• Midterm will be next Wednesday-Thursday.
• See Piazza for study guide/practice questions.
• The quiz will include additional questions to help with revision.
• There will be class on Tuesday but not on Thursday.
Last Class: Finished Up Johnson-Lindenstrauss Lemma

- Completed the proof of the Distributional JL lemma.
- Discussed an application to $k$-means clustering.
- Started discussion of high-dimensional geometry.
Last Class: Finished Up Johnson-Lindenstrauss Lemma

- Completed the proof of the Distributional JL lemma.
- Discussed an application to $k$-means clustering.
- Started discussion of high-dimensional geometry.

This Class: High-Dimensional Geometry

- Bizarre phenomena in high-dimensional space.
- Connections to JL lemma and random projection.
- Balls and Bins as a way to summarize some of the material from this section.
What is the largest set of mutually orthogonal unit vectors in $d$-dimensional space? Answer: $d$. 

An exponentially large set of random vectors will be nearly pairwise orthogonal with high probability!
What is the largest set of mutually orthogonal unit vectors in $d$-dimensional space? Answer: $d$.

What is the largest set of unit vectors in $d$-dimensional space that have all pairwise dot products $|\langle \vec{x}, \vec{y} \rangle| \leq \epsilon$? (think $\epsilon = .01$) Answer: $2^{\Theta(\epsilon^2 d)}$. 

An exponentially large set of random vectors will be nearly pairwise orthogonal with high probability!
What is the largest set of mutually orthogonal unit vectors in $d$-dimensional space? Answer: $d$.

What is the largest set of unit vectors in $d$-dimensional space that have all pairwise dot products $|\langle \vec{x}, \vec{y} \rangle| \leq \epsilon$? (think $\epsilon = .01$)
Answer: $2^{\Theta(\epsilon^2 d)}$.

An exponentially large set of random vectors will be nearly pairwise orthogonal with high probability!
Claim: In $d$-dimensional space, a set of $2^{\Theta(\epsilon^2 d)}$ random unit vectors have all pairwise dot products at most $\epsilon$ (think $\epsilon = .01$)

Implies: $\|\vec{x}_i - \vec{x}_j\|_2^2 = \|\vec{x}_i\|_2^2 + \|\vec{x}_j\|_2^2 - 2\vec{x}_i^T\vec{x}_j \in [1.98, 2.02]$.

Even with an exponential number of random vector samples, we don’t see any nearby vectors.

- One version of the ’curse of dimensionality’.
- If all your distances are roughly the same, distance based methods ($k$-means clustering, nearest neighbors, SVMS, etc.) aren’t going to work well.
- Distances are only meaningful if we have lots of structure and our data isn’t just independent random vectors.
CURSE OF DIMENSIONALITY

Distances for MNIST Digits:

Distances for Random Images:
Recall: The Johnson Lindenstrauss lemma states that if \( \Pi \in \mathbb{R}^{m \times d} \) is a random matrix (linear map) with \( m = O \left( \frac{\log n}{\epsilon^2} \right) \), for \( \vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d \) with high probability, for all \( i, j \):

\[
(1 - \epsilon) \| \vec{x}_i - \vec{x}_j \|_2^2 \leq \| \Pi \vec{x}_i - \Pi \vec{x}_j \|_2^2 \leq (1 + \epsilon) \| \vec{x}_i - \vec{x}_j \|_2^2.
\]
Recall: The Johnson Lindenstrauss lemma states that if \( \Pi \in \mathbb{R}^{m \times d} \) is a random matrix (linear map) with \( m = O \left( \frac{\log n}{\epsilon^2} \right) \), for \( \vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d \) with high probability, for all \( i, j \):

\[
(1 - \epsilon) \| \vec{x}_i - \vec{x}_j \|_2^2 \leq \| \Pi \vec{x}_i - \Pi \vec{x}_j \|_2^2 \leq (1 + \epsilon) \| \vec{x}_i - \vec{x}_j \|_2^2.
\]

