Algebraic Methods in Data Science: Lesson 1

Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology

Dan Garber
https://dangar.net.technion.ac.il/

Winter Semester 2020-2021

Introduction

Course staff:

1. Lecturer: Dan Garber (dangar@technion.ac.il)
2. TA: Ido Botzer (idobotzer@campus.technion.ac.il)

Grade:

1. Homework: - 15% of grade - TAKEF
   - 7 assignments (best 6 out of 7)
   - Mostly theoretical questions but also some programming in Python
2. Final exam: 85% of grade

Introduction

Topics:

1. Complementary material in linear algebra
2. The Singular Value Decomposition, algorithms and applications
3. Linear Systems and the Least Squares Problem, algorithms and applications

Importance: the material in this course is fundamental to data science, from some of the most basic models and algorithms to the most advanced ones.
It is hard to over estimate its importance to modern DS/ML/AI.

Word of advice: this is a very challenging mathematical / algorithmic course. You will have to be proficient in linear algebra to succeed.
Make as much effort as possible to keep up with it during the semester.
WORK HARD ON YOUR HOMEWORK.

Part I - Complementary material in linear algebra

Notions such as **distance** between points, **angles** between lines, or the concept of two lines being **perpendicular (orthogonal)** to each other are well familiar from plane geometry learned in high school.

We will first develop similar notions for the more general and more abstract **linear (vector) spaces** and in particular for $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$.

This will in turn lead to the theory of eigenvalues and eigenvectors for real matrices and (basically) to everything we will do in this course.

Norms

Norms generalize the notion of distance form plane geometry to linear spaces.

A norm is a function that assigns a strictly positive length (or size) to each vector in a linear space, except for the zero vector, which is assigned a length of zero.

In the following, let $\mathcal{X}$ be a linear space.

Definition

A function $\|\cdot\| : \mathcal{X} \to \mathbb{R}$ is a norm, if

1. $\forall \mathbf{x} \in \mathcal{X}\ \|\mathbf{x}\| \geq 0$, and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$ (positivity)
2. $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}\ \|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (triangle inequality)
3. $\forall \alpha \in \mathbb{R}, \mathbf{x} \in \mathcal{X} :\ \|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$ (homogeneity)

Norms - Example (p-norms)

Consider $\mathcal{X} = \mathbb{R}^n$. The family of ℓ_p norms is defined as follows:

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |\mathbf{x}_i|^p \right)^{1/p}, \quad 1 \leq p \leq \infty.$$

In particular, for $p = 2$ we get the standard Euclidean distance

$$\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^{n} \mathbf{x}_i^2}.$$

For $p = 1$ we obtain the sum-of-absolute-values length (Manhattan distance)

$$\|\mathbf{x}\|_1 := \sum_{i=1}^{n} |\mathbf{x}_i|.$$

The limit $p = \infty$ exits, in this case we get the max-absolute-value norm

$$\|\mathbf{x}\|_\infty := \lim_{p \to \infty} \|\mathbf{x}\|_p = \max_{i \in \{1,\ldots,n\}} |\mathbf{x}_i|.$$

Norms - Example (p-norms)

Theorem

Fix $\mathcal{X} = \mathbb{R}^n$. For any $p \in [1, \infty]$, $\|\mathbf{x}\|_p := \left(\sum_{i=1}^n |\mathbf{x}_i|^p\right)^{1/p}$ is a norm.

Recall that in order to prove theorem we need to show:

1. $\forall \mathbf{x} \in \mathbb{R}^n \ \|\mathbf{x}\|_p \geq 0$, and $\|\mathbf{x}\|_p = 0$ if and only if $\mathbf{x} = \mathbf{0}$ (positivity)
2. $\forall \mathbf{x}, \mathbf{y} \mathbb{R}^n \ \|\mathbf{x} + \mathbf{y}\|_p \leq \|\mathbf{x}\| + \|\mathbf{y}\|_p$ (triangle inequality)
3. $\forall \alpha \in \mathbb{R}, \mathbf{x} \in \mathbb{R}^n : \ \|\alpha \mathbf{x}\|_p = |\alpha| \cdot \|\mathbf{x}\|_p$ (homogeneity)

Note positivity holds trivially.
Similarly, homogeneity holds since

$$\|\alpha \mathbf{x}\|_p = \left(\sum_{i=1}^n |\alpha \mathbf{x}_i|^p\right)^{1/p} = \left(\sum_{i=1}^n |\alpha|^p |\mathbf{x}_i|^p\right)^{1/p}$$
$$= |\alpha| \left(\sum_{i=1}^n |\mathbf{x}_i|^p\right)^{1/p} = |\alpha| \|\mathbf{x}\|_p.$$

It remains to prove that the triangle inequality holds.

Proof of triangle inequality for p-norms

We begin with a warmup.
The case $p = 1$: for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ we have

$$\|\mathbf{x} + \mathbf{y}\|_1 = \sum_{i=1}^n |\mathbf{x}_i + \mathbf{y}_i| \underset{(1)}{\leq} \sum_{i=1}^n (|\mathbf{x}_i| + |\mathbf{y}_i|) = \|\mathbf{x}\|_1 + \|\mathbf{y}\|_1,$$

where (1) follows from the usual triangle inequality for scalars.
The case $p = 2$: simply the Euclidean-norm (basic geometry).
The case $p = \infty$:

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|_\infty &= \max_{i \in [n]} |\mathbf{x}_i + \mathbf{y}_i| \underset{(1)}{\leq} \max_{i \in [n]} (|\mathbf{x}_i| + |\mathbf{y}_i|) \\ &\leq \max_i |\mathbf{x}_i| + \max_j |\mathbf{y}_j| = \|\mathbf{x}\|_\infty + \|\mathbf{y}\|_\infty, \end{aligned}$$

where again, (1) follows from the triangle inequality for scalars.

Lets get to proving the general case, i.e., for all $p \in [1, \infty]$.

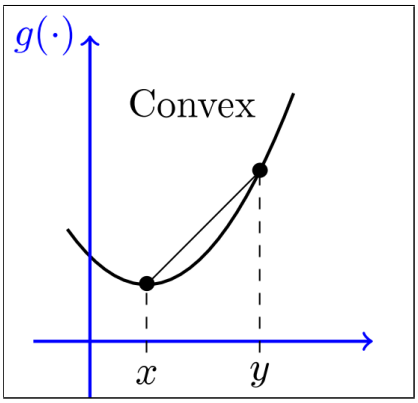
Proof of triangle inequality for p-norms

First, note that if $\mathbf{x} = \mathbf{0}$ or $\mathbf{y} = \mathbf{0}$ then the proof is trivial, since if w.l.o.g. $\mathbf{x} = \mathbf{0}$ we have $\|\mathbf{x} + \mathbf{y}\|_p = \|\mathbf{y}\|_p = \|\mathbf{y}\|_p + 0 = \|\mathbf{y}\|_p + \|\mathbf{x}\|_p$.

Consider now the case that $\|\mathbf{x}\|_p + \|\mathbf{y}\|_p = 1$. Then it suffices to show that

$$\|\mathbf{x} + \mathbf{y}\|_p^p \leq 1 = (\|\mathbf{x}\|_p + \|\mathbf{y}\|_p)^p.$$

Definition: a function $g(x) : \mathbb{R} \to \mathbb{R}$ is **convex** on interval (a, b) if for any $x, y \in (a, b)$, $\lambda \in [0, 1]$ we have $g(\lambda x + (1 - \lambda)y) \leq \lambda g(x) + (1 - \lambda)g(y)$.



Proof of triangle inequality for p-norms

First, note that if $\mathbf{x} = \mathbf{0}$ or $\mathbf{y} = \mathbf{0}$ then the proof is trivial, since if w.l.o.g. $\mathbf{x} = \mathbf{0}$ we have $\|\mathbf{x} + \mathbf{y}\|_p = \|\mathbf{y}\|_p = \|\mathbf{y}\|_p + 0 = \|\mathbf{y}\|_p + \|\mathbf{x}\|_p$.

Consider now the case that $\|\mathbf{x}\|_p + \|\mathbf{y}\|_p = 1$. Then it suffices to show that

$$\|\mathbf{x} + \mathbf{y}\|_p^p \leq 1 = (\|\mathbf{x}\|_p + \|\mathbf{y}\|_p)^p.$$

Definition: a function $g(x) : \mathbb{R} \to \mathbb{R}$ is **convex** on interval (a, b) if for any $x, y \in (a, b)$, $\lambda \in [0, 1]$ we have $g(\lambda x + (1 - \lambda)y) \leq \lambda g(x) + (1 - \lambda)g(y)$.

Fact: the scalar function $f(x) = |x|^p$ is *convex* on $(-\infty, \infty)$. That is, for any $x, y \in \mathbb{R}$ and $\lambda \in [0, 1]$ it holds that

$$|\lambda x + (1 - \lambda)y|^p \leq \lambda |x|^p + (1 - \lambda)|y|^p.$$

Proof of triangle inequality for p-norms

We assume $\mathbf{x}, \mathbf{y} \neq \mathbf{0}$, $\|\mathbf{x}\|_p + \|\mathbf{y}\|_p = 1$ and need to prove $\|\mathbf{x}+\mathbf{y}\|_p^p \leq 1$.

Convexity of $|x|^p$: $\forall x, y, \lambda \in [0,1]$: $|\lambda x + (1-\lambda)y|^p \leq \lambda |x|^p + (1-\lambda)|y|^p$.

Let us denote $\lambda = \|\mathbf{x}\|_p$. Note that $0 < \lambda < 1$ and that $\|\mathbf{y}\|_p = 1 - \lambda$ (since $\mathbf{x}, \mathbf{y} \neq \mathbf{0}$ and $\|\mathbf{x}\|_p < \|\mathbf{x}\|_p + \|\mathbf{y}\|_p = 1$).

$$\|\mathbf{x}+\mathbf{y}\|_p^p = \sum_{i=1}^n |\mathbf{x}_i + \mathbf{y}_i|^p = \sum_{i=1}^n \left|\lambda\left(\frac{\mathbf{x}_i}{\lambda}\right) + (1-\lambda)\left(\frac{\mathbf{y}_i}{1-\lambda}\right)\right|^p.$$

Applying convexity of $|x|^p$ for every $i \in [n]$ we have that

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|_p^p &\leq \sum_{i=1}^n \lambda\left|\frac{\mathbf{x}_i}{\lambda}\right|^p + (1-\lambda)\left|\frac{\mathbf{y}_i}{1-\lambda}\right|^p = \lambda^{1-p}\sum_{i=1}^n |\mathbf{x}_i|^p + (1-\lambda)^{1-p}\sum_{i=1}^n |\mathbf{y}_i|^p \\ &= \lambda^{1-p}\|\mathbf{x}\|_p^p + (1-\lambda)^{1-p}\|\mathbf{y}\|_p^p = \|\mathbf{x}\|_p^{1-p}\|\mathbf{x}\|_p^p + \|\mathbf{y}\|_p^{1-p}\|\mathbf{y}\|_p^p \\ &= \|\mathbf{x}\|_p + \|\mathbf{y}\|_p = 1.\end{aligned}$$

And so we have proved the claim for the case $\mathbf{x}, \mathbf{y} \neq \mathbf{0}$, $\|\mathbf{x}\|_p + \|\mathbf{y}\|_p = 1$.

Proof of triangle inequality for p-norms

Finally, we need to consider the case $\mathbf{x}, \mathbf{y} \neq 0$ and $\|\mathbf{x}\|_p + \|\mathbf{y}\|_p \neq 1$.

In this case let us denote $M = \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$. Using the homogeneity of $\|\cdot\|_p$ (which we already proved) we have

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|_p^p \leq (\|\mathbf{x}\|_p + \|\mathbf{y}\|_p)^p \quad &\Leftrightarrow \quad M^p \|\frac{\mathbf{x}}{M} + \frac{\mathbf{y}}{M}\|_p^p \leq M^p \\ &\Leftrightarrow \quad \|\frac{\mathbf{x}}{M} + \frac{\mathbf{y}}{M}\|_p^p \leq 1.\end{aligned}$$

Observe also that $\|\frac{\mathbf{x}}{M}\|_p + \|\frac{\mathbf{y}}{M}\|_p = \frac{1}{M}(\|\mathbf{x}\|_p + \|\mathbf{y}\|_p) = \frac{M}{M} = 1$.

Thus, we are back at the previous case.

Inner Product Spaces

Inner product is a function that associates any two vectors in a linear space with a scalar value. It will be important to generalize familiar concepts from plane geometry such as angles, or orthogonality and much more, to abstract linear spaces.

Definition

An inner product on a (real) vector space $\mathcal{X}$ is a function which maps any pair $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ into a real scalar denoted by $\langle \mathbf{x}, \mathbf{y} \rangle$, which satisfies the following axioms for any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$ and scalar $\alpha \in \mathbb{R}$:

1. $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$, and $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if and only if $\mathbf{x} = 0$ (positivity)
2. $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$ (additivity)
3. $\langle \alpha \mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle$ (homogeneity)
4. $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ (symmetry)

A vector space equipped with an inner product is called an *inner product space*.

