Introduction to High Dimensional Data

Scribe: Nickolas Fularczyk

1 Introduction

Let $X := \{x_1 \dots x_n\} \subset \mathbb{R}^d$ where $d \in \mathbb{N}$ and the coordinates of each point are randomly generated from a normal pdf with $\mu = 0,\ \sigma = 1$. Let's consider X as our dataset. A natural question to ask is: when d is large, how does high dimensional data compare with our intuition of 2D and 3D spaces? In learning about high dimensional data, we will discover the following observations:

- For d large, points are essentially equally spaced.
- Second, the volume of the unit ball, U_d, in $\mathbb{R}^d$ goes to 0 as $d \to \infty$.

The main idea we are going to use is applying the Law of Large Numbers. The law of large numbers says with high probability that the average of the samples, $|x-y|^2 = \sum_{i=1}^{d} |x_i - y_i|^2$ where $x, y \in \mathbb{R}^d$, will be close to the expectation of the random variable as d gets large. The precise statement is given in the following theorem, and the proof can be found in [1].

Theorem 1 (Law of Large Numbers). *Let $x_1 \dots x_n$ be n independent samples of a random variable x. Then,*

$$P(|\frac{1}{n}\sum_{i=1}^{n} x_i - \mathbb{E}(x)| > \varepsilon) \leq \frac{var(x)}{n\varepsilon^2}$$

Let's give a couple examples of applications of the law of large numbers.

Example 1. Let z be a random variable whose coordinates are taken from $\mathcal{N}(\mu = 0, \sigma = \frac{1}{2\pi})$ where $\mathcal{N}$ denotes a Gaussian distribution. With this choice of parameters, the pdf is 1 at the origin. Note, that the probability density function is bounded below by a constant over the unit ball.
By the law of large numbers,

$$|z-0|^2 = \sum_{i=1}^{d} z_i^2 \approx const\ d.$$

This shows that there is a vanishingly small probability that $z \in B(0,1)$.

Example 2. Let y, z be d-dimensional random variables whose coordinates are drawn from a normal pdf with $\mu = 0$, $\sigma = 1$. By the previous example, $|z|^2 \approx d$, $|y|^2 \approx d$. Observe that:

$$\begin{aligned} |y-z|^2 &= \sum_{i=1}^{d} |y_i - z_i|^2 \\ &\approx \mathbb{E}((y_i - z_i)^2)d \\ &= (\mathbb{E}(y_i^2) + \mathbb{E}(z_i^2) - 2\mathbb{E}(y_i z_i))d \\ &= 2d. \end{aligned}$$

By the Pythagorean theorem, y and z are essentially orthogonal.

2 Geometry of d-balls

Claim. Most of the volume is concentrated near the surface.

Given a set $A \subset \mathbb{R}^d$, we define

$$(1-\varepsilon)A := \{(1-\varepsilon)x : x \in A\}$$

If we shrink the set A by $1-\varepsilon$, then the resulting set is $(1-\varepsilon)A$. See Figure 1 for a visualization in two dimensions.

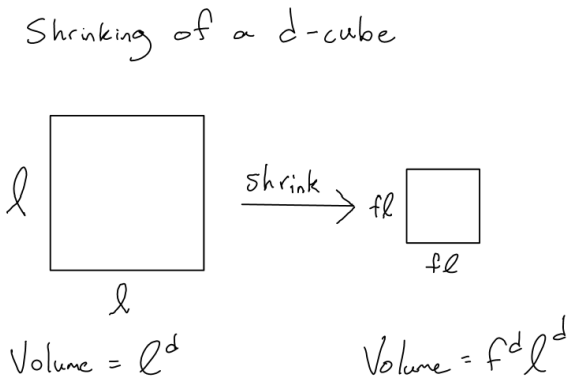

Fig. 1 Shrinking by a size length f implies a volume change by f^d

Observe that,

$$\frac{volume((1-\varepsilon)A)}{volume(A)} = (1-\varepsilon)^d$$
$$\leq e^{-\varepsilon d} \qquad [\,since\ (1-\varepsilon) \leq e^{-\varepsilon}]$$

For ε fixed, this suggests that the ratio goes to 0 and $d \to \infty$, showing that most of the volume of A is contained near the surface. To make this more precise, let U_d be the unit ball in $\mathbb{R}^d$. At least a $(1-e^{-\varepsilon d})$ fraction of its volume is concentrated in $\frac{U_d}{(1-\varepsilon)U_d}$, that is, on a shell of width ε near the surface. See the figure below from [1] for a visualization when $\varepsilon = \frac{1}{d}$.

