Brunn–Minkowski Inequality and Its Applications

Geometric Methods in Computer Science

Instructor: Yury Makarychev

1 Minkowski Sum

In these lecture notes, we prove the Brunn–Minkowski inequality and the closely related Prékopa–Leindler inequality, and then use them to derive isoperimetric and measure concentration results. We start by defining standard arithmetic operations – scalar multiplication and addition – on sets.

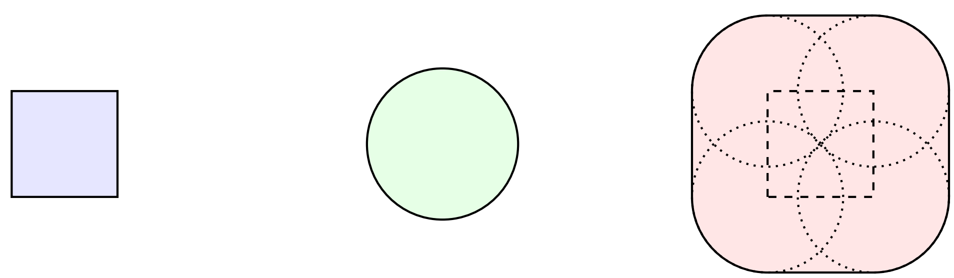
Definition 1.1. *We define scalar multiplication and addition of sets as follows.*

- ***Scalar Multiplication.*** *Consider a set $A \subseteq \mathbb{R}^d$ and a real number c. Define*

$$cA = \{cx : x \in A\}.$$

- ***Minkowski Sum*** *Consider two sets $A, B \subseteq \mathbb{R}^d$. Define*

$$A + B = \{a + b : a \in A, b \in B\}.$$



a square a circle their Minkowski sum

Exercise 1. *Check if the following identities hold. If any of them does not hold, find additional conditions under which it holds.*

1. $A + B = B + A$?
2. $(A + B) + C = A + (B + C)$?

3. $c(A+B) = cA + cB$?

4. $A + A = 2A$?

Definition 1.2. *We will denote the volume of a set A by* $\mathrm{vol}(A)$.

Exercise 2. *(1) Verify that* $\mathrm{vol}(cA) = |c|^d \,\mathrm{vol}(A)$. *(2) Let T be a linear transformation on $\mathbb{R}^d$. Check that* $\mathrm{vol}(TA) = |\det T| \cdot \mathrm{vol}(A)$.

2 Brunn–Minkowski Inequality

Theorem 2.1 (Brunn–Minkowski Inequality)**.** *Consider two non-empty measurable sets A and B in $\mathbb{R}^d$. Assume that $A+B$ is also measurable. Then, the following inequality holds.*

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

Remark 2.2. *We will not discuss measurability and related issues in this course. Instead, we will prove the inequality for closed sets A and B, in which case $A+B$ is an F_σ set and therefore measurable.*

Before we proceed with the proof, we will obtain an equivalent "multiplicative" or "dimensionless" form of the Brunn–Minkowski inequality. Assume that (1) holds. Let $\lambda \in [0,1]$. Then

$$\begin{aligned}\mathrm{vol}(\lambda A + (1-\lambda)B)^{1/d} &\geq \mathrm{vol}(\lambda A)^{1/d} + \mathrm{vol}((1-\lambda)B)^{1/d} = \lambda\,\mathrm{vol}(A)^{1/d} + (1-\lambda)\,\mathrm{vol}(B)^{1/d} \\ &\geq \mathrm{vol}(A)^{\lambda/d}\,\mathrm{vol}(B)^{(1-\lambda)/d}\end{aligned}$$

here we used a weighted variant of the AM-GM inequality

$$\lambda x + (1-\lambda) y \geq x^\lambda y^{1-\lambda}$$

applied to $x = \mathrm{vol}(A)^{1/d}$ and $y = \mathrm{vol}(B)^{1/d}$. The inequality immediately follows from the concavity of function $\ln x$. We conclude that,

$$\mathrm{vol}(\lambda A + (1-\lambda)B) \geq \mathrm{vol}(A)^\lambda\,\mathrm{vol}(B)^{1-\lambda}. \tag{2}$$

We now observe that (1) follows from (2) and therefore formulations (1) and (2) are equivalent. Let $a = \mathrm{vol}(A)^{1/d}$ and $b = \mathrm{vol}(B)^{1/d}$. Set $\lambda = a/(a+b)$; then $1-\lambda = b/(a+b)$. We have,

$$\begin{aligned}\mathrm{vol}(A+B) &= \mathrm{vol}\left(\lambda \cdot \frac{A}{\lambda} + (1-\lambda)\cdot\frac{B}{1-\lambda}\right) \geq \mathrm{vol}\left(\frac{A}{\lambda}\right)^\lambda \mathrm{vol}\left(\frac{B}{1-\lambda}\right)^{1-\lambda} \\ &= \frac{\mathrm{vol}(A)^\lambda\,\mathrm{vol}(B)^{1-\lambda}}{\lambda^{\lambda d}(1-\lambda)^{(1-\lambda)d}} = \frac{a^{\lambda d} b^{(1-\lambda)d}}{\left(\frac{a}{a+b}\right)^{\lambda d}\left(\frac{b}{a+b}\right)^{(1-\lambda)d}} = (a+b)^d.\end{aligned}$$

Raising both sides to the power $1/d$ yields (1).