Example - the standard inner product defined in $\mathbb{R}^n$

The standard inner product defined in $\mathbb{R}^n$ is the "row-column" product of two vectors

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{y}_i.$$

It is not difficult to show (try it for yourself) that indeed satisfy the inner product properties:

1. $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$, and $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if and only if $\mathbf{x} = 0$ (positivity)
2. $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$ (additivity)
3. $\langle \alpha \mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle$ (homogeneity)
4. $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ (symmetry)

The Cauchy-Schwarz Inequality

Theorem (Cauchy-Schwarz inequality)

For any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$: $|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle \cdot \langle \mathbf{y}, \mathbf{y} \rangle}$.

Proof: First, consider the case $\langle \mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{y}, \mathbf{y} \rangle = 1$. Using the inner-product properties:

$$
\begin{aligned}
0 \leq \langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle &= \langle \mathbf{x}, \mathbf{x} - \mathbf{y} \rangle + \langle -\mathbf{y}, \mathbf{x} - \mathbf{y} \rangle && \text{//positivity, additivity} \\
&= \langle \mathbf{x} - \mathbf{y}, \mathbf{x} \rangle - \langle \mathbf{x} - \mathbf{y}, \mathbf{y} \rangle && \text{//symmetry, homogeneity} \\
&= \langle \mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{y}, \mathbf{x} \rangle - \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle && \text{//additivity, homogeneity} \\
&= \langle \mathbf{x}, \mathbf{x} \rangle - 2\langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle && \text{//symmetry} \\
&= 2 - 2\langle \mathbf{x}, \mathbf{y} \rangle && \text{//assumption that } \langle \mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{y}, \mathbf{y} \rangle = 1
\end{aligned}
$$

Rearranging we indeed get: $\langle \mathbf{x}, \mathbf{y} \rangle \leq 1 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle \cdot \langle \mathbf{y}, \mathbf{y} \rangle}$.

The Cauchy-Schwarz Inequality

Recall we want to prove: $\mathbf{x}, \mathbf{y} \in \mathcal{X}$: $|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle \cdot \langle \mathbf{y}, \mathbf{y} \rangle}$.
Proof cont.: we have proved the theorem for the case $\langle \mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{y}, \mathbf{y} \rangle = 1$.
Let us now get to the remaining cases.
First, in case either $\mathbf{x} = \mathbf{0}$ or $\mathbf{y} = \mathbf{0}$, the theorem trivially, because $\langle \mathbf{0}, \mathbf{y} \rangle = \langle 0 \cdot \mathbf{0}, \mathbf{y} \rangle = 0 \cdot \langle \mathbf{0}, \mathbf{y} \rangle = 0$, and $\langle \mathbf{0}, \mathbf{0} \rangle = 0$.

Assume now that both $\mathbf{x} \neq \mathbf{0}, \mathbf{y} \neq \mathbf{0}$. Consider the normalized-vectors:

$$\bar{\mathbf{x}} = \frac{\mathbf{x}}{\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}}, \quad \bar{\mathbf{y}} = \frac{\mathbf{y}}{\sqrt{\langle \mathbf{y}, \mathbf{y} \rangle}}.$$

Clearly, $\langle \bar{\mathbf{x}}, \bar{\mathbf{x}} \rangle = \langle \bar{\mathbf{y}}, \bar{\mathbf{y}} \rangle = 1$. Then, using our result we have

$$\left| \left\langle \frac{\mathbf{x}}{\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}}, \frac{\mathbf{y}}{\sqrt{\langle \mathbf{y}, \mathbf{y} \rangle}} \right\rangle \right| = |\langle \bar{\mathbf{x}}, \bar{\mathbf{y}} \rangle| \leq 1.$$

Since $\left| \left\langle \frac{\mathbf{x}}{\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}}, \frac{\mathbf{y}}{\sqrt{\langle \mathbf{y}, \mathbf{y} \rangle}} \right\rangle \right| = \frac{1}{\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}} \frac{1}{\sqrt{\langle \mathbf{y}, \mathbf{y} \rangle}} |\langle \mathbf{x}, \mathbf{y} \rangle|$, rearranging we get the result.

Inner Products Induce Norms

Theorem

Let $\mathcal{X}$ be an inner product space. Then, the function $\|\cdot\| : \mathcal{X} \to \mathbb{R}$ given by $\|\mathbf{x}\| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ is a norm.

Recall we need to show:

1. $\forall \mathbf{x} \in \mathbb{R}^n \ \|\mathbf{x}\| \geq 0$, and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$ (positivity)
2. $\forall \mathbf{x}, \mathbf{y} \mathbb{R}^n \ \|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (triangle inequality)
3. $\forall \alpha \in \mathbb{R}, \mathbf{x} \in \mathbb{R}^n : \ \|\alpha \mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$ (homogeneity)

The fact that $\forall \mathbf{x}$: $\|\mathbf{x}\| \geq 0$ and $\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = 0$, follows directly from the first property of inner products.
To prove the homogeneity, fix some $\mathbf{x} \in \mathcal{X}$ and scalar $\alpha \in \mathbb{R}$. We have

$$\|\alpha \mathbf{x}\| = \sqrt{\langle \alpha \mathbf{x}, \alpha \mathbf{x} \rangle} \underset{(1)}{=} \sqrt{\alpha^2 \langle \mathbf{x}, \mathbf{y} \rangle} = |\alpha| \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = |\alpha| \|\mathbf{x}\|,$$

where (1) follows from homogeneity and symmetry of the inner product.
It remains to prove the triangle inequality.

Inner Products Induce Norms

Theorem

Let $\mathcal{X}$ be an inner product space. Then, the function $\|\cdot\| : \mathcal{X} \to \mathbb{R}$ given by $\|\mathbf{x}\| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ is a norm.

To show $\|\cdot\|$ satisfies the triangle inequality, take $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. Now, using properties of the inner-product we have:

$$\|\mathbf{x} + \mathbf{y}\|^2 \quad = \quad \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle + 2\langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle.$$

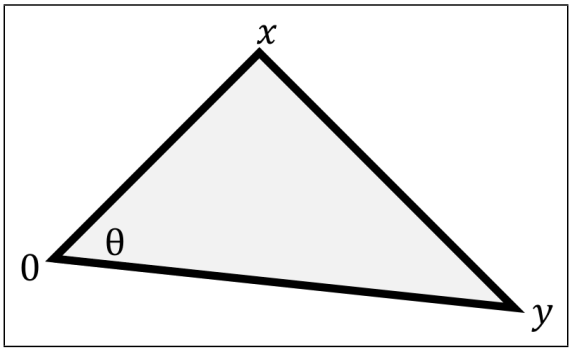
Using the CS-inequality we have $\langle \mathbf{x}, \mathbf{y} \rangle \leq \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle \langle \mathbf{y}, \mathbf{y} \rangle}$.
Thus,

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|^2 &\leq \langle \mathbf{x}, \mathbf{x} \rangle + 2\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle \langle \mathbf{y}, \mathbf{y} \rangle} + \langle \mathbf{y}, \mathbf{y} \rangle \\ &= \|\mathbf{x}\|^2 + 2\|\mathbf{x}\|\|\mathbf{y}\| + \|\mathbf{y}\|^2 \\ &= (\|\mathbf{x}\| + \|\mathbf{y}\|)^2. \end{aligned}$$

Hence, $\|\cdot\|$ satisfies the triangle inequality.

Standard Inner Product in $\mathbb{R}^n$ and Angles Between Vectors

The standard inner product in $\mathbb{R}^n$ ($\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^n \mathbf{x}_i \mathbf{y}_i$) is related to the notion of angle between two vectors.
For any two non-zero vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, consider the triangle whose vertices are the points $(\mathbf{0}, \mathbf{x}, \mathbf{y})$, and denote by θ the angle between the edges $\mathbf{x} - \mathbf{0}$ and $\mathbf{y} - \mathbf{0}$.



Recall the **cosine theorem** from plane geometry:

$$\begin{aligned} \|\mathbf{x} - \mathbf{y}\|_2^2 &= \|\mathbf{x} - \mathbf{0}\|_2^2 + \|\mathbf{y} - \mathbf{0}\|_2^2 - 2\|\mathbf{x} - \mathbf{0}\|_2 \|\mathbf{y} - \mathbf{0}\|_2 \cos\theta \\ &= \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - 2\|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos\theta. \end{aligned}$$

Standard Inner Product in $\mathbb{R}^n$ and Angles Between Vectors

The standard inner product in $\mathbb{R}^n$ ($\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^n \mathbf{x}_i \mathbf{y}_i$) is related to the notion of angle between two vectors.
For any two non-zero vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, consider the triangle whose vertices are the points $(\mathbf{0}, \mathbf{x}, \mathbf{y})$, and denote by θ the angle between the edges $\mathbf{x} - \mathbf{0}$ and $\mathbf{y} - \mathbf{0}$.
Recall the **cosine theorem** from plane geometry:

$$\begin{aligned} \|\mathbf{x} - \mathbf{y}\|_2^2 &= \|\mathbf{x} - \mathbf{0}\|_2^2 + \|\mathbf{y} - \mathbf{0}\|_2^2 - 2\|\mathbf{x} - \mathbf{0}\|_2 \|\mathbf{y} - \mathbf{0}\|_2 \cos\theta \\ &= \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - 2\|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos\theta. \end{aligned}$$

Also,

$$\begin{aligned} \|\mathbf{x} - \mathbf{y}\|_2^2 &= (\mathbf{x} - \mathbf{y})^\top (\mathbf{x} - \mathbf{y}) = \mathbf{x}^\top \mathbf{x} - \mathbf{x}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{x} + \mathbf{y}^\top \mathbf{y} \\ &= \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - 2\mathbf{x}^\top \mathbf{y}. \end{aligned}$$

Combining we have, $\mathbf{x}^\top \mathbf{y} = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos\theta$.

The angle between $\mathbf{x}$ and $\mathbf{y}$ is therefore given by $\cos\theta = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}$.

Standard Inner Product in $\mathbb{R}^n$ and Angles Between Vectors

We have seen that for the standard inner product in $\mathbb{R}^n$ it holds for any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ with angle θ between them that

$$\mathbf{x}^\top \mathbf{y} = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos\theta, \qquad \cos\theta = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}$$

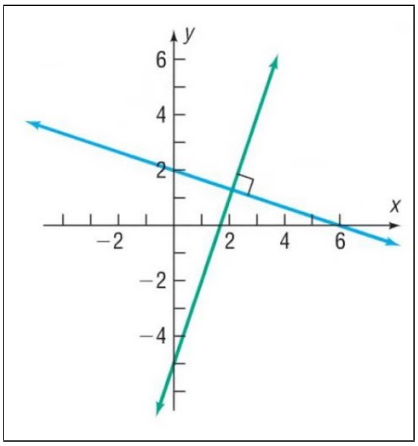
Note that since $\cos\theta \in [-1, 1]$ the following implies that

$$|\mathbf{x}^\top \mathbf{y}| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2.$$

Thus, we have reproved the Cauchy-Schwartz inequality for the special case of the standard inner product in $\mathbb{R}^n$.

Orthogonality

Orthogonality generalizes to notion of two perpendicular linear from plane geometry to abstract inner product spaces. It will be central to everything we will do in this course.



Definition

Given an inner product space $\mathcal{X}$ and vectors $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we say that $\mathbf{x}, \mathbf{y}$ are orthogonal if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$, and we write $\mathbf{x} \perp \mathbf{y}$.

Orthogonality

Definition

Given an inner product space $\mathcal{X}$ and vectors $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we say that $\mathbf{x}, \mathbf{y}$ are orthogonal if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$, and we write $\mathbf{x} \perp \mathbf{y}$.

Theorem (Pythagorean theorem)

Let $\mathcal{X}$ be an inner product space and let $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ such that $\mathbf{x} \perp \mathbf{y}$. Then

$$\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2,$$

where $\|\cdot\|$ is the norm induced by the inner product.

Proof: Using properties of the inner product we have

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|^2 &= \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle + 2\langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle \\ &= \|\mathbf{x}\|^2 + 2\langle \mathbf{x}, \mathbf{y} \rangle + \|\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2, \end{aligned}$$

where that last equality follows since $\mathbf{x} \perp \mathbf{y}$.

Orthogonality

Definition

Given an inner product space $\mathcal{X}$ and vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ in $\mathcal{X}$, all are non-zero, we say that $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ are mutually orthogonal if and only if $\langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \rangle = 0$ for all $i \neq j$.

Theorem

Given an inner product space $\mathcal{X}$, any mutually orthogonal vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ are linearly independent.

Recall $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ are linearly independent if and only if

$$\sum_{i=1}^{n} \alpha_i \mathbf{x}^{(i)} = \mathbf{0} \quad \Leftrightarrow \quad \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0.$$

Orthogonality

Theorem

Given an inner product space $\mathcal{X}$, any mutually orthogonal vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ are linearly independent.

Proof: Suppose by contradiction that $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ are linearly dependent. Assume w.l.o.g. that $\mathbf{x}^{(1)} = \sum_{i=2}^{n} \alpha_i \mathbf{x}^{(i)}$, and that $\alpha_j \neq 0$ for some $j \in \{2, \ldots, n\}$. Then, since $\mathbf{x}^{(1)}, \mathbf{x}^{(j)}$ are orthogonal we have that

$$0 = \langle \mathbf{x}^{(1)}, \mathbf{x}^{(j)} \rangle = \langle \sum_{i=2}^{n} \alpha_i \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \rangle = \sum_{i=2}^{n} \alpha_i \langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \rangle \quad \{\mathbf{x}^{(1)}, \mathbf{x}^{(j)} \text{ orthogonal}\}$$

$$= \sum_{i=2, i \neq j}^{n} \alpha_i \langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \rangle + \alpha_j \langle \mathbf{x}^{(j)}, \mathbf{x}^{(j)} \rangle$$

$$= \alpha_j \langle \mathbf{x}^{(j)}, \mathbf{x}^{(j)} \rangle \neq 0 \quad \{\mathbf{x}^{(i)}, \mathbf{x}^{(j)} \text{ are orthogonal, } \alpha_j \neq 0\}$$

We have arrived at a contradiction, and the vectors must be linearly ind.