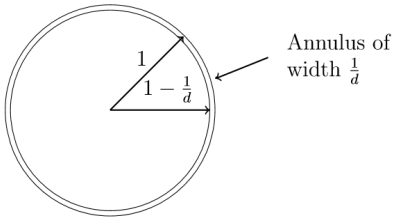

Figure 2.2: Most of the volume of the d-dimensional ball of radius r is contained in an annulus of width $O(r/d)$ near the boundary.

Fig. 2 Reprinted from *Foundations of Data Science* (p. 17), by A. Blum, J. Hopcroft and R. Kannan, 2020, Cambridge: Cambridge University Press. Copyright 2020 by Cambridge University Press.

Proposition 1. *As $d \to \infty$, $vol(U_d) \to 0$*

Proof. To prove this, lets first write the volume of U_d in polar coordinates.
That is,

$$vol(U_d) = \int_{S^d} d\Omega \int_{r=0}^{1} r^{d-1} dr = \frac{A(\Omega)}{d}$$

where $A(\Omega)$ is the surface of the d-sphere.
Consider,

$$I(d) = \int_{\mathbb{R}} \dots \int_{\mathbb{R}} e^{-x_1^2 - x_2^2 \dots - x_d^2} dx_1 \dots dx_d$$
$$= (\int_{\mathbb{R}} e^{-u^2} du)^d$$
$$= \pi^{\frac{d}{2}}$$

Moreover,

$$\begin{aligned} I(d) &= \int_{S^d} d\Omega \int_0^\infty e^{-r^2} r^{d-1} dr \\ &= A(\Omega) \int_0^\infty e^{-t} t^{\frac{d-1}{2}} (\frac{1}{2} t^{-\frac{1}{2}}) dt \\ &= A(\Omega) \frac{1}{2} \int_0^\infty e^{-t} t^{\frac{d}{2}-1} dt \\ &= A(\Omega) \frac{1}{2} \Gamma(\frac{d}{2}). \end{aligned}$$

It follows that

$$A(\Omega) = \frac{\pi^{\frac{d}{2}}}{\frac{1}{2}\Gamma(\frac{d}{2})}.$$

Thus,

$$vol(U_d) = \frac{2}{d} \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \to 0 \; as \; d \to \infty.$$

3 Most of the volume of U_d is near the equator

Let's first choose a vector x_1 to be the North pole. Let H denote the upper hemisphere of U_d, and let A be the fraction of H defined by $x_1 \geq \frac{c}{\sqrt{d-1}}$. This can be seen in the figure below from [1].

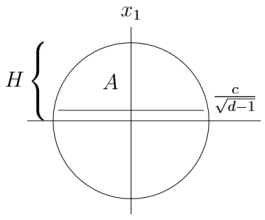

Figure 2.3: Most of the volume of the upper hemisphere of the d-dimensional ball is below the plane $x_1 = \frac{c}{\sqrt{d-1}}$.

Fig. 3 Reprinted from *Foundations of Data Science* (p. 21), by A. Blum, J. Hopcroft and R. Kannan, 2020, Cambridge: Cambridge University Press. Copyright 2020 by Cambridge University Press.

Theorem 2. *For $c \geq 1$ and $d \geq 3$, at least $\frac{2}{c} e^{-\frac{c^2}{2}}$ fraction of the volume of U_d has $|x_1| \leq \frac{c}{\sqrt{d-1}}$*

Proof. We will show that at most a $\frac{2}{c}e^{-\frac{c^2}{2}}$ fraction of the volume of H has $x_1 \geq \frac{c}{\sqrt{d-1}}$.
Specifically, we will show

$$\frac{vol(A)}{vol(H)} \leq \frac{2}{c}e^{-\frac{c^2}{2}}.$$

The surface area of a d-1 dimensional ball with radius $\sqrt{1-x_1^2}$ is $A_{d-1} = r^{d-1}V_{d-1}$. Integrating this in cylindrical coordinates gives,

