Today

1) Metric embeddings
2) Random Projections
3) Johnson-Lindenstrauss Lemma
4) Applications of (3)

Recall

The Metric Framework

i) Find a metric in your problem
ii) Find structure in your metric
iii) Use metric structure to solve problem

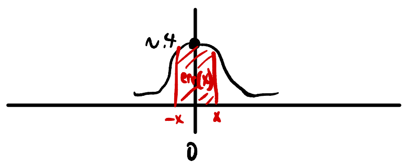
The Concentration Framework

a) Show (*) true if all RVs near $\mathbb{E}$
b) Concentration: each RV at $\mathbb{E}$ ($\pm \log n$) whP
c) Union bound: all RVs "

Fact: $\operatorname{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

Gaussians RV

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$



$$Z \sim N(\underset{\mathbb{E}}{0}, \underset{\text{var}}{1}) \rightarrow \text{"Standard Gaussian"}$$

Amazing Fact 1: $\int_{-\infty}^{\infty} \varphi(x) = 1$ → More generally, $\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}$

Amazing Fact 2: Rotational Symmetry: $(N(0,1), N(0,1), \ldots)$ in a uniformly random direction

Dfn. $N(\mu, \sigma^2) = \mu + \sigma Z$ where $Z \sim N(0,1)$ is a "non-Standard Gaussian"

Amazing Fact 3: Sum of Gaussians is a Gaussian

If $X \sim N(0,1)$, $Y \sim N(0,1)$, X, Y independent, $a, b \in \mathbb{R}$

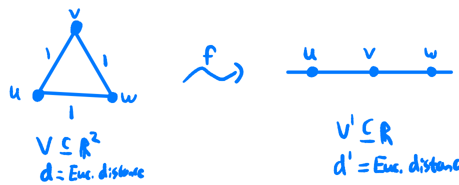
then $aX + bY \sim N(0, a^2 + b^2)$

Metric Embeddings: how can we approximate a complex metric w/ a simple one ((ii) today)

An **embedding** of metric space (V,d) into metric space (V',d') is a function $f: V \to V'$

(V,d) and (V',d') are **isometric** iff $\exists$ embedding $f: V \to V'$ s.t. $d(u,v) = d'(f(u), f(v))$ $\forall u,v \in V$
and vice versa

Isometry not always possible, e.g.



$d(u,v), d(v,w) \geq 1$
so $d(u,w) \geq 2 \neq 1$

f has **distortion** α if $d(u,v) \leq d'(f(u), f(v)) \leq \alpha \cdot d(u,v)$ $\forall u,v \in V$

E.g. $\alpha = 2$ above

Ideally: α small + (V',d') simpler/more structured than (V,d)

Today: linear embedding of (V, d) into (V', d) for $V \subseteq \mathbb{R}^n$, $V' \subseteq \mathbb{R}^k$, $k << n$ ($\approx \log n$)

$|V| = m$ Euc. distance

w/ distortion $\alpha = \sqrt{\frac{1+\varepsilon}{1-\varepsilon}}$ for small $\varepsilon > 0$

Reduction: for goal, suffices to give linear f s.t. $\|f(w)\|^2 \in [1-\varepsilon, 1+\varepsilon]$ for $\leq m^2$ unit w

g as in goal

$\Updownarrow$

$\|u-v\| \leq \|g(u) - g(v)\| \leq \alpha \|u-v\|$ $\forall u,v \in V$

$\Updownarrow$

$\|u-v\| \leq \|g(u-v)\| \leq \alpha \cdot \|u-v\|$

$\Updownarrow$

$\|z\| \leq \|g(z)\| \leq \alpha \cdot \|z\|$ $\forall z \in V-V$

$\Updownarrow$

$1 \leq \|g(\frac{z}{\|z\|})\| \leq \alpha$ $\forall z \in V-V$

$\Updownarrow$

$1 \leq \|g(w)\|^2 \leq \alpha^2$ $\forall w = \frac{z}{\|z\|}$ $\forall z \in V-V$

Let $f := \sqrt{(1-\varepsilon)} \cdot g$ and $\alpha = \sqrt{\frac{1+\varepsilon}{1-\varepsilon}}$

goal

$\Updownarrow$

$1-\varepsilon \leq (1-\varepsilon)\|g(w)\|^2 \leq 1+\varepsilon$ $\forall w = \frac{z}{\|z\|}$ $\forall z \in V-V$

$\Updownarrow$

$1-\varepsilon \leq \|\sqrt{1-\varepsilon}\, g(w)\|^2 \leq 1+\varepsilon$ $\forall w = \frac{z}{\|z\|}$ $\forall z \in V-V$

$\Updownarrow$

$\|f(w)\|^2 \in [1-\varepsilon, 1+\varepsilon]$ $\forall w = \frac{z}{\|z\|}$ $\forall z \in V-V$

