Lecture 9: Dimension Reduction

Rafael Oliveira

University of Waterloo
Cheriton School of Computer Science

rafael.oliveira.teaching@gmail.com

June 8, 2021

Overview

- Introduction
  - Why Reduce Dimensions?
  - Background: Continuous Probability Distributions
- Main Problem
  - Johnson-Lindenstrauss Lemma
- Acknowledgements

Why do we want low-dimensional objects?

When dealing with high-dimensional data, often times want to reduce dimension so that our algorithms run faster

In *smaller dimension*, things generally *run faster*, need *less storage space*, *easier to communicate*.

Why do we want low-dimensional objects?

When dealing with high-dimensional data, often times want to reduce dimension so that our algorithms run faster

In *smaller dimension*, things generally *run faster*, need *less storage space*, *easier to communicate*.

- Nearest Neighbor Search
- Large Scale Regression Problems
- Minimum Enclosing Ball
- Numerical linear algebra on large matrices
- Clustering

What properties do we want to preserve?

What do we want to preserve?

What properties do we want to preserve?

What do we want to preserve?

- distances between points

What properties do we want to preserve?

What do we want to preserve?

- distances between points
- angles between vectors

What properties do we want to preserve?

What do we want to preserve?

- distances between points
- angles between vectors
- volumes of subsets of the input

What properties do we want to preserve?

What do we want to preserve?

- distances between points
- angles between vectors
- volumes of subsets of the input
- optimal solutions to optimization problems

What properties do we want to preserve?

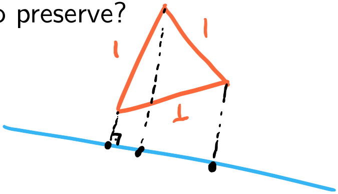
What do we want to preserve?

- distances between points
- angles between vectors
- volumes of subsets of the input
- optimal solutions to optimization problems

To preserve *distances*, need to allow some *distortion* (approximate guarantees).

What properties do we want to preserve?

What do we want to preserve?



- distances between points
- angles between vectors
- volumes of subsets of the input
- optimal solutions to optimization problems

To preserve *distances*, need to allow some *distortion* (approximate guarantees).

- Cannot compress simplex while preserving all distances.

Continuous Probability Distributions

So far we have only dealt with *discrete random variables*. Today, we will use *continuous random variables*.

Continuous Probability Distributions

So far we have only dealt with *discrete random variables*. Today, we will use *continuous random variables*.

How can we define random variables/probabilities over a *continuous* (infinite) set?

Continuous Probability Distributions

So far we have only dealt with *discrete random variables*. Today, we will use *continuous random variables*.

How can we define random variables/probabilities over a *continuous* (infinite) set?

Say we have a real-valued random variable - that is, X takes values in $\mathbb{R}$.

Definition (Probability Density Function)

A *probability density function* $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a function such that

- f is integrable over $\mathbb{R}$
- $\int_{-\infty}^{\infty} f(x)dx = 1$

non - negative probabilities

integral of f over R

"the sum of all probabilities is 1"

Continuous Probability Distributions

So far we have only dealt with *discrete random variables*. Today, we will use *continuous random variables*.

How can we define random variables/probabilities over a *continuous* (infinite) set?

Say we have a real-valued random variable - that is, X takes values in $\mathbb{R}$.

Definition (Probability Density Function)

A *probability density function* $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a function such that

- f is integrable over $\mathbb{R}$
- $\int_{-\infty}^{\infty} f(x)dx = 1$

- Probability density function $f(x)$ *intuitively* gives us *relative likelihood* that $X = x$.

Continuous Probability Distributions

So far we have only dealt with *discrete random variables*. Today, we will use *continuous random variables*.

How can we define random variables/probabilities over a *continuous* (infinite) set?

Say we have a real-valued random variable - that is, X takes values in $\mathbb{R}$.

Definition (Probability Density Function)

A *probability density function* $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a function such that

- f is integrable over $\mathbb{R}$
- $\int_{-\infty}^{\infty} f(x)dx = 1$

- Probability density function $f(x)$ *intuitively* gives us *relative likelihood* that $X = x$.
- $$\Pr[a \leq X \leq b] = \int_a^b f(x)dx$$

Gaussian Random Variables (Normal Random Variables)

Definition

A real-valued random variable X has the *normal distribution* with

- mean μ
- variance σ^2,

denoted $X \sim \mathcal{N}(\mu, \sigma^2)$, if the probability density function of X, denoted $f_X : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is:

$$f_X(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Calculus problem: prove that $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$

Gaussian Random Variables (Normal Random Variables)

Definition

A real-valued random variable X has the *normal distribution* with

- mean μ
- variance σ^2,

denoted $X \sim \mathcal{N}(\mu, \sigma^2)$, if the probability density function of X, denoted $f_X : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is:

$$f_X(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2} \cdot \left(\frac{x - \mu}{\sigma}\right)^2\right)$$

Remark

When $\mu = 0$ and $\sigma = 1$ we say that X has *standard normal distribution*.