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

Using $1-x \leq e^{-x}$, we get

$$vol(A) \leq \int_{\frac{c}{\sqrt{d-1}}}^{\infty} e^{-(\frac{d-1}{2})x_1^2} V_{d-1} dx_1.$$

By Hypothesis,

$$x_1 \geq \frac{c}{\sqrt{d-1}} \implies 1 \leq \frac{x_1\sqrt{d-1}}{c}.$$

Therefore,

$$\begin{aligned} vol(A) &\leq \int_{\frac{c}{\sqrt{d-1}}}^{\infty} \frac{x_1\sqrt{d-1}}{c} e^{-(\frac{d-1}{2})x_1^2} V_{d-1} dx_1 \\ &= V_{d-1}\frac{\sqrt{d-1}}{c}\int_{\frac{c}{\sqrt{d-1}}}^{\infty} x_1 e^{-(\frac{d-1}{2})x_1^2} dx_1 \\ &= V_{d-1}\frac{\sqrt{d-1}}{c}\frac{1}{d-1}e^{-\frac{c^2}{2}}. \end{aligned}$$

Thus,

$$vol(A) \leq \frac{V_{d-1}}{c\sqrt{d-1}}e^{-\frac{c^2}{2}}.$$

Next, we estimate $vol(H)$. The idea is to find a lower bound for $vol(H)$.
We have

$$\begin{aligned} vol(H) &\geq vol(\textit{cylinder of height } \frac{1}{\sqrt{d-1}} \textit{ with radius } \sqrt{1-\frac{1}{d-1}} \\ &= V_{d-1}(1-\frac{1}{d-1})^{\frac{d-1}{2}}\frac{1}{\sqrt{d-1}} \end{aligned}$$

where $V_{d-1}(1-\frac{1}{d-1})^{d-1}$ is the surface area of a ball of dimension d-1 with radius $\sqrt{1-\frac{1}{d-1}}$ and $\frac{1}{\sqrt{d-1}}$ is the height.
Note:

$$\begin{aligned}(1-\frac{1}{d-1})^{\frac{d-1}{2}} &\geq 1-(\frac{d-1}{2})(\frac{1}{d-1}) \\ &= 1-\frac{1}{2} \\ &= \frac{1}{2}.\end{aligned}$$

Using if $\alpha \geq 1$ then $(1-x)^{\alpha} \geq 1-\alpha x$ and that $d \geq 3$, we have

$$vol(H) \geq \frac{V_{d-1}}{2\sqrt{d-1}}.$$

Thus,

$$\begin{aligned}\frac{vol(A)}{vol(H)} &\leq \frac{\frac{V_{d-1}}{c\sqrt{d-1}}e^{-\frac{c^2}{2}}}{\frac{V_{d-1}}{2\sqrt{d-1}}} \\ &= \frac{2}{c}e^{-\frac{c^2}{2}}.\end{aligned}$$

4 Near orthogonality

Idea: 2 points drawn at random from U_d are almost orthogonal with high probability.

Theorem 3. *Let $x_1,\ldots,x_n$ be drawn at random from $U_d \subset \mathbb{R}^d$ Then, with probability 1-$\mathcal{O}(\frac{1}{n})$,*

1. $|x_i| \geq 1-2ln(\frac{n}{d}) \quad \forall i = 1,\ldots,n$
2. $|x_i - x_j| \leq \frac{\sqrt{6ln(n)}}{\sqrt{d-1}}$ *when* $i \neq j$

Proof. (1): Recall,

$$\frac{vol(1-\varepsilon)A}{vol(A)} \leq e^{-\varepsilon d}.$$

Choose $\varepsilon = \frac{2ln(n)}{d}$ then,

$$Prob(|x_i| < 1-\frac{2ln(n)}{d}) \leq e^{-(\frac{2ln(n)}{d})d} = n^{-2}.$$

Using the prior bound,

$$P(|x_1| < 1-\frac{2ln(n)}{d} \text{ or } \ldots \text{ or } |x_n| < \frac{2ln(n)}{d}) \leq 1/n.$$

It follows that

$$P(|x_i| > 1-\frac{2ln(n)}{d}) \geq 1-\frac{1}{n}.$$

(2): We are examining dot products $x_i \cdot x_j$, $i \neq j$. There are $\binom{n}{2}$ such pairs. For each pair, fix x_i as the north pole. We use Theorem 2 with $c = \sqrt{6ln(n)}$. This gives,

$$P(|x_i \cdot x_j| \leq \frac{\sqrt{6ln(n)}}{d-1}) \geq 1 - \frac{2}{\sqrt{6ln(n)}} e^{-3ln(n)}$$
$$= 1 - \frac{2}{\sqrt{6ln(n)}} n^{-3}.$$

