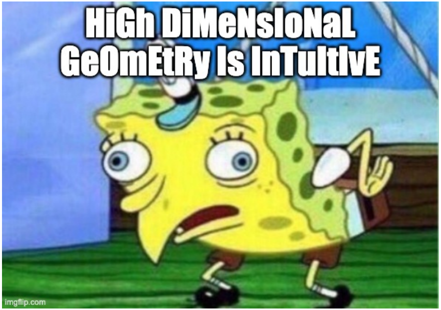
Foundations of Data Science

Avrim Blum, John Hopcroft, and Ravindran Kannan

Christian Howard

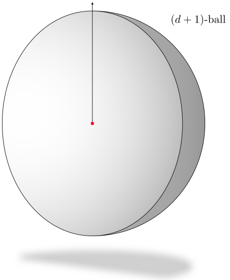
June 13, 2020

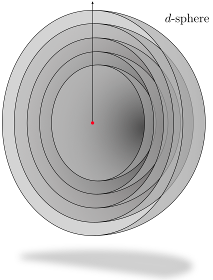
Main Message



- Things get weird in high dimensions
  - Volume and surface area of d-ball goes to 0 as $d \to \infty$
  - Majority of volume of d-ball is near surface and along "equators" of the ball
  - Any two random points in d-ball are (almost) orthogonal

Volume of Unit Ball - Section 2.4.1



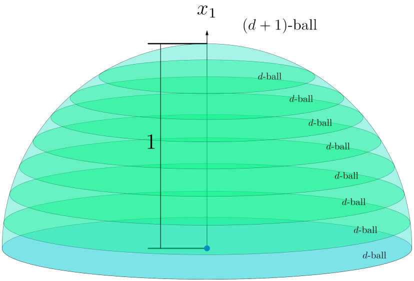


- $V_d(r) :=$ volume of d-ball with radius r
  - $V_d(r) = r^d V_d(1)$
- $S_d(r) :=$ area of d-sphere (surface area of $(d+1)$-ball) with radius r
  - $S_d(r) = r^d S_d(1)$
- View $V_d(r)$ as union of $(d-1)$-spheres with radii from 0 to r
  - $V_d(r) = \int_0^r S_{d-1}(x)dx$
  - Notice that $\frac{dV_d}{dr}(r) = S_{d-1}(r)$

Volume of Unit Ball - Section 2.4.1

- Since $V_d(r) = r^d V_d(1)$, need to just find $V_d(1)$
- $V_{d+1}(1)$ can be viewed as union of d-balls with radii between 0 and 1

$$
\begin{aligned}
V_{d+1}(1) &\overset{1}{=} 2\int_0^1 V_d((1-x^2)^{\frac{1}{2}})dx \\
&\overset{2}{=} 2V_d(1)\int_0^1 (1-x^2)^{\frac{d}{2}}dx \\
&\overset{3}{=} V_d(1)\int_0^1 u^{-\frac{1}{2}}(1-u)^{\frac{d}{2}}du \\
&\overset{4}{=} V_d(1)B\left(\frac{1}{2}, \frac{d}{2}+1\right)
\end{aligned}
$$



The Beta and Gamma functions

Gamma Function Facts

- $\Gamma(s) := \int_0^\infty x^{s-1}e^{-x}dx$
- $\Gamma(s+1) = s\Gamma(s)$
- $\Gamma(s+1) = s!$ for $s \in \mathbb{N}$
- $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$
- $\Gamma\left(\frac{3}{2}\right) = \frac{\sqrt{\pi}}{2}$

Beta Function Facts

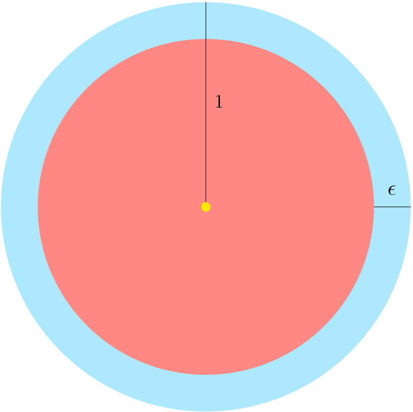
- $B(s,t) := \int_0^1 x^{s-1}(1-x)^{t-1}dx$
- $B(s,t) = B(t,s)$
- $B(s,t) = \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)}$