χ^2 Distributions + Concentration

X is a **Chi-squared RV** w/ k degrees of freedom if

$$X = \sum_{i=1}^{k} z_i^2$$

where each $z_i \sim N(0,1)$ independently

Notated $X \sim \chi_k^2$

Claim: $\mathbb{E}[X] = k$ for $X \sim \chi_k^2$

Have $\mathbb{E}[z_i^2] = Var(z_i) - \mathbb{E}[z_i]^2 = 1 - 0$

Claim follows by Lo$\mathbb{E}$

Can't apply Chernoff for concentration b/c not $\in \{0,1\}$ (or even bounded)

Nonetheless, similar proof works

Can get $-O(\varepsilon)$ in exponent instead of $-O(\varepsilon^2)$ $\forall \varepsilon$

Skip, come back if time

Claim: $\Pr(|X - \mathbb{E}[X]| \geq \varepsilon \cdot k) \leq 2 \cdot \exp(-\varepsilon^2 \cdot k/8)$ for $X \sim \chi_k^2$ $(\varepsilon \in (0,1))$

Will prove upper tail; lower tail symmetric; Claim follows by union bound

Let $t = \varepsilon/4$ so $t \in (0, \frac{1}{4})$ and $3t + \varepsilon t = \frac{3}{4}\varepsilon + \frac{1}{4}\varepsilon^2 \geq \varepsilon^2/8 \ \forall \varepsilon \in (0,1)$

$$\Pr(X - \mathbb{E}[X] \geq \varepsilon \cdot k) = \Pr(X \geq (1+\varepsilon) \cdot k) \underset{e^{tx}\text{ non-decreasing}}{\leq} \Pr\left(e^{t \cdot X} \geq e^{(1+\varepsilon) \cdot tk}\right) \underset{\text{Markov}}{\leq} \frac{\mathbb{E}[e^{t \cdot X}]}{e^{(1+\varepsilon) \cdot tk}} \underset{z_i \text{ ind.}}{=} \frac{\prod_i \mathbb{E}[e^{t \cdot z_i^2}]}{e^{(1+\varepsilon) \cdot tk}}$$

Now calculate
$$\mathbb{E}[e^{t \cdot z_i^2}] = \frac{1}{\sqrt{2\pi}} \int_a e^{ta^2} \cdot e^{-a^2/2} \, da = \frac{1}{\sqrt{2\pi}} \int_a e^{-a(\frac{1}{2} - t)} \underset{\text{Gaussian integral}}{=} \frac{1}{\sqrt{2(\frac{1}{2} - t)}} = \frac{1}{\sqrt{1-2t}}$$

Plugging $\mathbb{E}[e^{t \cdot z_i^2}]$ in, get $\Pr(X \geq (1+\varepsilon) \cdot k) \leq \left(\frac{1}{e^{(1+\varepsilon) \cdot t} \sqrt{1-2t}}\right)^k$

But
$$\frac{1}{e^{(1+\varepsilon)t} \sqrt{1-2t}} = \exp\left(-t - \tfrac{1}{2}\ln(1-2t) - \varepsilon t\right) \leq \exp(-3t - \varepsilon t) \leq \exp(-\varepsilon^2/8)$$

$e^{-x} \leq 1 - .5x$ for $x \in (0,1)$ so $-\ln(1-2t) \leq -4t$

Above choice of t

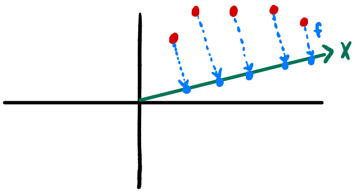
Combining above gives result

Random Projection

Simplest case: $k=1$

Let $X=(X_1, X_2, \dots, X_m)$ where $X_i \sim N(0,1)$ $\forall i$ be a vector in a uniformly random direction

Let $f:\mathbb{R}^m \to \mathbb{R}$ be $f(w)=\langle w, X\rangle$



Pick random direction X

w mapped to how far along X $\mathrm{Proj}(w \to x)$ is

Claim: $\|f(w)\|^2 \sim \chi_1^2$ $\forall w$ s.t. $\|w\|=1$

$\|w\|=1 \to \|w\|^2=1 \to \sum_i w_i^2 = 1$

So $\|f(w)\| = \underbrace{\sum_i w_i X_i}$

$\sim N(0, w_1^2+w_2^2+\dots+w_m^2)$ by $\sum_i$ Gaussians = Gaussian

$= N(0,1)$ b/c $\|w\|=1$

Corollaries:

1) $\mathbb{E}\left[\|f(w)\|^2\right]=1$ $\forall w$ s.t. $\|w\|=1$ → "Random projection preserves unit vector length in $\mathbb{E}$"

2) $\Pr\left(\left|\|f(w)\|^2-1\right| \ge \varepsilon\right) \le 2 \cdot \exp(-\varepsilon^2/8)$ → Vector length even concentrates under random projection

