Algorithms for Big Data (Fall 25)

Lecture 11

Final Notes on JL and Subspace Embedding

Ali Vakilian (vakilian@vt.edu)

Virginia Tech

Algorithms for Big Data Fall 2025

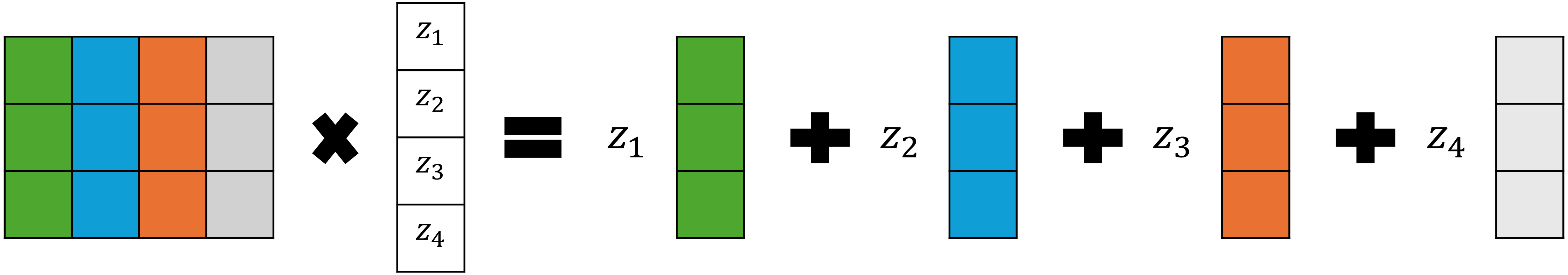
Dimensionality Reduction: Motivations

- **A main motivation**
  - reduce the dimensionality of the input with the hope to solve the problem faster!
  ($X \subset \mathbb{R}^d$; map X down in $\mathbb{R}^m$ for $m \ll d$ using a map f)
  - but, how fast can we compute the map f?

1. For the original construction of JL [JL84] requires $O(md)$ time.
2. Achlioptas [Achlioptas03] gave a sparser matrix $\Pi \in \mathbb{R}^{m \times d}$ with similar guarantees (entries are independently chosen at random; equal to $\frac{1}{\sqrt{s}}$ w.p. $\frac{1}{3}$, $\frac{-1}{\sqrt{s}}$ w.p. $\frac{1}{3}$; 0 otherwise)
   - $s = m/3$
   - $\Pi z = \sum_i z_i \Pi^i$ (where Π^i is the i-th column in Π)

Dimensionality Reduction: Motivations

1. For the original construction of JL [JL84] requires $O(md)$ time.
2. Achlioptas [Achlioptas03] gave a sparser matrix $\Pi \in \mathbb{R}^{m \times d}$ with similar guarantees (entries are independently chosen at random; equal to $\frac{1}{\sqrt{s}}$ w.p. $\frac{1}{3}$, $\frac{-1}{\sqrt{s}}$ w.p. $\frac{1}{3}$; 0 otherwise)
   - $s = m/3$
   - $\Pi z = \sum_i z_i \Pi^i$ (where Π^i is the i-th column in Π)



Dimensionality Reduction: Motivations

1. For the original construction of JL [JL84] requires $O(md)$ time.
2. Achlioptas [Achlioptas03] gave a sparser matrix $\Pi \in \mathbb{R}^{m \times d}$ with similar guarantees
3. Fast JL Transform: main idea is to pick a **sampling matrix** $S \in \mathbb{R}^{m \times d}$
   - S has a single 1 in a random location per row (zero elsewhere in the row); Rows are chosen at random
   - Computing $z \mapsto \Pi z$ is fast (takes $O(m)$ time)
   - $\mathbb{E}\left[\left\|\frac{1}{\sqrt{m}}\Pi z\right\|_2^2 = \|z\|_2^2\right]$; however, the variance might be quite high
     - if z has its mass concentrated in one or few coordinates
   - Apply a pre-conditioning operation R (for a certain orthogonal matrix R) s.t. $\frac{\|Rz\|_\infty}{\|Rz\|_2}$ is small w.h.p.
     - Rz is "well-spread", with no coordinate having too much mass
   - $\frac{1}{\sqrt{m}}SRz$ has roughly the same norm as z; the runtime to compute $\frac{1}{\sqrt{m}}SRz$ is $O(d \log d + m^3)$

