathbb{Z}_+$, ja viimeinen kohta funktion f mediaanin määritelmästä 2.5.

Samalla tavalla voidaan todistaa toinen epäyhtälö:

$$\begin{aligned} f > \text{med}(f) &\Rightarrow \text{med}(f) + \frac{1}{k} \leq f < \text{med}(f) + \frac{1}{k-1} \\ &\Rightarrow B_k = \left\{\text{med}(f) + \frac{1}{k} < f < \text{med}(f) + \frac{1}{k-1}\right\}, \end{aligned}$$

ja

$$\begin{aligned} P[f > \text{med}(f)] = P[\bigcup_{k=1}^{\infty} B_k] &= \sum_{k=1}^{\infty} P[\text{med}(f) + \frac{1}{k} \leq f < \text{med}(f) + \frac{1}{k-1}] \\ &\leq \sup_{k\geq 1} P[f \geq \text{med}(f) + \frac{1}{k}] \leq \frac{1}{2} \end{aligned}$$

vastaavalla tavalla. ■

Seuraavaksi osoitetaan haluttu ominaisuus Lipschitz-funktioille $f\colon S^{n-1} \to \mathbb{R}$. Yksinkertaisuuden vuoksi osoitetaan Lévyn apualuse vain 1-Lipschitz-funktiolle [9, s. 338]. Kyseinen apulause voidaan muotoilla seuraavasti:

Apulause 3.6 (Lévyn apulause).
Olkoon $f\colon S^{n-1} \to \mathbb{R}$ 1-Lipschitz-funktio. Tällöin

$$P[f > \text{med}(f) + t] \leq e^{-t^2 \cdot n/2} \text{ ja } P[f < \text{med}(f) - t] \leq e^{-t^2 \cdot n/2}$$

kaikilla $t \in [0, 1]$.

Todistus (Lévyn apulause).
Olkoon $f\colon S^{n-1} \to \mathbb{R}$ 1-Lipschitz-funktio. Määritellään joukko $A = \{x \in S^{n-1}\colon f(x) \leq \text{med}(f)\} \subset S^{n-1}$.

Apulauseen 3.5 perusteella joukon A mitta on $P[A] = P[f \leq \text{med}(f)] = 1 - P[f > \text{med}(f)] \geq 1 - \frac{1}{2} = \frac{1}{2}$. Koska funktio f on 1-Lipschitz-funktio, eli $|f(x) - f(y)| \leq ||x - y||_2$ kaikilla $x, y \in S^{n-1}$, ja määritellyn joukon t-ympäristö on $A_t = \{x \in S^{n-1}\colon \text{dist}(x, A) < t\}$, saadaan:

$$f(x) - \text{med}(f) = |f(x) - \text{med}(f)| \leq \text{dist}(x, A) \leq t \Leftrightarrow f(x) \leq \text{med}(f) + t \tag{39}$$

kaikilla $x \in A_t$ ja $t \in [0, 1]$. Yhtälössä (39) ensimmäinen kohta seuraa joukon A määritelmästä, toinen funktion 1-Lipschitz -ominaisuudesta ja viimeinen joukon A_t määritelmästä. Yhtälön (39) perusteella voidaan määrittää ympäristön A_t todennäköisyysmitaksi $P[A_t] = P[f(x) \leq \text{med}(f) + t]$, ja sillä $P[A] \geq \frac{1}{2}$ saadaan teoreeman 3.4 perusteella:

$$P[f(x) > \text{med}(f) + t] = 1 - P[f \leq \text{med}(f) + t] = 1 - P[A_t] \leq e^{-t^2 \cdot n/2}. \tag{40}$$

Vastaavalla tavalla voidaan osoittaa teoreeman toinen epäyhtälö. Olkoon $B = \{x \in S^{n-1}\colon f(x) \geq \text{med}(f)\}$, ja apulauseen 3.5 perusteella:

$$P[B] = P[f \geq \text{med}(f)] = 1 - P[f < \text{med}(f)] \geq 1/2.$$

Koska f on 1-Lipschitz-funktio, pätee joukon t-ympäristön B_t pisteille:

$$f(x) \geq \text{med}(f) - t \ \forall x \in B_t$$

kaikilla $t \in [0, 1]$. Koska $P[B] \geq \frac{1}{2}$, saadaan teoreeman 3.4 perusteella:

$$P[f < \text{med}(f) - t] = 1 - P[f \geq \text{med}(f) - t] = 1 - P[B_t] \leq e^{-t^2 \cdot n/2},$$

mikä osoittaa toisen epäyhtälön Lévyn apulauseesta. ■

Lévyn apulauseelle on monta hyödyllistä ominaisuutta, joita hyödynnetään monissa lauseissa, kuten äärellisten korkeaulotteisten joukkojen redusoimisessa pienempään ulottuvuuteen. Lévyn apulausetta hyödynnetään päätuloksen osoittamisessa seuraavassa kohdassa.

3.5 Johnson-Lindenstrauss-teoreema

Osoitetaan päätulos hyödyntämällä Lévyn apulausetta [9, s. 358–360] sekä Dasguptan ja Guptan [3, s. 62] tulkintaa.

