

CSCI B609:

“Foundations of Data Science”

Lecture 3/4: High-Dimensional Space

Slides at <http://grigory.us/data-science-class.html>

Grigory Yaroslavtsev

<http://grigory.us>

Geometry of High Dimensions

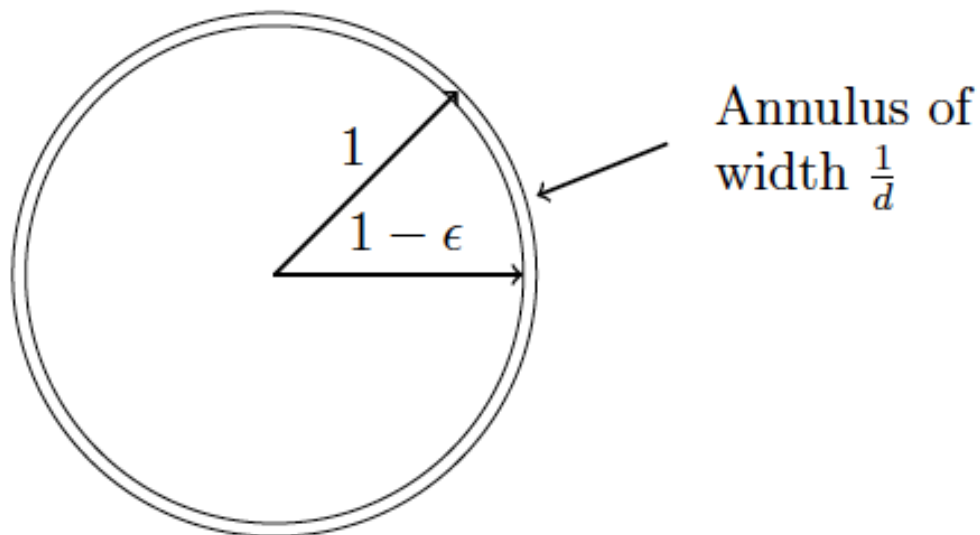
- Almost all volume near the surface:
 - Take arbitrary body $A \in \mathbb{R}^d$
 - Shrink to $(1 - \epsilon)A = \{(1 - \epsilon)x | x \in A\}$
 - Volume change:
$$\frac{\text{volume}((1 - \epsilon)A)}{\text{volume}(A)} = (1 - \epsilon)^d \leq e^{-\epsilon d}$$
 - Proof of $=$: partition into infinitesimal cubes

Today

- Geometry of High Dimensions (Sec 2.3 – 2.4)
 - Volume is near the surface
 - Volume of d -dimensional unit ball
 - Most of the volume is near equator
 - Near orthogonality of random vectors

Geometry of High Dimensions

- Let B_d = unit d -dimensional ball
- At least $1 - e^{-\epsilon d}$ fraction of its volume is in the annulus of width ϵ
- $\epsilon = O\left(\frac{1}{d}\right)$: most of the volume in the annulus



Volume of d-dimensional unit ball

- $V(\mathbf{d})$ = volume of \mathbf{d} -dimensional unit ball
- $S^{\mathbf{d}}$ = \mathbf{d} -dimensional unit sphere
- In spherical coordinates:
 - \mathbf{r} = radius
 - Ω = solid angle

$$V(\mathbf{d}) = \int_{\Omega \in S^{\mathbf{d}}} \int_{r=0}^1 \mathbf{r}^{\mathbf{d}-1} d\mathbf{r} d\Omega = \int_{\Omega \in S^{\mathbf{d}}} d\Omega \int_{r=0}^1 \mathbf{r}^{\mathbf{d}-1} d\mathbf{r}$$