We now turn to the proof of the Brunn–Minkowski inequality. Our goal is to prove it by induction on d, though, as we will see later, this approach will require a slight revision.

2.1 Proof of the Brunn–Minkowski Inequality in dimension 1

We begin with the base case $d = 1$. By truncating sets A and B and then taking a limit, we may assume that they are compact. Further, note that the volumes of sets A, B and $A+B$ remain unchanged if we shift A by some a and B by some b. Shifting A and B left by $a = \max A$ and $b = \min B$, respectively, we may assume that $A \subset (-\infty, 0]$ and $B \subset [0, \infty)$, and further $A \cap B = \{0\}$. Then $A+B \supset \{0\}+B = B$ and $A+B \subset A+\{0\} = A$. Therefore, $\mathrm{vol}(A+B) \geq \mathrm{vol}(A \cup B) = \mathrm{vol}(A) + \mathrm{vol}(B)$. We proved (1) for $d = 1$.

We now need to prove the inductive step. As is often the case, it is easier to do so using a *stronger* induction hypothesis. In fact, we will prove the following Prékopa–Leindler inequality

Theorem 2.3 (Prékopa–Leindler Inequality). *Consider integrable functions f, g, and $h : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ that satisfy*

$$f(\lambda x + (1-\lambda y)) \geq g(x)^\lambda h(y)^{1-\lambda}$$

for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0,1]$. Then

$$\int_{\mathbb{R}^d} f(z)dz \geq \left(\int_{\mathbb{R}^d} g(x)dx\right)^\lambda \left(\int_{\mathbb{R}^d} h(y)dy\right)^{1-\lambda}$$

If we apply the Prékopa–Leindler inequality to the indicator functions of sets $A+B$, A, and B, we will obtain the Brunn–Minkowski inequality.

Proof. We prove the Prékopa–Leindler inequality by induction on d. We show that the base case $d = 1$ follows from that for the Brunn–Minkowski inequality. Define

$$A_t = \mathcal{I}_{g(x)\geq t} = \{x : g(x) \geq t\} \quad \text{and} \quad B_t = \mathcal{I}_{h(y)\geq t} = \{y : h(y) \geq t\}.$$

Note that

$$\int_{\mathbb{R}} g(x)dx = \int_0^\infty \mathrm{vol}(A_t)dt \quad \text{and} \quad \int_{\mathbb{R}} h(y)dy = \int_0^\infty \mathrm{vol}(B_t)dt$$

Now we check that $A_t + B_t \subseteq \mathcal{I}_{f(z)\geq t}$: using the assumption of the theorem, we get for $z \in \lambda A_t + (1-\lambda)B_t$,

$$f(z) = f(\lambda x + (1-\lambda)y) \geq g(x)^\lambda h(y)^{1-\lambda} \geq t^\lambda t^{1-\lambda} = t.$$

From the Brunn–Minkowski inequality for $d = 1$, we get

$$\begin{aligned}\int f(z)dz &\geq \int_0^\infty \mathrm{vol}(\lambda A_t + (1-\lambda)B_t)dt \geq \int_0^\infty \lambda \,\mathrm{vol}(A_t) + (1-\lambda)\,\mathrm{vol}(B_t)dt \\ &= \lambda \int g(x)dx + (1-\lambda)\int h(y)dy \geq \left(\int g(x)dx\right)^\lambda \left(\int h(y)dy\right)^{1-\lambda}.\end{aligned}$$

Now we prove the induction step. We reinterpret functions f, g, and h as functions of two arguments: the first argument is a point in $\mathbb{R}^{d-1}$ and the second argument is the last coordinate. Further, we first fix first arguments $\hat{z} = \lambda x + (1-\lambda)y$, $\hat{x}$ and $\hat{y}$ of f, g, and h respectively, and then consider $f(\hat{z}, z_d)$, $g(\hat{x}, x_d)$, and $h(\hat{y}, y_d)$ as functions of their second arguments only. The functions satisfy the assumptions of the Prékopa–Leindler inequality. Applying the one-dimensional Prékopa–Leindler inequality to them, we get

$$\int_{\mathbb{R}} f(\hat{z}, z_d) dz_d \geq \left(\int_{\mathbb{R}} g(\hat{x}, x_d) dx_d\right)^{\lambda} \left(\int_{\mathbb{R}} h(\hat{y}, y_d) dy_d\right)^{1-\lambda}. \tag{3}$$