Algebraic Methods in Data Science: Lesson 2

Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology

Dan Garber
https://dangar.net.technion.ac.il/

Winter Semester 2020-2021

Recap

Definition (Norm)

A function $\|\cdot\| : \mathcal{X} \to \mathbb{R}$ is a norm, if

1. $\forall \mathbf{x} \in \mathcal{X}\ \|\mathbf{x}\| \geq 0$, and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$ (positivity)
2. $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}\ \|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (triangle inequality)
3. $\forall \alpha \in \mathbb{R}, \mathbf{x} \in \mathcal{X}:\ \|\alpha \mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$ (homogeneity)

Definition (Inner product)

An inner product on a (real) vector space $\mathcal{X}$ is a function which maps any pair $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ into a real scalar denoted by $\langle \mathbf{x}, \mathbf{y} \rangle$, which satisfies the following axioms for any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$ and scalar $\alpha \in \mathbb{R}$:

1. $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$, and $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if and only if $\mathbf{x} = 0$ (positivity)
2. $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$ (additivity)
3. $\langle \alpha \mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle$ (homogeneity)
4. $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ (symmetry)

Recap

Theorem

Let $\mathcal{X}$ be an inner product space. Then, the function $\|\cdot\| : \mathcal{X} \to \mathbb{R}$ given by $\|\mathbf{x}\| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ is a norm.

Example: for $\mathcal{X} = \mathbb{R}^n$ the function $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is the standard inner product, and the induced norm is simply the Euclidean norm $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^\top \mathbf{x}} = \sqrt{\sum_{i=1}^n \mathbf{x}_i^2}$.

Recap

Definition (Orthogonal vectors)

Given an inner product space $\mathcal{X}$ and vectors $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we say that $\mathbf{x}, \mathbf{y}$ are orthogonal if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$, and we write $\mathbf{x} \perp \mathbf{y}$.

Definition

Given an inner product space $\mathcal{X}$ and vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ in $\mathcal{X}$, all are non-zero, we say that $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ are mutually orthogonal if and only if $\langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \rangle = 0$ for all $i \neq j$.

Theorem

Given an inner product space $\mathcal{X}$, any mutually orthogonal vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ are linearly independent.

Orthonormal Vectors and Matrices

Definition (Orthonormal vectors)

A set of vectors $S = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ in an inner product space $\mathcal{X}$ is said to be **orthonormal** if, for all $i, j = 1, \ldots, n$,

$$\langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \rangle = \begin{cases} 0 & \text{if } i \neq j; \\ 1 & \text{if } i = j. \end{cases}$$

Since from the last theorem orthonormal vectors are linearly independent, we have that a set of orthonormal vectors S forms an **orthonormal basis** for the linear span of S.

Example: for $\mathcal{X} = \mathbb{R}^n$ with the standard inner product, the set of vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$, where $\forall i, j \in \{1, \ldots, n\}$ $i \neq j$, $\mathbf{e}_i(j) = 0$ and $\mathbf{e}_i(i) = 1$, forms an orthonormal basis to $\mathbb{R}^n$.

Orthonormal Vectors and Matrices

Definition (Orthonormal matrix)

A matrix $\mathbf{X} \in \mathbb{R}^{n \times m}, m \leq n$ is said to be orthonormal if its columns are orthonormal vectors.

Corollary

A square $n \times n$ orthonormal matrix $\mathbf{X}$ is invertible and $\mathbf{X}^{-1} = \mathbf{X}^\top$.

Proof: From theorem on orthogonal vectors we have that columns of $\mathbf{X}$ are linearly independent, i.e., $\mathbf{X}$ is full rank and thus it is invertible.

Let $\mathbf{X}_i$ denote the ith column of $\mathbf{X}$. Now, from the fact that columns of $\mathbf{X}$ are orthonormal:

$$\forall i \neq j: \ [\mathbf{X}^\top \mathbf{X}]_{i,j} = \mathbf{X}_i^\top \mathbf{X}_j = 0, \quad [\mathbf{X}^\top \mathbf{X}]_{i,i} = \mathbf{X}_i^\top \mathbf{X}_i = 1$$

That is, $\mathbf{X}^\top \mathbf{X} = \mathbf{I}$. Since $\mathbf{X}^{-1}$ exists, using the above we can deduce:

$$\mathbf{X}\mathbf{X}^{-1} = \mathbf{I} \Longrightarrow \mathbf{X}^\top \mathbf{X}\mathbf{X}^{-1} = \mathbf{X}^\top \Longrightarrow \mathbf{X}^{-1} = \mathbf{X}^\top.$$

Orthogonal Decomposition of Linear Spaces

Definition (orthogonal complement)

A vector $\mathbf{x}$ is said to be orthogonal to a subset S of an inner product space $\mathcal{X}$, if $\mathbf{x} \perp \mathbf{y}$ for all $\mathbf{y} \in S$. The set of vectors in $\mathcal{X}$ that are orthogonal to S is called the **orthogonal complement** of S, and it is denoted by $S^{\perp}$.

Observation

The orthogonal complement $S^{\perp}$ is always a subspace.

Example: Let $\mathcal{X} = \mathbb{R}^n$ with the standard inner product and consider the subspace $S := \{\mathbf{x} \in \mathbb{R}^n \mid \forall i > 1 : \ \mathbf{x}_i = 0\}$.
It is not hard to see that the orthogonal complement is given by $S^{\perp} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x}_1 = 0\}$.

Orthogonal Decomposition of Linear Spaces

Definition (orthogonal complement)

A vector $\mathbf{x}$ is said to be orthogonal to a subset S of an inner product space $\mathcal{X}$, if $\mathbf{x} \perp \mathbf{y}$ for all $\mathbf{y} \in S$. The set of vectors in $\mathcal{X}$ that are orthogonal to S is called the **orthogonal complement** of S, and it is denoted by $S^{\perp}$.

Definition (direct sum)

A vector space $\mathcal{X}$ is said to be the **direct sum** of two subspaces $\mathcal{A}, \mathcal{B}$ if any element $\mathbf{x} \in \mathcal{X}$ can be written in a **unique** way as $\mathbf{x} = \mathbf{a} + \mathbf{b}$, with $\mathbf{a} \in \mathcal{A}$ and $\mathbf{b} \in \mathcal{B}$, and we write $\mathcal{X} = \mathcal{A} \oplus \mathcal{B}$.

Theorem (Orthogonal decomposition of the space)

If $\mathcal{S}$ is a subspace of an inner product space $\mathcal{X}$, then any vector $\mathbf{x} \in \mathcal{X}$ can be written in a **unique** *way as the sum of one element in $\mathcal{S}$ and one in the orthogonal complement $\mathcal{S}^{\perp}$. That is $\mathcal{X} = \mathcal{S} \oplus \mathcal{S}^{\perp}$ for any subspace $\mathcal{S} \subseteq \mathcal{X}$.*

Orthogonal Decomposition of Linear Spaces

Theorem (Orthogonal decomposition of the space)

If $\mathcal{S}$ is a subspace of an inner product space $\mathcal{X}$, then any vector $\mathbf{x} \in \mathcal{X}$ can be written in a **unique** *way as the sum of one element in $\mathcal{S}$ and one in the orthogonal complement $\mathcal{S}^{\perp}$. That is $\mathcal{X} = \mathcal{S} \oplus \mathcal{S}^{\perp}$ for any subspace $\mathcal{S} \subseteq \mathcal{X}$.*

Proof: First note that $\mathcal{S} \cap \mathcal{S}^{\perp} = \{0\}$, since if $\mathbf{v} \in \mathcal{S} \cap \mathcal{S}^{\perp}$, then by definition $\|\mathbf{v}\|^2 = \langle \mathbf{v}, \mathbf{v} \rangle = 0$.

Denote $\mathcal{W} = \mathcal{S} + \mathcal{S}^{\perp}$. That is, $\mathcal{W} = \{\mathbf{a} + \mathbf{b} \mid \mathbf{a} \in \mathcal{S},\ \mathbf{b} \in \mathcal{S}^{\perp}\}$. We will first prove that $\mathcal{W} = \mathcal{X}$ and then we will prove the uniqueness.

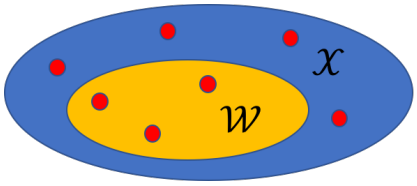
Observe that clearly $\mathcal{W} \subseteq \mathcal{X}$. Thus it remains to be shown that $\mathcal{X} \subseteq \mathcal{W}$.

Assume by contradiction that $\mathcal{X} \nsubseteq \mathcal{W}$. Consider an **orthonormal basis** of $\mathcal{W}$ and extend it (by adding additional elements) to an orthonormal basis of $\mathcal{X}$ (later on we show that for every subspace it is possible to construct an orthonormal basis via the **Gram-Schmidt procedure**), denote this basis by B.

Orthogonal Decomposition of Linear Spaces (proof cont.)

Denote $\mathcal{W} = \mathcal{S} + \mathcal{S}^{\perp}$. That is, $\mathcal{W} = \{\mathbf{a} + \mathbf{b} \mid \mathbf{a} \in \mathcal{S},\ \mathbf{b} \in \mathcal{S}^{\perp}\}$. Need to show $\mathcal{W} = \mathcal{X}$. We know $\mathcal{W} \subseteq \mathcal{X}$. Assume by contradiction that $\mathcal{X} \nsubseteq \mathcal{W}$.

Consider an orthonormal basis of $\mathcal{W}$ and extend it (by adding additional elements) to an orthonormal basis of $\mathcal{X}$ and denote this basis by B (red points in picture).



Since $\mathcal{X} \nsubseteq \mathcal{W}$ there must be a basis element $\mathbf{z} \in B$ such that $\mathbf{z} \notin \mathcal{W}$. In particular $\mathbf{z}$ is orthogonal to $\mathcal{W}$.

Since $\mathcal{S} \subseteq \mathcal{W}$, $\mathbf{z}$ is orthogonal to $\mathcal{S}$ as well $\implies \mathbf{z} \in \mathcal{S}^{\perp}$.

However, $\mathcal{S}^{\perp} \subseteq \mathcal{W}$ and hence $\mathbf{z} \in \mathcal{W}$, which results in a contradiction. Hence we proved that $\mathcal{W} = \mathcal{S} + \mathcal{S}^{\perp} = \mathcal{X}$ (each element $\mathbf{x} \in \mathcal{X}$ can be written as the sum of one element in $\mathcal{S}$ and one element in $\mathcal{S}^{\perp}$).

Orthogonal Decomposition of Linear Spaces (proof cont.)

Theorem (Orthogonal decomposition)

If $\mathcal{S}$ is a subspace of an inner product space $\mathcal{X}$, then any vector $\mathbf{x} \in \mathcal{X}$ can be written in a **unique** *way as the sum of one element in $\mathcal{S}$ and one in the orthogonal complement $\mathcal{S}^\perp$. That is $\mathcal{X} = \mathcal{S} \oplus \mathcal{S}^\perp$ for any subspace $\mathcal{S} \subseteq \mathcal{X}$.*

It remains to prove uniqueness. Suppose for the purpose of contradiction that uniqueness does not hold. Then, there exist $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}$ and $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{S}^\perp$, $\mathbf{x}_1 \neq \mathbf{x}_2$, $\mathbf{y}_1 \neq \mathbf{y}_2$ such that

$$\mathbf{x}_1 + \mathbf{y}_1 = \mathbf{x} \quad \text{and} \quad \mathbf{x}_2 + \mathbf{y}_2 = \mathbf{x}.$$

However, taking the difference of the two equations will give

$$\mathbf{x}_1 - \mathbf{x}_2 = \mathbf{y}_2 - \mathbf{y}_1.$$

Since $(\mathbf{x}_1 - \mathbf{x}_2) \in \mathcal{S}$ and $(\mathbf{y}_1 - \mathbf{y}_2) \in \mathcal{S}^\perp$ it follows that $\mathbf{x}_1 - \mathbf{x}_2 = \mathbf{y}_2 - \mathbf{y}_1 = \mathbf{0}$ (recall $S \cap S^\perp = \{\mathbf{0}\}$). However, this means that $\mathbf{x}_1 = \mathbf{x}_2$ and $\mathbf{y}_1 = \mathbf{y}_2$, and thus we arrive at a contradiction.

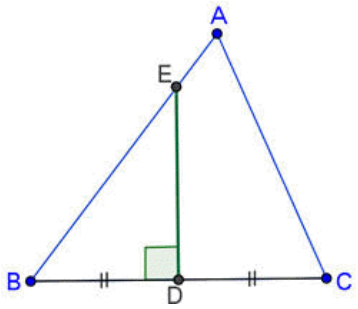
Projections onto Subspaces

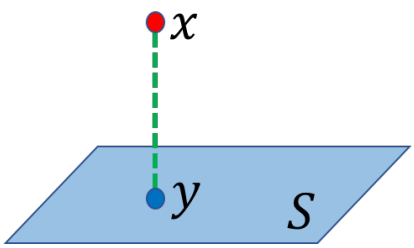
Projection is the problem of finding a point on a given set that is closest (in norm) to a given point.

