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 $\boldsymbol{A} \in \mathbb{R}^{\boldsymbol{d}}$
  - Shrink to $(1-\epsilon)\boldsymbol{A} = \{(1-\epsilon)x | x \in \boldsymbol{A}\}$
  - Volume change:

$$\frac{volume\big((1-\epsilon)\boldsymbol{A}\big)}{volume(\boldsymbol{A})} = (1-\epsilon)^{\boldsymbol{d}} \leq e^{-\epsilon \boldsymbol{d}}$$

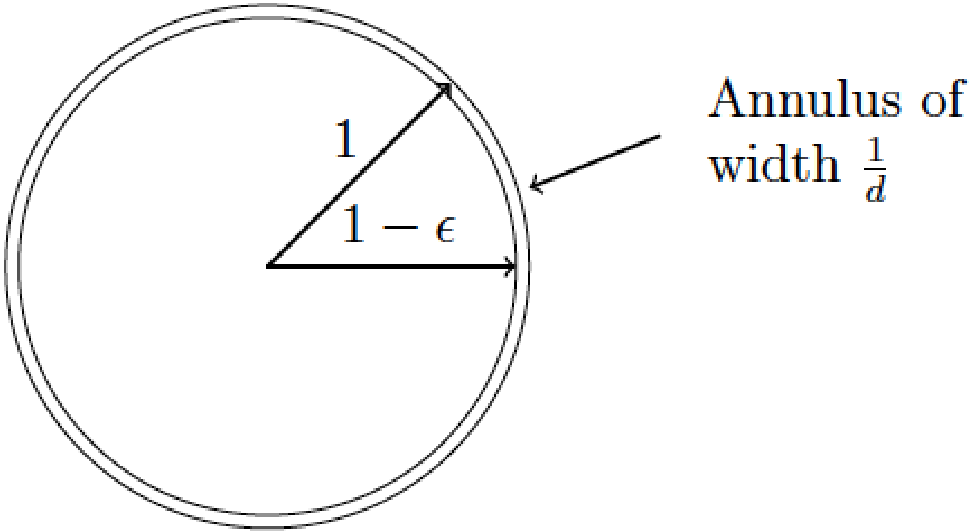
  - Proof of $=$: partition into infinitesimal cubes

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



Volume of d-dimensional unit ball

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

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

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

Volume of d-dimensional unit ball

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

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

- In spherical coordinates:

$$I(\boldsymbol{d}) = \int_{\Omega \in S^{\boldsymbol{d}}} d\Omega \int_{r=0}^{\infty} e^{-r^2} r^{\boldsymbol{d}-1} dr$$

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

Volume of d-dimensional unit ball

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

$Let\ \boldsymbol{r}^2 = \boldsymbol{t}$ (so $d\boldsymbol{t} = 2\boldsymbol{r}d\boldsymbol{r} \Rightarrow d\boldsymbol{r} = \frac{1}{2}\boldsymbol{t}^{-\frac{1}{2}}d\boldsymbol{t}$)

- $\int_{r=0}^{\infty} e^{-\boldsymbol{r}^2} \boldsymbol{r}^{\boldsymbol{d}-1} d\boldsymbol{r} = \int_{r=0}^{\infty} e^{-\boldsymbol{t}} \boldsymbol{t}^{\frac{\boldsymbol{d}-1}{2}} \left(\frac{1}{2}\boldsymbol{t}^{-\frac{1}{2}}d\boldsymbol{t}\right)$

$$= \frac{1}{2}\int_{r=0}^{\infty} e^{-\boldsymbol{t}}\, \boldsymbol{t}^{\frac{\boldsymbol{d}}{2}-1} d\boldsymbol{t} = \frac{1}{2}\,\Gamma\left(\frac{\boldsymbol{d}}{2}\right)$$

- $\Gamma(x)$ = 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: $I(\boldsymbol{d}) = (\sqrt{\pi})^{\boldsymbol{d}} = \frac{A(\boldsymbol{d})}{2}\,\Gamma\left(\frac{\boldsymbol{d}}{2}\right) \Rightarrow$

$$A(\boldsymbol{d}) = \frac{2\,\pi^{\frac{\boldsymbol{d}}{2}}}{\Gamma\left(\frac{\boldsymbol{d}}{2}\right)}; \qquad \mathrm{V}(\boldsymbol{d}) = \frac{2\,\pi^{\frac{\boldsymbol{d}}{2}}}{\boldsymbol{d}\,\Gamma\left(\frac{\boldsymbol{d}}{2}\right)}$$

