Announcements

1. Website updated with
   - lecture videos from the week before Spr Bk.
   - typeset notes from that week

2. Homework 5 updated with solution strategy for Problem (3).
   (Reminder: due Apr 15)

Dimensionality reduction

- Data set represented by n vectors in \( \mathbb{R}^d \).
  Assume \( n, d \) both large, e.g. 10^9.

Assume \( \| x - y \|_2 \) is a meaningful measure of similarity of data points \( x, y \).

E.g. \( \mathbb{R}^d \) in this example might be the output layer of a neural net that takes raw data and maps to a representation where \( l_2 \) dist. is a meaningful similarity measure.
Task. Try to find a set of points $x'_1, x'_2, \ldots, x'_n \in \mathbb{R}^k$ where $k \ll n$, such that

$$
\forall i, j \quad (1 - \epsilon) \|x'_i - x'_j\|_2 \leq \|x_i - x_j\|_2 \leq (1 + \epsilon) \|x'_i - x'_j\|_2
$$

Plan. Use a random linear transformation $\mathbb{R}^d \longrightarrow \mathbb{R}^k$ represented by a matrix $R$ with independent, Gaussian entries having distribution $\mathcal{N}(0, \frac{1}{k})$.

It will turn out that $k = 4 \ln(n/\delta)/\epsilon^2$ suffices, but we'll just keep calling dimension $k$ for now.
To analyze a \( \mathbb{R} \) random matrix \( k \times d \) with entries \( R_{ij} \sim N(0, \frac{1}{k}) \)

\[
y = x_i - x_j \in \mathbb{R}^d
\]

Want to know if \( \|Ry\|_2 \) is likely close to \( \|y\|_2 \)?

**Step 1:** Let \( Y = \|Ry\|_2^2 \). What is \( \mathbb{E}[Y] \)?

\[
R = \begin{bmatrix}
    - & r_1^T \\
    - & \_ \_ \_ \_ \_ \\
    - & \_ \_ \_ \_ \_ \\
    - & \_ \_ \_ \_ \_ \\
    - & r_k^T \\
\end{bmatrix} \quad R_y = \begin{bmatrix}
    \langle r_1, y \rangle \\
    \langle r_2, y \rangle \\
    \cdots \\
    \langle r_k, y \rangle \\
\end{bmatrix}
\]

\[
y = \|Ry\|_2^2 = \sum_{i=1}^k \langle r_i, y \rangle^2 = \sum_{i=1}^k x_i^2 \quad (x_i = \langle r_i, y \rangle)
\]

\( X_1, \ldots, X_k \) independent, identically distributed

\[
X_i \sim N(0, \frac{1}{k}\|y\|_2^2)
\]

\[
\mathbb{E}[x_i^2] = \frac{1}{k^2} \|y\|_2^2
\]

\[
\mathbb{E}[Y] = \sum_{i=1}^k \mathbb{E}[x_i^2] = k \cdot \frac{1}{k} \|y\|_2^2 = \|y\|_2^2
\]
Lemma 4.4 (Notes on Probability) tells us

\[ \forall \ 0 < \varepsilon < 1 \]

\[ \Pr \left( Y < (1 - \varepsilon)^2 \mathbb{E}Y \right) < e^{-\frac{k}{2} \varepsilon^2} \]

\[ \Pr \left( Y > (1 + \varepsilon)^2 \mathbb{E}Y \right) < e^{-\frac{k}{2} \varepsilon^2} \]

We want \( \| R(x_i - x_j) \|_2 \) to be close to \( \| x_i - x_j \|_2 \)
for \((n)\) distinct pairs \(ij\).

The analysis above with \( Y = \| R(x_i - x_j) \|_2^2 \)
shows that the approximation for
one single pair \(ij\) can fail with
probability \(< 2e^{-\frac{1}{2}k \varepsilon^2} \).

\underline{Union Bound:} Probability there is any pair
\(ij\) with \( \| R(x_i - x_j) \|_2 \) lying outside
interval \( [(1-\varepsilon) \| x_i - x_j \|_2, (1+\varepsilon) \| x_i - x_j \|_2] \)
is less than \( 2(\binom{n}{2}) e^{-\frac{1}{2}k \varepsilon^2} \)

Now make this less than 8...
\[ 2 \left( \frac{n}{2} \right) e^{-\frac{1}{2} k \varepsilon^2} < \delta \]

\[ e^{-\frac{1}{2} k \varepsilon^2} < \frac{\delta}{n(n-1)} \]

\[ e^{\frac{1}{2} k \varepsilon^2} > \frac{n(n-1)}{\delta} \]

\[ \frac{1}{2} k \varepsilon^2 > \ln \left( \frac{n(n-1)}{\delta} \right) \]

\[ k > 2 \ln \left( \frac{n(n-1)}{\delta} \right) / \varepsilon^2 \]

This inequality will hold when \( k > 4 \ln \left( \frac{n}{\delta} \right) / \varepsilon^2 \).

To see that \( \ln(n) / \varepsilon^2 \) dimensions are really needed, suppose \( x_1, \ldots, x_n \) are the standard basis vectors in \( \mathbb{R}^n \).

(\( \text{So } d = n \).)