Ensiksi voidaan todeta, ettei isometristä upotusta n-ulotteiselle joukolle V euklidisessa avaruudessa pienempään ulottuvuuteen k ole olemassa kuin vain harvinaisissa tapauksissa. Tässä mielessä joukko V on todellakin n-ulotteinen. Tilanne muuttuu, jos ei vaadita suoranaista isometriaa ulottuvuuksien välillä, eli sallitaan jonkinlainen pisteiden etäisyyksien välinen muutos [9, s. 358]. Kyseinen ilmiö voidaan kuvata seuraavasti:

Teoreema 3.7 (**Johnson-Lindenstrauss-teoreema**).
Olkoon X äärellinen N:n pisteen joukko euklidisessa avaruudessa $\mathbb{R}^n$ ja $\epsilon \in (0, 1]$ annettu. Tällöin joukolle X on olemassa $(1+\epsilon)$-upotus avaruuteen $\mathbb{R}^k$, jossa $k = O(\epsilon^{-2}\ln(N))$.[4]

Teoreeman 3.7 perusteella jokainen euklidisessa avaruudessa oleva N:n pisteen joukko $V \subset \mathbb{R}^n$ voidaan kuvata joukkona avaruudessa $\mathbb{R}^{O(\epsilon^{-2}\cdot\ln(N))}$, jos sallitaan pisteiden välisissä etäisyyksissä jonkin verran vääristymää esimerkiksi $\epsilon = 15$ %:n verran. Käytännöllisesti katsoen, jos halutaan kuvata N:n datapisteen etäisyydet, taustatiedon perusteella tarvitaan $\binom{N}{2}$ arvoa. Teoreeman 3.7 perusteella kuitenkin voidaan konstruoida kyseiset etäisyydet vain suuruusluokkaa $O(N \cdot \ln(N))$ arvoilla sallimalla pienen informaation menetyksen.

Johnson-Lindenstrauss-teoreemalla 3.7 löytyy monia erilaisia todistuksia kirjallisuudessa. Todistetaan kyseinen teoreema 3.7 todistamalla ensiksi hyödyllinen apulause liittyen projisointifunktioihin [9, s. 359–360].

Apulause 3.8 (**Projektion pituuden keskittyminen**).
Olkoon funktio

$$f(x) = \sqrt{x_1^2 + x_2^2 + ... + x_k^2} \tag{41}$$

yksikkövektorin $x \in S^{n-1}$ projektion p pituus aliavaruuteen L_0, jonka virittää k ensimmäistä koordinaattia. Oletetaan, että $x \in S^{n-1}$ on valittu satunnaisesti. Tällöin $f(x)$ on tiukasti keskittynyt jonkin luvun m $(n > m \geq k)$ ympärille siten, että

$$P[f(x) \geq m + t] \leq e^{-t^2\cdot n/2} \text{ ja } P[f(x) \leq m - t] \leq e^{-t^2\cdot n/2}, \tag{42}$$

missä P on yhtenäinen todennäköisyysmitta pallon pinnalla S^{n-1}. Lisäksi, jos n on tarpeeksi suuri ja $k \geq 10\ln(n)$, niin $m \geq \frac{1}{2}\sqrt{\frac{k}{n}}$.

Apulauseen 3.8 mukaisesti k-ulotteinen aliavaruus L_0 on kiinnitetty ja yksikkövektori x on valittu satunnaisesti. Vastaavanlaisesti voidaan kiinnittää yksikkövektori x ja valita k-ulotteinen aliavaruus L satunnaisesti, kuten Lévyn apulauseessa 3.6. Kyseinen valinta noudattaa samoja ehtoja apulauseessa 3.8, jota voidaan hyödyntää apulauseen todistamisessa.

Todistus (Apulause 3.8).
Olkoon $p\colon \mathbb{R}^n \to \mathbb{R}^k$ kohtisuora projektio siten, että $(x_1, ..., x_n) \mapsto (x_1, ..., x_k)$, joka on helposti osoitettavissa olevan 1-Lipschitz-funktio:

$$\|p(x) - p(y)\|_2 \leq \|x - y\|_2,$$

sillä neliöjuuri on kasvava funktio, ja epäyhtälön oikealla puolella on positiivisia arvoja enemmän kuin projektiossa. Koska p on 1-Lipschitz-funktio, täytyy funktion f olla myös 1-Lipschitz-funktio, sillä f voidaan kuvata muodossa $f(x) = ||p(x)||_2$. Tästä seuraa

[4]Teoreemassa 3.7 merkintä O tunnetaan iso-O notaationa, joka kuvaa suuruusluokkaa. Tässä tapauksessa k on siis suuruusluokaltaan $\epsilon^{-2}\ln(N)$, eli jollain vakiolla kerrottuna.

$$\left|f(x) - f(y)\right| = \left|\|p(x)\|_2 - \|p(y)\|_2\right| \leq \|p(x) - p(y)\|_2 \leq \|x - y\|_2,$$

missä ensimmäinen epäyhtälö seuraa kolmioepäyhtälöstä.