- $\int_{r=0}^1 \mathbf{r}^{\mathbf{d}-1} d\mathbf{r} = \frac{1}{\mathbf{d}} \Rightarrow V(\mathbf{d}) = \frac{1}{\mathbf{d}} \int_{\Omega \in S^{\mathbf{d}}} d\Omega = \frac{A(\mathbf{d})}{\mathbf{d}}$
 - For $\mathbf{d} = 2$: $A(\mathbf{d}) = 2\pi \Rightarrow V(\mathbf{d}) = \pi$
 - For $\mathbf{d} = 3$: $A(\mathbf{d}) = 4\pi \Rightarrow V(\mathbf{d}) = \frac{4\pi}{3}$

Volume of d-dimensional unit ball

- $A(\mathbf{d}) = \int_{\Omega \in S^{\mathbf{d}}} d\Omega = ?$
- $I(\mathbf{d}) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} e^{-(x_1^2 + \dots + x_{\mathbf{d}}^2)} dx_1 \dots dx_{\mathbf{d}}$
- In Cartesian coordinates:

$$I(\mathbf{d}) = \left[\int_{-\infty}^{+\infty} e^{-x^2} dx \right]^{\mathbf{d}} = (\sqrt{\pi})^{\mathbf{d}} = \pi^{\frac{\mathbf{d}}{2}}$$

- In spherical coordinates:

$$\begin{aligned} I(\mathbf{d}) &= \int_{\Omega \in S^{\mathbf{d}}} d\Omega \int_{r=0}^{\infty} e^{-r^2} r^{\mathbf{d}-1} dr \\ &= A(\mathbf{d}) \int_{r=0}^{\infty} e^{-r^2} r^{\mathbf{d}-1} dr \end{aligned}$$

Volume of d-dimensional unit ball

- $I(\mathbf{d}) = A(\mathbf{d}) \int_{r=0}^{\infty} e^{-r^2} r^{\mathbf{d}-1} dr$

Let $r^2 = t$ (so $dt = 2rdr \Rightarrow dr = \frac{1}{2} t^{-\frac{1}{2}} dt$)

- $$\begin{aligned} \int_{r=0}^{\infty} e^{-r^2} r^{\mathbf{d}-1} dr &= \int_{r=0}^{\infty} e^{-t} t^{\frac{\mathbf{d}-1}{2}} \left(\frac{1}{2} t^{-\frac{1}{2}} dt \right) \\ &= \frac{1}{2} \int_{r=0}^{\infty} e^{-t} t^{\frac{\mathbf{d}}{2}-1} dt = \frac{1}{2} \Gamma\left(\frac{\mathbf{d}}{2}\right) \end{aligned}$$

- $\Gamma(x)$ = Gamma-function (generalized factorial)

- $\Gamma(x) = (x-1)\Gamma(x-1)$; $\Gamma(1) = \Gamma(2) = 1$; $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

- We have: $I(\mathbf{d}) = (\sqrt{\pi})^{\mathbf{d}} = \frac{A(\mathbf{d})}{2} \Gamma\left(\frac{\mathbf{d}}{2}\right) \Rightarrow$

$$A(\mathbf{d}) = \frac{2 \pi^{\frac{\mathbf{d}}{2}}}{\Gamma\left(\frac{\mathbf{d}}{2}\right)}; \quad V(\mathbf{d}) = \frac{2 \pi^{\frac{\mathbf{d}}{2}}}{\mathbf{d} \Gamma\left(\frac{\mathbf{d}}{2}\right)}$$

Volume of d-dimensional unit ball

- $A(\mathbf{d}) = \frac{2 \pi^{\frac{\mathbf{d}}{2}}}{\Gamma(\frac{\mathbf{d}}{2})}$; $V(\mathbf{d}) = \frac{2 \pi^{\frac{\mathbf{d}}{2}}}{\mathbf{d} \Gamma(\frac{\mathbf{d}}{2})}$
- $\mathbf{d} = 2$:
 - $A(2) = \frac{2\pi}{\Gamma(1)} = 2\pi$; $V(2) = \frac{\pi}{\Gamma(1)} = \pi$
- $\mathbf{d} = 3$:
 - $V(3) = \frac{2 \pi^{3/2}}{3 \Gamma(3/2)} = \frac{4 \pi^{3/2}}{3 \Gamma(1/2)} = \frac{4}{3} \pi$
 - $A(3) = \frac{2\pi^{3/2}}{\Gamma(3/2)} = 4\pi$
- $\Gamma(\frac{\mathbf{d}}{2})$ grows as a factorial of \mathbf{d} : $\lim_{\mathbf{d} \rightarrow \infty} V(\mathbf{d}) = 0$