Dimensionality Reduction: Motivations

1. For the original construction of JL [JL84] requires $O(md)$ time.
2. Achlioptas [Achlioptas03] gave a sparser matrix $\Pi \in \mathbb{R}^{m \times d}$ with similar guarantees
3. Fast JL Transform: main idea is to pick a **sampling matrix** $S \in \mathbb{R}^{m \times d}$
   - While fast, it does not utilize the sparsity of z (when they're).
4. Sparse JL Transform: If Π has s non-zero per column, Πx can be multiplied in $O(s \cdot \|z\|_0)$ time
   - Then, the name of the game is to make m and s as small as possible.
   - CountSketch provides DJL with $m = O(1/(\varepsilon^2 \delta))$ and $s = 1$.
   - [KN'14]: similar to CountSkecth with $s > 1$. Improves to $m = O(\log(1/\delta)/\varepsilon^2)$ and $s = \varepsilon m$.

Sparse JL Transform

$\Pi \in \mathbb{R}^{m \times d}$ s.t. $\Pi_{r,i} = (\eta_{r,i} \cdot \sigma_{r,i}) / \sqrt{s}$, where $\sigma_{r,i}$ are independent Rademacher and $\eta_{r,i}$ are Bernoulli random variable satisfying:

- For all r, i, $\mathbb{E}[\eta_{r,i}] = s/m$
- For any i, $\sum_{1 \le r \le m} \eta_{r,i} = s$: I.e., each column of Π has exactly s non-zero entries.
- $\eta_{r,i}$ are negatively correlated: for any $S \subset [m] \times [d]$, $\mathbb{E}\left[\Pi_{(r,i) \in S} \eta_{r,i}\right] \le \Pi_{(r,i) \in S} \mathbb{E}[\eta_{r,i}] \le \left(\frac{s}{m}\right)^{|S|}$

Theorem. If $m = O(\log(1/\delta)/\varepsilon^2)$ and $s = \varepsilon m$, then for all z of unit norm, $\Pr_{\Pi}\left(\left| \|\Pi z\| - 1 \right| > \varepsilon\right) \le \delta$

What more? Fast JL by Ailon and Chazelle [AC'09] where Πx can be computed in $O(d \log d)$ time

- $\Pi = \frac{1}{\sqrt{m}} SHD$:
  - $S_{m \times d}$ is a sampling matrix
  - H is Hadamard matrix, and
  - D is a diagonal matrix with independent Rademacher

Dimensionality Reduction

JL Lemma and Subspace Embedding

Distributional Johnson-Lindenstrauss Lemma

Distributional JL Lemma. Fix $x \in \mathbb{R}^d$, and let $\Pi \in \mathbb{R}^{k \times d}$ be a matrix whose entries are chosen independently according to standard normal distribution $\mathcal{N}(0, 1)$. If $k = \Omega(\varepsilon^{-2} \log(1/\delta))$, then with probability at least $1 - \delta$,

$$\left\| \frac{1}{\sqrt{k}} \Pi x \right\|_2 = (1 \pm \varepsilon) \|x\|_2$$

Sum of Independent Normal Distribution

Lemma. Let X and Y be independent random variables.
Suppose $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. Let $Z = X + Y$. Then,

$$\boldsymbol{Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)}$$

Corollary. Let X and Y be independent random variables. Suppose $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$. Let $Z = aX + bY$ where a, b are arbitrary real numbers. Then, $\boldsymbol{Z \sim \mathcal{N}(0, a^2 + b^2)}$

Normal distribution is a *stable distribution:* adding two indep. r.v. within the same class gives a distribution inside the class. Other exist and useful in $\boldsymbol{F_p}$ estimation for $\boldsymbol{p \in (0, 2)}$.

Random Gaussian Vector

One can consider higher dimensional normal distributions, also called multivariate Gaussian (or Normal) distributions.

Random Gaussian vector: $Z = (Z_1, \dots, Z_k)$ if $Z_i \sim \mathcal{N}(0,1)$ for each i, and $Z_1, \dots, Z_k$ are independent.

