

Lecture 8:

Dimension Reduction



Plan

- Pick up PS1 at the end of the class
- PS2 out

- Dimension Reduction
- Fast Dimension Reduction

- Scriber?

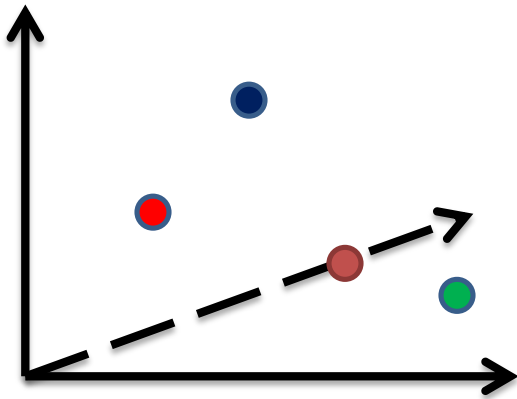
High-dimensional case

- Exact algorithms degrade rapidly with the dimension d

<i>Algorithm</i>	<i>Query time</i>	<i>Space</i>
Full indexing	$O(\log n \cdot d)$	$n^{O(d)}$ (Voronoi diagram size)
No indexing – linear scan	$O(n \cdot d)$	$O(n \cdot d)$

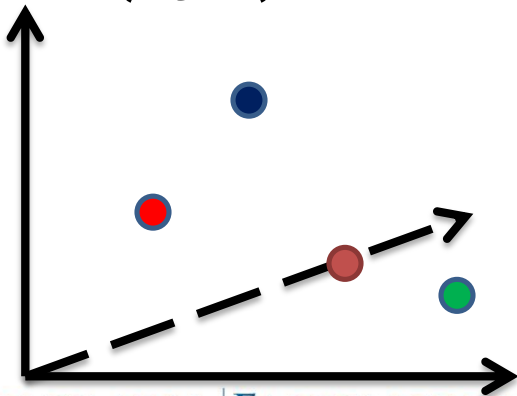
Dimension Reduction

- Reduce high dimension?!
 - “flatten” dimension d into dimension $k \ll d$
- Not possible in general: packing bound
- But can if: for a **fixed subset** of \mathfrak{R}^d



Johnson-Lindenstrauss Lemma

- **[JL84]:** There is a randomized linear map $F: \ell_2^d \rightarrow \ell_2^k$, $k \ll d$, that preserves distance between two vectors x, y
 - up to $1 + \epsilon$ factor:
$$\|x - y\| \leq \|F(x) - F(y)\| \leq (1 + \epsilon) \cdot \|x - y\|$$
 - with $1 - e^{-C\epsilon^2 k}$ probability (C some constant)
- Preserves distances between n points for $k = O\left(\frac{\log n}{\epsilon^2}\right)$ with probability at least $1 - 1/n$

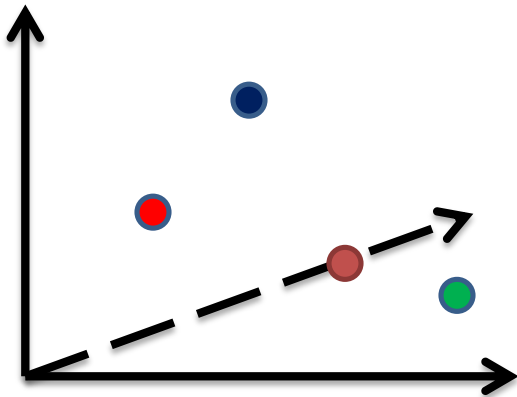


Dim-Reduction for NNS

- **[JL84]**: There is a randomized linear map $F: \ell_2^d \rightarrow \ell_2^k$, $k \ll d$, that preserves distance between two vectors x, y
 - up to $1 + \epsilon$ factor:
$$\|x - y\| \leq \|F(x) - F(y)\| \leq (1 + \epsilon) \cdot \|x - y\|$$
 - with $1 - e^{-C\epsilon^2 k}$ probability (C some constant)
- Application: NNS in ℓ_2^d
 - Trivial scan: $O(n \cdot d)$ query time
 - Reduce to $O(n \cdot k) + T_{dim-red}$ time after using dimension reduction
 - where $T_{dim-red}$ time to reduce dimension of the query point
 - Important that F is *oblivious* !
- Have we seen something similar to **JL84** in class?