Let us denote three integrals above by $F(\hat{z})$, $G(\hat{x})$, and $H(\hat{y})$, respectively:

$$\begin{aligned} F(\hat{z}) &= \int_{\mathbb{R}} f(\hat{z}, z_d) dz_d, \\ G(\hat{x}) &= \int_{\mathbb{R}} g(\hat{x}, x_d) dx_d, \\ H(\hat{y}) &= \int_{\mathbb{R}} h(\hat{y}, y_d) dy_d. \end{aligned}$$

Inequality (3) states precisely that functions F, G, and H satisfy the assumptions of the Prékopa–Leindler inequality in dimension $d-1$:

$$F(\lambda \hat{x} + (1-\lambda)\hat{y}) \geq G(\hat{x})^{\lambda} H(\hat{y})^{1-\lambda}.$$

Applying the inequality, we get

$$\begin{aligned} \int_{\mathbb{R}^d} f(\hat{z}) = \int_{\mathbb{R}^{d-1}} F(\hat{z}) d\hat{z} &\geq \left(\int_{\mathbb{R}^{d-1}} G(\hat{x}) d\hat{x}\right)^{\lambda} \left(\int_{\mathbb{R}^{d-1}} H(\hat{y}) d\hat{y}\right)^{1-\lambda} \\ &= \left(\int_{\mathbb{R}^d} g(x) dx\right)^{\lambda} \left(\int_{\mathbb{R}^d} h(y) dy\right)^{1-\lambda}. \end{aligned}$$

□

3 Log-concave measures

Consider a measure μ on $\mathbb{R}^d$ with density $p(x)$. We say that μ is log-concave if $\log p(x)$ is a concave function or, equivalently,

$$p(\lambda x + (1-\lambda)y) \geq p(x)^{\lambda} p(y)^{1-\lambda}.$$

Observe that this condition resembles the one in the assumption of the Prékopa–Leindler inequality: we obtain this condition by setting $f(x) = p(x)$, $g(x) = p(x)$, and $h(y) = p(y)$ in the assumption of the Prékopa–Leindler inequality.

Many probability distributions are log-concave. Two important examples are the uniform distribution on a convex set the multivariate normal distribution.

Applying the Prékopa–Leindler inequality to functions

$$\begin{aligned} f(z) &= \mathbf{1}_{\lambda A+(1-\lambda)B}(z) \cdot p(z), \\ g(x) &= \mathbf{1}_A(x) \cdot p(x), \\ h(y) &= \mathbf{1}_B(y) \cdot p(y), \end{aligned}$$

we get the following variant of the Brunn-Minkowski inequality for log-concave measures.

Corollary 3.1 (Brunn–Minkowski Inequality for Log-concave Measures)**.** *Consider two non-empty measurable sets A and B in $\mathbb{R}^d$. Let μ be a log-concave measure. Then for every $\lambda \in [0,1]$, we have*

$$\mu(\lambda A + (1-\lambda)B) \geq \mu(A)^\lambda \mu(B)^{1-\lambda}.$$

(We assume here that $\lambda A + (1-\lambda)B$ is measurable.)

Exercise 3. *Check that the additive version of the Brunn–Minkowski inequality does not necessarily hold for log-concave measures. Specifically, consider the Gaussian measure γ on $\mathbb{R}$ with density $p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Present two sets A and B such that $\gamma(A+B) < \gamma(A)+\gamma(B)$.*

4 Applications of the Brunn–Minkowski Inequality

4.1 Brunn's Principle

Consider a convex set S in $\mathbb{R}^d$ and a direction $u \neq 0$. For every $t \in \mathbb{R}$, let $H(t) = \{\langle x, u\rangle = t\|u\|^2\}$ be the affine hyperplane orthogonal to u that passes through tu. Define

$$S(t) = S \cap H(t).$$

to be the section of S by the hyperplane $H(t)$. Finally, let $f(t)$ denote the $(d-1)$-dimensional volume of $S(t)$ (see Figure 4.1).