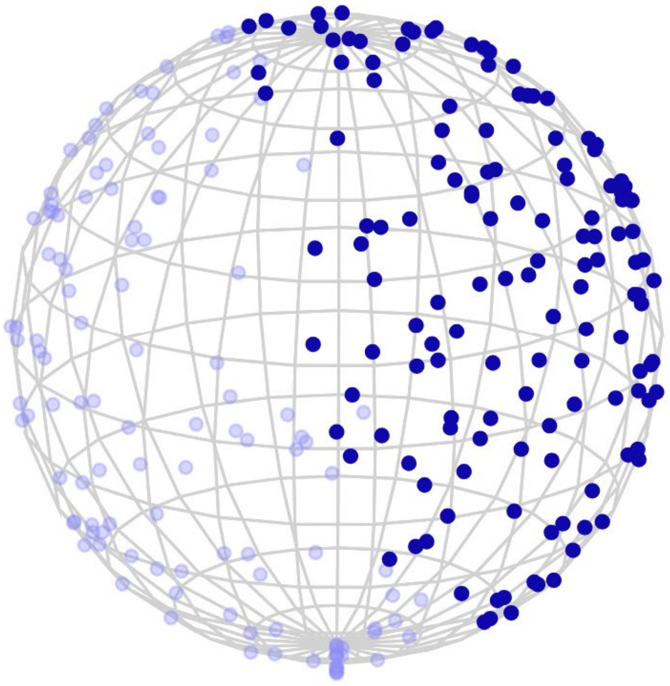
- Density function is $f(y_1, \dots, y_k) = \left(\frac{1}{\sqrt{2\pi}}\right)^k \exp\left(-\frac{y_1^2 + \dots + y_k^2}{2}\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^k e^{-\|y\|_2/2}$
- Only depends on $\|y\|_2$
- The distribution is **centrally symmetric**. (can be used to generate a random unit vector in $\mathbb{R}^k$). $U = \frac{Z}{\|Z\|}$ is uniform on the unit sphere.
- $\mathbb{E}\left[\|Z\|_2^2\right] = \sum_i \mathbb{E}\left[Z_i^2\right] = k$. Length is concentrated around k.

Random Gaussian Vector

One can consider higher dimensional normal distributions, also called multivariate Gaussian (or Normal) distribution

Random Gaussian vector: $Z = (Z_1, \dots, Z_k)$ if Z
and $Z_1, \dots, Z_k$ are independent.

- Density function is $f(y_1, \dots, y_k) = \left(\frac{1}{\sqrt{2\pi}}\right)^k \exp\left(-\frac{y_1^2}{\right.$
- Only depends on $\|y\|_2$
- The distribution is **centrally symmetric**. (can be use vector in $\mathbb{R}^k$). $U = \frac{Z}{\|Z\|}$ is uniform on the unit sphere.
- $\mathbb{E}\left[\|Z\|_2^2\right] = \sum_i \mathbb{E}\left[Z_i^2\right] = k$. Length is concentrated ar



Concentration of sum of squares of normally distributed variables

$\chi^2(k)$ distribution: distribution of sum of squares of k independent standard normally distributed random variables,

$$Y = \sum_{1 \le i \le k} Z_i^2 \text{ where each } Z_i \sim \mathcal{N}(0,1)$$

Lemma. Let $Z_1, \ldots, Z_k$ be independent $\mathcal{N}(0,1)$ r.v.s. and let $Y = \sum_i Z_i^2$. Then, for $\varepsilon \in (0,1/2)$, there is a constant c such that,

$$\Pr\left[(1-\varepsilon)^2 k \le Y \le (1+\varepsilon)^2 k\right] \ge 1 - 2e^{-c\varepsilon^2 k}$$

- Recall Chernoff for bounded independent non-negative rv. Z_i^2 are not bounded, however, Chernoff bounds extend to sums of random variables with exponentially decaying tails.

Distributional Johnson-Lindenstrauss Lemma

Distributional JL Lemma. Fix $x \in \mathbb{R}^d$ and let $\Pi \in \mathbb{R}^{k \times d}$ be a matrix whose entries are chosen independently according to standard normal distribution $\mathcal{N}(0,1)$. If $k = \Omega(\varepsilon^{-2} \log(1/\delta))$, then with probability at least $1 - \delta$,

$$\left\| \frac{1}{\sqrt{k}} \Pi x \right\|_2 = (1 \pm \varepsilon) \|x\|_2$$

Can we guarantee this property for all $x \in \mathbb{R}^d$?

Not possible. **Why? No!** Since Π maps an n-dimension to a d-dimension space, some non-zero vectors must be mapped to zero under Π.

Subspace Embedding

Question. Suppose $E \subset \mathbb{R}^n$ is a linear subspace of dimension d. Can we find a projection $\Pi: \mathbb{R}^d \to \mathbb{R}^k$ such that for *every* $x \in E$, $\left\| \frac{1}{\sqrt{k}} \boldsymbol{\Pi x} \right\|_2 = (\mathbf{1} \pm \boldsymbol{\varepsilon}) \|\boldsymbol{x}\|_2$?

Not possible if $k < d$.

Possible if $k = d$. Why? Pick Π to be an orthonormal basis for E.