Valitaan seuraavaksi $m = \text{med}(f)$. Koska $k \leq m < n$, ja $f\colon S^{n-1} \to \mathbb{R}$ on 1-Lipschitz-funktio, Lévyn apulauseen 3.6 perusteella saadaan haluttu tulos (42). Täten riittää vain todistaa arvolle m annettu alaraja, joka voidaan osoittaa alkeellisesti löytämällä yksinkertainen alue pinnalla S^{n-1}. Tämä voidaan kuitenkin välttää hyödyntämällä yleistä mitan keskittymisen tulosta. Olkoon siis $x \in S^{n-1}$ satunnainen. Koska x on yksikkövektori, tiedetään, että $||x||_2^2 = 1$. Tästä seuraa

$$1 = \mathbb{E}[1] = \mathbb{E}[\|x\|_2^2] = \sum_{i=1}^{n} \mathbb{E}[x_i^2], \tag{43}$$

missä viimeinen kohta seuraa odotusarvon lineaarisuudesta. Pallon pinnan S^{n-1} symmetriasta sekä yhtälöstä (43) seuraa $\mathbb{E}[x_i^2] = \dfrac{1}{n}$ kaikilla $i = 1, 2, 3, \ldots$ Merkitään $f^2 = f(x)^2 = \|p(x)\|_2^2 = \sum_{i=1}^{k} x_i^2$. Aikaisemman tuloksen perusteella tällöin f^2 odotusarvoksi saadaan

$$\mathbb{E}[f^2] = \sum_{i=1}^{k} \mathbb{E}[x_i^2] = \sum_{i=1}^{k} \frac{1}{n} = \frac{k}{n}. \tag{44}$$

Koska f on tiukasti keskittynyt, voidaan näyttää, ettei $\mathbb{E}[f^2]$ voi olla paljon suurempi kuin m^2, joten m ei ole liian pieni. Arvoa $\mathbb{E}[f^2]$ voidaan estimoida hyödyntämällä todennäköisyyslaskentaa diskreeteille arvoille. Tällöin saadaan jokaiselle $t \geq 0$ arvolle

$$\begin{aligned}
\frac{k}{n} &\leq \mathbb{E}[f^2] = P[f \leq m + t] \cdot (m + t)^2 + P[f > m + t] \cdot (m + t)^2 \\
&\leq P[f \leq m + t] \cdot (m + t)^2 + P[f > m + t] \cdot \max_{x \in S^{n-1}} \{f(x)^2\} \\
&\leq 1 \cdot (m + t)^2 + e^{-t^2 \cdot n/2} \cdot 1 = (m + t)^2 + e^{-t^2 \cdot n/2},
\end{aligned}$$

missä käytettiin Lévyn apulausetta, todennäköisyysmitan P ominaisuutta ja Lipschitz ominaisuutta: $\max_{x \in S^{n-1}} \{f(x)^2\} \leq \|x\|_2^2 = 1$.

Seuraavaksi asetetaan $t = \sqrt{\dfrac{k}{5n}}$. Jos oletetaan, että $k \geq 10 \ln(n)$, niin:

$$e^{-t^2 \cdot n/2} = e^{-k/10} \leq e^{-10 \ln(n)/10} = \frac{1}{n}. \tag{45}$$

Yhdistämällä yhtälö (45) aikaisempaan tulokseen saadaan

$$\frac{k}{n} \leq (m + t)^2 + e^{-t^2 \cdot n/2} \leq (m + t)^2 + \frac{1}{n} \Leftrightarrow m \geq \sqrt{\frac{k-1}{n}} - t \geq \frac{1}{2}\sqrt{\frac{k}{n}},$$

missä viimeiset kohdat voidaan johtaa kyseisten oletuksien avulla. Tämä on haluttu alaraja arvolle m. ■

Seuraavaksi osoitetaan Johnson-Lindenstrauss-teoreema 3.7 hyödyntämällä apulausetta 3.8.

Todistus (Johnson-Lindenstrauss-teoreema).
Voidaan olettaa, että n on tarpeeksi suuri ja $\epsilon > 0$ pieni. Olkoon $X \subset \mathbb{R}^n$ annettu N:n pisteen joukko. Asetetaan $k = 200\epsilon^{-2}\ln(N)$. Jos $k \geq n$, ei ole mitään todistettavaa, joten oletetaan, että $k < n$. Olkoon L satunnainen k-ulotteinen lineaarinen aliavaruus avaruudesta $\mathbb{R}^n$, joka on saatu satunnaisesta avaruuden L_0 kierrosta.

Valittu L on kopio avaruudesta $\mathbb{R}^k$. Olkoon funktio $p\colon \mathbb{R}^n \to L$ ortogonaalinen projektio avaruuteen L ja m arvo, mihin funktio $f(x) = \|p(x)\|_2$ on keskittynyt apulauseen 3.8 mukaisesti. Pyritään näyttämään, että jokaiselle eri pisteille $x, y \in \mathbb{R}^n$ upotus

$$(1 - \frac{\epsilon}{3})m\|x - y\|_2 \leq \|p(x) - p(y)\|_2 \leq (1 + \frac{\epsilon}{3})m\|x - y\|_2, \tag{46}$$

rikkoontuu enintään todennäköisyydellä $2N^{-2}$.
Upotuksessa (46) projisointi p on T-upotus joukosta X avaruuteen $\mathbb{R}^k$ arvolla

$T = \dfrac{(1 + \frac{\epsilon}{3})m}{(1 - \frac{\epsilon}{3})m} < 1 + \epsilon$, kun $\epsilon \in (0, 1]$.