Hence,

$$P(|x_i \cdot x_j| > \frac{\sqrt{6ln(n)}}{d-1}) < \frac{2}{\sqrt{6ln(n)}} n^{-3} = \mathcal{O}(n^{-3})$$

for one point. This condition is violated with probability at most $\mathcal{O}((\frac{n}{2})n^{-3}) = \mathcal{O}(n^{-1})$.

Remarks:
Using $c = 2\sqrt{ln(d)}$ in Theorem 1, it gives that the fraction of the volume of the ball satisfying $|x_1| > \frac{c}{\sqrt{d-1}} = \frac{2\sqrt{ln(d)}}{\sqrt{d-1}}$ is at most $\frac{2}{c} e^{\frac{-c^2}{2}} = (\ldots) = \frac{d^{-2}}{\sqrt{ln(d)}} < d^{-2}$. This estimate holds for any of the coordinates, so that at most $\mathcal{O}(d^{-1})$ of the volume of the ball is outside a cube of side length $\frac{2c}{\sqrt{d-1}} = 4\frac{\sqrt{ln(d)}}{\sqrt{d-1}}$. Hence, the unit ball has value at most twice the volume of the cube, that is,

$$vol(U_d) \geq (\frac{16ln(d)}{d-1})^{\frac{d}{2}} \to 0 \; as \; d \to \infty.$$

We recover the result that $vol(U_d) \to 0 \; as \; d \to \infty$.

5 Relationship between a ball and cube in $\mathbb{R}^d$

For large d, we expect to find most points of the unit ball U_d near the surface and at the same time, within a cube of side length

$$\mathcal{O}(\frac{ln(d)}{d-1}).$$

Note: on the surface of U_d a point satisfies

$$x_1^2 + \ldots + x_d^2 = 1,$$

so that typically $x_i \approx \pm\frac{1}{\sqrt{d}}$ for each i. Therefore, a randomly drawn point from the surface of U_d is of the form $(\pm\frac{1}{\sqrt{d}}, \ldots, \pm\frac{1}{\sqrt{d}})$. Look at the following figure from [1] to get an idea for the relative geometry of U_d and a cube inside U_d.

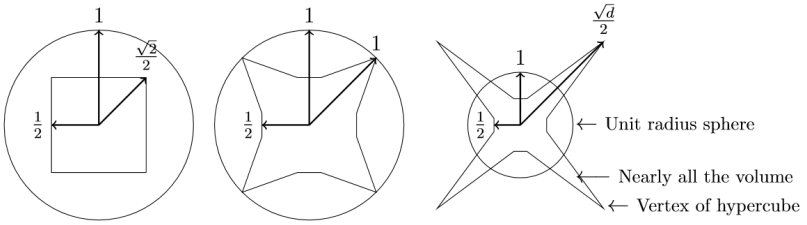

Figure 2.4: Illustration of the relationship between the sphere and the cube in 2, 4, and d-dimensions.

Fig. 4 Reprinted from *Foundations of Data Science* (p. 22), by A. Blum, J. Hopcroft and R. Kannan, 2020, Cambridge: Cambridge University Press. Copyright 2020 by Cambridge University Press.

6 Generating Points Uniformly at Random on the surface of a sphere

In applications, it is useful to be able to generate points uniformly at random on the surface of a sphere. The question is how to implement this. Let's examine one possible approach for the case when $d = 2$. To generate a point (x,y), we start by generating x and y coordinates chosen uniformly at random from $[-1,1]$ to form the point (x,y). We repeat this process to get a collection of points that have been generated random uniformly in $[-1,1]$. We then project these points onto the circumference where we discard points outside the disk since rays have nonuniform length. By only keeping points inside the disk and projecting on the circumference, we generate a collection of samples uniformly at random on the circumference of the sphere.