Formally, given a vector $\mathbf{x}$ in an inner product space $\mathcal{X}$ and a closed set $\mathcal{S} \subseteq \mathcal{X}$ (i.e., $\mathcal{S}$ contains its boundary), the projection of $\mathbf{x}$ onto $\mathcal{S}$, denoted by $\Pi_{\mathcal{S}}(\mathbf{x})$, is defined as the point in $\mathcal{S}$ at **minimal distance** from $\mathbf{x}$:

$$\Pi_{\mathcal{S}}(\mathbf{x}) = \arg\min_{\mathbf{y} \in \mathcal{S}} \|\mathbf{y} - \mathbf{x}\|,$$

where the norm used is the one induced by the inner product.





Warmup: Projection onto a One-dimension Subspace

Given a non-zero vector $\mathbf{v} \in \mathcal{X}$, where $\mathcal{X}$ is an inner-product space, let $\mathcal{S}_\mathbf{v}$ denote the subspace spanned by $\mathbf{v}$, i.e., $\mathcal{S}_\mathbf{v} = \{\lambda \mathbf{v}, \lambda \in \mathbb{R}\}$.
Given a vector $\mathbf{x} \in \mathcal{X}$, we seek $\Pi_{\mathcal{S}_\mathbf{v}}(\mathbf{x}) = \arg\min_{\mathbf{y} \in \mathcal{S}_\mathbf{v}} \|\mathbf{y} - \mathbf{x}\|$.

We will show the projection is characterized by the fact that the difference $(\mathbf{x} - \Pi_{\mathcal{S}_\mathbf{v}}(\mathbf{x}))$ is orthogonal to $\mathbf{v}$.

Let $\mathbf{x}_\mathbf{v}$ be a point in $\mathcal{S}_\mathbf{v}$ such that $(\mathbf{x} - \mathbf{x}_\mathbf{v}) \perp \mathbf{v}$ (note from the subspace decomposition theorem we can always write $\mathbf{x} = \mathbf{x}_v + \mathbf{z}$ for some $\mathbf{z} \perp \mathcal{S}_\mathbf{v}$).

Consider an arbitrary vector $\mathbf{y} \in \mathcal{S}_\mathbf{v}$. Note that $(\mathbf{y} - \mathbf{x}_\mathbf{v}) \perp (\mathbf{x} - \mathbf{x}_\mathbf{v})$.

Thus, by the Pythagoras theorem we have

$$\|\mathbf{y} - \mathbf{x}\|^2 = \|(\mathbf{y} - \mathbf{x}_\mathbf{v}) - (\mathbf{x} - \mathbf{x}_\mathbf{v})\|^2 = \|\mathbf{y} - \mathbf{x}_\mathbf{v}\|^2 + \|\mathbf{x} - \mathbf{x}_\mathbf{v}\|^2.$$

Since the first term is always non-negative, it follows the minimum over $\mathbf{y}$ is obtained by taking $\mathbf{y} = \mathbf{x}_\mathbf{v}$, which proves that $\mathbf{x}_\mathbf{v}$ is indeed the projection we sought. Note in particular the projection is **unique**.

Warmup: Projection onto a One-dimension Subspace

Given a non-zero vector $\mathbf{v} \in \mathcal{X}$, where $\mathcal{X}$ is an inner-product space, let $\mathcal{S}_\mathbf{v}$ denote the subspace spanned by $\mathbf{v}$, i.e., $\mathcal{S}_\mathbf{v} = \{\lambda \mathbf{v}, \lambda \in \mathbb{R}\}$. Given a vector $\mathbf{x} \in \mathcal{X}$, we seek $\Pi_{\mathcal{S}_\mathbf{v}}(\mathbf{x}) = \arg\min_{\mathbf{y} \in \mathcal{S}_\mathbf{v}} \|\mathbf{y} - \mathbf{x}\|$.

We showed the projection $\mathbf{x}_\mathbf{v} = \Pi_{\mathcal{S}_\mathbf{v}}(\mathbf{x})$ is characterized by the orthogonality condition $\langle \mathbf{x} - \mathbf{x}_\mathbf{v}, \mathbf{v} \rangle = 0$.
Let's now find an explicit expression for $\mathbf{x}_\mathbf{v}$.

Since $\mathbf{x}_\mathbf{v} \in \mathcal{S}_\mathbf{v}$ it follows that $\mathbf{x}_\mathbf{v} = \lambda \mathbf{v}$ for some $\lambda \in \mathbb{R}$. We have that

$$0 = \langle \mathbf{x} - \mathbf{x}_\mathbf{v}, \mathbf{v} \rangle = \langle \mathbf{x} - \lambda \mathbf{v}, \mathbf{v} \rangle = \langle \mathbf{x}, \mathbf{v} \rangle - \lambda \langle \mathbf{v}, \mathbf{v} \rangle = \langle \mathbf{x}, \mathbf{v} \rangle - \lambda \|\mathbf{v}\|^2.$$

Hence, we obtain $\lambda = \frac{\langle \mathbf{x}, \mathbf{v} \rangle}{\|\mathbf{v}\|^2}$, which results in the projection

$$\mathbf{x}_\mathbf{v} = \Pi_{\mathcal{S}_\mathbf{v}}(\mathbf{x}) = \frac{\langle \mathbf{x}, \mathbf{v} \rangle}{\|\mathbf{v}\|^2} \mathbf{v}.$$

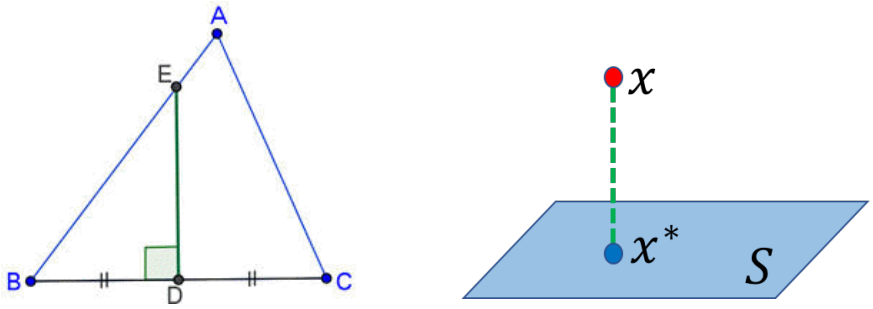
$\mathbf{x}_\mathbf{v}$ is usually called the *component* of $\mathbf{x}$ along the direction $\mathbf{v}$.
In particular, if $\|\mathbf{v}\| = 1$ then $\mathbf{x}_\mathbf{v} = \langle \mathbf{x}, \mathbf{v} \rangle \mathbf{v}$.

Projection onto Arbitrary Subspaces

We now extend the previous result to the case when $\mathcal{S}$ is not necessarily one-dimensional. In this case also orthogonality plays a key role.

Theorem (projection theorem)

Let $\mathcal{X}$ be an inner product space, let $\mathbf{x}$ be a given element in $\mathcal{X}$, and let $\mathcal{S}$ be a subspace of $\mathcal{X}$. Then, there exists a unique vector $\mathbf{x}^* \in \mathcal{S}$ which is the solution to the problem $\min_{\mathbf{y}\in\mathcal{S}} \|\mathbf{y}-\mathbf{x}\|$.
Moreover, a necessary and sufficient condition for $\mathbf{x}^*$ being the optimal solution for this problem is that $\mathbf{x}^* \in \mathcal{S}, \quad (\mathbf{x}-\mathbf{x}^*) \perp \mathcal{S}$.



Projection onto Arbitrary Subspaces

Theorem (projection theorem)

Let $\mathcal{X}$ be an inner product space, let $\mathbf{x}$ be a given element in $\mathcal{X}$, and let $\mathcal{S}$ be a subspace of $\mathcal{X}$. Then, there exists a unique vector $\mathbf{x}^* \in \mathcal{S}$ which is the solution to the problem $\min_{\mathbf{y}\in\mathcal{S}} \|\mathbf{y}-\mathbf{x}\|$.
Moreover, a necessary and sufficient condition for $\mathbf{x}^*$ being the optimal solution for this problem is that $\mathbf{x}^* \in \mathcal{S}, \quad (\mathbf{x}-\mathbf{x}^*) \perp \mathcal{S}$.

Proof: by the subspace decomposition theorem $\mathbf{x}$ can be written in a unique way as $\mathbf{x} = \mathbf{u} + \mathbf{z},\ \mathbf{u} \in \mathcal{S},\ \mathbf{z} \in \mathcal{S}^\perp$. Hence, for any $\mathbf{y} \in \mathcal{S}$, since $(\mathbf{y}-\mathbf{u}) \in \mathcal{S}$ and $\mathbf{z} \in \mathcal{S}^\perp$ and using the Pythagoras theorem, we have:

$$\|\mathbf{y}-\mathbf{x}\|^2 = \|(\mathbf{y}-\mathbf{u})-\mathbf{z}\|^2 = \|\mathbf{y}-\mathbf{u}\|^2 + \|\mathbf{z}\|^2.$$

It follows that the unique minimizer is $\mathbf{x}^* = \mathbf{u}$. Indeed, with this choice we a have $(\mathbf{x}-\mathbf{x}^*) = \mathbf{z} \perp \mathcal{S}$.

Projection onto Vector Span

Suppose now we have a basis for a subspace $\mathcal{S} \subseteq \mathcal{X}$, that is $\mathcal{S} = \text{span}(\mathbf{x}_1, \ldots, \mathbf{x}_m)$. Given $\mathbf{x} \in \mathcal{X}$, the projection theorem tells us that the unique projection $\mathbf{x}^*$ of $\mathbf{x}$ onto $\mathcal{S}$ is characterized by the orthogonality condition $(\mathbf{x} - \mathbf{x}^*) \perp \mathcal{S}$.
Since $\mathbf{x}^* \in \mathcal{S}$, we can write $\mathbf{x}^*$ as some (unknown) linear combination of the basis elements $\mathbf{x}_1, \ldots, \mathbf{x}_m$, that is $\mathbf{x}^* = \sum_{i=1}^m \alpha_i \mathbf{x}_i$, where the scalar coefficients $\alpha_1, \ldots, \alpha_m$ are unknown. Note that

$$\begin{aligned}(\mathbf{x} - \mathbf{x}^*) \perp \mathcal{S} &\Leftrightarrow \forall i \in \{1, \ldots, m\} : \langle \mathbf{x} - \mathbf{x}^*, \mathbf{x}_i \rangle = 0 \\ &\Leftrightarrow \forall i \in \{1, \ldots, m\} : \langle \mathbf{x}^*, \mathbf{x}_i \rangle = \langle \mathbf{x}, \mathbf{x}_i \rangle.\end{aligned}$$

Plugging $\mathbf{x}^* = \sum_{i=1}^m \alpha_i \mathbf{x}_i$, we arrive at the following system of m linear equations in m unknowns (the scalars $\alpha_1, \ldots, \alpha_m$):

$$\sum_{i=1}^m \alpha_i \langle \mathbf{x}_i, \mathbf{x}_k \rangle = \langle \mathbf{x}, \mathbf{x}_k \rangle, \quad k = 1, \ldots m.$$

The solution provides the coefficients $\alpha_1, \ldots, \alpha_m$, and hence the projection $\mathbf{x}^*$.

Projection onto Span of Orthonormal Vectors

Recall the projection is given by $\mathbf{x}^* = \sum_{i=1}^m \alpha_i \mathbf{x}_i$ where $\alpha_1, \ldots, \alpha_m$ are the result of solving the linear system:

$$\sum_{i=1}^m \alpha_i \langle \mathbf{x}_i, \mathbf{x}_k \rangle = \langle \mathbf{x}, \mathbf{x}_k \rangle, \quad k = 1, \ldots m.$$

Consider the previous case with the difference that now the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are **orthonormal** (recall we can always construct an orthonormal basis via the Gram-Schmidt procedure, to be detailed later). Now, we get

$$\alpha_k = \sum_{i=1}^m \alpha_i \langle \mathbf{x}_i, \mathbf{x}_k \rangle = \langle \mathbf{x}, \mathbf{x}_k \rangle, \quad k = 1, \ldots, m,$$

which gives the closed-form solution to the projection:

$$\mathbf{x}^* = \sum_{i=1}^m \langle \mathbf{x}, \mathbf{x}_i \rangle \mathbf{x}_i.$$

Projection onto Span of Orthonormal Vectors

In case $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are orthonormal, the projection of $\mathbf{x}$ onto $\mathcal{S} = \mathrm{span}(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ is given by

$$\mathbf{x}^* = \sum_{i=1}^{m} \langle \mathbf{x}, \mathbf{x}_i \rangle \mathbf{x}_i.$$

In the **special case** in which the linear space is $\mathcal{X} = \mathbb{R}^n$ with the standard inner product, i.e., $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$, the above can be written in the following matrix form:

$$\mathbf{x}^* = \sum_{i=1}^{m} \mathbf{x}^\top \mathbf{x}_i \mathbf{x}_i = \sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{x} = \left(\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^\top \right) \mathbf{x} = \mathbf{P}\mathbf{P}^\top \mathbf{x},$$

where $\mathbf{P}$ is the $n \times m$ matrix whose columns are exactly the basis vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$.
$\mathbf{P}\mathbf{P}^\top$ is called the **projection matrix** onto $\mathrm{span}(\mathbf{x}_1, \ldots, \mathbf{x}_m)$.