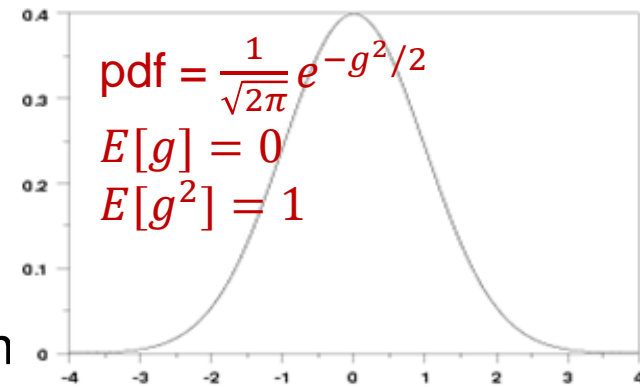
Idea:

- Project onto a *random* subspace of dimension k !
- In general, F linear:
 - $F(x) - F(y) = F(x - y)$
 - Ok to prove that for $z = x - y$
 - $F(z) \approx ||z||$



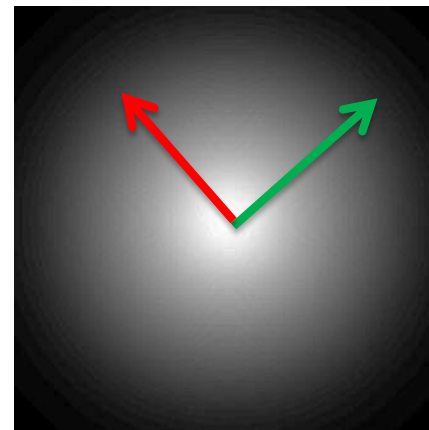
1D embedding

- Map $f: \ell_2^d \rightarrow \mathbb{R}$
 - $f(x) = \sum_i g_i \cdot x_i$,
 - where g_i are iid normal (Gaussian) ran



- Why Gaussian?
 - **Stability property**: $\sum_i g_i \cdot x_i$ is distributed as $\|x\| \cdot g$, where g is also Gaussian
 - Proof: $\langle g_1, \dots, g_d \rangle$ is centrally distributed, i.e., has random direction, and projection on random direction depends only on length of x
 - Hence, enough to consider $x = e_1$

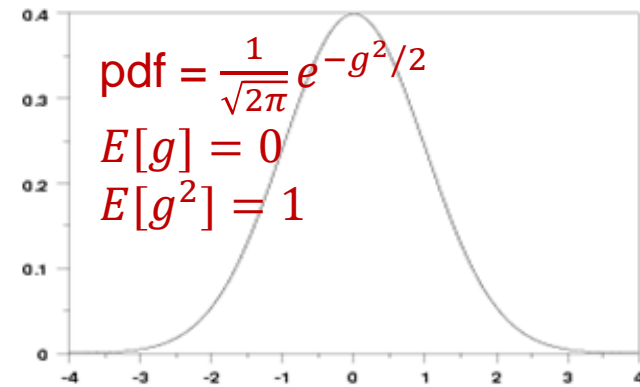
$$\begin{aligned} P(a) \cdot P(b) &= \\ &= \frac{1}{\sqrt{2\pi}} e^{-a^2/2} \frac{1}{\sqrt{2\pi}} e^{-b^2/2} \\ &= \frac{1}{2\pi} e^{-(a^2+b^2)/2} \end{aligned}$$



1D embedding

- Map $f(x) = \sum_i g_i \cdot x_i$,
 - for any x , $f(x) \sim \|x\| \cdot g$
 - Linear
- Want: $|f(x)| \approx \|x\|$
- **Claim:** for any $x \in \mathbb{R}^d$, we have
 - Expectation: $E[|f(x)|^2] = \|x\|^2$
 - Standard deviation:
 - $\sigma[|f(x)|^2] = O(\|x\|^2)$
- **Proof:**

- Expectation = $E \left[(f(x))^2 \right] = E[\|x\|^2 \cdot g^2]$
 $= \|x\|^2$



Full dimension reduction

- Just repeat the 1D embedding k times
 - $F(x) = (g_1 \cdot x, g_2 \cdot x, \dots, g_k \cdot x) / \sqrt{k} = \frac{1}{\sqrt{k}} Gx$
 - where G is a $k \times d$ random Gaussian matrix
- Again, want to prove that
 - $F(z) = (1 \pm \epsilon) \cdot \|z\|$
 - For fixed z
 - With probability $1 - e^{-\Omega(\epsilon^2 k)}$

Concentration

- $F(z)$ is distributed as
 - $\frac{1}{\sqrt{k}} (\|z\| \cdot a_1, \|z\| \cdot a_2, \dots, \|z\| \cdot a_k)$
 - where each a_i is distributed as Gaussian
- Norm $\|F(z)\|^2 = \|z\|^2 \cdot \frac{1}{k} \sum_i a_i^2$
 - $\sum_i a_i^2$ is called chi-squared distribution with k degrees
- **Fact:** chi-squared very well concentrated:
 - Equal to $1 + \epsilon$ with probability $1 - e^{-\Omega(\epsilon^2 k)}$
 - Akin to central limit theorem



Johnson Lindenstrauss: wrap-up

- $F(x) = (g_1 \cdot x, g_2 \cdot x, \dots, g_k \cdot x) / \sqrt{k} = \frac{1}{\sqrt{k}} Gx$
- $\|F(x)\| = (1 \pm \epsilon)\|x\|$ with high probability
- Contrast to Tug-Of-War:
 - $F(x) = \frac{1}{\sqrt{k}} Rx$ for R contained of ± 1
 - Only proved 90% probability
 - Would apply median to get high probability
 - Can also prove high probability [[Achlioptas'01](#)]
 - Gaussians have geometric interpretation