Implies: If \( \vec{x}_1, \ldots, \vec{x}_n \) are nearly orthogonal unit vectors in \( d \)-dimensions (with pairwise dot products bounded by \( \epsilon/8 \)), then \( \frac{\Pi \vec{x}_1}{\| \Pi \vec{x}_1 \|_2}, \ldots, \frac{\Pi \vec{x}_n}{\| \Pi \vec{x}_n \|_2} \) are nearly orthogonal unit vectors in \( m \)-dimensions (with pairwise dot products bounded by \( \epsilon \)).
**Claim 1:** $n$ nearly orthogonal unit vectors in any dimension $d$ can be projected to $m = O \left( \frac{\log n}{\epsilon^2} \right)$ dimensions and still be nearly orthogonal.

**Claim 2:** In $m$ dimensions, there can be at most $2^{O(\epsilon^2 m)}$ nearly orthogonal unit vectors.
Claim 1: $n$ nearly orthogonal unit vectors in any dimension $d$ can be projected to $m = O\left(\frac{\log n}{\epsilon^2}\right)$ dimensions and still be nearly orthogonal.

Claim 2: In $m$ dimensions, there can be at most $2^{O(\epsilon^2 m)}$ nearly orthogonal unit vectors.

• For both of these to hold it must be that $n \leq 2^{O(\epsilon^2 m)}$. 
Claim 1: $n$ nearly orthogonal unit vectors in any dimension $d$ can be projected to $m = O \left( \frac{\log n}{\epsilon^2} \right)$ dimensions and still be nearly orthogonal.

Claim 2: In $m$ dimensions, there can be at most $2^{O(\epsilon^2 m)}$ nearly orthogonal unit vectors.

- For both of these to hold it must be that $n \leq 2^{O(\epsilon^2 m)}$.
- I.e., $n = 2^{\log n} \leq 2^{O(\epsilon^2 m)}$ and so $m \geq O \left( \frac{\log n}{\epsilon^2} \right)$. 
Claim 1: $n$ nearly orthogonal unit vectors in any dimension $d$ can be projected to $m = O\left(\frac{\log n}{\epsilon^2}\right)$ dimensions and still be nearly orthogonal.

Claim 2: In $m$ dimensions, there can be at most $2^{O(\epsilon^2 m)}$ nearly orthogonal unit vectors.

- For both of these to hold it must be that $n \leq 2^{O(\epsilon^2 m)}$.
- I.e., $n = 2^{\log n} \leq 2^{O(\epsilon^2 m)}$ and so $m \geq O\left(\frac{\log n}{\epsilon^2}\right)$.
- Tells us that the JL lemma is optimal up to constants.
Let $\mathcal{B}_d$ be the unit ball in $d$ dimensions. $\mathcal{B}_d = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$. 

What percentage of the volume of $\mathcal{B}_d$ falls within $\epsilon$ distance of its surface?

Answer: all but a $(1 - \epsilon^d)^d \leq e^{-\epsilon^d}$ fraction. Exponentially small in the dimension $d$!
Let $B_d$ be the unit ball in $d$ dimensions. $B_d = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$.

What percentage of the volume of $B_d$ falls within $\epsilon$ distance of its surface?
Let $B_d$ be the unit ball in $d$ dimensions. $B_d = \{ x \in \mathbb{R}^d : \|x\|_2 \leq 1 \}$.

What percentage of the volume of $B_d$ falls within $\epsilon$ distance of its surface?

Volume of a radius $R$ ball is $\frac{\pi^{d/2}}{(d/2)!} \cdot R^d$. 
Let $B_d$ be the unit ball in $d$ dimensions. $B_d = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$.

What percentage of the volume of $B_d$ falls within $\epsilon$ distance of its surface? Answer: all but a $(1 - \epsilon)^d \leq e^{-\epsilon d}$ fraction. Exponentially small in the dimension $d$!