Volume of Unit Ball - Section 2.4.1

By unrolling the recursion for $V_d(1)$, and recognizing the base case $V_0(1) = 1$, we have

$$\begin{aligned} V_d(1) &\overset{1}{=} V_{d-1}(1) B\left(\frac{1}{2}, \frac{d-1}{2}+1\right) \\ &\overset{2}{=} V_0(1) \prod_{j=0}^{d-1} B\left(\frac{1}{2}, \frac{j}{2}+1\right) \\ &\overset{3}{=} V_0(1) \Gamma\left(1/2\right)^d \frac{\Gamma\left(1\right)}{\Gamma\left(\frac{3}{2}\right)} \frac{\Gamma\left(\frac{3}{2}\right)}{\Gamma\left(2\right)} \cdots \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}+1\right)} \\ &\overset{4}{=} \frac{2\pi^{\frac{d}{2}}}{d\Gamma\left(\frac{d}{2}\right)} \end{aligned}$$

Implies that $V_d(r) = \frac{2\pi^{\frac{d}{2}} r^d}{d\Gamma\left(\frac{d}{2}\right)}$ and $S_d(r) = \frac{dV_{d+1}}{dr}(r) = \frac{2\pi^{\frac{d+1}{2}} r^d}{\Gamma\left(\frac{d+1}{2}\right)}$. ■



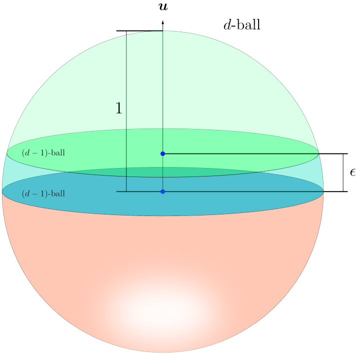
Define a unit d-ball with radius r as $B_d(r)$. For any fixed $\epsilon > 0$, we have

$$\Pr\{\|\boldsymbol{p}\| < 1 - \epsilon\} = \frac{\mathrm{vol}\,(B_d(1-\epsilon))}{\mathrm{vol}\,(B_d(1))} = \frac{V_d\,(1-\epsilon)}{V_d(1)} = \frac{V_d(1)(1-\epsilon)^d}{V_d(1)} \leq e^{-\epsilon d}$$

Volume Near the Equator - Section 2.4.2

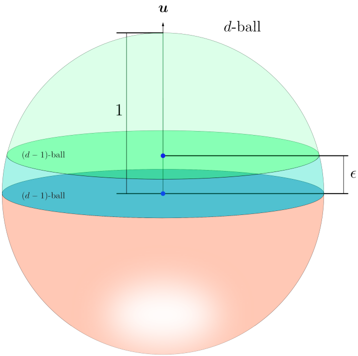
- For any fixed unit vector $\boldsymbol{u}$, *most* of the volume in a d-ball is made of points $\boldsymbol{p}$ where
$$|\boldsymbol{u} \cdot \boldsymbol{p}| = O\left(1/\sqrt{d}\right)$$
- Implies most points in the ball are nearly orthogonal to $\boldsymbol{u}$

Volume Near the Equator - Section 2.4.2



- Fix some arbitrary unit vector $\boldsymbol{u}$
- $H \subseteq B_d(1)$ such that any $\boldsymbol{p} \in H$ satisfies $|\boldsymbol{u} \cdot \boldsymbol{p}| \geq \epsilon$ for some fixed $\epsilon > 0$.
- Consider finding $\Pr\{|\boldsymbol{u} \cdot \boldsymbol{p}| \geq \epsilon\}$ for a point $\boldsymbol{p}$ uniformly at random chosen from $B_d(1)$