Otetaan joukon X kaksi pistettä x ja y, jotka ovat kiinnitettyjä, ja merkitään $u = x - y$. Koska projektiot ovat lineaarisia kuvauksia, saadaan $p(x) - p(y) = p(x - y) = p(u)$, mistä seuraa, että yhtälö (46) voidaan kuvata muodossa

$$(1 - \frac{\epsilon}{3})m\|u\|_2 \leq \|p(u)\|_2 \leq (1 + \frac{\epsilon}{3})m\|u\|_2. \tag{47}$$

Yhtälö (47) on muuttumaton skaalauksen suhteen, joten voidaan olettaa $||u||_2 = 1$. Tällöin ehto (47) saadaan muotoon:

$$\Big|\|p(u)\|_2 - m\Big| \leq \frac{\epsilon}{3}m \tag{48}$$

avaamalla epäyhtälö (47), erottamalla molemmin puolin arvo m ja ottamalla itseisarvo molemmin puolin.

Apulauseen 3.8 perusteella, kun u on kiinnitetty ja L valittu satunnaisesti, ehto (48) rikkoontuu suurimmillaan todennäköisyydellä:

$$\begin{aligned}
&P[\Big|\|p(u)\|_2 - m\Big| \leq \frac{\epsilon}{3}m] = P[-\frac{\epsilon}{3}m \leq \|p(u)\|_2 - m \leq \frac{\epsilon}{3}m] = \\
&P[-\frac{\epsilon}{3}m \leq \|p(u)\|_2 - m] + P[\|p(u)\|_2 - m \leq \frac{\epsilon}{3}m] = \\
&P[\|p(u)\|_2 \geq m - \frac{\epsilon}{3}m] + P[\|p(u)\|_2 \leq m + \frac{\epsilon}{3}m] \leq \\
&e^{-\frac{(-\frac{\epsilon}{3}m)^2}{2}n} + e^{-\frac{(-\frac{\epsilon}{3}m)^2}{2}n} = 2e^{-\frac{\epsilon^2 m^2}{18}n} \leq 2e^{-\frac{\epsilon^2 \cdot \frac{1}{4} \cdot \frac{k}{n}}{18} \cdot n} = \\
&2e^{-\frac{\epsilon^2 k}{72}} = 2e^{-\epsilon^2 \cdot \frac{200}{72}\epsilon^{-2}\ln(N)} = 2N^{-\frac{200}{72}} < 2N^{-2},
\end{aligned}$$

missä hyödynnettiin arvoja $k = 200\epsilon^{-2}\ln(N)$ ja $m \geq \dfrac{1}{2}\sqrt{\dfrac{k}{n}}$.

Täten joukon X upotuksessa kahden pisteen välinen etäisyyden muutos välillä $[1-\epsilon, 1+\epsilon]$ rikkoontuu suurimmillaan todennäköisyydellä $2N^{-2}$. Koska joukossa X on N pistettä, pisteiden välisiä etäisyyksiä on aikaisemman perusteella $\binom{N}{2}$ kappaletta.

Tällöin joukossa X pisteiden väliset etäisyyksien vääristymät eivät ole halutulla välillä suurimmillaan todennäköisyydellä $\binom{N}{2} \cdot 2N^{-2} = 1 - \frac{1}{N}$ kaikilla $N \geq 2$ arvoilla [3, s. 62]. Toisin sanoen todennäköisyys, että $(1+\epsilon)$-upotus toteutuu satunnaiseen tasoon L nähden on vähintään $\frac{1}{N} > 0$ kaikilla arvoilla $N \geq 2$. Tämä osoittaa upotuksen olevan mahdollinen, joten joukolle X on olemassa $(1+\epsilon)$-upotus aliavaruuteen $\mathbb{R}^k$. ■

Teoreemassa 3.7 joukon X pisteiden välinen etäisyyksien suurin sallittu vääristymä saadaan määritettyä T-upotuksesta (46), jossa $T < 1 + \epsilon$. Tämän perusteella joukon X ja upotettujen pisteiden välinen etäisyyksien muutos on

$$(1-\epsilon)m < \frac{\|p(x) - p(y)\|_2}{\|x - y\|_2} < (1+\epsilon)m.$$

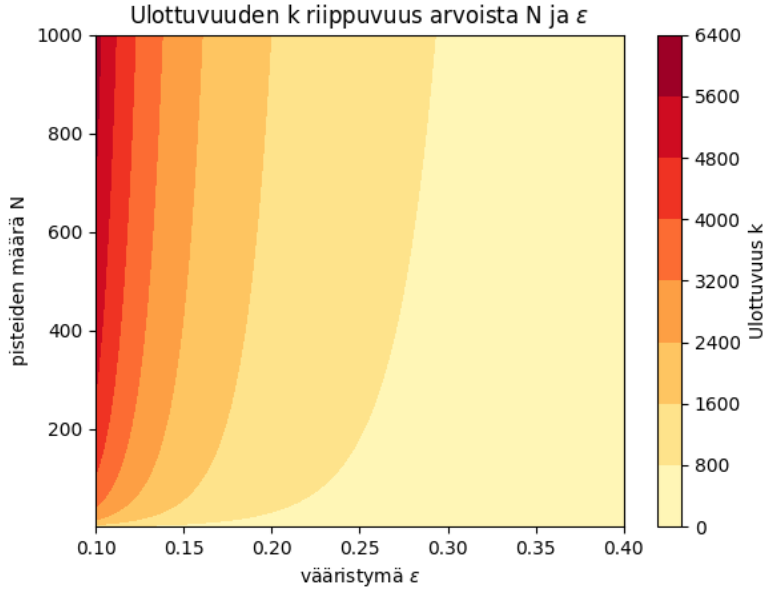
Todistuksessa ulottuvuuden k suuruusluokkaa voidaan pienentää entisestään [9]. Kirjallisuuden [3, s. 61] perusteella nykyisin paras käytetty arvo on

$$k = \frac{4}{e^2/2 - e^3/3} \ln(N). \tag{49}$$

Hyödynnetään kaavaa (49) arvolle k sovelluksissa, sillä suurin osa laskennallisista arvoista hyödyntää kyseistä arvoa. Johnson-Lindenstrauss-teoreema 3.7 on todettu olevan tehokas työkalu moneen käyttötarkoitukseen. Seuraavassa osiossa käydään läpi kyseisen teoreeman merkitystä, seurauksia ja sovelluksia.