Problem: Upper tail useful but lower tail not → $\pm \log n$ from $\mathbb{E}$ = bad distortion

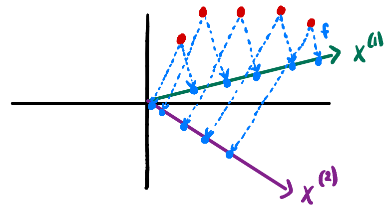
Solution: bring $\mathbb{E}$ up to $\approx \log n$ by repeated trials

Johnson–Lindenstrauss Lemma

Let $X^{(i)} \sim (X_1^{(i)}, X_2^{(i)}, \dots, X_m^{(i)})$ where each $X_j^{(i)} \sim N(0,1)$

Let $f_i(w) := \langle X^{(i)}, w \rangle$

Let $F(w) := (f_1(w), f_2(w), \dots, f_k(w))$ → Equivalent to Aw where $A_{ij} \sim N(0,1)$
So F linear



Claim: $\|F(w)\|^2 \sim \chi_k^2$ $\forall w$ s.t. $\|w\| = 1$

$\|F(w)\|^2 = \sum_i (f_i(w))^2 = \sum_i \|f_i(w)\|^2$ and $\|f_i(w)\|^2 \sim \chi_1^2$ by previous Claim

Corollaries:

1) $\mathbb{E}[\|F(w)\|^2] = k$ $\forall w$ s.t. $\|w\| = 1$ → So, $k \approx \log n$ → good Concentration

2) $\Pr(|\|F(w)\|^2 - k| \geq \epsilon \cdot k) \leq 2 \cdot \exp(-\epsilon^2 \cdot k / 8)$ $\forall w$ s.t. $\|w\| = 1$

Let $\tilde{F}(w) := \frac{1}{\sqrt{k}} \cdot F(w)$
→ Scale down so unit vector mapped to unit vector
→ Trivially Poly-time computable (even w/o knowing points)

JL Lemma: For any m^2 unit vectors $\bar{W}$ and $k = 24 \cdot \frac{\ln m}{\epsilon^2}$ $\tilde{F}: \mathbb{R}^M \to \mathbb{R}^k$ satisfies

$$\|\tilde{F}(w)\|^2 \in [1-\epsilon, 1+\epsilon] \quad \forall w \in \bar{W}$$

except w/ $\Pr \leq 1 - \frac{2}{n}$

Note $\mathbb{E}[\|\tilde{F}(w)\|^2] = \frac{1}{k} \mathbb{E}[\|F(w)\|^2] = 1$ (a)

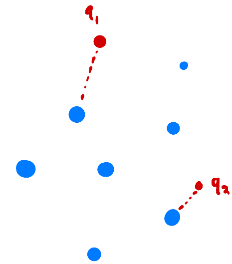
$\|\tilde{F}(w)\|^2 \notin [1-\epsilon, 1+\epsilon]$ only if $\|F(w)\|^2 \notin [(1-\epsilon)\cdot k, (1+\epsilon)k]$ only if $|\|F(w)\|^2 - k| \geq \epsilon \cdot k$

But $\Pr(|\|F(w)\|^2 - k| \geq \epsilon k) \leq 2 \cdot \exp(-3 \ln n) = \frac{2}{n^3}$ (b)

By union bound, $\tilde{F}(w) \in [1-\epsilon, 1+\epsilon]$ $\forall n^2$ w except w/ $\Pr \leq \frac{2}{n}$ (c)

Application of JL

$\sqrt{\frac{1+\varepsilon}{1-\varepsilon}}$ - NN - Search: initially given $X \subseteq \mathbb{R}^n$, w/ $|X| = m$ (i)



repeatedly given queries $q_1, q_2, \ldots$

$\forall\, q_i$, return $x \in X$ s.t. $d(q_i, x) \leq \sqrt{\frac{1+\varepsilon}{1-\varepsilon}} \cdot d(q_i, X)$
as fast as possible

Naive solution: $O(n \cdot m)$ time per query

JL solution: $O\left(\frac{\log n}{\varepsilon^2} \cdot m\right)$ per query

let $\hat{F}: \mathbb{R}^n \to \mathbb{R}^k$ be linear embedding w/ distortion $\sqrt{\frac{1+\varepsilon}{1-\varepsilon}}$ from JL

for $k = O\left(\frac{\log n}{\varepsilon^2}\right)$ so

Let $Y := \{\hat{F}(x) : x \in X\}$ (ii)

For query q_i,

return $x \in X$ s.t. $\hat{F}(x) = y$ where $y = \arg\min_{y \in Y} d(y, q_i)$ (iii)

Easy to verify result is (approximately) correct

Takes $O\left(\frac{\log n}{\varepsilon^2} \cdot m\right)$ per query

↳ Can even reduce to $\tilde{O}\left(\frac{1}{\varepsilon^2}\right)$ w/ a little more work!