The Gram-Schmidt procedure (briefly)

We now complement the above results by showing that given any basis $(\mathbf{x}_1, \ldots \mathbf{x}_m)$ for a subspace $\mathcal{S} \subseteq \mathcal{X}$, one can construct an **orthonormal basis** for $\mathcal{S}$. This could be done via the Gram-Schmidt procedure.

The process takes any basis $(\mathbf{x}_1, \ldots \mathbf{x}_m)$ and generates an **orthonormal basis** $(\mathbf{y}_1, \ldots \mathbf{y}_m)$ such that $\mathrm{span}(\mathbf{y}_1, \ldots \mathbf{y}_m) = \mathrm{span}(\mathbf{x}_1, \ldots \mathbf{x}_m)$.

The actual proof is by induction, showing that for all $i \in \{1, \ldots, m\}$, $\mathrm{span}(\mathbf{y}_1, \ldots, \mathbf{y}_i) = \mathrm{span}(\mathbf{x}_1, \ldots \mathbf{x}_i)$.

The process works in m steps, each generating a new basis member $\mathbf{y}_i$.

The Gram-Schmidt procedure (briefly)

For the first step we simply take $\mathbf{y}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|}$. Clearly the induction holds for $i = 1$, since $\text{span}(\mathbf{y}_1) = \text{span}(\mathbf{x}_1)$.

Suppose the induction holds for some $i \geq 1$ and denote $\mathcal{S}_i = \text{span}(\mathbf{y}_1, \ldots, \mathbf{y}_i) = \text{span}(\mathbf{x}_1, \ldots, \mathbf{x}_i)$.

Let $\mathbf{z}_{i+1} = \mathbf{x}_{i+1} - \Pi_{\mathcal{S}_i}(\mathbf{x}_{i+1})$ and $\mathbf{y}_{i+1} = \frac{\mathbf{z}_{i+1}}{\|\mathbf{z}_{i+1}\|}$ (recall $\Pi_{\mathcal{S}_i}(\mathbf{x}_{i+1})$ is the projection of $\mathbf{x}_{t+1}$ onto $S_i = \text{span}(\mathbf{y}_1, \ldots, \mathbf{y}_i)$).
From the projection theorem: $\mathbf{x}_{i+1} - \Pi_{\mathcal{S}_i}(\mathbf{x}_{i+1}) \perp \text{span}(\mathbf{y}_1, \ldots, \mathbf{y}_i)$.
Thus, $\mathbf{y}_{i+1} \perp \text{span}(\mathbf{y}_1, \ldots, \mathbf{y}_i)$ and $\|\mathbf{y}_{t+1}\| = 1$.

Note that since $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ is a basis, we have that $\mathbf{x}_{i+1} \notin \mathcal{S}_i$ and hence $\mathbf{z}_{i+1} \neq \mathbf{0}$, and therefore $\mathbf{y}_{i+1}$ is well defined.

Moreover, it clearly holds that $\mathbf{x}_{i+1} \in \text{span}(\mathbf{y}_1, \ldots, \mathbf{y}_{i+1})$ and hence the induction holds for $i + 1$ as well, i.e.,
$\mathcal{S}_i = \text{span}(\mathbf{y}_1, \ldots, \mathbf{y}_{i+1}) = \text{span}(\mathbf{x}_1, \ldots, \mathbf{x}_{i+1})$.

Did you notice the circular logic bug in today's lecture?

The Gram-Schmidt procedure (briefly)

Did you notice the circular logic bug in today's lecture?

We have used Gram-Schmidt to prove the Orthogonal Decompositon of Linear Spaces theorem (to construct an orthonormal basis).
But we have proved Gram-Schmidt using the Projection theorem whose proof uses the Orthogonal Decompositon theorem.

Not a real problem: in the tutorial you will see a more detailed version and proof of Gram-Schmidt which **DOES NOT** rely on the Projection theorem.

Algebraic Methods in Data Science: Lesson 3

Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology

Dan Garber
https://dangar.net.technion.ac.il/

Winter Semester 2020-2021

Range and Nullspace of Matrices

We focus our attention to real-valued $m \times n$ matrices. Let us begin with recalling some of the most basic definition.

Fix a $m \times n$ real matrix $\mathbf{A}$. We denote the **range** (or image) and **nullspace** (kernel) as

$$\begin{aligned} \mathcal{R}(\mathbf{A}) &= \mathrm{Im}(\mathbf{A}) := \{\mathbf{A}\mathbf{x} \ : \ \mathbf{x} \in \mathbb{R}^n\}; \\ \mathcal{N}(\mathbf{A}) &= \mathrm{Ker}(\mathbf{A}) := \{\mathbf{x} \in \mathbb{R}^n \ : \ \mathbf{A}\mathbf{x} = 0\}. \end{aligned}$$

Recall $\mathcal{R}(\mathbf{A})$ is a subspace of $\mathbb{R}^m$ and $\mathcal{N}(\mathbf{A})$ is a subspace of $\mathbb{R}^n$. Recall the rank-nullity theorem:

$$\begin{aligned} \dim \mathcal{R}(\mathbf{A}) + \dim \mathcal{N}(\mathbf{A}) &= n; \\ \dim \mathcal{R}(\mathbf{A}^\top) + \dim \mathcal{N}(\mathbf{A}^\top) &= m. \end{aligned}$$

Matrix Inner Products and Norms

Let $\mathcal{X} = \mathbb{R}^{m\times n}$. The standard matrix inner product is defined (similarly to the standard inner product for vectors) as:

$$\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i\in[m], j\in[n]} \mathbf{A}_{i,j}\mathbf{B}_{i,j}.$$

It is also common to write $\mathbf{A} \bullet \mathbf{B}$. It is not hard to show that (HW):

$$\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}(\mathbf{A}^\top \mathbf{B}).$$

The Euclidean norm for $\mathbb{R}^{m\times n}$ is defined similarly to the vector case, and as in the vector case it is induced by the standard inner product.
It is called the **Frobenius norm** and it is given by

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} \mathbf{A}_{i,j}^2} = \sqrt{\mathrm{Tr}(\mathbf{A}^\top \mathbf{A})}.$$

Matrix Inner Products and Norms

Note that the ℓ_p norms defined for vectors, also naturally extend to matrices, i.e., the following are norms

$$\|\mathbf{A}\|_{(p)} = \Big(\sum_{i,j} |\mathbf{A}_{i,j}|^p\Big)^{1/p} \qquad 1 \le p \le \infty.$$

Theorem (HW)

For every $p, q \in [1,\infty] \times [1,\infty]$, the following function is a norm

$$\|\mathbf{A}\|_{p\to q} = \max_{\mathbf{x}\in\mathbb{R}^n : \|\mathbf{x}\|_p = 1} \|\mathbf{A}\mathbf{x}\|_q.$$

$\|\mathbf{A}\|_{p\to q}$ measures how much can $\mathbf{A}$ "amplify" in q-norm, a vector of unit length with respect to the p-norm.

In particular, we will use the notation $\|\mathbf{A}\|_p := \|\mathbf{A}\|_{p\to p}$.

Matrix Inner Products and Norms

Examples (HW): Denote by $\mathbf{A}^j$ the jth column and by $\mathbf{A}_j$ the jth row. Then,

$$\begin{aligned}\|\mathbf{A}\|_1 &:= \|\mathbf{A}\|_{1\to 1} = \max_{\|\mathbf{x}\|_1=1} \|\mathbf{A}\mathbf{x}\|_1 = \max_{j=1,\ldots,n} \|\mathbf{A}^j\|_1,\\ \|\mathbf{A}\|_\infty &:= \|\mathbf{A}\|_{\infty\to\infty} = \max_{\|\mathbf{x}\|_\infty=1} \|\mathbf{A}\mathbf{x}\|_\infty = \max_{j=1,\ldots,m} \|\mathbf{A}_j\|_1.\end{aligned}$$

Of particular interest (as we'll discuss in the sequel) is the so-called **spectral norm** which is given by

$$\|\mathbf{A}\|_2 := \|\mathbf{A}\|_{2\to 2} = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2.$$

In particular, as we shall see, this norm is related to the eigenvalues of $\mathbf{A}^\top\mathbf{A}$, since it holds that

$$\|\mathbf{A}\|_2 = \sqrt{\max_i \lambda_i(\mathbf{A}^\top\mathbf{A})}.$$

Recap on Eigenvalues and Eigenvectors of Square Matrices

Definition

Given a real square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ we say λ is an eigenvalue of $\mathbf{A}$, if there exists a vector $\mathbf{u} \neq \mathbf{0}$ such that $\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$.
We say $\mathbf{u}$ is an eigenvector corresponding to eigenvalue λ.

Recall that even when $\mathbf{A}$ is real, both eigenvalues and eigenvectors need not be real (i.e., can be complex).

The eigenvalues of a square real matrix $\mathbf{A}$ are the roots of the characteristic polynomial: $p_{\mathbf{A}}(\lambda) = \det(\lambda\mathbf{I} - \mathbf{A})$.

Recall this is a polynomial of degree n with real coefficients.

Theorem (Fundamental theorem of algebra)

Any matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ has n (not necessarily real) eigenvalues, counting multiplicities.

Recap on Eigenvalues and Eigenvectors of Square Matrices

Theorem (Fundamental theorem of algebra)

Any matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ has n (not necessarily real) eigenvalues, counting multiplicities.

Recall that for any eigenvalue λ of a real matrix $\mathbf{A}$, the eigenvectors of $\mathbf{A}$ that correspond to the eigenvalue λ, are the set of all solutions to the linear system

$$(\lambda \mathbf{I} - \mathbf{A})\mathbf{v} = \mathbf{0},$$

or equivalently, it is the nullspace

$$\mathcal{N}(\lambda \mathbf{I} - \mathbf{A}) = \mathrm{Ker}(\lambda \mathbf{I} - \mathbf{A}).$$

Hence, the set of all eigenvectors associated with an eigenvalue λ is in particular a subspace of $\mathbb{C}^n$.

Recap on Eigenvalues and Eigenvectors of Square Matrices

For any eigenvalue λ, denoting by μ its number of appearances as a root of the characteristic polynomial $\det(\lambda \mathbf{I} - \mathbf{A})$ (algebraic multiplicity) and by v the dimension of the eigenvectors subspace, i.e., $v = \dim(\mathcal{N}(\lambda \mathbf{I} - \mathbf{A}))$ (the geometric multiplicity). We have $v \leq \mu$.

Theorem

Let $\lambda_i, i = 1, \ldots, k$ be the distinct eigenvalues of a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, and let $\phi_i = Ker(\lambda_i \mathbf{I} - \mathbf{A}), i = 1, \ldots, k$ be the corresponding eigenspaces. Then, any k nonzero vectors $\mathbf{u}_i \in \phi_i, i = 1, \ldots, k$, are linearly independent.

Recap on Eigenvalues and Eigenvectors of Square Matrices

Theorem (Diagnolization via eigenvectors)

Let λ_i, $i = 1, \ldots, k \leq n$, be the distinct eigenvalues of a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, let $\mu_i, i = 1, \ldots, k$, denote the corresponding algebraic multiplicities, and let $\phi_i = \mathcal{N}(\lambda_i \mathbf{I}_n - \mathbf{A})$. Let further $\mathbf{U}^{(i)} = (\mathbf{u}_1^{(i)} \cdots \mathbf{u}_{v_i}^{(i)})$ be a matrix containing by columns a basis of ϕ_i where $v_i = \dim(\phi_i)$, $i = 1..k$. If $v_i = \mu_i$ for all $i = 1, \ldots, k$, then the matrix $\mathbf{U} = (\mathbf{U}^{(1)} \cdots \mathbf{U}^{(k)}) \in \mathbb{R}^{n \times n}$ is invertible, and $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^{-1}$, where

$$\Lambda = \begin{pmatrix} \lambda_1 \mathbf{I}_{\mu_1} & 0 & \cdots & 0 \\ 0 & \lambda_2 \mathbf{I}_{\mu_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \vdots & 0 & \lambda_k \mathbf{I}_{\mu_K} \end{pmatrix}.$$

Here, $\mathbf{I}_{\mu_j}$ denotes the $\mu_j \times \mu_j$ identity matrix.

Recap on Eigenvalues and Eigenvectors of Square Matrices

Theorem (Diagnolization via eigenvectors)

Let λ_i, $i = 1, \ldots, k \leq n$, be the distinct eigenvalues of a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, let $\mu_i, i = 1, \ldots, k$, denote the corresponding algebraic multiplicities, and let $\phi_i = \mathcal{N}(\lambda_i \mathbf{I}_n - \mathbf{A})$. Let further $\mathbf{U}^{(i)} = (\mathbf{u}_1^{(i)} \cdots \mathbf{u}_{v_i}^{(i)})$ be a matrix containing by columns a basis of ϕ_i where $v_i = \dim(\phi_i)$, $i = 1..k$. If $v_i = \mu_i$ for all $i = 1, \ldots, k$, then the matrix $\mathbf{U} = (\mathbf{U}^{(1)} \cdots \mathbf{U}^{(k)}) \in \mathbb{R}^{n \times n}$ is invertible, and $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^{-1}$.