4 Seuraukset

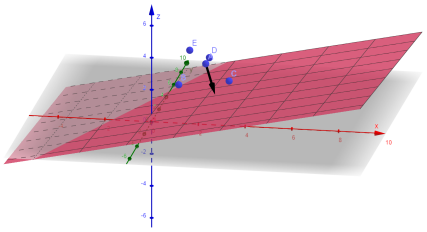
Teoreemasta 3.7 voidaan havaita, että pienimpään mahdolliseen ulottuvuuteen k upottaminen riippuu vain joukon X pisteiden määrästä N ja sallitun vääristymän ϵ arvosta, eikä lainkaan pisteiden ulottuvuuden arvosta n. Liitteen A mukaisesti voidaan kuvata dimension k riippuvuussuhdetta lukuihin N ja ϵ.



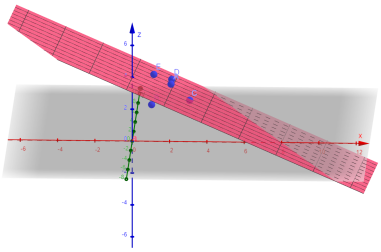
Kuva 4: Ulottuvuuden k riippuvuussuhde joukon X pisteiden määrästä N ja sallitusta vääristymästä ϵ.

Kuvasta 4 havaitaan, että pienin mahdollinen ulottuvuus k, johon joukko X voidaan upottaa, riippuu sallitun vääristymän ϵ ja joukossa X olevien pisteiden N määrästä. Kuvan 4 perusteella etenkin vääristymä ϵ vaikuttaa paljon arvoon k. Jos $\epsilon \to 0$, arvo k kasvaa rajatta, joka tarkoittaa, ettei upotus tällöin ole mahdollinen. Tämä vahvistaa sen, ettei täydellisen isometrian säilyttäminen joukon X upotuksissa ole mahdollista kuin vain hyvin poikkeuksellisissa tapauksissa. Huomattavaa on myös, että joukon pisteiden lukumäärän N kasvaessa upottaminen pienempään mahdolliseen ulottuvuuteen k myös kasvaa. Tätä voidaan pitää järkevänä, sillä jos pisteet upotetaan satunnaisesti k-ulotteiseen tasoon, suuren joukon X pisteiden väliset etäisyydet ovat vaikeampia säilyttää samoina.

Tämä voi vaikuttaa aluksi hyvin epäselvältä, joten havainnollistetaan äärellisen pistejoukon X upotusta hieman esimerkillä pienessä ulottuvudessa, vaikka teoreema 3.7 ei toimi pienissä ulottuvuuksissa.



(a) Joukon X pisteiden upotus satunnaiseen tasoon nähden.



(b) Joukon X pisteiden upotus optimaaliseen tasoon nähden.

Kuva 5: Pistejoukon $X \subset \mathbb{R}^n$ upotus tasoon $L \subset \mathbb{R}^k$.

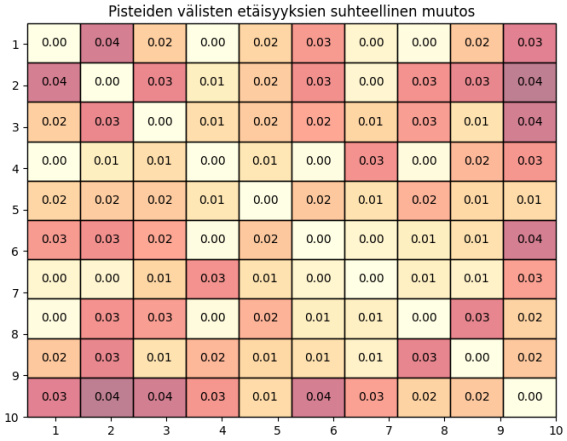
Olkoon äärellinen pistejoukko X kuvien 5a ja 5b mukainen. Kuvassa 5a pisteet projisoidaan satunnaiseen tasoon nähden, ja geometrisesti tarkasteltuna pisteiden väliset etäisyydet eivät luultavasti muutu oleellisesti. Vastaavasti kuvassa 5b esiintyy optimaalinen taso, johon pisteet kannattaa projisoida. Tämän havaitaan olevan pisteitä lähellä oleva taso. Teoreema 3.7 ei ota kantaa, mihin tasoon nähden joukon X pisteet upotetaan, sillä projisointi valitaan satunnaisesti. Täten taso, johon pisteet upotetaan ei välttämättä ole paras. Teoreeman 3.7 mukaan, jos pisteet upotetaan satunnaiseen tasoon nähden, niin todennäköisesti pisteiden väliset etäisyydet eivät muutu suuresti[5].

