Algorithms for Big Data (Fall 25)

Lecture 10

Subspace Embedding

Ali Vakilian (vakilian@vt.edu)

Virginia Tech

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 $1 - \delta$,

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

i. We can instead choose entries from $\{-1, +1\}$ as well.

ii. Unlike AMS sketch, entries of Π are independent.

Basically, we've projected x from $\mathbb{R}^d$ into $\mathbb{R}^k$ while preserving length to a $(1 \pm \varepsilon)$-factor.

Dimensionality Reduction

Metric JL Lemma. Let $v_1, \dots, v_n$ be n points in $\mathbb{R}^d$. For any $\varepsilon \in (0, \frac{1}{2})$, there is a linear map $f: \mathbb{R}^d \to \mathbb{R}^k$ where $k \leq 8\varepsilon^{-2} \ln n$, such that for all $i \neq j \in [n]$,

$$(\mathbf{1} - \boldsymbol{\varepsilon}) \left\| \boldsymbol{v_i} - \boldsymbol{v_j} \right\|_{\mathbf{2}} \leq \left\| \boldsymbol{f}(\boldsymbol{v_i}) - \boldsymbol{f}(\boldsymbol{v_j}) \right\|_{\mathbf{2}} \leq (\mathbf{1} + \boldsymbol{\varepsilon}) \left\| \boldsymbol{v_i} - \boldsymbol{v_j} \right\|_{\mathbf{2}}$$

- The linear map is simply given the random matrix Π; i.e., $\boldsymbol{f}(\boldsymbol{v}) = \boldsymbol{\Pi v}$
- The mapping is oblivious (to data)

Proof. Apply DJL with $\delta = n^{-2}$, and union bound over the $\binom{n}{2}$ vectors $\boldsymbol{v_i} - \boldsymbol{v_j}$, for all pairs $\boldsymbol{i} \neq \boldsymbol{j} \in [\boldsymbol{n}]$.

More on JL

Questions.

- Are the bounds achieved by the lemmas tight or can we do better?
- How about non-linear maps?

- Essentially optimal modulo constant factors for worst-case point sets.

Fast JL and Sparse JL

- The described projection matrix Π is dense and takes $\Theta(kd)$ to compute.
- Can we find Π to improve time bound?

- Each entry of Π is either -1/0/1 with similar probability
- Sparse JL: Each column is s-sparse for $s = O(\varepsilon^{-1} \log(1/\delta))$ / CountSketch

Oblivious 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 $1 - \delta$,

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

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 $x \in E$,

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

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 $x \in E$,

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

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

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

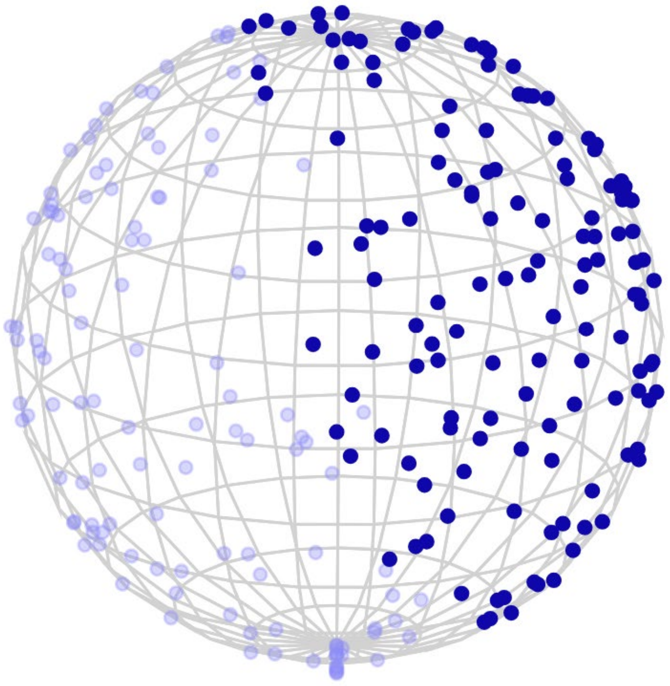
- Density function is $f(y_1, \ldots, y_k) = \left(\frac{1}{\sqrt{2\pi}}\right)^k \exp\left(-\frac{y_1^2+\cdots+y_k^2}{2}\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^k e^{-\|y\|_2/2}$
- Only depends on $\|y\|_2$
- The distribution is **centrally symmetric**. (can be used to generate a random unit vector 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

$\chi^2(\boldsymbol{k})$ **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 $Z_1, \dots, Z_k$ 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,

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

- 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.

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.