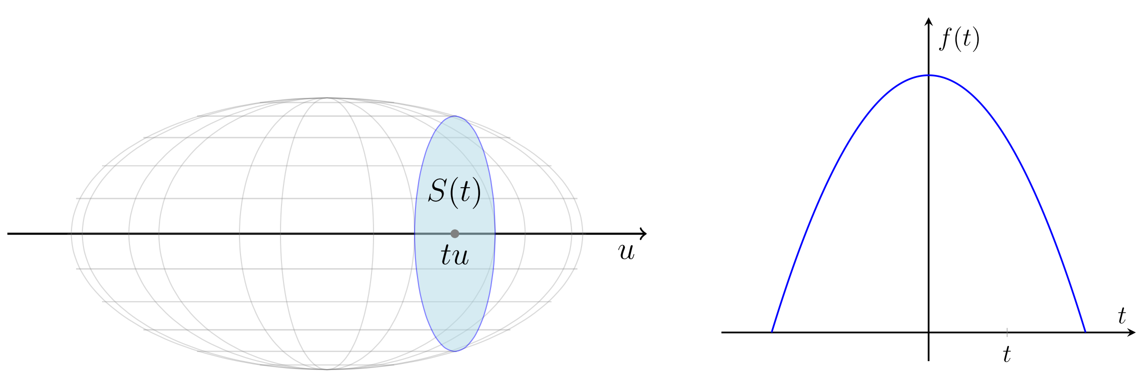


Figure 1: A convex body, its section $S(t)$, and the function $f(t) = \mathrm{vol}_{d-1}(S(t))$. The volume of the section $f(t)$ is a log-concave function.

Theorem 4.1 (Brunn's Principle)**.** *Function $f(t)$ is log-concave. That is, for every $s, t \in \mathbb{R}$ and $\lambda \in [0,1]$, we have*

$$f(\lambda s + (1-\lambda)t) \geq f(s)^\lambda f(t)^{1-\lambda}.$$

Proof. First, observe that the convexity of S implies that $S = \lambda S + (1-\lambda)S$ and therefore

$$S(\lambda s + (1-\lambda)t) \supseteq \lambda S(s) + (1-\lambda)S(t).$$

We apply the Brunn–Minkowski inequality for $(d-1)$-dimensional bodies to the sets $S(s)$ and $S(t)$ and obtain

$$f(\lambda s+(1-\lambda)t) \geq \mathrm{vol}_{d-1}(\lambda S(s)+(1-\lambda)S(t)) \geq \mathrm{vol}_{d-1}(S(s))^\lambda \,\mathrm{vol}_{d-1}(S(t))^{1-\lambda} = f(s)^\lambda f(t)^{1-\lambda}.$$

□

Brunn's Principle can be easily generalized to k-dimensional sections of a convex body S. Fix a k-dimensional linear subspace L of $\mathbb{R}^d$. For every point $v \in L$, consider the $(d-k)$-dimensional affine subspace of $\mathbb{R}^d$ orthogonal to L that passes through v, and let $S(v)$ denote the section of S by this subspace. Define the function $f(v) = \mathrm{vol}_{d-k}(S(v))$. Then f is a log-concave function on L.

4.2 Isoperimetric Inequality

In this section, we prove the isoperimetric inequality, which asserts that among all sets with a given volume, the ball minimizes surface area.

Definition 4.2. *Consider a set A in $\mathbb{R}^d$. Let B be the unit ball centered at the origin. We define the surface area of A as*

$$\mathrm{area}(A) = \lim_{t\to 0} \frac{\mathrm{vol}(A+tB) - \mathrm{vol}(A)}{t}.$$

Equivalently, let $f(t) = \mathrm{vol}(A+tB)$. Then

$$\mathrm{area}(A) = f'(0).$$

Theorem 4.3 (Isoperimetric Inequality)**.** *Let A be a non-empty set in $\mathbb{R}^d$ and B_r be the ball of the same measure. Then*

$$\mathrm{area}(A) \geq \mathrm{area}(B_r).$$

Proof. From the Brunn–Minkowski inequality, we get

$$(f(t)^{1/d})'|_{t=0} = \lim_{t\to 0} \frac{\mathrm{vol}(A+tB)^{1/d} - \mathrm{vol}(A)}{t}^{1/d} \geq \lim_{t\to 0} \frac{\mathrm{vol}(tB)^{1/d}}{t} = \mathrm{vol}(B)^{1/d}.$$

On the other hand,

$$(f(t)^{1/d})'|_{t=0} = f'(0)\cdot \frac{f(0)^{\frac{1}{d}-1}}{d} = \frac{\mathrm{area}(A)}{d\cdot \mathrm{vol}(A)^{1-1/d}}.$$

We get,

$$\text{area}(A) \geq d \cdot \text{vol}(A)^{1-1/d} \cdot \text{vol}(B)^{1/d}.$$

Let us compare that with the surface area of the ball B_r of volume $\text{vol}(A)$. Solving for r, we get $r = \left(\frac{\text{vol}(A)}{\text{vol}(B)}\right)^{1/d}$. The surface area of B_r is

$$\text{vol}(B_{r+t})'|_{t=0} = (r^d \, \text{vol}(B))' = dr^{d-1} \, \text{vol}(B) = d \left(\frac{\text{vol}(A)}{\text{vol}(B)}\right)^{\frac{d-1}{d}} \text{vol}(B) = d \, \text{vol}(A)^{1-1/d} \, \text{vol}(B)^{1/d}.$$

We conclude that

$$\text{area}(A) \geq \text{area}(B_r),$$

as required. □

5 Measure Concentration

In this section, we will prove measure concentration inequalities for the unit sphere $S^{d-1} \equiv \{u : \|u\|_2 = 1\}$ in $\mathbb{R}^d$, equipped with the uniform probability measure μ on it.