- This requires knowing E and computing orthonormal basis which is slow.

Goal. Find an oblivious subspace embedding; JL based on random projections

You can think of E as column space of $n \times d$ matrix A

Then, one has to show $\|\boldsymbol{SAx}\|_2 = (\mathbf{1} \pm \boldsymbol{\varepsilon}) \|\boldsymbol{Ax}\|_2$ for all $x \in \mathbb{R}^d$

Oblivious Subspace Embedding

Theorem. Suppose $E \subset \mathbb{R}^n$ is a linear subspace of dimension d. Let $\Pi \in \mathbb{R}^{k\times n}$ with $k = O\left(\frac{d}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right)$ rows. Then with probability $(1-\delta)$, for every $x \in E$,

$$\left\|\frac{\mathbf{1}}{\sqrt{\boldsymbol{k}}}\boldsymbol{\Pi x}\right\|_{\mathbf{2}} = (\mathbf{1} \pm \boldsymbol{\varepsilon})\|\boldsymbol{x}\|_{\mathbf{2}}$$

In other words, JL Lemma extends from one dimension to arbitrary number of dimensions in a smoothly.

Proof Challenges

How do we prove that Π works for all $x \in E$ which is an **infinite set**?

In particular, union bound doesn't work as is.

Useful Idea. Net Argument

- Choose a large but finite set of vectors T carefully (the net)
- Prove that Π preserves length of vectors in T (via union bound)
- Argue that any vector $x \in E$ is sufficiently close to a vector in T; hence, Π also preserves the length of x

Net Argument

Observation. It is sufficient to focus on unit vectors in E. **Why?**

Theorem. Suppose $E \subset \mathbb{R}^n$ is a linear subspace of dimension d. Let $\Pi \in \mathbb{R}^{k \times n}$ with $k = O\left(\frac{d}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)\right)$ rows. Then with probability $(1-\delta)$, for every $x \in E$,

$$\left\| \frac{\mathbf{1}}{\sqrt{\boldsymbol{k}}} \boldsymbol{\Pi x} \right\|_{\mathbf{2}} = (\mathbf{1} \pm \boldsymbol{\varepsilon}) \|\boldsymbol{x}\|_{\mathbf{2}}$$

Net Argument

Observation. It is sufficient to focus on unit vectors in E. **Why?**

Without loss of generality, lets assume that $\boldsymbol{E}$ is the subspace formed by the first d coordinate in the standard basis.

Claim 1. There is a net T of size $e^{O(d)}$ such that preserving lengths of vectors in T suffices.

Use DJL with $k = O(\frac{d}{\varepsilon^2} \log(1/\delta))$ and union bound to show that all vectors in T are preserved in length up to $(1 \pm \varepsilon)$-factor.

Net Argument

Observation. It is sufficient to focus on unit vectors in E. **Why?**

Without loss of generality, lets assume that $\boldsymbol{E}$ is the subspace formed by the first d coordinate in the standard basis.

Claim 1. There is a net of size T of size $e^{O(d)}$ such that preserving lengths of vectors in T suffices.

Definition (ε-net). A subset T is an ε-net for a space S if for every point $\boldsymbol{p} \in S$, there is a point $\boldsymbol{x}$ in the net T such that

- In ℓ_2 space: $\|\boldsymbol{x} - \boldsymbol{p}\|_2 \leq \varepsilon$, or
- In ℓ_∞ space: $\|\boldsymbol{x} - \boldsymbol{p}\|_\infty \leq \varepsilon$, or

Net Argument

Observation. It is sufficient to focus on unit vectors in E. **Why?**

Without loss of generality, lets assume that $\boldsymbol{E}$ is the subspace formed by the first d coordinate in the standard basis.

Claim 1. There is a net of size T of size $e^{O(d)}$ such that preserving lengths of vectors in T suffices.

A weaker ε-net construction.

- For $[-1,1]^d$, make a grid of length (ε/d)
- Number of grid points is $(2d/\varepsilon)^d$
- Better net constructions exist too.