Proof: The fact that $\mathbf{U}$ is invertible follows since its columns correspond to eigenvectors of different eigenvalues. Since by previous theorem we have that eigenvectors of different eigenvalues are linearly independent, we have that the columns of $\mathbf{U}$ are linearly independent and hence it has full rank and thus it is invertible.

Recap on Eigenvalues and Eigenvectors of Square Matrices

Theorem (Diagnolization via eigenvectors)

Let λ_i, $i = 1, \ldots, k \leq n$, be the distinct eigenvalues of a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, let $\mu_i, i = 1, \ldots, k$, denote the corresponding algebraic multiplicities, and let $\phi_i = \mathcal{N}(\lambda_i \mathbf{I}_n - \mathbf{A})$. Let further $\mathbf{U}^{(i)} = (\mathbf{u}_1^{(i)} \cdots \mathbf{u}_{v_i}^{(i)})$ be a matrix containing by columns a basis of ϕ_i where $v_i = \dim(\phi_i)$, $i = 1..k$. If $v_i = \mu_i$ for all $i = 1, \ldots, k$, then the matrix $\mathbf{U} = (\mathbf{U}^{(1)} \cdots \mathbf{U}^{(k)})$ is invertible, and $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^{-1}$.

Proof cont.: It follows from simple calculation that $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^{-1}$:

$$\begin{aligned}
\mathbf{AU} &= (\mathbf{AU}^{(1)}, \mathbf{AU}^{(2)}, ..., \mathbf{AU}^{(k)}) \\
&= (\mathbf{Au}_1^{(1)}, \ldots, \mathbf{Au}_{v_1}^{(1)}, \mathbf{Au}_1^{(2)}, \ldots, \mathbf{Au}_{v_2}^{(2)}, \ldots, \mathbf{Au}_1^{(k)}, \ldots, \mathbf{Au}_{v_k}^{(k)}) \\
&= (\lambda_1 \mathbf{u}_1^{(1)}, \ldots, \lambda_1 u_{v_1}^{(1)}, \lambda_2 \mathbf{u}_1^{(2)}, \ldots, \lambda_2 \mathbf{u}_{v_2}^{(2)}, \ldots, \lambda_k \mathbf{u}_1^{(k)}, \ldots, \lambda_k \mathbf{u}_{v_k}^{(k)}) \\
&= (\lambda_1 \mathbf{U}^{(1)}, \lambda_2 \mathbf{U}^{(2)}, ..., \lambda_k \mathbf{U}^{(k)}) = (\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, ..., \mathbf{U}^{(k)})\Lambda = \mathbf{U}\Lambda.
\end{aligned}$$

Now, multiplying on both sides with $\mathbf{U}^{-1}$ from the right, we get the result.

Eigendecomposition of Real Symmetric Matrices

We turn to discuss real **symmetric** matrices (recall $\mathbf{A}$ is symmetric if $\mathbf{A}_{i,j} = \mathbf{A}_{j,i}$). We shall see that these matrices have a unique structure (w.r.t. eigenvalues and eigenvectors) which will be the basis to much of the material in this course.

We denote by $\mathbb{S}^n$ the space of real $n \times n$ symmetric matrices.

Theorem (Eigendecomposition of a symmetric matrix)

Let $\mathbf{A} \in \mathbb{S}^n$, let λ_i, $i = 1, \ldots, k \leq n$, be the distinct eigenvalues of $\mathbf{A}$. Let further μ_i denote the algebraic multiplicity of λ_i, and let $\phi_i = Ker(\lambda_i \mathbf{I}_n - \mathbf{A})$. Then, for all $i = 1, \ldots, k$:

1. $\lambda_i \in \mathbb{R}$ *and corresponding eigenvectors can alway be chosen to be in* $\mathbb{R}^n$,
2. $\phi_i \perp \phi_j$ $(\forall\ \mathbf{u}_i \in \phi_i, \mathbf{u}_j \in \phi_j,\ i \neq j : \mathbf{u}_i^\top \mathbf{u}_j = 0)$,
3. $\dim \phi_i = \mu_i$.

Proof of the Eigen-decomposition Theorem

Proof of part 1 - real eigenvalues and eigenvectors:
Let $\lambda, \mathbf{u}$ be any eigenvalue/eigenvector pair for $\mathbf{A}$, i.e., $\mathbf{Au} = \lambda\mathbf{u}$.

By taking the **conjugate transpose** of both sides we have $\mathbf{u}^*\mathbf{A}^* = \lambda^*\mathbf{u}^*$.

Multiplying the first blue equation on the left by $\mathbf{u}^*$ and the second blue equation on the right by $\mathbf{u}$, we have

$$\mathbf{u}^*\mathbf{Au} = \lambda\mathbf{u}^*\mathbf{u}, \quad \mathbf{u}^*\mathbf{A}^*\mathbf{u} = \lambda^*\mathbf{u}^*\mathbf{u}. \tag{1}$$

Since $\mathbf{u}^*\mathbf{u} = \|\mathbf{u}\|_2^2 \neq 0$ (recall from tutorial we know $\mathbf{u}^*\mathbf{v}$ is inner product over $\mathbb{C}^n$), recalling that $\mathbf{A}$ is real implies that $\mathbf{A}^* = \mathbf{A}^\top$, and subtracting the two equalities in (1), it follows that

$$\mathbf{u}^*\mathbf{Au} - \mathbf{u}^*\mathbf{A}^*\mathbf{u} = \mathbf{u}^*(\mathbf{A} - \mathbf{A}^\top)\mathbf{u} = (\lambda - \lambda^*)\|\mathbf{u}\|_2^2.$$

Now, since $\mathbf{A} - \mathbf{A}^\top = \mathbf{0}$, it must hold that $\lambda - \lambda^* = 0$, which implies that λ must be a real number.

Proof of the Eigen-decomposition Theorem

Proof of part 1 - real eigenvalues and eigenvectors:
Let $\lambda, \mathbf{u}$ be any eigenvalue/eigenvector pair for $\mathbf{A}$, i.e., $\mathbf{Au} = \lambda\mathbf{u}$.
We have shown λ must be real.
Let us now show that without loss of generality, we can always choose $\mathbf{u}$ to be real, i.e., $\mathbf{u} \in \mathbb{R}^n$.
Suppose $\mathbf{u} \neq \mathbf{0}$ is complex and suppose $\mathrm{Re}(\mathbf{u}) \neq \mathbf{0}$ (if $\mathrm{Re}(\mathbf{u}) = \mathbf{0}$ we can alway take $i\mathbf{u}$ instead and then $i\mathbf{u} \in \mathbb{R}^n$).
On one hand we have:

$$\mathrm{Re}(\mathbf{Au}) = \mathrm{Re}(\lambda\mathbf{u}) = \lambda\mathrm{Re}(\mathbf{u}).$$

On the other hand we also have that

$$\mathrm{Re}(\mathbf{Au}) = \mathrm{Re}(\mathbf{A}(\mathrm{Re}(\mathbf{u}) + i \cdot \mathrm{Im}(\mathbf{u})) = \mathrm{Re}(\mathbf{A}\mathrm{Re}(\mathbf{u})) = \mathbf{A}\mathrm{Re}(\mathbf{u}).$$

Thus, we have $\mathbf{A}\mathrm{Re}(\mathbf{u}) = \lambda\mathrm{Re}(\mathbf{u})$, which means that $\mathrm{Re}(\mathbf{u})$ is an eigenvector of $\mathbf{A}$ associated with λ.

Proof of the Eigen-decomposition Theorem

Proof of part 2 - $\phi_i \perp \phi_j$: Recall $\phi_i = \mathrm{Ker}(\lambda_i \mathbf{I}_n - \mathbf{A})$.

Let $\mathbf{v}_i \in \phi_i, \mathbf{v}_j \in \phi_j, i \neq j$.

Since $\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$ we have

$$\mathbf{v}_j^\top \mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_j^\top \mathbf{v}_i. \tag{2}$$

Since $\mathbf{A}\mathbf{v}_j = \lambda_j \mathbf{v}_j$ we have

$$\mathbf{v}_j^\top \mathbf{A}\mathbf{v}_i = \mathbf{v}_i^\top \mathbf{A}^\top \mathbf{v}_j = \mathbf{v}_i^\top \mathbf{A}\mathbf{v}_j = \lambda_j \mathbf{v}_i^\top \mathbf{v}_j = \lambda_j \mathbf{v}_j^\top \mathbf{v}_i. \tag{3}$$

Subtracting the Eq. (3) from (2) we obtain

$$0 = (\lambda_i - \lambda_j)\mathbf{v}_j^\top \mathbf{v}_i.$$

Since $\lambda_i \neq \lambda_j$ it must hold that $\mathbf{v}_j^\top \mathbf{v}_i = 0$.

Proof of part 3 of Theorem: $\dim \phi_i = \mu_i$

Recall $\phi_i = \mathrm{Ker}(\lambda_i \mathbf{I}_n - \mathbf{A})$. Fix eigenvalue λ. We will prove this by constructing an **orthonormal basis** for $\phi = \phi_i$ composed of $\mu = \mu_i$ elements.

Lemma (Auxiliary Lemma)

Let $\mathbf{B} \in \mathbb{S}^m$ and let λ be an eigenvalue of $\mathbf{B}$. Then, there exists an **orthogonal matrix** $\mathbf{U} = [\mathbf{u}\ \mathbf{Q}] \in \mathbb{R}^{m \times m}$, $\mathbf{Q} \in \mathbb{R}^{m \times (m-1)}$, *such that*

$$\mathbf{B}\mathbf{u} = \lambda \mathbf{u},\ \|\mathbf{u}\|_2 = 1, \quad \mathbf{U}^\top \mathbf{B}\mathbf{U} = \begin{pmatrix} \lambda & 0 \\ 0 & \mathbf{Q}^\top \mathbf{B}\mathbf{Q} \end{pmatrix}, \quad \mathbf{Q}^\top \mathbf{B}\mathbf{Q} \in \mathbb{S}^{m-1}.$$

Recall an orthogonal matrix is a square matrix whose columns are orthonormal vectors.

We will prove the lemma later on.

Proof of part 3 of Theorem: $\dim \phi_i = \mu_i$

Lemma (Auxiliary Lemma)

Let $\mathbf{B} \in \mathbb{S}^m$ and let λ be an eigenvalue of $\mathbf{B}$. Then, there exists an **orthogonal matrix** $\mathbf{U} = [\mathbf{u}\ \mathbf{Q}] \in \mathbb{R}^{m\times m}$, $\mathbf{Q} \in \mathbb{R}^{m\times(m-1)}$, *such that*

$$\mathbf{Bu} = \lambda \mathbf{u},\ \|\mathbf{u}\|_2 = 1, \quad \mathbf{U}^\top \mathbf{B}\mathbf{U} = \begin{pmatrix} \lambda & 0 \\ 0 & \mathbf{Q}^\top \mathbf{B}\mathbf{Q} \end{pmatrix}, \quad \mathbf{Q}^\top \mathbf{B}\mathbf{Q} \in \mathbb{S}^{m-1}.$$

Throughout the proof we fix some eigenvalue λ of $\mathbf{A}$.

We first apply the lemma to $\mathbf{A} \in \mathbb{S}^n$: since $\mu \geq 1$, there exists an orthogonal matrix $\mathbf{U}_1 = [\mathbf{u}_1\ \mathbf{Q}_1] \in \mathbb{R}^{n\times n}$, such that $\mathbf{A}\mathbf{u}_1 = \lambda \mathbf{u}_1$, and

$$\mathbf{U}_1^\top \mathbf{A}\mathbf{U}_1 = \begin{pmatrix} \lambda & 0 \\ 0 & \mathbf{A}_1 \end{pmatrix}, \quad \mathbf{A}_1 = \mathbf{Q}_1^\top \mathbf{A}\mathbf{Q}_1 \in \mathbb{S}^{n-1}.$$

Now, if $\mu = 1$ we have finished the proof, since we found a subspace of ϕ of dimension one (the subspace is $\mathrm{span}(\mathbf{u}_1)$).

Proof of part 3 of Theorem: $\dim \phi_i = \mu_i$

Suppose now $\mu > 1$.
Recall that there exist orthogonal $\mathbf{U}_1 = [\mathbf{u}_1\ \mathbf{Q}_1] \in \mathbb{R}^{n\times n}$, such that $\mathbf{A}\mathbf{u}_1 = \lambda \mathbf{u}_1$, and

$$\mathbf{U}_1^\top \mathbf{A}\mathbf{U}_1 = \begin{pmatrix} \lambda & 0 \\ 0 & \mathbf{A}_1 \end{pmatrix}, \quad \mathbf{A}_1 = \mathbf{Q}_1^\top \mathbf{A}\mathbf{Q}_1 \in \mathbb{S}^{n-1}.$$

Note that since $\mathbf{U}_1$ is orthogonal, we have $\mathbf{U}_1^{-1} = \mathbf{U}_1^\top$.

Thus, the matrices $\mathbf{A}$ and $\mathbf{U}_1^\top \mathbf{A}\mathbf{U}_1$ are **similar!** In particular they have the same eigenvalues (including algebraic multiplicities).