5.1 Levy's isoperimetric inequality for the sphere

For a non-empty subset $A \subset S^{d-1}$, let A_t denote the set of points on the sphere that are within Euclidean distance at most t from A:

$$A_t = \{x \in S^{d-1} : d(x, A) \leq t\}.$$

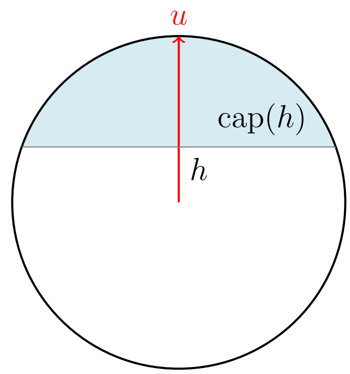
Distances are measured using the Euclidean metric rather than the geodesic metric on the sphere. Alternatively, we could use the geodesic metric instead; the rest of the discussion would be the same.

Definition 5.1. *A spherical cap is a set of the form* $\text{cap}(h) = \{x \in S^{d-1} : \langle x, u \rangle \geq h\}$ *for some direction* $u \in S^{d-1}$ *and parameter* $h \in [-1, 1]$.

Theorem 5.2 (Levy's isoperimetric inequality)**.** *Let A be a non-empty measurable subset of S^{d-1} and $C = \text{cap}(h)$ be a spherical cap of the same measure. Then for every $t \geq 0$, we have*

$$\mu(A_t) \geq \mu(C_t).$$

Note that C_t is also a spherical cap.



This theorem turns out to be quite useful in applications. While we will not prove it here, its proof is not particularly difficult; for instance, it can be proved using the two-point symmetrization transformation. Instead, we will prove a slightly weaker but more convenient variant of the theorem, which follows from the Brunn–Minkowski inequality.

Warm-up Discussion Consider a random point $u = (u_1, \dots, u_d)$ on S^{d-1} (in other words, u is a random unit vector in $\mathbb{R}^d$).

Q: What is the distribution of u_i for a fixed i, approximately?

To answer this question, imagine that we want to sample a unit vector. How can we do that? One of the easiest approaches is as follows. Let us generate a random Gaussian vector $g = (g_1, \dots, g_n)$ with i.i.d. $g_i \sim \mathcal{N}(0,1)$ and then normalize it. The distribution of u_i is the same as that of $g_i/\|g\|_2$. By the law of large numbers,

$$\frac{g_1^2 + \cdots + g_d^2}{d} \overset{\text{a.s.}}{\to} \mathbb{E}\left[g_1^2\right] = 1.$$

Therefore, $\|g\|/\sqrt{d} \to 1$ a.s. and $\sqrt{d}u_i = \frac{\sqrt{d}g_i}{\|g\|_2} \to g_i$ a.s. That is, $\sqrt{d}u_i$ is approximately distributed as $\mathcal{N}(0,1)$ when $d \to \infty$.

In particular, we can get an estimate for the measure of $\text{cap}(t/\sqrt{d})$:

$$\mu(\text{cap}(t/\sqrt{d})) = \Pr(u_i \geq t/\sqrt{d}) = \Pr(\sqrt{d}u_i \geq t) \to \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2} dx.$$

Plugging in known tail bounds for Gaussian random variables, we get that for $t > 0$,

$$\frac{e^{-t^2/2}}{\sqrt{2\pi}(t + \min(1, 1/t))} \leq \lim_{t\to\infty} \Pr\left(u_i \geq \frac{t}{\sqrt{d}}\right) = \lim_{t\to\infty} \Pr\left(u_i \leq -\frac{t}{\sqrt{d}}\right) \leq \frac{e^{-t^2/2}}{\sqrt{2\pi}\, t}.$$

Let us say A has measure $1/2$, the same measure as a half-sphere cap(0). Let $\delta = t/\sqrt{d}$. From Levy's isoperimetric inequality, we get that

$$\mu(S^{d-1} \setminus A_\delta) \leq \mu(S^{t-1} \setminus \text{cap}(0)_\delta) \approx \mu(C_\delta) \approx \frac{e^{-\delta^2 d/2}}{\sqrt{2\pi}(\delta\sqrt{d} + 1)}$$

Unfortunately, this bound provides only an asymptotic estimate of the spherical cap measure for a fixed t, which significantly limits its applicability.