Proof via Net Argument Analysis

Theorem. Let $E \subset \mathbb{R}^n$ be a linear subspace of dimension d. Let $\Pi \in \mathbb{R}^{k \times n}$ with $k = O\left(\frac{d}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)\right)$ rows. Then with probability $(1-\delta)$, for every $x \in E$,

$$\left\| \frac{1}{\sqrt{k}} \Pi x \right\|_2 = (1 \pm \varepsilon)\|x\|_2$$

Fix any $x \in E$ such that $\|x\|_2 = 1$

- $\exists$ a grid point $y \in T$ s.t. $\|y\|_2 \leq 1$ and $\|x - y\|_\infty \leq \frac{\varepsilon}{d}$. Let $z = x - y$

$$\begin{aligned} \|\Pi x\|_2 = \|\Pi(y + (x - y))\|_2 &\leq \|\Pi x\|_2 + \|\Pi z\|_2 \\ &\leq (1+\varepsilon) + (1+\varepsilon) \textstyle\sum_{i \in [d]} |z_i| \\ &\leq (1+\varepsilon) + (1+\varepsilon)\varepsilon \leq 1 + 3\varepsilon. \end{aligned}$$

Similarly, $\|\Pi x\|_2 \geq 1 - O(\varepsilon)$

Proof of Subspace Embedding

Subspace Embedding

A $(1 \pm \varepsilon)$ ℓ_2-subspace embedding for column space of an $n \times d$ matrix A is a matrix S for which for all $x \in \mathbb{R}^d$

$$||SAx||_2^2 = (1 \pm \varepsilon)||Ax||_2^2$$

S is also an ℓ_2-subspace embedding for U, where U is an orthonormal basis for column space of A. So,

$$||SUx||_2^2 = (1 \pm \varepsilon)||Ux||_2^2$$

Subspace Embedding

A $(1 \pm \varepsilon)$ ℓ_2-subspace embedding for column space of an $n \times d$ matrix A is a matrix S for which for all $x \in \mathbb{R}^d$

$$\|SUx\|_2^2 = (1 \pm \varepsilon)\|Ux\|_2^2$$

- If it holds for all unit vectors y, then it is satisfied for all vectors x by scaling.
- Consider an ε-net N over the sphere $\mathcal{S}^{d-1}$.
  - **Definition:** For all $x \in \mathcal{S}^{d-1}$, there exists y such that $\|x - y\|_2 \leq \varepsilon$
  - **Greedy Construction:** While $\exists\ x \in \mathcal{S}^{d-1}$ of distance larger than ε from N; include x in N.
  - **Size Analysis:**
    - Consider a ball of radius $\varepsilon/2$ around every point in N. By construction they are disjoint.
    - All are contained in a ball of radius $(1 + \varepsilon/2)$ around the origin.
    - $\Rightarrow |N| \leq \frac{(1+\varepsilon/2)^d}{(\varepsilon/2)^d} = \left(1 + \frac{\varepsilon}{2}\right)^d$

Subspace Embedding

A $(1 \pm \varepsilon)$ ℓ_2-subspace embedding for column space of an $n \times d$ matrix A is a matrix S for which for all $x \in \mathbb{R}^d$

$$\|SUx\|_2^2 = (1 \pm \varepsilon)\|Ux\|_2^2$$

- If it holds for all unit vectors y, then it is satisfied for all vectors x by scaling.
- Consider an ε-net N over the sphere $\mathcal{S}^{d-1}$.
- Let $M = \{Ux \mid x \in N\}$

Claim. For every $x \in \mathcal{S}^{d-1}$, there is a $y \in M$ for which $\|Ux - y\|_2 \leq \varepsilon$.

Proof. Let $x' \in \mathcal{S}^{d-1}$ be s.t. $\|x' - x\|_2 \leq \varepsilon$. Then $\|Ux - Ux'\|_2 = \|x' - x\|_2 \leq \varepsilon$

Set $y = Ux'$.

Subspace Embedding (Net Argument)

Claim I. For every $x \in \mathcal{S}^{d-1}$, there is a $y' \in M$ for which $\|Ux - y'\|_2 \le \varepsilon$.

- Let $y = Ax$ for an arbitrary $x \in \mathcal{S}^{d-1}$
- By **Claim I**, there exists $y_1 \in M$ s.t. $\|y - y_1\|_2 \le \varepsilon$.
- Let α be s.t. $\|\alpha(y - y_1)\|_2 = 1$. In particular, $\alpha \le 1/\varepsilon$
- By **Claim I**, there exists $y_2' \in M$ s.t. $\|\alpha(y - y_1) - y_2'\|_2 \le \varepsilon$.
  - Then, $\|y - y_1' - (y_2')/\alpha\|_2 \le \varepsilon/\alpha \le \varepsilon^2$
  - Set $y_2 = y_2'/\alpha$
- Repeat the process to obtain $y_1, y_2, y_3, \cdots$ s.t. for all i,

$$\|y - y_1 - y_2 - \cdots - y_i\|_2 \le \varepsilon^i$$

- By triangle inequality, for all i, $\|y_i\|_2 \le \varepsilon^{i-1} + \varepsilon^i \le 2\varepsilon^{i-1}$