Because of the block diagonal structure of $\mathbf{U}_1^\top \mathbf{A}\mathbf{U}_1$, λ is an eigenvalue of $\mathbf{A}_1$ of multiplicity $\mu - 1$ (in particular, note that the charcharistic polynomial of $\mathbf{A}$ is given by $\rho_{\mathbf{A}}(\sigma) = (\sigma - \lambda) \cdot \rho_{\mathbf{A}_1}(\sigma)$).

Proof of part 3 of Theorem: $\dim \phi_i = \mu_i$

Suppose now $\mu > 1$.
Recall that there exist orthogonal $\mathbf{U}_1 = [\mathbf{u}_1 \ \mathbf{Q}_1] \in \mathbb{R}^{n \times n}$, such that $\mathbf{A}\mathbf{u}_1 = \lambda \mathbf{u}_1$, and

$$\mathbf{U}_1^\top \mathbf{A} \mathbf{U}_1 = \begin{pmatrix} \lambda & 0 \\ 0 & \mathbf{A}_1 \end{pmatrix}, \quad \mathbf{A}_1 = \mathbf{Q}_1^\top \mathbf{A} \mathbf{Q}_1 \in \mathbb{S}^{n-1}.$$

We showed λ is an eigenvalue of $\mathbf{A}_1$ of multiplicity $\mu - 1$.

We hence apply the same reasoning to the symmetric matrix $\mathbf{A}_1 \in \mathbb{S}^{n-1}$: there exists an orthogonal matrix $\mathbf{U}_2 = [\tilde{\mathbf{u}}_2 \ \mathbf{Q}_2] \in \mathbb{R}^{(n-1)\times(n-1)}$ such that $\mathbf{A}_1 \tilde{\mathbf{u}}_2 = \lambda \tilde{\mathbf{u}}_2$, $\|\tilde{\mathbf{u}}_2\|_2 = 1$, and

$$\mathbf{U}_2^\top \mathbf{A}_1 \mathbf{U}_2 = \begin{pmatrix} \lambda & 0 \\ 0 & \mathbf{A}_2 \end{pmatrix}, \quad \mathbf{A}_2 = \mathbf{Q}_2^\top \mathbf{A}_1 \mathbf{Q}_2 \in \mathbb{S}^{n-2}.$$

Proof of part 3 of Theorem: $\dim \phi_i = \mu_i$

We next show that the vector $\mathbf{u}_2 = \mathbf{U}_1 \begin{pmatrix} 0 \\ \tilde{\mathbf{u}}_2 \end{pmatrix}$ is a unit-norm eigenvector of $\mathbf{A}$ corresponding to the eigenvalue λ, and it is orthogonal to $\mathbf{u}_1$. Indeed,

$$\mathbf{A}\mathbf{u}_2 = \left(\mathbf{U}_1 \begin{pmatrix} \lambda & \mathbf{0}_{n-1} \\ \mathbf{0}_{n-1} & \mathbf{A}_1 \end{pmatrix} \mathbf{U}_1^\top\right) \mathbf{U}_1 \begin{pmatrix} 0 \\ \tilde{\mathbf{u}}_2 \end{pmatrix} = \mathbf{U}_1 \begin{pmatrix} \lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_1 \end{pmatrix} \begin{pmatrix} 0 \\ \tilde{\mathbf{u}}_2 \end{pmatrix}$$
$$= \mathbf{U}_1 \begin{pmatrix} 0 \\ \mathbf{A}_1 \tilde{\mathbf{u}}_2 \end{pmatrix} = \mathbf{U}_1 \begin{pmatrix} 0 \\ \lambda \tilde{\mathbf{u}}_2 \end{pmatrix} = \lambda \mathbf{u}_2.$$

Moreover,

$$\|\mathbf{u}_2\|_2^2 = \mathbf{u}_2^\top \mathbf{u}_2 = \begin{pmatrix} 0 \\ \tilde{\mathbf{u}}_2 \end{pmatrix}^\top \mathbf{U}_1^\top \mathbf{U}_1 \begin{pmatrix} 0 \\ \tilde{\mathbf{u}}_2 \end{pmatrix} = \|\tilde{\mathbf{u}}_2\|_2^2 = 1,$$

and

$$\mathbf{u}_1^\top \mathbf{u}_2 = \mathbf{u}_1^\top \mathbf{U}_1 \begin{pmatrix} 0 \\ \tilde{\mathbf{u}}_2 \end{pmatrix} = \mathbf{u}_1^\top [\mathbf{u}_1 \ \mathbf{Q}_1] \begin{pmatrix} 0 \\ \tilde{\mathbf{u}}_2 \end{pmatrix} = [1 \ \mathbf{0}_{n-1}] \begin{pmatrix} 0 \\ \tilde{\mathbf{u}}_2 \end{pmatrix} = 0.$$

If $\mu = 2$, then the proof is finished, since we have found an orthonormal basis of dimension two for ϕ (the vectors $\mathbf{u}_1, \mathbf{u}_2$).

Proof of part 3 of Theorem: $\dim \phi_i = \mu_i$

Otherwise, if $\mu > 2$, we iterate the same reasoning on matrix $\mathbf{A}_2 \in \mathbb{R}^{(n-2)\times(n-2)}$: we find an eigenvector $\mathbf{u}_3$ orthogonal to $\mathbf{u}_1, \mathbf{u}_2$. We do this, by taking a unit-norm eigenvector $\tilde{\mathbf{u}}'_3$ satisfying $\mathbf{A}_2\tilde{\mathbf{u}}'_3 = \lambda\tilde{\mathbf{u}}'_3$.

Then, by re-iterating the above arguments, we have that $\tilde{\mathbf{u}}_3 = \mathbf{U}_2 \begin{pmatrix} 0 \\ \tilde{\mathbf{u}}'_3 \end{pmatrix}$ is a unit-length eigenvector of $\mathbf{A}_1 \in \mathbb{R}^{(n-1)\times(n-1)}$ corresponding to eigenvalue λ and orthogonal to $\tilde{\mathbf{u}}_2$ (previously found eigenvector, of $\mathbf{A}_1$).

Using the same argument once more, we have that the vector $\mathbf{u}_3 = \mathbf{U}_1 \begin{pmatrix} 0 \\ \tilde{\mathbf{u}}_3 \end{pmatrix}$ is a unit-length eigenvector of $\mathbf{A}$ corresponding to eigenvalue λ, and orthogonal to both $\mathbf{u}_1, \mathbf{u}_2$.

We can continue this process until we reach the actual value of μ (notice that by the above, for each $\mathbf{A}_i$, λ is an eigenvalue of algebraic multiplicity $\mu - i$) and at this point we exit the procedure with an orthonormal basis of ϕ composed of exactly μ vectors.

Eigendecomposition of Real Symmetric Matrices

Theorem (Eigendecomposition of a symmetric matrix)

Let $\mathbf{A} \in \mathbb{S}^n$, let λ_i, $i = 1, \ldots, k \leq n$, be the distinct eigenvalues of $\mathbf{A}$. Let further μ_i denote the algebraic multiplicity of λ_i, and let $\phi_i = Ker(\lambda_i \mathbf{I}_n - \mathbf{A})$. Then, for all $i = 1, \ldots, k$:

1. $\lambda_i \in \mathbb{R}$ *and corresponding eigenvectors can alway be chosen to be in* $\mathbb{R}^n$,
2. $\phi_i \perp \phi_j$ $(\forall\ \mathbf{u}_i \in \phi_i, \mathbf{u}_j \in \phi_j,\ i \neq j : \mathbf{u}_i^\top \mathbf{u}_j = 0)$,
3. $\dim \phi_i = \mu_i$.

The Spectral Theorem / Eigen-decompositon

Combining the Eigen-decomposition theorem and the diagonalization-via-eigenvectors theorem we have:

Theorem (Spectral theorem)

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be symmetric, let $\lambda_i \in \mathbb{R}$, $i = 1, \ldots, n$, be the eigenvalues of $\mathbf{A}$ (counting multiplicities). Then, there exists a set of orthonormal vectors $\mathbf{u}_i \in \mathbb{R}^n$, $i = 1, \ldots, n$, such that $\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i$. Equivalently, there exists an orthogonal matrix $\mathbf{U} = [\mathbf{u}_1 \cdots \mathbf{u}_n]$ (i.e., $\mathbf{U}\mathbf{U}^\top = \mathbf{U}^\top\mathbf{U} = \mathbf{I}_n$) such that $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^\top = \sum_{i=1}^n \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$, $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$

In particular, any symmetric matrix can be decomposed as a weighted sum of simple rank-one matrices of the form $\mathbf{u}_i\mathbf{u}_i^\top$, where the weights are given by the eigenvalues λ_i.

Convention: from now on we consider the eigenvalues in non-increasing order, i.e., $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n$, and the eigenvectors as an orthonormal basis of $\mathbb{R}^n$.

Proof of the Auxiliary Lemma

Recall that in order to prove that $\mu_i = \dim \phi_i$ we have used the following lemma:

Lemma

Let $\mathbf{B} \in \mathbb{S}^m$ and let λ be an eigenvalue of $\mathbf{B}$. Then, there exists an **orthogonal matrix** $\mathbf{U} = [\mathbf{u}\ \mathbf{Q}] \in \mathbb{R}^{m\times m}$, $\mathbf{Q} \in \mathbb{R}^{m\times(m-1)}$, *such that*

$$\mathbf{B}\mathbf{u} = \lambda\mathbf{u}, \quad \mathbf{U}^\top\mathbf{B}\mathbf{U} = \begin{pmatrix} \lambda & 0 \\ 0 & \mathbf{Q}^\top\mathbf{B}\mathbf{Q} \end{pmatrix}, \quad \mathbf{Q}^\top\mathbf{B}\mathbf{Q} \in \mathbb{S}^{m-1}.$$

Proof: Let $\mathbf{u}$ be any unit-norm eigenvector of $\mathbf{B}$ associated with λ. We can now take $\mathbf{Q}$ to be matrix whose columns are an orthonormal basis to the subspace orthogonal to $\mathbf{u}$. Hence, $\mathbf{U} = [\mathbf{u}\ \mathbf{Q}]$ is orthogonal by construction.

Proof of the Auxiliary Lemma

Lemma

Let $\mathbf{B} \in \mathbb{S}^m$ and let λ be an eigenvalue of $\mathbf{B}$. Then, there exists an **orthogonal matrix** $\mathbf{U} = [\mathbf{u}\ \mathbf{Q}] \in \mathbb{R}^{m \times m}$, $\mathbf{Q} \in \mathbb{R}^{m \times (m-1)}$, *such that*

$$\mathbf{B}\mathbf{u} = \lambda \mathbf{u}, \quad \mathbf{U}^\top \mathbf{B} \mathbf{U} = \begin{pmatrix} \lambda & 0 \\ 0 & \mathbf{Q}^\top \mathbf{B} \mathbf{Q} \end{pmatrix}, \quad \mathbf{Q}^\top \mathbf{B} \mathbf{Q} \in \mathbb{S}^{m-1}.$$

Proof cont.: By calculation:

$$\begin{aligned} \mathbf{U}^\top \mathbf{B} \mathbf{U} &= [\mathbf{u}\ \mathbf{Q}]^\top \mathbf{B} [\mathbf{u}\ \mathbf{Q}] = \begin{pmatrix} \mathbf{u}^\top \mathbf{B}\mathbf{u} & \mathbf{u}^\top \mathbf{B}\mathbf{Q} \\ \mathbf{Q}^\top \mathbf{B}\mathbf{u} & \mathbf{Q}^\top \mathbf{B}\mathbf{Q} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{u}^\top \mathbf{B}\mathbf{u} & (\mathbf{Q}^\top \mathbf{B}\mathbf{u})^\top \\ \mathbf{Q}^\top \mathbf{B}\mathbf{u} & \mathbf{Q}^\top \mathbf{B}\mathbf{Q} \end{pmatrix} = \begin{pmatrix} \mathbf{u}^\top (\lambda \mathbf{u}) & (\mathbf{Q}^\top \lambda \mathbf{u})^\top \\ \mathbf{Q}^\top \lambda \mathbf{u} & \mathbf{Q}^\top \mathbf{B}\mathbf{Q} \end{pmatrix} \\ &= \begin{pmatrix} \lambda & 0 \\ 0 & \mathbf{Q}^\top \mathbf{B}\mathbf{Q} \end{pmatrix}, \end{aligned}$$

where the last equality follows since the columns of $\mathbf{Q}$ are orthogonal to $\mathbf{u}$.

Some Applications of the Spectral Theorem

Theorem (inverse and matrix power (HW))

Let $\mathbf{A} \in \mathbb{S}^n$, and write its eigen-decomposition $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^\top$. Then

1. *if $\mathbf{A}$ is invertible then $\mathbf{A}^{-1} = \mathbf{U}\Lambda^{-1}\mathbf{U}^\top$,*
2. *for any $k \in \mathbb{N}_+$: $\mathbf{A}^k = \mathbf{U}\Lambda^k\mathbf{U}^\top$.*

Observation

Let $\mathbf{A} \in \mathbb{S}^n$. Let $\lambda_1, \ldots, \lambda_n$ denote its eigenvalues and let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ denote the corresponding eigenvectors. Then,

$$\mathbf{A} = \sum_{i \in [n]} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top = \sum_{i \in [n] : \lambda_i \neq 0} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top.$$

Observation

Let $\mathbf{A} \in \mathbb{S}^n$. $rank(\mathbf{A})$ is equal to number of non-zero eigenvalues of $\mathbf{A}$.