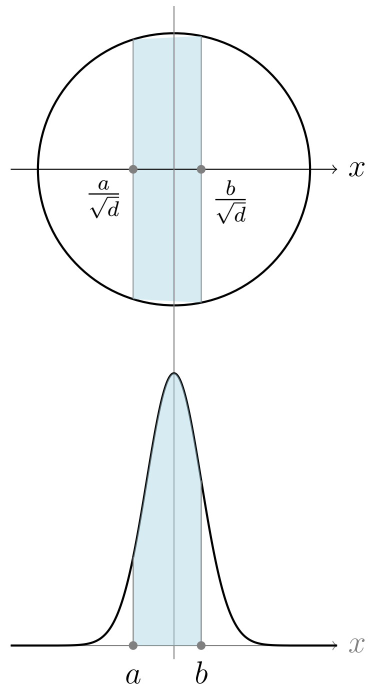


Figure 2: In high dimensions, the Gaussian measure of the interval $[a, b]$ closely approximates the area of the slab of the unit sphere between $a/\sqrt{d}$ and $b/\sqrt{d}$.

5.2 Quantitative Isoperimetric Inequality on the Sphere

Theorem 5.3. *Let μ be the uniform measure on the unit sphere S^{d-1} and $\delta \in (0, \sqrt{2})$. Consider a subset X of S^{d-1} of positive measure. Then*

$$\mu(\{y : d(y, X) \geq \delta\}) \leq \frac{e^{-\delta^2 d/4}}{\mu(X)}.$$

Proof. Define $Y = \{y : d(y, X) \geq \delta\}$. Consider cones $C(X)$ and $C(Y)$ over X and Y, respectively:

$$C(X) = \{\lambda x : x \in X \text{ and } \lambda \in [0, 1]\} \quad \text{and} \quad C(Y) = \{\lambda y : y \in Y \text{ and } \lambda \in [0, 1]\}.$$

Note that the volume of a cone is proportional to the volume of its base, and Thus

$$\begin{aligned} \mathrm{vol}(C(X)) &= \mu(X)\,\mathrm{vol}(B) \\ \mathrm{vol}(C(Y)) &= \mu(Y)\,\mathrm{vol}(B). \end{aligned}$$

Let $Z = \frac{C(X)+C(Y)}{2}$. Consider a point $z \in Z$. Then $z = \frac{\lambda_1 x + \lambda_2 y}{2}$ for some $x \in X$, $y \in Y$, and

$\lambda_1, \lambda_2 \in [0,1]$. Let $\varphi = \arccos\langle x, y\rangle$ denote the angle between x and y. Then

$$\begin{aligned}\|z\|_2^2 = \frac{\lambda_1^2 + \lambda_2^2 + 2\lambda_1\lambda_2\cos\varphi}{4} &\le \max\left(\frac{1+1+2\cos\varphi}{4}, \frac{1+1+0}{4}\right)\\ &\le \max\left(\frac{1+\cos\varphi}{2}, \frac{1}{2}\right) = \max\left(\cos^2\frac{\varphi}{2}, \frac{1}{2}\right).\end{aligned}$$

On the other hand, $\delta \le \|x-y\|_2 = 2\sin\frac{\varphi}{2}$. Using that $\sin^2\frac{\varphi}{2} + \cos^2\frac{\varphi}{2} = 1$, we get that

$$\|z\|_2^2 \le \max\left(1 - \frac{\delta^2}{4}, \frac{1}{2}\right) = 1 - \frac{\delta^2}{4}.$$

We conclude that $Z \subseteq \left(1 - \frac{\delta^2}{4}\right) B$, where B is the unit ball. Applying the Brunn–Minkowski inequality, we get that

$$\operatorname{vol}(Z) \ge \sqrt{\operatorname{vol}(C(X)) \cdot \operatorname{vol}(C(Y))} = \sqrt{(\operatorname{vol}(B)\mu(X)) \cdot (\mu(Y)\operatorname{vol}(B))} = \sqrt{\mu(X)\mu(Y)}\operatorname{vol}(B).$$

Thus,

$$\mu(Y) \le \frac{\operatorname{vol}(Z)^2}{\operatorname{vol}(B)^2\mu(X)} \le \frac{1}{\mu(X)}\left(1 - \frac{\delta^2}{4}\right)^d \le \frac{e^{-\delta^2 d/4}}{\mu(X)}.$$

□

5.3 Measure Concentration for Lipschitz Functions

Theorem 5.4 (Measure Concentration for Lipschitz Functions)**.** *Consider a 1-Lipschitz function f on the unit sphere S^{d-1}. That is, assume that $|f(x) - f(y)| \le \|x-y\|_2$ for all $x, y \in S^{d-1}$. Let m_f be the median of f. Then*

$$\Pr(f(X) \ge m_f + \delta) \le 2e^{-\delta^2 d/4} \quad \textit{and} \quad \Pr(f(X) \le m_f - \delta) \le 2e^{-\delta^2 d/4}$$

where X is sampled uniformly at random from S^{d-1}.