4.1 Sovellukset

Teoreemalle löytyy silti monta hyödyllistä käyttökohdetta esimerkiksi koneoppimisessa. Voidaan käydä kokeellisesti läpi teoreeman 3.7 toimivuus. Teoreeman 3.7 todistuksen perusteella vähintään todennäköisyydellä $\frac{1}{N} > 0$ haluttu upotus satunnaiseen tasoon L löytyy yhdellä yrityksellä. Kyseistä todennäköisyyttä voidaan kasvattaa toistamalla kyseinen prosessi $O(N)$ kertaa, jolloin haluttu upotus löytyy joistakin näistä kerroista suuremalla todennäköisyydellä.

Liitteen A mukaisesti generoidaan satunnainen data, eli pistejoukko X, jonka pisteet $N = 10$ ovat $n = 100\ 000$ ulottuvuudessa, ja pisteiden koordinaatit saavat arvoja satunnaisesti väliltä $[0, 10\ 000]$. Seuraavaksi lasketaan teoreeman 3.7 mukaisesti kaavalla (49) pienin mahdollinen ulottuvuus, johon pisteet voidaan upottaa niin, etteivät pisteiden väliset etäisyydet muutu sallittua vääristymää $\epsilon = 10\ \%$ enempää. Tämän havaitaan olevan $k = 1\ 973$. Onnistuneen satunnaisen upotuksen jälkeen saadaan määritettyä upotetun joukon Y pisteiden sijannit, ja voidaan määrittää pisteiden väliset etäisyydet molemmissa joukoissa X ja Y. Lopulta lasketaan, kuinka paljon etäisyydet suhteellisesti muuttuivat upotuksessa.

[5]Koska pisteiden väliset etäisyydet eivät luultavasti joukon X upotuksessa muutu paljon, joukon X avaruudellinen rakenne, eli isometria, upotettavaan tasoon nähden pysyy melkein samana.



Kuva 6: Pisteiden välisten etäisyyksien suhteellinen muutos satunnaisen projisoinnin jälkeen.

Kuvan 6 perusteella pisteiden välisten etäisyyksien suhteellinen muutos satunnaisessa upotuksessa oli suurimmillaan 4 %, joka on vähemmän kuin sallitun vääristymän $\epsilon = 10$ % suuruus. Voidaan siis todeta, että kyseinen teoreema 3.7 toimii käytännössä. Kyseistä ominaisuutta on hyödynnetty muutamissa sovelluksissa, joita käydään seuraavaksi läpi.

Klusterointi

Koneoppimisessa pyritään yleisesti ratkaisemaan luokittelongelmia, joissa erilaiset luokat erotellaan määrittämällä parametri kyseiselle luokalle esimerkiksi numeroiden. Eräs koneoppimisen ratkaisumalli on klusterointi. Datajoukon X klusterointia voidaan ajatella joukon X osittamisena siten, että jokaisen osajoukon $X_i \subset X$ elementit ovat luokaltaan samoja ja eroavat toisista osajoukoista [4, s. 7–10]. Klusterointimenetelmiä on monia, mutta käsitellään yksinkertaisuuden vuoksi klassista klusterointia, eli k:n keskiarvon klusterointia. k:n keskiarvon klustreoinnissa tarkoitus on osittaa datajoukko X juuri k kappaletta vastaaviin osajoukkoihin. Osajoukkojen, eli klustereiden, keskipisteet $c_1, c_2, ..., c_k$ valitaan siten, että ne pienentävät osajoukkojen pisteiden väliset etäisyydet keskipisteisiin nähden:

$$\text{argmin}_{c_1,c_2,...,c_k} \sum_{x \in X} \min_i ||x - c_i||_2^2. \tag{50}$$

Klustereiden keskipisteiden $c_1, c_2, ..., c_k$ optimaalisen sijainnin määrittäminen yhtälön (50) mukaisesti on todettu olevan haastava ongelma jopa pienillä k arvoilla. Ongelmaan on löydetty erilaisia ratkaisuja, joista yleisin on Lloydin algoritmi. Algoritmi etsii optimaaliset sijainnit keskipisteille iteraatiivisesti alkuarvausten perusteella. Jos algoritmi vaatii t iteraatiota, on algoritmin suoritusaika suuruusluokaltaan $O(tknN)$. Suoritusaika siis riippuu datapisteiden määrästä N ja niiden ulottuvuudesta n. Täten halutun ratkaisun saamiseksi algoritmin suoritusaika voi olla hyvin

suuri, jos datan määrä on valtava. Suoritusaikaa voidaan parantaa ensiksi hyödyntämällä Johnson-Lindenstrauss-teoreemaa upottamalla alkuperäinen datajoukko X pienemmäksi datajoukoksi Y. Upotettua joukkoa voidaan pitää lähes vastaavanlaisena datajoukkona kuin X, sillä etäisyydet pisteiden välillä eivät muuttuneet sallittua vääristymää ϵ enempää. Tällöin algoritmin ajamisaika klusteroinnissa on $O(N \cdot n \ln(n) + tkN\epsilon^{-2} \ln(N))$, mikä vie ajallisesti vähemmän aikaa kuin alkuperäistä joukkoa X käyttämällä [4, s. 7–10].