Most of the volume is near equator

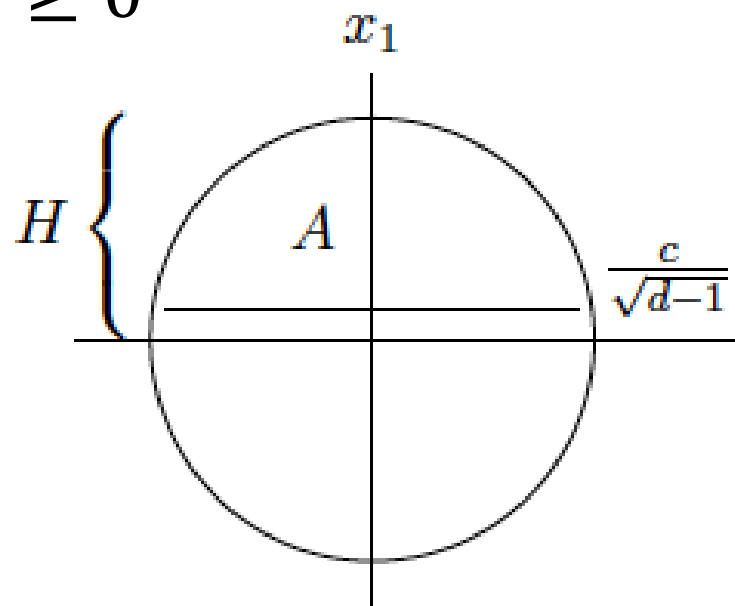
- x_1 = arbitrary coordinate
- Most of the volume has $|x_1| = O\left(\frac{1}{\sqrt{d}}\right)$
- For $c \geq 1$ and $d \geq 3$ at least $1 - \frac{2}{c} e^{-\frac{c^2}{2}}$ fraction of the volume of the d -dimensional unit ball has

$$|x_1| \leq \frac{c}{\sqrt{d-1}}$$

Most of the volume is near equator

- Will show: $\leq \left(\frac{2}{\textcolor{red}{c}} e^{-\frac{\textcolor{red}{c}^2}{2}} \right)$ -fraction of volume of hemisphere $x_1 \geq 0$ has $x_1 \geq \frac{\textcolor{red}{c}}{\sqrt{\textcolor{blue}{d}-1}}$
- A = portion with $x_1 \geq \frac{\textcolor{red}{c}}{\sqrt{\textcolor{blue}{d}-1}}$
- H = entire upper hemisphere $x_1 \geq 0$
- Will show:

$$\frac{\text{vol}(A)}{\text{vol}(H)} \leq \frac{\text{upper bound vol}(A)}{\text{lower bound vol}(H)}$$



Upper bound on vol(A)

- $\text{vol}(A)$: integrate volume of the disk of width dx_1 with face = ($d - 1$)-dim. ball of radius $\sqrt{1 - x_1^2}$
- Surface area of the disk = $(1 - x_1^2)^{\frac{d-1}{2}} V(d - 1)$

- $\text{vol}(A) = \int_{\frac{c}{\sqrt{d-1}}}^1 (1 - x_1^2)^{\frac{d-1}{2}} V(d - 1) dx_1$

- Use $(1 - x) \leq e^{-x}$ and $\frac{x_1 \sqrt{d-1}}{c} \geq 1$:

$$\text{vol}(A) \leq \int_{\frac{c}{\sqrt{d-1}}}^{\infty} \frac{x_1 \sqrt{d-1}}{c} e^{-\frac{d-1}{2} x_1^2} V(d - 1) dx_1 = \dots$$

$$= V(d - 1) \frac{\sqrt{d-1}}{c} \times \frac{1}{d-1} e^{-\frac{c^2}{2}} = V(d - 1) \frac{e^{-\frac{c^2}{2}}}{c \sqrt{d-1}}$$