Proof. Let $X = \{x \in S^{d-1} : f(x) \le m_f\}$. By the definition of the median, $\mu(X) \ge 1/2$. If y is at distance at most δ from X, then $f(y) \le m_f + \delta$. Therefore,

$$\mu(\{y : f(y) \ge m_f + \delta\}) \le \mu(\{y : d(y, X) \ge \delta\}) \le \frac{e^{-\delta^2 d/4}}{\mu(X)} \le 2e^{-\delta^2 d/4}.$$

Similarly,

$$\mu(\{y : f(y) \le m_f - \delta\}) \le 2e^{-\delta^2 d/4}.$$

□

Corollary 5.5.

$$\Pr\left(|f(X) - m_f| \ge \delta\right) \le 4e^{-\delta^2 d/4}.$$

Exercise 4. *1. Use Corollary 5.5 to show*

$$\left|\mathbb{E}\left[f(X)\right] - m_f\right| \le \sqrt{\frac{4\pi}{d}}.$$

Hint: Use that

$$\mathbb{E}\left[f(X)\right] - m_f \le \int_0^\infty \Pr(f(X) \ge m_f + t)dt.$$

2. Show that

$$\Pr\left(\left|f(X) - \mathbb{E}\left[f(X)\right]\right| \ge \delta\right) \le 6e^{-\delta^2 d/16}.$$

5.4 Johnson–Lindenstrauss Lemma

In this section, we will prove the Johnson–Lindenstrauss lemma. The lemma states that a subset of n points in a high-dimensional space $\mathbb{R}^d$ can be embedded into a space of dimension roughly $\frac{\log n}{\varepsilon^2}$, while preserving pairwise distances within a factor of $1 \pm \varepsilon$. The lemma has numerous applications in machine learning, theoretical computer science, and analysis.

For example, consider a set of n items represented as feature vectors in $\mathbb{R}^d$, where the similarity between any two items is given by the Euclidean distance between the corresponding points. In general, the number of features can be very large, requiring storage of nd floating-point numbers. Alternatively, one could store the matrix of pairwise distances between the points, but this requires $O(n^2)$ space, which is also substantial.

The Johnson–Lindenstrauss lemma allows us to reduce the number of features to $k = O\left(\frac{\log n}{\varepsilon^2}\right)$, and thus the storage requirement to $O(nk) = O\left(\frac{n \log n}{\varepsilon^2}\right)$ floating-point numbers. Furthermore, working with this lower-dimensional representation can speed up algorithms that process the data – for example, clustering algorithms.

Theorem 5.6 (Johnson–Lindenstrauss Lemma)**.** *Consider a set of n points $X \subset \mathbb{R}^d$. Let $\varepsilon \in (0,1)$ and $k > C \ln n/\varepsilon^2$ (where C is a sufficiently large absolute constant). Then there exists an embedding φ of X into ℓ_2^k s.t.*

$$(1-\varepsilon) \le \frac{\|\varphi(x) - \varphi(y)\|}{\|x - y\|_2} \le (1+\varepsilon). \tag{4}$$

(that is, the embedding φ is "almost" isometric). Moreover, we can find such embedding in randomized polynomial time.

We define the embedding φ simply as an appropriately normalized projection on a random k-dimensional subspace of $\mathbb{R}^d$. Specifically, let H be a random k-dimensional subspace S of $\mathbb{R}^d$ and P_S be an orthogonal projection on P_S. Then $\varphi = \sqrt{\frac{d}{k}} P_S$ satisfies the required properties with high probability. We remark that that there are other ways to construct the random embedding φ; for example, we can use an appropriately normalized $k \times n$ matrix with i.i.d. Gaussian entries or, alternatively, ± 1 entries. The analysis relies on the following claim.

Claim 5.7. *Let v be an arbitrary unit vector. Then $\|\varphi(v)\|_2 \in (1-\varepsilon, 1+\varepsilon)$ with probability at least $1-\delta/n^2$, where $\delta \to 0$ as $C \to \infty$.*

Before we prove this claim, let us show that it implies the Johnson–Lindenstrauss lemma. Suppose that $X = \{v_1, \dots, v_n\}$. For $i < j$, let $z_{ij} = \frac{v_i - v_j}{\|v_i - v_j\|_2}$. According to the claim, $\|\varphi(z_{ij})\|_2 \in (1-\varepsilon, 1+\varepsilon)$ with probability at least δ/n^2. Consequently,

$$\begin{aligned} \|\varphi(v_i) - \varphi(v_j)\|_2 &= \|\varphi(v_i - v_j)\|_2 = \big\|\varphi\big(\|v_i - v_j\|_2 z_{ij}\big)\big\|_2 && (5)\\ &= \|v_i - v_j\|_2 \cdot \|\varphi(z_{ij})\|_2 \in (1 \pm \varepsilon)\|v_i - v_j\| && (6) \end{aligned}$$

with probability at least $1 - \delta/n^2$. Applying the union bound, we get that (5) holds for all pairs i and j simultaneously with probability at least $1 - \delta$, as required.