Properties of Gaussians

Proposition (Sums of Gaussians)

If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ *and* $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ *are independent Gaussians, then*

$$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

Sum of Gaussian is also a Gaussian

Properties of Gaussians

Proposition (Sums of Gaussians)

If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ *and* $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ *are independent Gaussians, then*

$$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

Proposition (Multiplication by scalar)

If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$*, then*

$$\sigma \cdot X \sim \mathcal{N}(\sigma\mu_X, (\sigma \cdot \sigma_X)^2).$$

Properties of Gaussians

Proposition (Sums of Gaussians)

If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ *and* $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ *are independent Gaussians, then*

$$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

Proposition (Multiplication by scalar)

If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$*, then*

$$\sigma \cdot X \sim \mathcal{N}(\sigma \mu_X, (\sigma \cdot \sigma_X)^2).$$

Proposition (General Linear Combinations)

If $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ *are independent random Gaussians, then*

$$\sum_{i=1}^{n} \alpha_i \cdot X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \alpha_i \cdot \mu_i \ , \ \sum_{i=1}^{n} (\alpha_i \cdot \sigma_i)^2\right).$$

χ^2 Random Variables

Definition

A real-valued random variable X has the χ^2 *distribution* with *k degrees of freedom*, denoted $X \sim \chi^2(k)$, if

$$X = Z_1^2 + \ldots + Z_k^2$$

where each $Z_i \sim \mathcal{N}(0, 1)$ is an independent standard normal random variable.

Concentration of χ^2 random variables

Lemma (Chernoff for $\chi^2(k)$)

If $Y = \sum_{i=1}^{k} X_i^2$ is a $\chi^2(k)$ random variable with k degrees of freedom (recall $X_i \sim \mathcal{N}(0,1)$), then

$$\Pr[Y > (1+\varepsilon)^2 \cdot k] \leq \exp\left(-\frac{3}{4} \cdot k\varepsilon^2\right)$$

$$\Pr[Y < (1-\epsilon)^2 \cdot k] \leq \exp\left(-\frac{3}{4} k \varepsilon^2\right)$$

Concentration of χ^2 random variables

Lemma (Chernoff for $\chi^2(k)$)

If $Y = \sum_{i=1}^{k} X_i^2$ is a $\chi^2(k)$ random variable with k degrees of freedom (recall $X_i \sim \mathcal{N}(0,1)$), then

$$\Pr[Y > (1+\varepsilon)^2 \cdot k] \leq \exp\left(-\frac{3}{4} \cdot k\varepsilon^2\right)$$

- Let $t \in (0, 1/2)$ be a parameter

Concentration of χ^2 random variables

Lemma (Chernoff for $\chi^2(k)$)

If $Y = \sum_{i=1}^k X_i^2$ is a $\chi^2(k)$ random variable with k degrees of freedom (recall $X_i \sim \mathcal{N}(0,1)$), then

$$\Pr[Y > (1+\varepsilon)^2 \cdot k] \leq \exp\left(-\frac{3}{4} \cdot k\varepsilon^2\right)$$

- Let $t \in (0, 1/2)$ be a parameter
- $$\Pr[Y > (1+\varepsilon)^2 \cdot k] = \Pr\left[e^{tY} > e^{t\cdot(1+\varepsilon)^2\cdot k}\right] \leq \frac{\mathbb{E}[e^{tY}]}{e^{t\cdot(1+\varepsilon)^2\cdot k}}$$

↑ Chernoff strategy (exp is increasing)

↑ Markov

Concentration of χ^2 random variables

Lemma (Chernoff for $\chi^2(k)$)

If $Y = \sum_{i=1}^k X_i^2$ is a $\chi^2(k)$ random variable with k degrees of freedom (recall $X_i \sim \mathcal{N}(0,1)$), then

$$\Pr[Y > (1+\varepsilon)^2 \cdot k] \leq \exp\left(-\frac{3}{4} \cdot k\varepsilon^2\right)$$

- Let $t \in (0, 1/2)$ be a parameter
- $$\Pr[Y > (1+\varepsilon)^2 \cdot k] = \Pr\left[e^{tY} > e^{t\cdot(1+\varepsilon)^2\cdot k}\right] \leq \frac{\mathbb{E}[e^{tY}]}{e^{t\cdot(1+\varepsilon)^2\cdot k}}$$
- By independence:

$$\mathbb{E}[e^{tY}] = \mathbb{E}\left[\exp\left(\sum_{i=1}^k t \cdot X_i^2\right)\right] = \prod_{i=1}^k \mathbb{E}[e^{tX_i^2}]$$

Concentration of χ^2 random variables

Need to compute $\mathbb{E}[e^{tX_i^2}]$, where $X_i \sim \mathcal{N}(0, 1)$

Concentration of χ^2 random variables

Need to compute $\mathbb{E}[e^{tX_i^2}]$, where $X_i \sim \mathcal{N}(0,1)$

- PDF of X_i:

$X_i \sim \mathcal{N}(0,1)$

$$f_{X_i}(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp(-x^2/2)$$

Concentration of χ^2 random variables

Need to compute $\mathbb{E}[e^{tX_i^2}]$, where $X_i \sim \mathcal{N}(0,1)$

- PDF of X_i:

$$f_{X_i}(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp(-x^2/2)$$

- Thus we know that

$$\int_{-\infty}^{\infty} f_{X_i}(x)dx = 1$$

Concentration of χ^2 random variables

Need to compute $\mathbb{E}[e^{tX_i^2}]$, where $X_i \sim \mathcal{N}(0,1)$

- PDF of X_i:

$$f_{X_i}(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp(-x^2/2)$$

- Thus we know that

$$\int_{-\infty}^{\infty} f_{X_i}(x)dx = 1$$

value of random variable $e^{tX_i^2}$

- $$\mathbb{E}[e^{tX_i^2}] = \int_{-\infty}^{\infty} f_{X_i}(x) \cdot e^{tx^2} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2} \cdot e^{tx^2} dx$$

"prob. that $X_i = x$"

$\exp\left(-\frac{(1-2t)x^2}{2}\right)$

Change of vars:
$z = \sqrt{1-2t}\, x$
$dz = \sqrt{1-2t}\, dx$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} \frac{dz}{\sqrt{1-2t}}$$

Concentration of χ^2 random variables

Need to compute $\mathbb{E}[e^{tX_i^2}]$, where $X_i \sim \mathcal{N}(0,1)$

- PDF of X_i:
$$f_{X_i}(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp(-x^2/2)$$
- Thus we know that
$$\int_{-\infty}^{\infty} f_{X_i}(x)dx = 1$$
- $$\mathbb{E}[e^{tX_i^2}] = \int_{-\infty}^{\infty} f_{X_i}(x) \cdot e^{tx^2} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2} \cdot e^{tx^2} dx$$
- Change of variables $z = x\sqrt{1-2t}$
$$\mathbb{E}[e^{tX_i^2}] = \frac{1}{\sqrt{2\pi} \cdot \sqrt{1-2t}} \cdot \int_{-\infty}^{\infty} e^{-z^2/2} dz = \frac{1}{\sqrt{1-2t}}$$

PDF ∴ =1

Concentration of χ^2 random variables

Putting everything together:

Concentration of χ^2 random variables

Putting everything together:

- $$\Pr[Y > (1+\varepsilon)^2 \cdot k] \leq \frac{\mathbb{E}[e^{tY}]}{e^{t\cdot(1+\varepsilon)^2\cdot k}}$$

Concentration of χ^2 random variables

Putting everything together:

- $$\Pr[Y > (1+\varepsilon)^2 \cdot k] \leq \frac{\mathbb{E}[e^{tY}]}{e^{t\cdot(1+\varepsilon)^2\cdot k}}$$
- $$\mathbb{E}[e^{tY}] = \prod_{i=1}^{k} \mathbb{E}[e^{tX_i^2}] = \left(\frac{1}{\sqrt{1-2t}}\right)^k$$

independence

computation from previous slide

Concentration of χ^2 random variables

Putting everything together:

- $$\Pr[Y > (1+\varepsilon)^2 \cdot k] \le \frac{\mathbb{E}[e^{tY}]}{e^{t\cdot(1+\varepsilon)^2\cdot k}}$$
- $$\mathbb{E}[e^{tY}] = \prod_{i=1}^{k} \mathbb{E}[e^{tX_i^2}] = \left(\frac{1}{\sqrt{1-2t}}\right)^k = (1-2t)^{-k/2}$$
- $$\Pr[Y > (1+\varepsilon)^2 \cdot k] \le e^{-t\cdot(1+\varepsilon)^2\cdot k} \cdot (1-2t)^{-k/2} = \left[(1+\varepsilon)^2 e^{1-(1+\varepsilon)^2}\right]^{k/2}$$
- Setting $t = (1/2)\cdot\left(1 - \frac{1}{(1+\varepsilon)^2}\right)$ above

$$\Updownarrow$$

$$1-2t = (1+\epsilon)^{-2}$$

$$-t(1+\epsilon)^2 = \frac{1}{2}\left[1-(1+\epsilon)^2\right]$$

$$(1-2t)^{k/2} = (1+\epsilon)^{-k}$$

Concentration of χ^2 random variables

Putting everything together:

- $$\Pr[Y > (1+\varepsilon)^2 \cdot k] \le \frac{\mathbb{E}[e^{tY}]}{e^{t\cdot(1+\varepsilon)^2\cdot k}}$$
- $$\mathbb{E}[e^{tY}] = \prod_{i=1}^{k} \mathbb{E}[e^{tX_i^2}] = \left(\frac{1}{\sqrt{1-2t}}\right)^k$$
- $$\Pr[Y > (1+\varepsilon)^2\cdot k] \le e^{-t\cdot(1+\varepsilon)^2\cdot k}\cdot(1-2t)^{-k/2} = \left[(1+\varepsilon)^2 e^{1-(1+\varepsilon)^2}\right]^{k/2}$$
- Setting $t = (1/2)\cdot\left(1 - \frac{1}{(1+\varepsilon)^2}\right)$ above
- Use $\ln(1+x) \le x - x^4/4$ for $x \in [0,1]$

$$\Pr[Y > (1+\varepsilon)^2 \cdot k] \le \exp(-(3/4)\cdot k\varepsilon^2)$$

Concentration of χ^2 random variables

Putting everything together:

- $$\Pr[Y > (1+\varepsilon)^2 \cdot k] \leq \frac{\mathbb{E}[e^{tY}]}{e^{t\cdot(1+\varepsilon)^2\cdot k}}$$
- $$\mathbb{E}[e^{tY}] = \prod_{i=1}^{k} \mathbb{E}[e^{tX_i^2}] = \left(\frac{1}{\sqrt{1-2t}}\right)^k$$
- $$\Pr[Y > (1+\varepsilon)^2\cdot k] \leq e^{-t\cdot(1+\varepsilon)^2\cdot k}\cdot(1-2t)^{-k/2} = \left[(1+\varepsilon)^2 e^{1-(1+\varepsilon)^2}\right]^{k/2}$$
- Setting $t = (1/2)\cdot\left(1 - \frac{1}{(1+\varepsilon)^2}\right)$ above
- Use $\ln(1+x) \leq x - x^4/4$ for $x \in [0,1]$
$$\Pr[Y > (1+\varepsilon)^2 \cdot k] \leq \exp(-(3/4)\cdot k\varepsilon^2)$$
- Similar result for $\Pr[Y < (1-\varepsilon)^2 \cdot k]$ - **Practice problem.**

Dimension Reduction

- **Input:** m points $x_1, \ldots, x_m \in \mathbb{R}^n$.
- **Output:** m points $y_1, \ldots, y_m \in \mathbb{R}^d$, where $d \ll n$ such that

Dimension Reduction

- **Input:** m points $x_1, \ldots, x_m \in \mathbb{R}^n$. $\epsilon > 0$ approximation parameter, $d \in \mathbb{N}$ target dimension
- **Output:** m points $y_1, \ldots, y_m \in \mathbb{R}^d$, where $d \ll n$ such that

$$\|y_a - y_b\|_2 \approx \|x_a - x_b\|_2 \quad \forall a, b \in [m]$$

$$(1-\epsilon)\|x_a - x_b\|_2 \le \|y_a - y_b\|_2 \le (1+\epsilon)\|x_a - x_b\|_2$$

Dimension Reduction

- **Input:** m points $x_1, \ldots, x_m \in \mathbb{R}^n$.
- **Output:** m points $y_1, \ldots, y_m \in \mathbb{R}^d$, where $d \ll n$ such that

$$\|y_a - y_b\|_2 \approx \|x_a - x_b\|_2 \quad \forall a, b \in [m]$$

Theorem (Johnson-Lindenstrauss Theorem)

Let $x_1, \ldots, x_m \in \mathbb{R}^n$ and $\varepsilon \in (0,1)$. For $d = O(\log(m)/\varepsilon^2)$ there exist points $y_1, \ldots, y_m \in \mathbb{R}^d$ such that:

$$(1-\varepsilon) \cdot \|x_a - x_b\|_2 \leq \|y_a - y_b\|_2 \leq (1+\varepsilon) \cdot \|x_a - x_b\|_2 \quad \forall a, b \in [m]$$

Moreover, the points $y_j = Lx_j$, where $L \in \mathbb{R}^{d \times n}$ is a matrix whose entries $L_{a,b} \sim \mathcal{N}(0,1)$, satisfies the above with probability $\geq 1 - 2/m$.

high prob.

Dimension Reduction

- **Input:** m points $x_1, \ldots, x_m \in \mathbb{R}^n$.
- **Output:** m points $y_1, \ldots, y_m \in \mathbb{R}^d$, where $d \ll n$ such that

$$\|y_a - y_b\|_2 \approx \|x_a - x_b\|_2 \quad \forall a, b \in [m]$$

Theorem (Johnson-Lindenstrauss Theorem)

Let $x_1, \ldots, x_m \in \mathbb{R}^n$ and $\varepsilon \in (0,1)$. For $d = O(\log(m)/\varepsilon^2)$ there exist points $y_1, \ldots, y_m \in \mathbb{R}^d$ such that:

$$(1-\varepsilon) \cdot \|x_a - x_b\|_2 \ \leq \ \|y_a - y_b\|_2 \ \leq \ (1+\varepsilon) \cdot \|x_a - x_b\|_2 \quad \forall a, b \in [m]$$

Moreover, the points $y_j = Lx_j$, where $L \in \mathbb{R}^{d \times n}$ is a matrix whose entries $L_{a,b} \sim \mathcal{N}(0,1)$, satisfies the above with probability $\geq 1 - 2/m$.