Lower bound on $\text{vol}(H)$

- $\text{vol}(H)$ = volume of hemisphere with $x_1 \leq \frac{c}{\sqrt{d-1}}$
- $\text{vol}(H) \geq$ volume of hemisphere with $x_1 \leq \frac{1}{\sqrt{d-1}}$
- $\text{vol}(H) \geq$ volume of cylinder with:
 - Height: $h = \frac{1}{\sqrt{d-1}}$
 - Radius: $R = \sqrt{1 - \frac{1}{d-1}}$
- Volume of cylinder = $h \times V(d-1)R^{d-1} =$

$$\frac{1}{\sqrt{d-1}} V(d-1) \left(1 - \frac{1}{d-1}\right)^{\frac{d-1}{2}} \geq \frac{V(d-1)}{2\sqrt{d-1}}$$
- Last inequality since $(1-x)^a \geq 1-ax$ (for $a \geq 1$)

Putting things together

$$\begin{aligned} \frac{\text{vol}(A)}{\text{vol}(H)} &\leq \frac{\text{upper bound } \text{vol}(A)}{\text{lower bound } \text{vol}(H)} \\ &\leq \frac{V(\textcolor{blue}{d} - 1) \frac{e^{-\textcolor{red}{c}^2/2}}{\textcolor{red}{c}\sqrt{\textcolor{blue}{d} - 1}}}{\frac{V(\textcolor{blue}{d} - 1)}{2\sqrt{\textcolor{blue}{d} - 1}}} = \frac{2 e^{-\textcolor{red}{c}^2/2}}{\textcolor{red}{c}} \end{aligned}$$

- **Q:** Why didn't we use $\text{vol}(H) = \frac{1}{2} V(\textcolor{blue}{d})$?

Today:

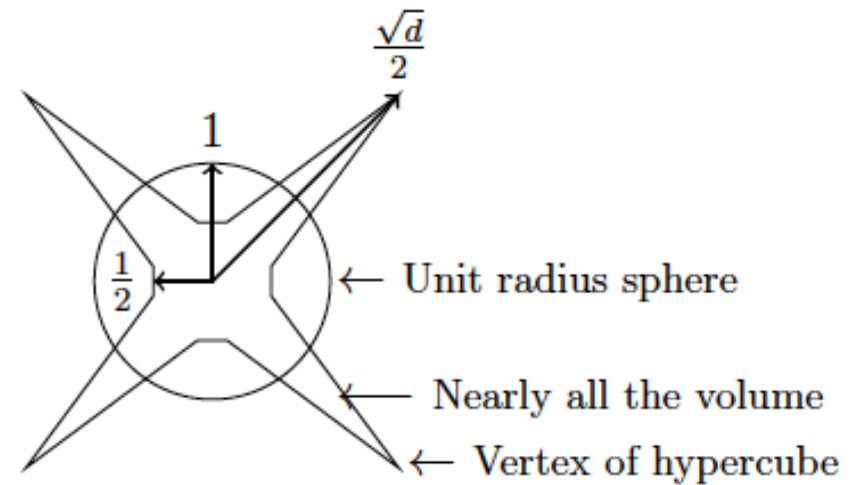
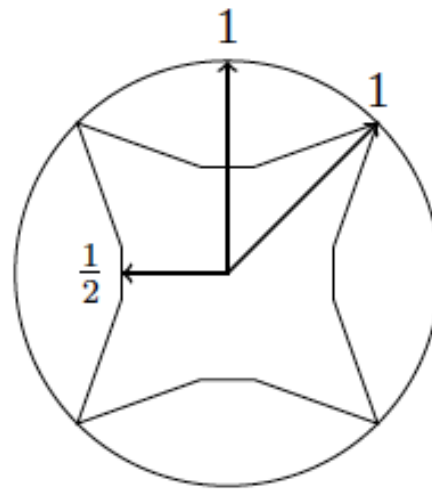
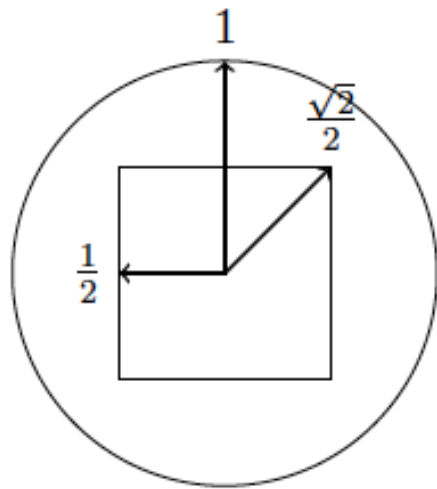
≈ Sec 2.4.2 – 2.7