To prove the claim, let us first fix a unit vector v and consider its projection $P_S v$ on a random k-dimensional subspace S. For the sake of analysis, we also consider the subspace K spanned by the first k standard basis vectors $e_1, \dots, e_k$.

Exercise 5. *Show that the following two random variables have the same distribution:*

- $\|P_S v\|_2$
- $\|P_K u\|_2$, *where u is a random unit vector uniformly distributed on the sphere S^{d-1}.*

Hint consider a random orthogonal transformation T and argue that (1) Tv and u have the same distribution; (2) S and $T^{-1}K$ have the same distribution.

In light of the exercise, we will analyze the random variable $\|P_K u\|_2$ instead of $\|P_S v\|_2$. Note that $P_K u = (u_1, \dots, u_k)$. We observe that $f(u) = \|P_K(u)\|_2$ is 1-Lipschitz:

$$\Big|\|P_K(u)\|_2 - \|P_K(v)\|_2\Big| \le \|P_K(u) - P_K(v)\|_2 = \|P_K(u - v)\| \le \|u - v\|_2.$$

Applying Theorem 5.5 to f, we get

$$\Pr\left(\big|\|f(u)\| - m_f\big| > \varepsilon\sqrt{\frac{k}{d}}\right) \le 4e^{-\frac{\varepsilon^2 k}{4}}. \tag{7}$$

We will now approximately compute the median m_f of $f(u)$. Directly computing m_f is difficult, so we will compute $\sqrt{\mathbb{E}\left[f(u)^2\right]} \approx m_f$ instead. Since u is a unit vector, we have $\sum_{i=1}^d u_i^2 = \|u\|_2^2 = 1$. Therefore, $\sum_{i=1}^d \mathbb{E}\left[u_i^2\right] = 1$. From symmetry, all terms $\mathbb{E}\left[u_i^2\right]$ are equal. Thus, $\mathbb{E}\left[u_i^2\right] = \frac{1}{d}$ for all i. Accordingly,

$$\mathbb{E}\left[\|P_K(u)\|_2^2\right] = \sum_{i=1}^k \frac{1}{d} = \frac{k}{d}.$$

We conclude that $\sqrt{\mathbb{E}\left[f(u)^2\right]} = \sqrt{\frac{k}{d}}$.

Exercise 6. *Check that* $\left|m_f - \sqrt{\mathbb{E}\left[f(u)^2\right]}\right| \le \frac{O(\sqrt{k})}{d}$*. To this end, write*

$$m_f^2 = \mathbb{E}\left[(f(u) + (m_f - f(u))^2\right] = \mathbb{E}\left[f(u)^2\right] + \mathbb{E}\left[(m_f - f(u))^2\right] + 2\,\mathbb{E}\left[f(u)(m_f - f(u))\right].$$

Then use the Cauchy–Schwarz inequality to bound the last term:

$$\mathbb{E}\left[f(u)(m_f - f(u))\right] \le \sqrt{\mathbb{E}\left[f(u)^2\right]} \cdot \sqrt{\mathbb{E}\left[(m_f - f(u))^2\right]}.$$

Finally, write

$$\left|m_f^2 - \mathbb{E}\left[f(u)^2\right]\right| \le \mathbb{E}\left[(m_f - f(u))^2\right] + \sqrt{\mathbb{E}\left[f(u)^2\right]} \cdot \sqrt{\mathbb{E}\left[(m_f - f(u))^2\right]}.$$

Now get an estimate for $\mathbb{E}\left[(m_f - f(u))^2\right]$ *using the approach from Exercise 4.*

From the exercise, we get that $m_f = \sqrt{\frac{k+O(\sqrt{k})}{d}} = \frac{\sqrt{k}+O(1)}{\sqrt{d}}$. Plugging in $k = C\log n/\varepsilon^2$ and $m_f = \sqrt{k+O(1)}/\sqrt{d}$ in (7), we get

$$\begin{aligned}\Pr\left(\left|\|\varphi(v)\|_2 - 1\right| > \varepsilon + O(1/\sqrt{k})\right) &= \Pr\left(\left|\|\varphi(v)\|_2 - (1 + O(1/\sqrt{k}))\right| > \varepsilon\right) \\ &= \Pr\left(\left|\|f(u)\| - \frac{\sqrt{k}+O(1)}{\sqrt{d}}\right| > \varepsilon\sqrt{\frac{k}{d}}\right) \le 4e^{-\frac{C\log n}{4}} = \frac{4}{n^C}.\end{aligned}$$

Noting that $\varepsilon + O(1/\sqrt{k}) = O(\varepsilon)$, we get the desired claim.