Volume of d-dimensional unit ball

- $A(\boldsymbol{d}) = \frac{2\,\pi^{\frac{\boldsymbol{d}}{2}}}{\Gamma\left(\frac{\boldsymbol{d}}{2}\right)}$; $\mathrm{V}(\boldsymbol{d}) = \frac{2\,\pi^{\frac{\boldsymbol{d}}{2}}}{\boldsymbol{d}\,\Gamma\left(\frac{\boldsymbol{d}}{2}\right)}$
- $\boldsymbol{d} = 2$:
  - $\mathrm{A}(2) = \frac{2\pi}{\Gamma(1)} = 2\pi$; $\mathrm{V}(2) = \frac{\pi}{\Gamma(1)} = \pi$
- $\boldsymbol{d} = 3$:
  - $\mathrm{V}(3) = \frac{2\,\pi^{3/2}}{3\,\Gamma(3/2)} = \frac{4\,\pi^{3/2}}{3\,\Gamma(1/2)} = \frac{4}{3}\pi$
  - $\mathrm{A}(3) = \frac{2\pi^{3/2}}{\Gamma(3/2)} = 4\pi$
- $\Gamma\left(\frac{\boldsymbol{d}}{2}\right)$ grows as a factorial of $\boldsymbol{d}$: $lim_{\boldsymbol{d}\to\infty}\mathrm{V}(\boldsymbol{d}) = 0$

Most of the volume is near equator

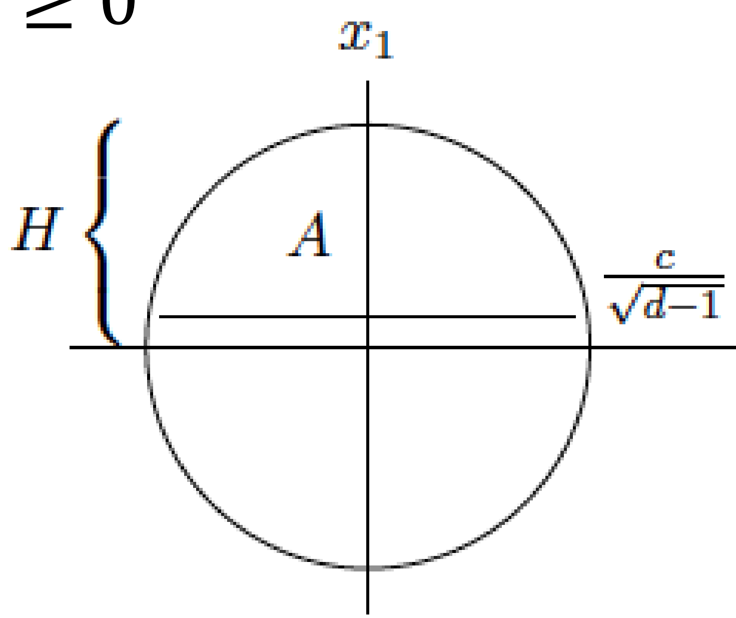
- x_1 = arbitrary coordinate
- Most of the volume has $|x_1| = O\left(\frac{1}{\sqrt{\boldsymbol{d}}}\right)$
- For $\boldsymbol{c} \geq 1$ and $\boldsymbol{d} \geq 3$ at least $1 - \frac{2}{\boldsymbol{c}}\, e^{-\frac{\boldsymbol{c}^2}{2}}$ fraction of the volume of the $\boldsymbol{d}$-dimensional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

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

Most of the volume is near equator

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

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



Upper bound on vol(A)

- vol(A): integrate volume of the disk of width dx_1 with face = ($\boldsymbol{d}-1$)-dim. ball of radius $\sqrt{1-x_1^2}$
- Surface area of the disk $= (1-x_1^2)^{\frac{\boldsymbol{d}-1}{2}} V(\boldsymbol{d}-1)$
- vol(A) $= \int_{\frac{\boldsymbol{c}}{\sqrt{\boldsymbol{d}-1}}}^{1} (1-x_1^2)^{\frac{\boldsymbol{d}-1}{2}} V(\boldsymbol{d}-1)\, dx_1$
- Use $(1-x) \le e^{-x}$ and $\frac{x_1\sqrt{\boldsymbol{d}-1}}{\boldsymbol{c}} \ge 1$:

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

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

Lower bound on vol(H)

- $vol(H) =$ volume of hemisphere with $x_1 \leq \frac{c}{\sqrt{d-1}}$
- $vol(H) \geq$ volume of hemisphere with $x_1 \leq \frac{1}{\sqrt{d-1}}$
- $vol(H) \geq$ volume of cylinder with:
  - Height: $\mathrm{h} = \frac{1}{\sqrt{d-1}}$
  - 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

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

$$\leq \frac{V(\boldsymbol{d}-1)\,\dfrac{e^{-\frac{\boldsymbol{c}^2}{2}}}{\boldsymbol{c}\sqrt{\boldsymbol{d}-1}}}{\dfrac{V(\boldsymbol{d}-1)}{2\sqrt{\boldsymbol{d}-1}}} = \frac{2\,e^{-\frac{\boldsymbol{c}^2}{2}}}{\boldsymbol{c}}$$

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

Today:

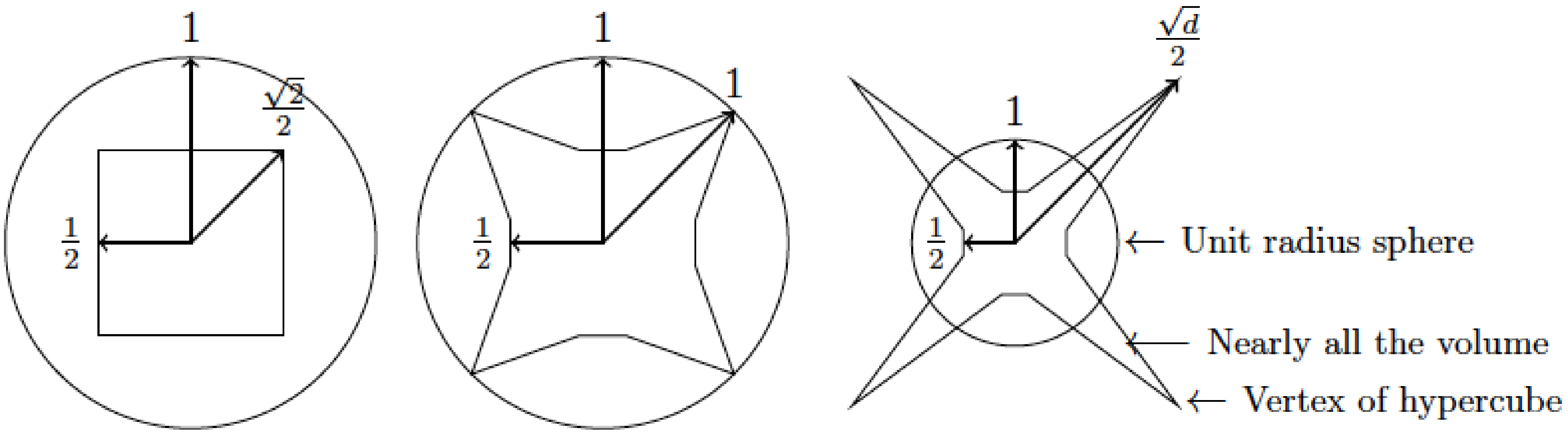
≈Sec 2.4.2 – 2.7

- Near orthogonality of random vectors
- Sampling Uniform Distribution over $B_{\boldsymbol{d}}$
- Gaussian Annulus Theorem (concentration)
- Nearest neighbor search & random projections

Near orthogonality

- Consider drawing n points $x_1, \ldots, x_n$ at random from the unit $\boldsymbol{d}$-dimensional ball
- **Thm:** With probability $1 - O(1/n)$:
  - For all i: $\left\| |x_i| \right\|_2 \geq 1 - \frac{2 \ln n}{\boldsymbol{d}}$
  - For all $i \neq j$: $\left| \langle x_i, x_j \rangle \right| \leq \frac{\sqrt{6 \ln n}}{\sqrt{\boldsymbol{d} - 1}}$
- $\Pr\left[\left\| |x_i| \right\|_2 < 1 - \frac{2 \ln n}{\boldsymbol{d}} \right] \leq e^{-\frac{2 \ln n}{\boldsymbol{d}} \boldsymbol{d}} = \frac{1}{n^2}$
- $\Pr\left[\left| \langle x_i, x_j \rangle \right| > \frac{\sqrt{6 \ln n}}{\sqrt{\boldsymbol{d} - 1}} \right] \leq O\left(e^{-\frac{6 \ln n}{2}} \right) = O(n^{-3})$
- + 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 $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent r.v. with normal distribution then $\boldsymbol{X} + \boldsymbol{Y}$ has normal distribution
  - $Var[c\boldsymbol{X}] = c^2 Var[\boldsymbol{X}]$
  - If $\boldsymbol{X}, \boldsymbol{Y}$ are independent, then:

$$Var[\boldsymbol{X} + \boldsymbol{Y}] = Var[\boldsymbol{X}] + Var[\boldsymbol{Y}]$$

Sampling Uniform Distribution over B_d

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

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

$$\frac{A(\boldsymbol{d})}{V(\boldsymbol{d})} = \boldsymbol{d}$$

Gaussian Annulus Theorem

- Gaussian in $\boldsymbol{d}$ dimensions $(N_{\boldsymbol{d}}(0^{\boldsymbol{d}}, 1))$:

$$\Pr[\boldsymbol{x} = (z_1, \ldots, z_{\boldsymbol{d}})] = (2\pi)^{-\frac{\boldsymbol{d}}{2}} e^{-\frac{z_1^2 + z_2^2 + \cdots + z_{\boldsymbol{d}}^2}{2}}$$

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

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

Gaussians in High Dimension

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

Gaussians in High Dimension

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

Thm 12.5. If $a \in \left[0, \sqrt{2}n\sigma^2\right]$ 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 $r = 3,4, \dots, s$

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

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

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