- Near orthogonality of random vectors
- Sampling Uniform Distribution over B_d
- Gaussian Annulus Theorem (concentration)
- Nearest neighbor search & random projections

Near orthogonality

- Consider drawing n points x_1, \dots, x_n at random from the unit d -dimensional ball
- **Thm:** With probability $1 - O(1/n)$:
 - For all i : $\|x_i\|_2 \geq 1 - \frac{2 \ln n}{d}$
 - For all $i \neq j$: $|\langle x_i, x_j \rangle| \leq \frac{\sqrt{6 \ln n}}{\sqrt{d-1}}$
- $\Pr \left[\|x_i\|_2 < 1 - \frac{2 \ln n}{d} \right] \leq e^{-\frac{2 \ln n}{d} d} = \frac{1}{n^2}$
- $\Pr \left[|\langle x_i, x_j \rangle| > \frac{\sqrt{6 \ln n}}{\sqrt{d-1}} \right] \leq O\left(e^{-\frac{6 \ln n}{2}}\right) = O(n^{-3})$
- + Union bound (over n vectors and $O(n^2)$ pairs)

Sphere vs. cube in 2, 4, d dimensions



Sampling Uniform Distribution over B_d

- How to sample uniformly from a unit ball?
- Sample uniformly from a unit cube
 - Output the sample if inside B_d
 - Repeat if outside B_d
- Number of repetitions to output a sample?

Normal Distribution

- Normal distribution $N(0,1)$
 - Range: $(-\infty, +\infty)$
 - Density: $\mu(x) = (2\pi)^{-\frac{1}{2}} e^{-\frac{x^2}{2}}$
 - Mean = 0, Variance = 1
- Basic facts:
 - If X and Y are independent r.v. with normal distribution then $X + Y$ has normal distribution
 - $Var[cX] = c^2 Var[X]$
 - If X, Y are independent, then:
$$Var[X + Y] = Var[X] + Var[Y]$$

Sampling Uniform Distribution over B_d

- Sample x_1, x_2, \dots, x_d i.i.d with $x_i \sim N(0,1)$
- $\Pr[x_i = z] = (2\pi)^{-\frac{1}{2}} e^{-\frac{z^2}{2}}$
- $\Pr[\mathbf{x} = (z_1, \dots, z_d)] = (2\pi)^{-\frac{d}{2}} e^{-\frac{z_1^2 + z_2^2 + \dots + z_d^2}{2}}$
- $\frac{\mathbf{x}}{\|\mathbf{x}\|_2} \sim U(S_d)$, how to make it $U(B_d)$?
- Scale by $\rho \in [0,1]$: $U(B_d) = \frac{\rho^{\mathbf{x}}}{\|\mathbf{x}\|_2}$ for $\rho(r) = dr^{d-1}$

$$V(d) = \int_0^1 A(d) r^{d-1} dr \Rightarrow 1 = \int_0^1 \frac{A(d)}{V(d)} r^{d-1} dr$$