Kyseinen upottaminen on hyödyllistä vain, jos joukon Y osittaminen vastaa joukon X osittamista. Voidaan osoittaa, että joukon X osittaminen upotuksen kautta noudattaa epäyhtälöä

$$k_d \leq (1 + 4\epsilon)(1 + \gamma)k_d^\star, \tag{51}$$

missä $k_d^\star$ vastaa joukon X optimaalista osittamisen hintaa, k_d joukon X osittamisen hintaa ja $(1 + \gamma)$ kuvaa virherajaa upotetun joukon Y osittamisesta [4, s. 7–10]. Epäyhtälön (51) perusteella datajoukon X osittaminen upotuksen kautta voi olla kannattavaa. Täten upottaminen klusteroinnissa voi olla ajallisesti tehokasta, jos datajoukon koko on suuri.

Lineaarinen optimointi

Johnson-Lindenstrauss-teoreemaa on myös hyödynnetty lineaarisissa optimoinnissa ongelmissa, jotka voidaan yleisesti kuvata matriisimuodossa

$$\begin{aligned} \max\ & z = c^T x \\ \text{s.e. } & Ax \leq b \\ & x \geq 0, \end{aligned} \tag{52}$$

missä z tunnetaan tavoitefunktiona, jota halutaan optimoida. Kaavassa (52) matriisi $A \in \mathbb{R}^{n \times d}$ ja vektori $b \in \mathbb{R}^n$ muodostavat mahdollisten ratkaisuiden alueen ja vektori $x \in \mathbb{R}^d$ sisältää valitut päätösmuuttujat, joita vaihtamalla pyritään optimoimaan arvoa z.

Vu, Poirion ja Liberti todistavat teoreeman hyödyllisyyden lineaarisissa optimointiongelmissa [12]. Yleisesti ottaen ongelma (52) voidaan ratkaista simplex-algoritmilla tai vastaavanlaisella metodilla, kun ratkaisualueen tiedetään olevan konveksi joukko. Jos kuitenkin matriisi A on suuri, eli rajoituksien määrä n on suuri, saattaa optimointi ajallisesti kestää kauan. Korkealla todennäköisyydellä voidaan pienentää rajoituksien määrää arvoon $k << n$ hyödyntämällä teoreemaa 3.7 ja saada silti lähes sama ratkaisu arvolle x redusoidussa ongelmassa [12, s. 1].

Johnson-Lindenstrauss-teoreemaa on yleisesti käytetty jo useissa numeerisissa sovelluksissa klusteroinnin ja optimointiongelmien lisäksi. Näitä ovat esimerkiksi lähin ympäristöetsintä sekä graafiteoreettiset ongelmat [9, 4]. Teoreemalle on löydetty myös käyttötarkoituksia neurotieteessä ja neuroverkkojen parantamisessa. Teoreemaa voidaan yleisesti hyödyntää kaikissa sovelluksissa, joissa käsitellään suuria määriä dataa.

Viitteet

[1] Pekka Alestalo. Integrointi n-ulotteisessa avaruudessa. 2019. URL https://mycourses.aalto.fi/pluginfile.php/930406/mod_resource/content/1/spherical.pdf.

[2] Keith Ball et al. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31(1-58):26, 1997. URL https://math.uchicago.edu/~shmuel/AAT-readings/Combinatorial%20Geometry,%20Concentration,%20Real%20Algebraic%20Geometry/ball.pdf.

[3] Sanjoy Dasgupta ja Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22 (1):60–65, 2003. URL https://www.cs.yale.edu/homes/aspnes/pinewiki/attachments/JohnsonLindenstraussTheorem/dasgupta-gupta.pdf.

[4] Casper Benjamin Freksen. An introduction to Johnson-Lindenstrauss transforms, 2021. URL https://arxiv.org/pdf/2103.00564.

[5] Penelope Gehring. The isoperimetric inequality: Proofs by convex and differential. *Rose-Hulman Undergraduate Mathematics Journal*, 20, 2019. URL https://scholar.rose-hulman.edu/cgi/viewcontent.cgi?article=1380&context=rhumj.

[6] Juha Kinnunen. Measure and integral. 2019. URL https://math.aalto.fi/~jkkinnun/files/measure_and_integral.pdf.

[7] Aleksis Koski. General topology. 2024. URL https://atkoski.fi/files/topology.pdf.

[8] Kalle Kytölä. Metric spaces. 2024. URL https://github.com/kkytola/Metric-spaces-materials/blob/main/MetSp-2024.pdf.

[9] Jiří Matoušek. *Lectures on Discrete Geometry,* Graduate Texts in Mathematics volume 212. Springer-Verlag, 2000.

[10] Gideon Schechtman. Concentration, results and applications. Teoksessa *Handbook of the geometry of Banach spaces*, volume 2, pages 1603–1634. Elsevier, 2003. URL https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=3cc8e8f867114c6e51e93bd9563835cfaf1e153b.

[11] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2024. URL https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.pdf.

[12] Ky Vu, Pierre-Louis Poirion, ja Leo Liberti. Using the Johnson-Lindenstrauss lemma in linear and integer programming. 2015. URL https://arxiv.org/pdf/1507.00990.