Volume of a radius $R$ ball is $\frac{\pi^{d/2}}{(d/2)!} \cdot R^d$. 
All but an $e^{-\epsilon^d}$ fraction of a unit ball’s volume is within $\epsilon$ of its surface.
All but an $e^{-\epsilon d}$ fraction of a unit ball’s volume is within $\epsilon$ of its surface. If we randomly sample points with $\|x\|_2 \leq 1$, nearly all will have $\|x\|_2 \geq 1 - \epsilon$. 

BIZARRE SHAPE OF HIGH-DIMENSIONAL BALLS
All but an $e^{-ed}$ fraction of a unit ball’s volume is within $\epsilon$ of its surface. If we randomly sample points with $\|x\|_2 \leq 1$, nearly all will have $\|x\|_2 \geq 1 - \epsilon$.

• **Isoperimetric inequality**: the ball has the minimum surface area/volume ratio of any shape.
All but an $e^{-\epsilon d}$ fraction of a unit ball’s volume is within $\epsilon$ of its surface. If we randomly sample points with $\|x\|_2 \leq 1$, nearly all will have $\|x\|_2 \geq 1 - \epsilon$.

- **Isoperimetric inequality**: the ball has the minimum surface area/volume ratio of any shape.

- If we randomly sample points from any high-dimensional shape, nearly all will fall near its surface.
All but an $e^{-\epsilon d}$ fraction of a unit ball’s volume is within $\epsilon$ of its surface. If we randomly sample points with $\|x\|_2 \leq 1$, nearly all will have $\|x\|_2 \geq 1 - \epsilon$.

- **Isoperimetric inequality**: the ball has the minimum surface area/volume ratio of any shape.

- If we randomly sample points from any high-dimensional shape, nearly all will fall near its surface.

- ‘All points are outliers.’
What fraction of the cubes are visible on the surface of the cube?

a) 80% b) 50% c) 25% d) 10%
What fraction of the cubes are visible on the surface of the cube?

a) 80%  b) 50%  c) 25%  d) 10%

\[
\frac{10^3 - 8^3}{10^3} = \frac{1000 - 512}{1000} = .488.
\]
SUMMARY OF FIRST SECTION
• **Probability Tools:** Linearity of Expectation, Linear of Variance of Independent Variables, Concentration Bounds (incl. Markov, Chebyshev, Bernstein, Chernoff), Union Bound, Median Trick.

• **Hash Tables and Bloom Filters:** Analyzing collisions. Building 2-level hash tables. Bloom filters and false positive rates.

• **Locality Sensitive Hashing:** MinHash for Jaccard Similarity and SimHash for Cosine Similarity. Nearest Neighbor. All-Pairs Similarity Search.

• **Small Space Data Stream Algorithms:** a) distinct items, b) frequent elements, c) frequency moments (homework).

• **Johnson Lindenstrauss Lemma:** Reducing dimension of vectors via a random projection such that pairwise differences are approximately preserved.
Most of the analysis of hash functions that we’ve considered can be abstracted as “balls and bins” problems: we throw $n$ balls and each ball is equally likely to land in one of $m$ bins.

Let $R_i$ be the number of balls that land in $i$th bin. We showed:

$$\mathbb{E}[R_i] = \frac{n}{m} \quad \mathbb{E}\left[\sum R_i^2\right] = \frac{n}{m} + \frac{n(n-1)}{m^2}$$

$$\Pr[\text{collisions}] = \Pr[\max(R_1, \ldots, R_m) > 1] \leq 1/8 \text{ if } m > 4n^2$$

and more generally

$$\Pr[\max(R_1, \ldots, R_m) \leq 2n/m] \leq \frac{m^2}{n}$$

In the exam, you’ll be expected to do calculations like these.
• Most of the analysis of hash functions that we’ve considered can be abstracted as “balls and bins” problems: we throw $n$ balls and each ball is equally likely to land in one of $m$ bins.

• Let $T$ be the number of bins where $R_i = 0$. We showed:

$$\mathbb{E}[T] = m(1 - 1/m)^n$$

and proved the variance in homework.

• The probability the next $k$ balls thrown all land in non-empty bins is

$$\left(1 - 1/T\right)^k$$

and this lets us analyze the false positive rate of a Bloom filter.