Volume Near the Equator - Section 2.4.2



- Geometrically,

$$\Pr\{|\boldsymbol{u} \cdot \boldsymbol{p}| \geq \epsilon\} = \frac{\text{vol}(H)}{\text{vol}(B_d(1))} = \frac{\text{vol}(H)}{V_d(1)}$$

- We have that

$$\text{vol}(H) = 2V_{d-1}(1) \int_{\epsilon}^{1} \left(1 - x^2\right)^{\frac{d-1}{2}} dx$$

Volume Near the Equator - Section 2.4.2

We can further obtain that

$$
\begin{aligned}
\mathrm{vol}(H) &\overset{1}{=} 2V_{d-1}(1)\int_{\epsilon}^{1}\left(1-x^2\right)^{\frac{d-1}{2}}dx \\
&\overset{2}{\leq} 2V_{d-1}(1)\int_{\epsilon}^{1} e^{-\frac{x^2(d-1)}{2}}dx && (1-s\leq e^{-s}) \\
&\overset{3}{\leq} 2\frac{V_{d-1}(1)}{\epsilon(d-1)}\int_{\epsilon}^{\infty}(d-1)xe^{-\frac{x^2(d-1)}{2}}dx && (\tfrac{1}{\epsilon}\geq\tfrac{x}{\epsilon}\geq 1) \\
&\overset{4}{=} \frac{2V_{d-1}(1)}{\epsilon(d-1)}e^{-\frac{\epsilon^2(d-1)}{2}}
\end{aligned}
$$

Volume Near the Equator - Section 2.4.2

The probability bound is then

$$
\begin{aligned}
\Pr\{|\boldsymbol{u}\cdot\boldsymbol{p}|\geq\epsilon\} &\overset{1}{=} \frac{\mathrm{vol}(H)}{V_d(1)} \\
&\overset{2}{\leq} \frac{2V_{d-1}(1)}{\epsilon(d-1)V_d(1)} e^{-\frac{\epsilon^2(d-1)}{2}} \\
&\overset{3}{=} \frac{2e^{-\frac{\epsilon^2(d-1)}{2}}}{\epsilon(d-1)B\left(\frac{1}{2},\frac{d-1}{2}+1\right)} \\
&\overset{4}{=} \frac{2e^{-\frac{a^2}{2}}}{a\sqrt{d-1}B\left(\frac{1}{2},\frac{d-1}{2}+1\right)} && \left(\epsilon=\frac{a}{\sqrt{d-1}}\right)
\end{aligned}
$$

Define $f(d) := \frac{\sqrt{d}}{2} B\left(\frac{1}{2}, \frac{d}{2}+1\right)$. Can show that $f(d)$ is monotonically increasing for $d \geq 0$ by taking derivative and seeing that $f'(d) \geq 0$ for $d \geq 0$. Thus for $d \geq 1$, we have that

$$\begin{aligned} f(d) &\overset{1}{=} \frac{\sqrt{d}}{2} B\left(\frac{1}{2}, \frac{d}{2}+1\right) \\ &\overset{2}{\geq} \frac{1}{2} B\left(\frac{1}{2}, \frac{1}{2}+1\right) \\ &\overset{3}{=} \frac{\Gamma(1/2)\Gamma(3/2)}{2\Gamma(2)} \\ &\overset{4}{=} \frac{\pi}{4} \qquad\qquad (\Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi}, \Gamma\left(\tfrac{3}{2}\right) = \tfrac{\sqrt{\pi}}{2}) \end{aligned}$$

Volume Near the Equator - Section 2.4.2

Using the previous result, we have that

$$\begin{aligned}
\Pr\left\{|\boldsymbol{u}\cdot\boldsymbol{p}| \geq \frac{a}{\sqrt{d-1}}\right\} &\overset{1}{\leq} \frac{2e^{-\frac{a^2}{2}}}{a\sqrt{d-1}B\left(\frac{1}{2}, \frac{d-1}{2}+1\right)} \\
&\overset{2}{\leq} \frac{4}{\pi a}e^{-\frac{a^2}{2}} \\
&\overset{3}{\leq} \frac{4}{\pi}e^{-\frac{a^2}{2}} \qquad (a \geq 1)
\end{aligned}$$