- If one of the points is 0 then approximate norm of vectors as well!
- Independent of the original dimension n

Johnson-Lindenstrauss Lemma

Lemma (Johnson-Lindenstrauss Lemma)

Let $v \in \mathbb{R}^n$ such that $\|v\|_2 = 1$, $\varepsilon \in (0,1)$ and $d = O(\log(m)/\varepsilon^2)$. Let $r_1, \ldots, r_d \in \mathbb{R}^n$ be such that $r_i \sim \mathcal{N}(0,1)^n$. If we let $f : \mathbb{R}^n \to \mathbb{R}^d$ s.t.

$$f(v) = (r_1^T v, r_2^T v, \ldots, r_d^T v)$$

norm is well approximated

Then

$$\Pr\left[(1-\varepsilon) \leq \frac{\|f(v)\|_2}{\sqrt{d}} \leq (1+\varepsilon)\right] \geq 1 - 2/m^3.$$

high probability

each entry of r_i is $\sim \mathcal{N}(0,1)$

Johnson-Lindenstrauss Lemma

Lemma (Johnson-Lindenstrauss Lemma)

Let $v \in \mathbb{R}^n$ *such that* $\|v\|_2 = 1$, $\varepsilon \in (0,1)$ *and* $d = O(\log(m)/\varepsilon^2)$. *Let* $r_1, \ldots, r_d \in \mathbb{R}^n$ *be such that* $r_i \sim \mathcal{N}(0,1)^n$. *If we let* $f : \mathbb{R}^n \to \mathbb{R}^d$ *s.t.*

$$f(v) = (r_1^T v, r_2^T v, \ldots, r_d^T v)$$

Then

$$\Pr\left[(1-\varepsilon) \leq \frac{\|f(v)\|_2}{\sqrt{d}} \leq (1+\varepsilon)\right] \geq 1 - 2/m^3.$$

Proof of theorem given lemma:

- Define linear map $L(v) = f(v)/\sqrt{d}$

Johnson-Lindenstrauss Lemma

Lemma (Johnson-Lindenstrauss Lemma)

Let $v \in \mathbb{R}^n$ such that $\|v\|_2 = 1$, $\varepsilon \in (0, 1)$ and $d = O(\log(m)/\varepsilon^2)$. Let $r_1, \ldots, r_d \in \mathbb{R}^n$ be such that $r_i \sim \mathcal{N}(0, 1)^n$. If we let $f : \mathbb{R}^n \to \mathbb{R}^d$ s.t.

$$f(v) = (r_1^T v, r_2^T v, \ldots, r_d^T v)$$

Then

$$\Pr\left[(1-\varepsilon) \leq \frac{\|f(v)\|_2}{\sqrt{d}} \leq (1+\varepsilon)\right] \geq 1 - 2/m^3.$$

Proof of theorem given lemma:

- Define linear map $L(v) = f(v)/\sqrt{d}$
- By lemma, for *any* $u \in \mathbb{R}^n$, we have
  $\Pr[(1-\varepsilon) \cdot \|u\|_2 \leq \|L(u)\|_2 \leq (1+\varepsilon) \cdot \|u\|_2] \geq 1 - 2/m^3$

Pr [norm of image is bad] $\leq \frac{2}{m^3}$

Johnson-Lindenstrauss Lemma

Lemma (Johnson-Lindenstrauss Lemma)

Let $v \in \mathbb{R}^n$ such that $\|v\|_2 = 1$, $\varepsilon \in (0,1)$ and $d = O(\log(m)/\varepsilon^2)$. Let $r_1, \ldots, r_d \in \mathbb{R}^n$ be such that $r_i \sim \mathcal{N}(0,1)^n$. If we let $f : \mathbb{R}^n \to \mathbb{R}^d$ s.t.

$$f(v) = (r_1^T v, r_2^T v, \ldots, r_d^T v)$$

Then

$$\Pr\left[(1-\varepsilon) \ \le \ \frac{\|f(v)\|_2}{\sqrt{d}} \le \ (1+\varepsilon)\right] \ge 1 - 2/m^3.$$

Proof of theorem given lemma:

$u = \|u\| \cdot v$

$\Pr[(1-\varepsilon) \le \|L(v)\|_2 \le 1+\varepsilon]$

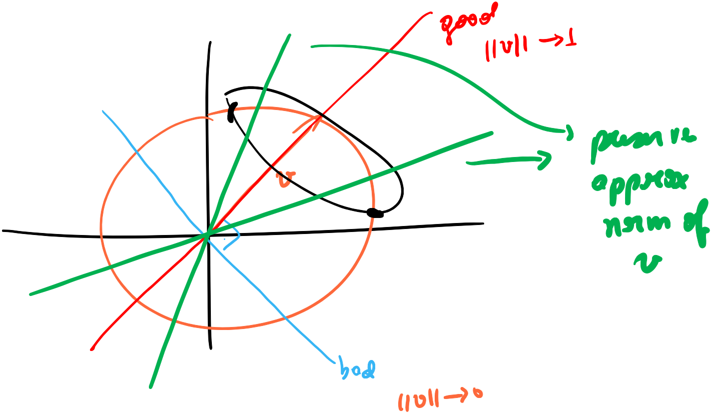
$\|Lu\| = \|u\| \cdot \|Lv\|$

- Define linear map $L(v) = f(v)/\sqrt{d}$
- By lemma, for *any* $u \in \mathbb{R}^n$, we have $\Pr[(1-\varepsilon)\cdot\|u\|_2 \ \le \ \|L(u)\|_2 \ \le \ (1+\varepsilon)\cdot\|u\|_2] \ge 1 - 2/m^3$
- Apply this result and union bound to all vectors $x_a - x_b$.
- Probability any failure on the norm $\le m^2 \cdot 2/m^3 = 2/m$.

pairs $x_a - x_b$

Intuition Please?

- JL Lemma essentially states that if we project a unit vector to a *uniformly random d-dimensional subspace* we can (almost) preserve the norm!



Intuition Please?

- JL Lemma essentially states that if we project a unit vector to a *uniformly random d-dimensional subspace* we can (almost) preserve the norm!
- One advantage of choosing random subspace is that we could *flip the randomness*: consider any d-dimensional space and take vector to be *uniformly random* unit vector

Intuition Please?

- JL Lemma essentially states that if we project a unit vector to a *uniformly random d-dimensional subspace* we can (almost) preserve the norm!
- One advantage of choosing random subspace is that we could *flip the randomness*: consider any d-dimensional space and take vector to be *uniformly random* unit vector
- So why not do that?

Intuition Please?

- JL Lemma essentially states that if we project a unit vector to a *uniformly random d-dimensional subspace* we can (almost) preserve the norm!
- One advantage of choosing random subspace is that we could *flip the randomness*: consider any d-dimensional space and take vector to be *uniformly random* unit vector
- So why not do that?
- A bit cumbersome to get random subspace (need to make L orthonormal - so need to use Gram-Schmidt)

Intuition Please?

- JL Lemma essentially states that if we project a unit vector to a *uniformly random d-dimensional subspace* we can (almost) preserve the norm!
- One advantage of choosing random subspace is that we could *flip the randomness*: consider any d-dimensional space and take vector to be *uniformly random* unit vector
- So why not do that?
- A bit cumbersome to get random subspace (need to make L orthonormal - so need to use Gram-Schmidt)
- Just taking Gaussians do the trick without Gram-Schmidt!
- More convenient algorithmically

Proof of Johnson-Lindenstrauss Lemma

Theorem (Johnson-Lindenstrauss Lemma)

Let $v \in \mathbb{R}^n$ *such that* $\|v\|_2 = 1$, $\varepsilon \in (0,1)$ *and* $d = O(\log(m)/\varepsilon^2)$. *Let* $r_1, \ldots, r_d \in \mathbb{R}^n$ *be such that* $r_i \sim \mathcal{N}(0,1)^n$. *If we let* $f : \mathbb{R}^n \to \mathbb{R}^d$ *s.t.*

$$f(v) = (r_1^T v, r_2^T v, \ldots, r_d^T v)$$

Then

$$\Pr\left[(1-\varepsilon) \ \leq \ \frac{\|f(v)\|_2}{\sqrt{d}} \leq \ (1+\varepsilon)\right] \geq 1 - 2/m^3.$$

Proof of Johnson-Lindenstrauss Lemma

Theorem (Johnson-Lindenstrauss Lemma)

Let $v \in \mathbb{R}^n$ *such that* $\|v\|_2 = 1$, $\varepsilon \in (0,1)$ *and* $d = O(\log(m)/\varepsilon^2)$. *Let* $r_1, \ldots, r_d \in \mathbb{R}^n$ *be such that* $r_i \sim \mathcal{N}(0,1)^n$. *If we let* $f : \mathbb{R}^n \to \mathbb{R}^d$ *s.t.*

$$f(v) = (r_1^T v, r_2^T v, \ldots, r_d^T v)$$

Then

$$\Pr\left[(1-\varepsilon) \ \leq \ \frac{\|f(v)\|_2}{\sqrt{d}} \leq \ (1+\varepsilon)\right] \geq 1 - 2/m^3.$$

Proof of upper tail: $\Pr[\|f(v)\|_2 > \sqrt{d} \cdot (1+\varepsilon)] < 1/m^3$

Proof of Johnson-Lindenstrauss Lemma

Theorem (Johnson-Lindenstrauss Lemma)

Let $v \in \mathbb{R}^n$ *such that* $\|v\|_2 = 1$, $\varepsilon \in (0, 1)$ *and* $d = O(\log(m)/\varepsilon^2)$. *Let* $r_1, \ldots, r_d \in \mathbb{R}^n$ *be such that* $r_i \sim \mathcal{N}(0, 1)^n$. *If we let* $f : \mathbb{R}^n \to \mathbb{R}^d$ *s.t.*

$$f(v) = (r_1^T v, r_2^T v, \ldots, r_d^T v)$$

Then

$$\Pr\left[(1-\varepsilon) \ \leq \ \frac{\|f(v)\|_2}{\sqrt{d}} \leq \ (1+\varepsilon)\right] \geq 1 - 2/m^3.$$