Subspace Embedding (Net Argument)

Claim I. For every $x \in \mathcal{S}^{d-1}$, there is a $y' \in I$

- Let $y = Ax$ for an arbitrary $x \in \mathcal{S}^{d-1}$
- There exist $y_1, y_2, y_3, \cdots$ s.t. $y = \sum_i y_i$, and

$$\|Sy\|_2^2 = \left\|S \sum_i y_i\right\|_2^2$$

$$\|Sy\|_2^2 = \sum_i \|Sy_i\|_2^2 + 2\sum_{i,j} \langle Sy_i, Sy_j \rangle$$

$$\|Sy\|_2^2 = \sum_i \|y_i\|_2^2 + 2\sum_{i,j} \langle y_i, y_j \rangle \pm O(\varepsilon) \sum_{i,j} \|y_i\|_2 \|y_j\|_2$$

For unit vectors $y, y' \in M$,

- $\|Sy\|_2^2 = (1 \pm \varepsilon)\|y\|_2^2$
- $\|Sy'\|_2^2 = (1 \pm \varepsilon)\|y'\|_2^2$
- $\|S(y - y')\|_2^2 = (1 \pm \varepsilon)\|y - y'\|_2^2$

$\|S(y - y')\|_2^2 = \|Sy\|_2^2 + \|Sy'\|_2^2 - 2\langle Sy, Sy' \rangle$

$\|y - y'\|_2^2 = \|y\|_2^2 + \|y'\|_2^2 - 2\langle y, y' \rangle$

$\Longrightarrow (1 \pm \varepsilon)\|y\|_2^2 + (1 \pm \varepsilon)\|y'\|_2^2 - 2\langle Sy, Sy' \rangle$

$\quad = (1 \pm \varepsilon)\|y\|_2^2 + (1 \pm \varepsilon)\|y'\|_2^2 - 2(1 \pm \varepsilon)\langle y, y' \rangle$

$\Longrightarrow \boldsymbol{\langle Sy, Sy' \rangle = \langle y, y' \rangle \pm O(\varepsilon)}$

$\Longrightarrow \boldsymbol{\langle \alpha Sy, \beta Sy' \rangle = \langle \alpha y, \beta y' \rangle \pm O(\varepsilon\alpha\beta)}$

Subspace Embedding (Net Argument)

Claim I. For every $x \in \mathcal{S}^{d-1}$, there is a $y' \in$

- Let $y = Ax$ for an arbitrary $x \in \mathcal{S}^{d-1}$
- There exist $y_1, y_2, y_3, \cdots$ s.t. $y = \sum_i y_i$, and

For unit vectors $y, y' \in M$,

- $\|Sy\|_2^2 = (1 \pm \varepsilon)\|y\|_2^2$
- $\|Sy'\|_2^2 = (1 \pm \varepsilon)\|y'\|_2^2$
- $\|S(y - y')\|_2^2 = (1 \pm \varepsilon)\|y - y'\|_2^2$

$\|S(y - y')\|_2^2 = \|Sy\|_2^2 + \|Sy'\|_2^2 - 2\langle Sy, Sy' \rangle$

$\|y - y'\|_2^2 = \|y\|_2^2 + \|y'\|_2^2 - 2\langle y, y' \rangle$

$\Longrightarrow (1 \pm \varepsilon)\|y\|_2^2 + (1 \pm \varepsilon)\|y'\|_2^2 - 2\langle Sy, Sy' \rangle$

$= (1 \pm \varepsilon)\|y\|_2^2 + (1 \pm \varepsilon)\|y'\|_2^2 - 2(1 \pm \varepsilon)\langle y, y' \rangle$

$\boldsymbol{\Longrightarrow \langle Sy, Sy' \rangle = \langle y, y' \rangle \pm O(\varepsilon)}$