Implies that for any fixed direction $\boldsymbol{u}$ and with probability at least $1 - \frac{4}{\pi a}e^{-\frac{a^2}{2}}$, we have for a random point $\boldsymbol{p}$ chosen from a d-ball for $d \geq 1$ that $|\boldsymbol{u}\cdot\boldsymbol{p}| \leq \frac{a}{\sqrt{d-1}}$. ■

Volume Near the Equator - Section 2.4.2

Theorem 1 (Properties of randomly sampled points on unit ball)

Suppose you randomly sample n points $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ i.i.d. from a unit d-ball. For some $k \geq 1$ and probability at least $1 - O(1/n^k)$, these points will satisfy both conditions:

1. *For all i,* $\|\boldsymbol{x}_i\| \geq 1 - \frac{(k+1)\ln n}{d}$
2. *For all $i \neq j$,* $|\boldsymbol{x}_i \cdot \boldsymbol{x}_j| \leq \frac{\sqrt{2(k+2)\ln n}}{\sqrt{d-1}}$

Volume Near the Equator - Section 2.4.2

Proof

For Condition 1, fix some point $\boldsymbol{x}_i$ and define $\mathcal{E}_i^{(1)}$ the error event that $\|\boldsymbol{x}_i\| < 1 - \frac{(k+1)\ln n}{d}$. We know from earlier that $\Pr\{\|\boldsymbol{x}_i\| < 1 - \epsilon\} \leq e^{-\epsilon d}$, implying for $\epsilon = \frac{(k+1)\ln n}{d}$ that

$$\Pr\left\{\mathcal{E}_i^{(1)}\right\} = \Pr\left\{\|\boldsymbol{x}_i\| < 1 - \frac{(k+1)\ln n}{d}\right\} \leq e^{-\frac{(k+1)\ln n}{d}d} = 1/n^{k+1}$$

Then, overall error probability $\Pr\left\{\exists i : \mathcal{E}_i^{(1)}\right\} \leq n\Pr\left\{\mathcal{E}_1^{(1)}\right\} = 1/n^k$.

Proof continued

For Condition 2, fix two points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ with $i < j$ and define $\mathcal{E}_{i,j}^{(2)}$ as the error event that $|\boldsymbol{x}_i \cdot \boldsymbol{x}_j| \geq \frac{\sqrt{2(k+2)\ln n}}{\sqrt{d-1}}$. Fix $\boldsymbol{u} = \boldsymbol{x}_i / \|\boldsymbol{x}_i\|$ as the direction of interest. We know from earlier for $d \geq 1$ and $\sqrt{2(k+2)\ln n} \geq 1$ that

$$\Pr\left\{\mathcal{E}_{i,j}^{(2)}\right\} \leq \Pr\left\{|\boldsymbol{u} \cdot \boldsymbol{x}_j| \geq \frac{\sqrt{2(k+2)\ln n}}{\sqrt{d-1}}\right\} \leq \frac{4}{\pi} e^{-\frac{2(k+2)\ln n}{2}} = \frac{4}{\pi n^{(k+2)}}$$

Then, overall error probability $\Pr\left\{\exists i < j : \mathcal{E}_{i,j}^{(2)}\right\} \leq \binom{n}{2} \Pr\left\{\mathcal{E}_{1,2}^{(2)}\right\} \leq \frac{4}{\pi n^k}$.

Proof continued

By union bound, we have that

$$\Pr\left\{\text{Condition 1 or 2 unsatisfied}\right\} \leq \frac{1}{n^k} + \frac{4}{\pi n^k} = O\left(\frac{1}{n^k}\right)$$

So the probability that Conditions 1 and 2 are **both** satisfied is at least $1 - O\left(1/n^k\right)$. ∎