Proof of upper tail: $\Pr[\|f(v)\|_2 > \sqrt{d} \cdot (1+\varepsilon)] < 1/m^3$

- Let $X_i = r_i^T v$ random variable for i^{th} coordinate of $f(v)$

Proof of Johnson-Lindenstrauss Lemma

Theorem (Johnson-Lindenstrauss Lemma)

Let $v \in \mathbb{R}^n$ such that $\|v\|_2 = 1$, $\varepsilon \in (0,1)$ and $d = O(\log(m)/\varepsilon^2)$. Let $r_1, \ldots, r_d \in \mathbb{R}^n$ be such that $r_i \sim \mathcal{N}(0,1)^n$. If we let $f : \mathbb{R}^n \to \mathbb{R}^d$ s.t.

$$f(v) = (r_1^T v, r_2^T v, \ldots, r_d^T v)$$

Then

$$\Pr\left[(1-\varepsilon) \leq \frac{\|f(v)\|_2}{\sqrt{d}} \leq (1+\varepsilon)\right] \geq 1 - 2/m^3.$$

Proof of upper tail: $\Pr[\|f(v)\|_2 > \sqrt{d} \cdot (1+\varepsilon)] < 1/m^3$

- Let $X_i = r_i^T v$ — random variable for i^{th} coordinate of $f(v)$
- $X_i \sim \mathcal{N}(0, \sum_{i=1}^n v_i^2) = \mathcal{N}(0,1)$ — sum of Gaussians

$r_i = \begin{pmatrix} r_{i1} \\ r_{i2} \\ \vdots \\ r_{in} \end{pmatrix}$

$r_{i\ell} \sim \mathcal{N}(0,1)$

$X_i = \sum_{\ell=1}^{n} v_\ell r_{i\ell}$ — sum of Gaussians!

$\mu(X_i) = \sum \mu(r_{i\ell}) \cdot v_\ell = 0$

$\sigma_{X_i}^2 = \sum v_\ell^2 \cdot \sigma_{r_{i\ell}}^2 = \|v\|_2^2 = 1$ (with $\sigma_{r_{i\ell}}^2 = 1$)

Proof of Johnson-Lindenstrauss Lemma

Theorem (Johnson-Lindenstrauss Lemma)

Let $v \in \mathbb{R}^n$ such that $\|v\|_2 = 1$, $\varepsilon \in (0,1)$ and $d = O(\log(m)/\varepsilon^2)$. Let $r_1, \ldots, r_d \in \mathbb{R}^n$ be such that $r_i \sim \mathcal{N}(0,1)^n$. If we let $f : \mathbb{R}^n \to \mathbb{R}^d$ s.t.

$$f(v) = (r_1^T v, r_2^T v, \ldots, r_d^T v)$$

Then

$$\Pr\left[(1-\varepsilon) \ \leq \ \frac{\|f(v)\|_2}{\sqrt{d}} \leq \ (1+\varepsilon)\right] \geq 1 - 2/m^3.$$

Proof of upper tail: $\Pr[\|f(v)\|_2 > \sqrt{d} \cdot (1+\varepsilon)] < 1/m^3$

- Let $X_i = r_i^T v$ random variable for i^{th} coordinate of $f(v)$
- $X_i \sim \mathcal{N}(0, \sum_{i=1}^n v_i^2) = \mathcal{N}(0,1)$ sum of Gaussians
-

$$\|f(v)\|_2^2 = \sum_{i=1}^d (r_i^T v)^2 = \sum_{i=1}^d X_i^2$$

$\chi^2(d)$ - random variable!

Proof of Johnson-Lindenstrauss Lemma

Theorem (Johnson-Lindenstrauss Lemma)

Let $v \in \mathbb{R}^n$ *such that* $\|v\|_2 = 1$, $\varepsilon \in (0,1)$ *and* $d = O(\log(m)/\varepsilon^2)$. *Let* $r_1, \ldots, r_d \in \mathbb{R}^n$ *be such that* $r_i \sim \mathcal{N}(0,1)$. *If we let* $f : \mathbb{R}^n \to \mathbb{R}^d$ *s.t.*

$$f(v) = (r_1^T v, r_2^T v, \ldots, r_d^T v)$$

Then

$$\Pr\left[(1-\varepsilon) \ \le \ \frac{\|f(v)\|_2}{\sqrt{d}} \le \ (1+\varepsilon)\right] \ge 1 - 2/m^3.$$

Proof of upper tail: $\Pr[\|f(v)\|_2 > \sqrt{d} \cdot (1+\varepsilon)] < 1/m^3$

- $\|f(v)\|_2^2 = \sum_{i=1}^{d} (r_i^T v)^2 = \sum_{i=1}^{d} X_i^2$ $\qquad$ $\chi^2(d)$ random variable!