$\boldsymbol{\Longrightarrow \langle \alpha Sy, \beta Sy' \rangle = \langle \alpha y, \beta y' \rangle \pm O(\varepsilon \alpha \beta)}$

$$\|Sy\|_2^2 = \left\|S \textstyle\sum_i y_i\right\|_2^2$$

$$\|Sy\|_2^2 = \textstyle\sum_i \|Sy_i\|_2^2 + 2\sum_{i,j} \langle Sy_i, Sy_j \rangle$$

$$\|Sy\|_2^2 = \textstyle\sum_i \|y_i\|_2^2 + 2\sum_{i,j} \langle y_i, y_j \rangle \pm O(\varepsilon) \sum_{i,j} \|y_i\|_2 \|y_j\|_2$$

$$\|Sy\|_2^2 = \left\|\textstyle\sum_i y_i\right\|_2^2 \pm O(\varepsilon)\left(\textstyle\sum_i 2\varepsilon^{i-1}\left(\sum_{j>i} 2\varepsilon^{j-1}\right)\right)$$

$$\|Sy\|_2^2 = \|y\|_2^2 \pm O(\varepsilon)\left(\textstyle\sum_i 4\varepsilon^{2i-1}/(1-\varepsilon)\right) = 1 \pm O(\varepsilon)$$

Applications of Subspace Embedding

Regression

Applications of Subspace Embedding

Faster algorithms for approximate

- matrix multiplication
- regression
- SVD

Basic idea. Want to perform operations on matrix A with n data columns (in a large dimension $\mathbb{R}^h$) with small actual rank d.

Our goal is to reduce to a matrix of size roughly $\mathbb{R}^{d\times d}$ by spending time proportional to the number of non-zero entries in A.

Regression: Linear Model Fitting

A classic problem in **data analysis**

- n data points in $a_1, \cdots, a_n \in \mathbb{R}^d$
- Each data point a_i is associated with a value $b_i \in \mathbb{R}$

What model should one use to explain the data?

Simplest model? Linear fitting:

- $b_i = w_0 + \sum_{1 \le j \le d} w_j \cdot a_{i,j}$ for a vector $w \coloneqq (w_0, \cdots, w_d)$
- However, usually data is noisy and won't be able to satisfy for all data points
- Without loss of generality, we can restrict to $w_0 = 0$ by lifting to $d + 1$ dimensions

Regression

Goal: want to choose $w_1, \cdots, w_d$ to estimate $b_i \sim \sum_{1 \le j \le d} w_j \cdot a_{i,j}$

Let A be matrix with one row per data point a_i. We write $x_1, \ldots, x_d$ as variables for finding $w_1, \ldots, w_d$.

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \ldots & a_{1,d} \\ a_{2,1} & a_{2,2} & a_{2,3} & \ldots & a_{2,d} \\ & & \vdots & & \\ a_{n,1} & a_{n,2} & a_{n,3} & \ldots & a_{n,d} \end{pmatrix}$$

Ideally: Find $x \in \mathbb{R}^d$ such that $Ax = b$

Best fit: Find $x \in \mathbb{R}^d$ to minimize $Ax - b$ under some norm

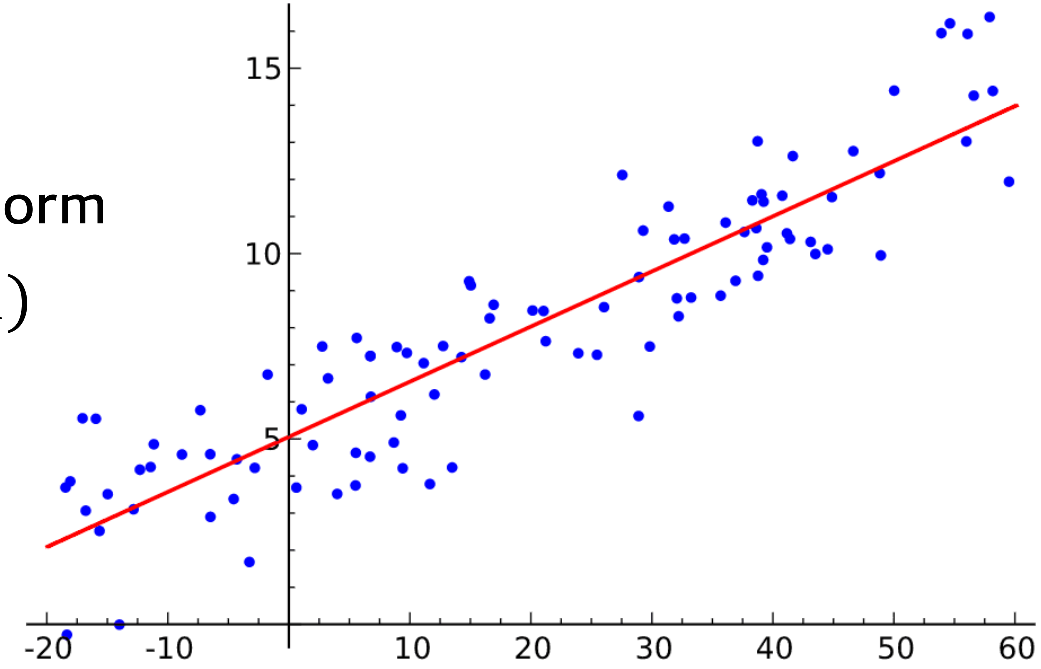
- $\|Ax - b\|_1, \|Ax - b\|_2, \|Ax - b\|_\infty$