$= d$

Gaussian Annulus Theorem

- Gaussian in d dimensions ($N_d(0^d, 1)$):

$$\Pr[\mathbf{x} = (z_1, \dots, z_d)] = (2\pi)^{-\frac{d}{2}} e^{-\frac{z_1^2 + z_2^2 + \dots + z_d^2}{2}}$$

Nearly all mass in annulus of radius \sqrt{d} and width $O(1)$:

- **Thm.** For any $\beta \leq \sqrt{d}$ all but $3e^{-c\beta^2}$ probability mass satisfies $\sqrt{d} - \beta \leq \|\mathbf{x}\|_2 \leq \sqrt{d} + \beta$ for constant c
- **Proof:** Let $\mathbf{y} = (y_1, \dots, y_d) \sim N_d(0^d, 1)$ and $r = \|\mathbf{y}\|_2$
 - $|r - \sqrt{d}| \geq \beta \Leftrightarrow |r^2 - d| \geq \beta(r + \sqrt{d}) \geq \beta\sqrt{d}$
 - Will bound $\Pr[|r^2 - d| \geq \beta\sqrt{d}]$

Gaussians in High Dimension

- Will bound $\Pr[|r^2 - \mathbf{d}| \geq \boldsymbol{\beta}\sqrt{\mathbf{d}}]$
- $r^2 - \mathbf{d} = (y_1^2 - 1) + \dots + (y_{\mathbf{d}}^2 - 1)$
- Let $x_i = y_i^2 - 1$, bound $\Pr[|\sum_{i=1}^{\mathbf{d}} x_i| \geq \boldsymbol{\beta}\sqrt{\mathbf{d}}]$
- $\mathbb{E}[x_i] = \mathbb{E}[y_i^2] - 1 = 0$
- Fix an integer $\mathbf{s} > 1$
 - For $|y_i| \leq 1$ we have $|x_i|^{\mathbf{s}} \leq 1$
 - For $|y_i| \geq 1$ we have $|x_i|^{\mathbf{s}} \leq |y_i|^{2\mathbf{s}}$
- $|\mathbb{E}[x_i^{\mathbf{s}}]| \leq \mathbb{E}[|x_i|^{\mathbf{s}}] \leq \mathbb{E}[1 + y_i^{2\mathbf{s}}] = 1 + \mathbb{E}[y_i^{2\mathbf{s}}]$
- $1 + \mathbb{E}[y_i^{2\mathbf{s}}] = 1 + \sqrt{2/\pi} \int_0^\infty y^{2\mathbf{s}} e^{-\frac{y^2}{2}} dy \leq 2^{\mathbf{s}} \mathbf{s}!$

Gaussians in High Dimension

Let $z = \sum_{i=1}^n z_i$ where z_i are i.i.d r.v.s: $\mathbb{E}[z_i] = 0$ and $\text{Var}[z_i] \leq \sigma^2$

Thm 12.5. If $a \in [0, \sqrt{2}n\sigma^2]$ and $a^2/(4n\sigma^2) \leq s \leq n\sigma^2/2$, s is an even integer and $|\mathbb{E}[z_i^r]| \leq \sigma^2 r!$ for all $r = 3, 4, \dots, s$

$$\Rightarrow \Pr \left[\left| \sum_{i=1}^n z_i \right| \geq a \right] \leq 3e^{-\frac{a^2}{12n\sigma^2}}$$

- Take $a = \beta\sqrt{d}$, $n = d$, scale $x_i \rightarrow w_i = \frac{x_i}{2}$
- $|\mathbb{E}[x_i^s]| \leq 2^s s! \Rightarrow |\mathbb{E}[w_i^s]| \leq s!$
- $\mathbb{E}[x_i] = 0 \Rightarrow \text{Var}[w_i] = \frac{1}{4} \text{Var}[x_i] = \frac{1}{4} \mathbb{E}[x_i^2] \leq \frac{2^2 2!}{4} = 2 = \sigma^2$

$$\Pr \left[\frac{1}{2} \left| \sum_{i=1}^n x_i \right| \geq \beta\sqrt{d} \right] \leq 3e^{-\beta^2}$$