A Liitteet

```python
# Johnson-Lindenstrauss teoreema esimerkki
import numpy as np
import pandas as pd
import math
import matplotlib.pyplot as plt
from sklearn import random_projection
from sklearn.random_projection import johnson_lindenstrauss_min_dim
from sklearn.metrics.pairwise import euclidean_distances
from matplotlib.table import Table
from mpl_toolkits.mplot3d import Axes3D
np.set_printoptions(threshold = np.inf)

# Parametrit:
d = 100000 # Joukon X pisteiden ulottuvuus
n = 10 # Joukon X pisteiden määrä
e = 0.1 # Sallittu vääristymä pisteiden välisissä etäisyyksissä. Esim. 10 %
# Joukon X muodostaminen:
# Joukon X pisteiden generointi satunnaisesti annetussa dimensiossa d
X = np.random.uniform(0, 10000, size=(n, d))
# Joukon X muoto; palauttaa ensiksi pisteiden määrän ja ulottuvuuden missä pisteet ovat
print(f"Joukon X muoto:\t {X.shape}")
print(f"Joukko X:\n {pd.DataFrame(X)}")

# Projisointi
# Lasketaan pienin mahdollinen ulottuvuuden joukon pisteet voidaan
# projisoida kyseisellä virhe-rajalla
# k = (4 * np.log(n) / (e**2 / 2 - e**3 / 3)).astype(np.int64)
k = johnson_lindenstrauss_min_dim(n_samples = n, eps = e)
# Satunnainen projektio pienimpään mahdolliseen ulottuvuuteen
transformer = random_projection.GaussianRandomProjection(n_components = k)
# Vastaavasti voidaan laskea kirjallisuudessa olevalla arvolla:
# k2 = (200*np.log(n)/(e**2)).astype(np.int64)
print(f"Pienin ulottuvuus johon JL mukaan voidaan projisoida: {k}")

# Joukon X projisoitu joukko Y
Y = transformer.fit_transform(X)
print(f"X joukon projisoitu joukon Y muoto:\t {Y.shape}")
print(f"Joukko Y:\n {pd.DataFrame(Y)}")

# Joukon X Euklidiset etäisyydet pisteiden välillä:
d1 = euclidean_distances(X, X)
print(f"\nJoukon X pisteiden väliset etäisyydet:\n {pd.DataFrame(d1)}")
# Joukon Y Euklidiset etäisyydet pisteiden välillä:
d2 = euclidean_distances(Y, Y)
print(f"\nJoukon Y pisteiden väliset etäisyydet:\n {pd.DataFrame(d2)}")

# Pisteiden välisten suhteelliset etäisyydet projisoinnin jälkeen, eli vääristymät
error = np.zeros((n, n))
for i in range(n):
    for j in range(n):
        if i == j:
```

```
            error[i, j] = 0
        else:
            error[i, j] = abs(d1[i, j] - d2[i, j]) / d1[i, j]

print(f"\nPisteiden väliset etäisyyksien vääristymät:\n {pd.DataFrame(error)}")
print(f"\nSuurin vääristymä: {error.max()}")

# Taulukoiden piirtäminen:
def plot_matrix(data, alpha=0.5, Title=''):  # Kuvaaja funktio
    fig, ax = plt.subplots(figsize=(8, 6))
    ax.set_frame_on(False)
    nrows, ncols = data.shape
    ax.set_xticks([i + 0.5 for i in range(ncols)])
    ax.set_yticks([i + 0.5 for i in range(nrows)])
    ax.set_xticklabels([str(i) for i in range(1, ncols + 1)])
    ax.set_yticklabels([str(i) for i in range(1, nrows + 1)])
    ax.invert_yaxis()
    ax.tick_params(axis='x', top=False, bottom=True, labeltop=False, labelbottom=True)
    ax.tick_params(axis='y', left=True, right=False, labelleft=True, labelright=False)
    table = Table(ax, bbox=[0, 0, 1, 1])
    width, height = 1.0/ncols, 1.0/nrows
    for i in range(nrows):
        for j in range(ncols):
            color = plt.cm.get_cmap('YlOrRd')(data[i, j] / data.max(), alpha=alpha)
            table.add_cell(i, j, width, height, text=f'{data[i, j]:.2f}',
            loc='center', facecolor=color)
    ax.add_table(table)
    plt.title(Title)
    plt.show()


# Taulukoiden kuvaajat:
plot_matrix(d1, alpha=0.5, Title = 'Pisteiden väliset etäisyydet joukossa X')
plot_matrix(d2, alpha=0.5, Title = 'Pisteiden väliset etäisyydet joukossa Y')
plot_matrix(error, alpha=0.5, Title = 'Pisteiden välisten etäisyyksien suhteellinen muutos')

# Kuvaajien teko pienemmän ulottuvuuden k riippuvuudesta virherajasta ja pisteiden määrästä:
e_values = np.linspace(0.1, 0.40, 100)          # Virherajojen/vääristymien rajoja
n_values = np.linspace(2, int(1e3), 10000)      # Pisteiden määrien arvoja

# Vektoreiden laskeminen ja k:n arvojen laskeminen:
E, N = np.meshgrid(e_values, n_values)
K = 4 * np.log(N) / ((1/2) * E**2 - (1/3) * E**3)

# 2-ulotteisen kuvajan tekeminen:
plt.contourf(E, N, K, cmap = 'YlOrRd')
cbar = plt.colorbar()
cbar.set_label('Ulottuvuus k')
plt.xlabel('vääristymä $\epsilon$')
plt.ylabel('pisteiden määrä N')
plt.title('Ulottuvuuden k riippuvuus arvoista N ja $\epsilon$')
plt.show()
```