What happens when d is large? The same idea does not work since the number of points inside the sphere would be negligible and likely outside the boundary. We need to use an alternative approach. Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^d$ where each coordinate x_i is chosen independently from $\mathcal{N}(\mu = 0, \sigma^2 = 1)$. As a result,

$$x \sim \textit{ spherically symmetric normal pdf with } \sigma^2 = 1.$$

By normalization, the element $\frac{x}{|x|}$ is a unit vector whose distribution is uniform over the surface of U_d.
Remark:

- To generate a random number from a given pdf uniformly at random over [0,1], let $x = P^{-1}(\mu)$ where P is the cumulative distribution function.

- To generate a point y uniformly at random *over a unit ball*, we need to scale a point $\frac{x}{|x|}$ on the surface of the sphere by a scalar $\rho \in [0,1]$. Clearly, $\rho = \rho(r)$, where r is the radius. For d=2, $\rho(r) \propto r$. In general, for general d, $\rho(r) \propto r^{d-1}$. Since the pdf has area 1,

$$1 = \int_{r=0}^{r=1} \rho(r)dr = \int_{r=0}^{r=1} cr^{d-1}dr$$

which implies c=d. Thus, $\rho(r) = dr^{d-1}$. In conclusion, to generate a point uniformly at random over a unit sphere, first generate $\frac{x}{|x|}$ using the spherical Gaussian pdf and then re-scale as $y = \rho\frac{x}{|x|}$.

7 Spherically Symmetric Gaussian in High Dimensions

Let's briefly compare the 1-dimensional Gaussian and the d-dimensional Gaussian before going into the details. Recall for d=1,

$$\rho(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}.$$

It is known that the area is concentrated near the origin.
In general for d dimensions,

$$\rho_\sigma(x) = \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d} e^{-\frac{|x|^2}{2\sigma^2}}$$

has very little volume near the origin. In fact, $\int_0^1 \rho_\sigma(x)dx$, $x \in \mathbb{R}^d$ is negligible for d large. Some intuition as to why is to observe that,

$$\mathbb{E}(|x|^2) = \sum_{i=1}^{d} \mathbb{E}(|x_i|^2) = \mathbb{E}(x_i^2)d = \sigma^2 d.$$

This shows that the mean square distance of a point from the origin is $\sigma\sqrt{d} = \mathcal{O}(\sqrt{d})$. In words, most of the points are about $\sqrt{d}$ away from the origin where we call $\sqrt{d}$ the radius of the Gaussian.

Theorem 4 (Gaussian Annulus theorem). *For a d-dimensional spherical Gaussian with $\mu = 0$, $\sigma = 1$, for any $\beta \leq \sqrt{d}$, we have*

$$\int_{\sqrt{d}-\beta \leq |x| \leq \sqrt{d}+\beta} \rho_\sigma(x)dx \geq 1 - 3e^{-c\beta^2}$$

for a fixed constant $c > 0$.

Proof. Let $y = (y_1, \ldots, y_d)$ and $r = |y|$. If $|r - \sqrt{d}| > \beta$ then

$$|r^2-d| = |r+\sqrt{d}||r-\sqrt{d}| \geq (r+\sqrt{d})\beta \geq \beta\sqrt{d} \tag{1}$$

Also, observe that

$$r^2-d = y_1^2+\ldots+y_d^2-d = (y_1^2-1)+\ldots+(y_d^2-1).$$

Set $x_i = y_i^2-1$ for each $i \in \{1,\ldots,d\}$ then (1) becomes

$$|x_1+\ldots+x_d| \geq \beta\sqrt{d}.$$

Notice that,

$$\mathbb{E}[x_i] = \mathbb{E}[y_i^2-1] = \mathbb{E}[y^2]-1 = 0$$

The next step is to apply the following theorem which is a theorem about tail bounds.

Theorem 5. *Let $x = x_1+\ldots+x_n$ where the x_i are mutually independent random variables with $\mu = 0$, and variance at most σ^2. Suppose $a \in [0, \sqrt{2}n\sigma^2]$ and $s \leq n\frac{\sigma^2}{2}$ is a positive even number and $\mathbb{E}[x_i^r] \leq \sigma^2 r!$ for $r = 3,\ldots,s$. Then,*

$$P(|x_1+\ldots+x_n| \geq a) \leq (\frac{2sn\sigma^2}{a^2})^{\frac{s}{2}}.$$

If, in addition, $s \geq \frac{a^2}{4n\sigma^2}$ then

$$P(|x_1+\ldots+x_n| \geq a) \leq 3e^{-\frac{a^2}{12n\sigma^2}}$$

Remark: The x_i need not be i.i.d only independent. To apply this theorem, we need to verify the bound on high order moments. Let s be a positive integer as above.