Least Squares Error Regression

Given $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^d$, find x to minimize $||Ax - b||_2$

Interesting when $n \gg d$; there is no solution to $Ax = b$ and want to find the best fit

- Ax is a linear combination of columns in A
- $z \in \text{colspace}(A)$ that is closest to b in ℓ_2-norm
- So, z is the projection of b onto $\text{colspace}(A)$



How to find it?

Least Squares Regression

Given $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^d$, find x to minimize $||Ax - b||_2$

- Closest vector to b is the projection of b onto $\mathrm{colspace}(A)$
  - Find orthonormal basis $z_1, \ldots, z_r$ for the columns of A
  - Compute projection c of b to $\mathrm{colspace}(A)$ which is $c = \sum_{1 \le j \le r} \langle b, z_j \rangle \, z_j$
- Back to our question, what is x?
  - $Ax = c$. We need to solve the linear system.
  - By solving normal equation: $x^* = (A^\top A)^- b^\top A$ (Moore-Penrose Pseudoinverse)
  - Naively requires $O(nd^2)$ time to compute

Can we speed up the process with some potential approximation?

LSE Regression via Subspace Embedding

Let E denote the subspace spanned by columns of A and b. It has dimension at most $d + 1$.

Use Subspace Embedding S on E with $k = O(d/\varepsilon^2)$ rows to reduce $\{A^{(1)}, A^{(2)}, \cdots, A^{(d)}, b\}$ to $\{A'^{(1)}, A'^{(2)}, \cdots, A'^{(d)}, b'\}$ which are in $\mathbb{R}^k$.

Solve $\min_{x' \in \mathbb{R}^d} \|A'x' - b'\|_2$

Lemma. With probability $1 - \delta$,

$$(1 - \varepsilon) \min_{x \in \mathbb{R}^d} \|Ax - b\|_2 \leq \min_{x' \in \mathbb{R}^d} \|A'x' - b'\|_2 \leq (1 + \varepsilon) \min_{x \in \mathbb{R}^d} \|Ax - b\|_2$$

LSE Regression via Subspace Embedding

Lemma. With probability $1-\delta$,

$$(1-\varepsilon)\min_{x\in\mathbb{R}^d}\|Ax-b\|_2 \le \min_{x'\in\mathbb{R}^d}\|A'x'-b'\|_2 \le (1+\varepsilon)\min_{x\in\mathbb{R}^d}\|Ax-b\|_2$$

With probability $(1-\delta)$, via subspace embedding guarantee, for all $z\in E$,

$$(1-\varepsilon)\|z\|_2 \le \|Sz\|_2 \le (1+\varepsilon)\|z\|_2$$

- Let x^*, y^* be respectively the optimal solution to $\min_{x\in\mathbb{R}^d}\|Ax-b\|_2$ and $\min_{x'\in\mathbb{R}^d}\|A'x'-b'\|_2$
- Let $z = Ax^* - b$. Since $z\in E$, $\|Sz\|_2 \le (1+\varepsilon)\|z\|_2$.
- Since x^* is a feasible solution to $\min_{x'\in\mathbb{R}^d}\|A'x'-b'\|_2$,

$$\|A'y^*-b'\|_2 \le \|A'x^*-b'\|_2 \le (1+\varepsilon)\|Ax^*-b\|_2$$

- Since for any $y\in\mathbb{R}^d$, $\|A'y-b'\|_2 = \|SAy-Sb\|_2 \le (1+\varepsilon)\|Ay-b\|_2$

$$\|Ay^*-b\|_2 \le (1+\varepsilon)\|A'y^*-b'\|_2 \le (1+\varepsilon)\|A'x^*-b'\|_2 \le (1+3\varepsilon)\|Ax^*-b\|_2$$

Running Time

- Reduce the problem for d vectors in $\mathbb{R}^n$ to d vectors in $\mathbb{R}^k$ with $k = O(d/\varepsilon^2)$.
- Computing SA and Sb can be done in $nnz(A)$ via sparse/fast JL
- The reduced problem can be solved in time $O(d^3/\varepsilon^2)$
- Useful when $n \gg d/\varepsilon^2$