Proof of Johnson-Lindenstrauss Lemma

Theorem (Johnson-Lindenstrauss Lemma)

Let $v \in \mathbb{R}^n$ such that $\|v\|_2 = 1$, $\varepsilon \in (0,1)$ and $d = O(\log(m)/\varepsilon^2)$. Let $r_1, \ldots, r_d \in \mathbb{R}^n$ be such that $r_i \sim \mathcal{N}(0,1)$. If we let $f : \mathbb{R}^n \to \mathbb{R}^d$ s.t.

$$f(v) = (r_1^T v, r_2^T v, \ldots, r_d^T v)$$

Then

$$\Pr\left[(1-\varepsilon) \leq \frac{\|f(v)\|_2}{\sqrt{d}} \leq (1+\varepsilon)\right] \geq 1 - 2/m^3.$$

Proof of upper tail: $\Pr[\|f(v)\|_2 > \sqrt{d} \cdot (1+\varepsilon)] < 1/m^3$

- $\|f(v)\|_2^2 = \sum_{i=1}^{d} (r_i^T v)^2 = \sum_{i=1}^{d} X_i^2$ — $\chi^2(d)$ random variable!

  $d = 4 \log(m)/\varepsilon^2$

- Chernoff:

$$\Pr[\|f(v)\|_2^2 > d \cdot (1+\varepsilon)^2] < \exp(-(3/4) \cdot d\varepsilon^2) < 1/m^3$$

What if I don't like Gaussians?

- Can we even sample from a Gaussian?
- Same results also hold if pick a random matrix with entries uniformly from $\{-1, 1\}$ (Rademacher random variables).
- Proof a little more involved (see Jelani's notes for a proof)

Remarks on JL Lemma

How tight is the JL lemma?

Remarks on JL Lemma

How tight is the JL lemma?

Very tight!

Theorem (Noga Alon)

Let $y_0, \ldots, y_n \in \mathbb{R}^d$ *such that* $1 \leq \|y_i - y_j\|_2 \leq 1 + \varepsilon$ *for all* $i \neq j$. *Then*

$$d = \Omega\left(\frac{\log n}{\varepsilon^2 \cdot \log 1/\varepsilon}\right)$$

JL $\mapsto$ ε-distortion with $d = O\left(\frac{\log n}{\varepsilon^2}\right)$

NA $\mapsto$ n pts in $\mathbb{R}^d$ then $d = \Omega\left(\frac{\log n}{\varepsilon^2 \log 1/\varepsilon}\right)$

Remarks on JL Lemma

How tight is the JL lemma?

Very tight!

Theorem (Noga Alon)

Let $y_0, \ldots, y_n \in \mathbb{R}^d$ *such that* $1 \leq \|y_i - y_j\|_2 \leq 1 + \varepsilon$ *for all* $i \neq j$. *Then*

$$d = \Omega\left(\frac{\log n}{\varepsilon^2 \cdot \log 1/\varepsilon}\right)$$

Can I also compress other norms?

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad \|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

Remarks on JL Lemma

How tight is the JL lemma?

Very tight!

Theorem (Noga Alon)

Let $y_0, \ldots, y_n \in \mathbb{R}^d$ *such that* $1 \leq \|y_i - y_j\|_2 \leq 1 + \varepsilon$ *for all* $i \neq j$. *Then*

$$d = \Omega\left(\frac{\log n}{\varepsilon^2 \cdot \log 1/\varepsilon}\right)$$

Can I also compress other norms?

- Answer is *NO* in general.

Remarks on JL Lemma

How tight is the JL lemma?

Very tight!

Theorem (Noga Alon)

Let $y_0, \ldots, y_n \in \mathbb{R}^d$ such that $1 \leq \|y_i - y_j\|_2 \leq 1 + \varepsilon$ for all $i \neq j$. Then

$$d = \Omega\left(\frac{\log n}{\varepsilon^2 \cdot \log 1/\varepsilon}\right)$$

Can I also compress other norms?

- Answer is *NO* in general.
- [Brinkman, Charikar 2005]: For the ℓ_1-norm, where $\|x\|_1 = \sum_{i=1}^n |x_i|$, if want distortion $(1+\varepsilon)$ dimension must be $\Omega(n^{1/(1+\varepsilon)^2})$

Acknowledgement

- Lecture based largely on Jelani Nelson's and Nick Harvey's notes.
- See Jelani's notes at
  http://web.mit.edu/minilek/www/jl_notes.pdf
- See Nick's notes at
  http://www.cs.ubc.ca/~nickhar/W12/Lecture6Notes.pdf

References I

Brinkman, Bo and Charikar, Moses (2005)
On the impossibility of dimension reduction in ℓ_1
Journal of the ACM 52(5), 766–788.

William B. Johnson and Joram Lindenstrauss (1984)
Extensions of Lipschitz mappings into a Hilbert space
Contemporary Mathematics, 26:189–206, 1984.