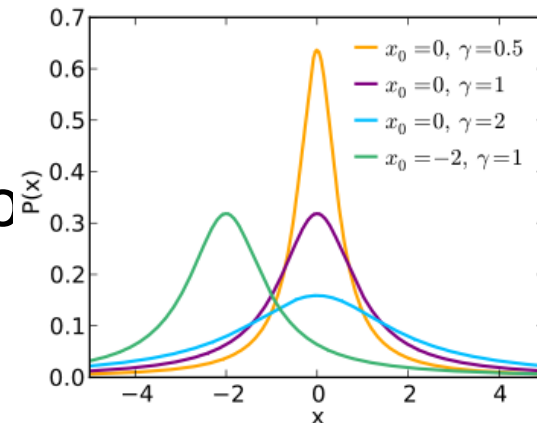
Dimension Reduction for ℓ_1

- Dimension reduction?
 - Essentially no [CS'02, BC'03, LN'04, JN'10...]
 - For n points, D approximation: between $n^{\Omega(1/D^2)}$ and $O(n/D)$ [BC03, NR10, ANN10...]
 - even if map depends on the dataset!
 - In contrast: [JL] gives $O(\epsilon^{-2} \log n)$, and doesn't depend on the dataset
 - No distributional dimension reduction either
 - But can sketch!

Sketch

- Can we do the “analog” of Euclidean projections?
- For ℓ_2 , we used: Gaussian distribution
 - has **stability property**:
 - $g_1z_1 + g_2z_2 + \dots + g_dz_d$ is distributed as $g \cdot \|z\|$
- Is there something similar for 1-norm?
 - **Yes**: Cauchy distribution!
 - 1-stable:
 - $c_1z_1 + c_2z_2 + \dots + c_dz_d$ is distributed as $c \cdot \|z\|_1$
- What’s wrong then?
 - Cauchy are **heavy-tailed**...
 - doesn’t even have finite expectation

$$pdf(s) = \frac{1}{\pi(s^2 + 1)}$$



Sketching for ℓ_1 [Indyk'00]

- Still, can consider map as before
 - $S(x) = (C_1x, C_2x, \dots, C_kx) = \mathbf{C}x$
- Consider $S(x) - S(y) = \mathbf{C}x - \mathbf{C}y = \mathbf{C}(x - y) = \mathbf{C}z$
 - where $z = x - y$
 - each coordinate distributed as $\|z\|_1 \times \text{Cauchy}$
 - Take 1-norm $\|\mathbf{C}z\|_1$?
 - does not have finite expectation, but...
- Can estimate $\|z\|_1$ by:
 - *Median* of absolute values of coordinates of $\mathbf{C}z$!
- **Correctness claim:** for each i
 - $\Pr[|C_i z| > \|z\|_1 \cdot (1 - \epsilon)] > 1/2 + \Omega(\epsilon)$
 - $\Pr[|C_i z| < \|z\|_1 \cdot (1 + \epsilon)] > 1/2 + \Omega(\epsilon)$

Estimator for ℓ_1

- Estimator: $\text{median}(|C_1 z|, |C_2 z|, \dots, |C_k z|)$
- **Correctness claim:** for each i
 - $\Pr[|C_i z| > \|z\|_1 \cdot (1 - \epsilon)] > 1/2 + \Omega(\epsilon)$
 - $\Pr[|C_i z| < \|z\|_1 \cdot (1 + \epsilon)] > 1/2 + \Omega(\epsilon)$
- **Proof:**
 - $|C_i z| = \text{abs}(C_i z)$ is distributed as $\text{abs}(\|z\|_1 c) = \|z\|_1 \cdot |c|$
 - Need to verify that
 - $\Pr[|c| > (1 - \epsilon)] > 1/2 + \Omega(\epsilon)$
 - $\Pr[|c| < (1 + \epsilon)] > 1/2 + \Omega(\epsilon)$



Estimator for ℓ_1

- Estimator: $\text{median}(|C_1 z|, |C_2 z|, \dots, |C_k z|)$

- **Correctness claim:** for each i

$L_i = 1$

if holds

$$\Pr[|C_i z| > \|z\|_1 \cdot (1 - \epsilon)] > 1/2 + \Omega(\epsilon)$$

$U_i = 1$

if holds

$$\Pr[|C_i z| < \|z\|_1 \cdot (1 + \epsilon)] > 1/2 + \Omega(\epsilon)$$

- Take $k = O(1/\epsilon^2)$
 - $E[L_i] \geq 1/2 + \Omega(\epsilon)$
 - Hence $\Pr\left[\sum_i L_i \leq \frac{k}{2}\right] < 0.05$ (by Chebyshev)
 - Similarly with U_i
- The above means that
 - $\text{median}(|C_1 z|, |C_2 z|, \dots, |C_k z|) \in (1 \pm \epsilon)\|z\|_1$
with probability at least 0.90

PS1

- Avg: 65.4
- Standard deviation: 20.5
- Max: 96

- By problems (average % points):
 - 1: 0.83
 - 2: 0.62
 - 3: 0.44