$$If\ |y_i| \leq 1, then\ |x_i|^s \leq 1.$$

$$If\ |y_i| > 1, then\ |x_i|^s \leq |y_i|^{2s}.$$

Hence $|x_i|^s \leq 1+y_i^{2s}$.
Let's find an upper bound for $|\mathbb{E}[x_i^s]|$,

$$\begin{aligned}|\mathbb{E}[x_i^s]| \leq \mathbb{E}(1+y_i^{2s}) &= 1+\mathbb{E}(yi^{2s}) \\ &= 1+\frac{2}{\pi}\int_0^\infty y^{2s}e^{-\frac{y^2}{2}}dy \\ &\leq 2^s s! \qquad [Upper\ Bound\ for\ the\ Gamma\ integral]\end{aligned}$$

Thus, $|\mathbb{E}[x_i^r]| \leq 2^r r!$ and $\mathbb{E}[x_i^2] \leq 2^2 2 = 8 = \sigma^2$. To get the right estimate to apply the tail bounds theorem, we need to use a change of variables. Let $w_i = \frac{x_i}{2}$.
Now,

$$var(w_i) \leq 2,\ \mathbb{E}[w_i^s] \leq 2s!.$$

We can now apply the theorem to obtain a bound on

$$P(|w_1 + \ldots + w_d| \geq \frac{\beta\sqrt{d}}{2}).$$

We derive that this bound is satisfied with probability less than or equal to $3e^{-\frac{\beta^2}{96}}$.

8 Random Projections

The material in this section is useful in nearest neighbor search algorithms.
Problem: We want to match a query point in $\mathbb{R}^d$ to a database.
For example, in facial recognition. This high dimensional problem can be computationally time consuming. To speed up the comparison, it is convenient to reduce the dimensionality of the problem. One thing to keep in mind is it is important to maintain the geometry when we reduce the dimensionality of the problem. That is, if points were close in $\mathbb{R}^d$ then the points should be close in $\mathbb{R}^k$. We consider the following approach. Let $u_1, \ldots, u_k$ be independent random vectors in $\mathbb{R}^d$ drawn from the spherical Gaussian: $(2\pi)^{-\frac{d}{2}} exp(-\frac{|x|^2}{2})$, $\sigma^2 = 1$.

Definition 1. For any $v \in \mathbb{R}^d$, we define the *random projection* $f : \mathbb{R}^d \to \mathbb{R}^k$ by

$$f(v) = (u_1 \cdot v, \ldots, u_k \cdot v).$$

Claim. With high probability $|f(v)| \approx \sqrt{k}|v|$.

It follows that if we want to measure $|v_1 - v_2|$, we can compute $|f(v_1) - f(v_2)| = |f(v_1 - v_2)| = \sqrt{k}|v_1 - v_2|$ The following theorem makes the claim precise.

Theorem 6 (Random Projection Theorem). *Let $v \in \mathbb{R}^d$ and f be defined as above. Then $\exists c > 0$ s.t for any $\varepsilon \in (0,1)$,*

$$P(|f(v) - \sqrt{k}|v|| \geq \varepsilon\sqrt{k}|v|) \leq 3e^{-ck\varepsilon^2}$$

where the probability is taken over the random draws of the vectors u_i.

Proof. We can divide both sides of the inequality inside $P(\cdot)$ by $|v|$. Hence, we can assume $|v| = 1$. Observe that for each $i = 1, \ldots, k$

$$u_i \cdot v = \sum_{j=1}^{d} u_{ij} v_j$$

Each u_{ij} has zero mean, hence $u_i \cdot v$ has zero mean. Since each u_{ij} has variance 1, the variance of $u_{ij}v_j$ is v_j^2. It follows that

$$var(u_i - v) = \sum v_j^2 = |v^2| = 1.$$

Since $u_1, \ldots, u_k$ are independent, we can apply the Gaussian Annulus Theorem with $\beta = \varepsilon\sqrt{k}$. This completes the proof.

References

1. Blum, Hopcroft, & Kannan, *Foundations of Data Science*, Cambridge University Press, Cambridge, 2020.