Computational Consequences of the Spectral Theorem

We now give some motivation why the eigen-decompistion (and related decompositions that we will see in the sequel) are so important in modern data-intensive applications.

For the following, think on the case that $\mathrm{rank}(\mathbf{A}) = k << n$ (this indeed holds for many matrices representing data that we see in real-life).

Let $\mathbf{A} \in \mathbb{S}^n$ and suppose that the eigen-decomposition $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^\top = \sum_{i=1}^k \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$ is given, where $\mathrm{rank}(\mathbf{A}) = k$. Then,

- Informally speaking, $\mathbf{A}$ can be stored in a computer's memory using only $k(1+n)$ memory units - storing the k non-zero eigenvalues and the k eigenvectors (assuming each memory unit can store a scalar).

  On the other hand, an explicit $n \times n$ symmetric matrix requires $n + \frac{n^2-n}{2} = \Theta(n^2)$ memory units (independent of k).

Computational Consequences of the Spectral Theorem

Let $\mathbf{A} \in \mathbb{S}^n$ and suppose that the eigen-decomposition $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^\top = \sum_{i=1}^k \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$ is given, where $\mathrm{rank}(\mathbf{A}) = k$. Then,

- Multiplying a vector $\mathbf{x} \in \mathbb{R}^n$ by $\mathbf{A}$ takes $O(kn)$ time:
  1. first, compute the scalars $\mathbf{u}_i^\top \mathbf{x}, i = 1, \ldots, k = O(kn)$ time
  2. then, compute the sum $\sum_{i=1}^k \lambda_i \mathbf{u}_i \mathbf{u}_i^\top \mathbf{x}$ - $O(kn)$ time.

  On the other hand, computing $\mathbf{A}\mathbf{x}$, when $\mathbf{A}$ is given explicitly (entry by entry) takes $O(n^2)$ time (n row-column inner products).

- Multiplying a matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$ with $\mathbf{A}$ takes $O(kn^2)$ time, instead of $O(n^3)$ time for an explicitly given $n \times n$ matrix (similar argument for the matrix-column product, but now apply for each of the n columns).

Algebraic Methods in Data Science: Lesson 4

Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology

Dan Garber
https://dangar.net.technion.ac.il/

Winter Semester 2020-2021

Variational Characterization of Eigenvalues

Given a matrix $\mathbf{A} \in \mathbb{S}^n$, we consider its eigenvalues in non-increasing order, i.e., $\lambda_{\max} = \lambda_1 \geq \lambda_2 \cdots \lambda_n = \lambda_{\min}$.

The extreme eigenvalues λ_1, λ_n are related to certain optimization problems - the minimum and maximum value of the quadratic form induced by $\mathbf{A}$ over the unit Euclidean sphere.

Towards this end, for any $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{0}$, the ratio

$$\frac{\mathbf{x}^\top \mathbf{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}}$$

is called the *Rayleigh quotient* of $\mathbf{x}$ w.r.t. $\mathbf{A}$.

Note that the RQ is independent of the norm of $\mathbf{x}$.

Variational Characterization of Eigenvalues

For any $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{0}$, the ratio

$$\frac{\mathbf{x}^\top \mathbf{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}}$$

is called the *Rayleigh quotient* of $\mathbf{x}$ w.r.t. $\mathbf{A}$.

Theorem (Rayleigh quotients)

Given $\mathbf{A} \in \mathbb{S}^n$, it holds that

$$\forall \mathbf{x} \neq \mathbf{0}: \quad \lambda_{\min}(\mathbf{A}) \leq \frac{\mathbf{x}^\top \mathbf{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}} \leq \lambda_{\max}(\mathbf{A}).$$

Moreover,

$$\lambda_{\max} = \max_{\mathbf{x}:\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x}, \quad \lambda_{\min} = \min_{\mathbf{x}:\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x},$$

and the maximum and minimum are attained for $\mathbf{u}_1$ and $\mathbf{u}_n$ respectively, where $\mathbf{u}_1, \ldots \mathbf{u}_n$ are the eigenvectors of $\mathbf{A}$.

Proof of Rayleigh's Theorem

First need to prove: $\forall \mathbf{x} \neq \mathbf{0}: \ \lambda_{\min}(\mathbf{A}) \leq \frac{\mathbf{x}^\top \mathbf{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}} \leq \lambda_{\max}(\mathbf{A})$.

Write the eig-decomp. $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^\top$.
Given $\mathbf{x} \neq \mathbf{0}$ define $\bar{\mathbf{x}} = \mathbf{U}^\top \mathbf{x}$. We have

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \mathbf{x}^\top \mathbf{U}\Lambda\mathbf{U}^\top \mathbf{x} = \bar{\mathbf{x}}^\top \Lambda \bar{\mathbf{x}} = \sum_{i=1}^n \lambda_i \bar{\mathbf{x}}_i^2.$$

Clearly,

$$\lambda_{\min} \sum_{i=1}^n \bar{\mathbf{x}}_i^2 \leq \lambda_i \sum_{i=1}^n \bar{\mathbf{x}}_i^2 \leq \lambda_{\max} \sum_{i=1}^n \bar{\mathbf{x}}_i^2.$$

Thus, we have that

$$\lambda_{\min} \sum_{i=1}^n \bar{\mathbf{x}}_i^2 \leq \mathbf{x}^\top \mathbf{A}\mathbf{x} \leq \lambda_{\max} \sum_{i=1}^n \bar{\mathbf{x}}_i^2.$$

Proof of Rayleigh's Theorem

First need to prove: $\forall \mathbf{x} \neq \mathbf{0}: \ \lambda_{\min}(\mathbf{A}) \leq \frac{\mathbf{x}^\top \mathbf{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}} \leq \lambda_{\max}(\mathbf{A})$.

We have shown that $\forall \mathbf{x} \neq 0: \ \lambda_{\min} \sum_{i=1}^n \bar{\mathbf{x}}_i^2 \leq \mathbf{x}^\top \mathbf{A}\mathbf{x} \leq \lambda_{\max} \sum_{i=1}^n \bar{\mathbf{x}}_i^2$.

It remains to show that $\sum_{i=1}^n \bar{\mathbf{x}}_i^2 = \mathbf{x}^\top \mathbf{x} = \|\mathbf{x}\|_2^2$.

This clearly follows since $\mathbf{U}$ has orthonormal columns and thus,

$$\|\bar{\mathbf{x}}\|_2^2 = \|\mathbf{U}^\top \mathbf{x}\|_2^2 = \mathbf{x}^\top \mathbf{U}\mathbf{U}^\top \mathbf{x} = \mathbf{x}^\top \mathbf{I}_n \mathbf{x} = \|\mathbf{x}\|_2^2.$$

For the second part of theorem we need to show $\mathbf{u}_1^\top \mathbf{A}\mathbf{u}_1 = \lambda_{\max}$ and $\mathbf{u}_n^\top \mathbf{A}\mathbf{u}_n = \lambda_{\min}$. Indeed:

$$\frac{\mathbf{u}_1^\top \mathbf{A}\mathbf{u}_1}{\mathbf{u}_1^\top \mathbf{u}_1} = \frac{\mathbf{u}_1^\top (\lambda_1 \mathbf{u}_1)}{\|\mathbf{u}_1\|_2^2} = \lambda_1 \mathbf{u}_1^\top \mathbf{u}_1 = \lambda_1 = \lambda_{\max}.$$

A similar reasoning gives the result for $\lambda_{\min}$.

Example: matrix gain and the spectral norm

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ (not necessarily symmetric) and consider some vector $\mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{0}$.

We would be interested in bounds on the gain $\frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$.
That is, how much $\mathbf{A}$ can scale (either up or down) a given vector.

Applying Rayleigh's theorem w.r.t. the matrix $\mathbf{A}^\top \mathbf{A} \in \mathbb{S}^n$ we have

$$\forall \mathbf{x} \neq \mathbf{0}: \quad \lambda_{\min}(\mathbf{A}^\top \mathbf{A}) \leq \frac{\mathbf{x}^\top \mathbf{A}^\top \mathbf{A}\mathbf{x}}{\|\mathbf{x}\|_2^2} \leq \lambda_{\max}(\mathbf{A}^\top \mathbf{A}).$$

Since $\mathbf{x}^\top \mathbf{A}^\top \mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|_2^2$ we get

$$\forall \mathbf{x} \neq \mathbf{0}: \qquad \sqrt{\lambda_{\min}(\mathbf{A}^\top \mathbf{A})} \leq \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \sqrt{\lambda_{\max}(\mathbf{A}^\top \mathbf{A})}.$$

Example: matrix gain and the spectral norm

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ (not necessarily symmetric) and consider some vector $\mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{0}$.

We would be interested in bounds on the gain $\frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$.

We saw that

$$\forall \mathbf{x} \neq \mathbf{0}: \qquad \sqrt{\lambda_{\min}(\mathbf{A}^\top \mathbf{A})} \leq \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \sqrt{\lambda_{\max}(\mathbf{A}^\top \mathbf{A})}.$$

Note also that by Rayleigh's theorem we also have that

$$\lambda_{\max}(\mathbf{A}^\top \mathbf{A}) = \max_{\mathbf{x}:\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}^\top \mathbf{A}\mathbf{x} = \max_{\mathbf{x}:\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2^2.$$

Recall our definition of the matrix norm $\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2$.

Now we see that indeed $\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}^\top \mathbf{A})}$.

This is why $\|\mathbf{A}\|_2$ is called the spectral norm.

Variational Characterization of Intermediate Eigenvalues

Rayleigh's theorem gives characterization of $\lambda_{\min}, \lambda_{\max}$ as optimization problems. We will now be interested in results of similar flavor also for intermediate eigenvalues $\lambda_2, \ldots, \lambda_{n-1}$.

Theorem (Poincaré inequality)

Let $\mathbf{A} \in \mathbb{S}^n$ and let $\mathcal{V}$ be any k-dimensional subspace of $\mathbb{R}^n$, $1 \leq k \leq n$. Then, there exist vectors $\mathbf{x}, \mathbf{y} \in \mathcal{V}$, with $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1$, such that

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} \leq \lambda_k(\mathbf{A}), \quad \mathbf{y}^\top \mathbf{A}\mathbf{y} \geq \lambda_{n-k+1}(\mathbf{A}).$$

Proof of Poincaré Inequality

Fix some $\mathcal{V}$ - a k-dimensional subspace $\mathbb{R}^n$.
We first prove that: $\exists \mathbf{x} \in \mathcal{V}, \|\mathbf{x}\|_2 = 1$: $\mathbf{x}^\top \mathbf{A}\mathbf{x} \leq \lambda_k(\mathbf{A})$.

Let $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^\top$ be the eigen-decomposition of $\mathbf{A}$ and let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ denote the eigenvectors (in order or eigenvalues $\lambda_1 \geq \lambda_2 \ldots$).

Denote $\mathcal{Q} = \mathcal{R}(\mathbf{U}_k) = \text{span}(\mathbf{u}_k, \ldots, \mathbf{u}_n)$, where $\mathbf{U}_k = (\mathbf{u}_k \cdots \mathbf{u}_n)$.

Since $\mathcal{Q}$ has dimension $n - k + 1$, the intersection $\mathcal{V} \cap \mathcal{Q}$ must be nonempty and of dimension at least 1 (since otherwise the direct sum $\mathcal{Q} \oplus \mathcal{V} \subseteq \mathbb{R}^n$ would have a dimension larger than n).

Take a unit-norm vector $\mathbf{x} \in \mathcal{V} \cap \mathcal{Q}$.

Since $\mathbf{x} \in \mathcal{Q}$, we can write $\mathbf{x} = \mathbf{U}_k\mathbf{z}$, for some $\mathbf{z} \in \mathbb{R}^{n-k+1}$ with $\|\mathbf{z}\|_2 = 1$ (since $\|\mathbf{U}_k\mathbf{z}\|_2^2 = \mathbf{z}^\top\mathbf{U}_k^\top\mathbf{U}_k\mathbf{z} = \mathbf{z}^\top\mathbf{I}_{n-k+1}\mathbf{z} = \|\mathbf{z}\|_2^2$).

Proof of Poincaré Inequality

Fix some $\mathcal{V}$ - a k-dimensional subspace $\mathbb{R}^n$.
We first prove that: $\exists \mathbf{x} \in \mathcal{V}, \|\mathbf{x}\|_2 = 1$: $\mathbf{x}^\top \mathbf{A}\mathbf{x} \leq \lambda_k(\mathbf{A})$.

Let $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^\top$ be the eigen-decomposition of $\mathbf{A}$ and let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ denote the eigenvectors (in order or eigenvalues $\lambda_1 \geq \lambda_2 \ldots$).

Denote $\mathcal{Q} = \mathcal{R}(\mathbf{U}_k) = \text{span}(\mathbf{u}_k, \ldots, \mathbf{u}_n)$, where $\mathbf{U}_k = (\mathbf{u}_k \cdots \mathbf{u}_n)$.

Take a unit-norm vector $\mathbf{x} \in \mathcal{V} \cap \mathcal{Q}$ and write it as $\mathbf{x} = \mathbf{U}_k\mathbf{z}$, for some $\mathbf{z} \in \mathbb{R}^{n-k+1}$ with $\|\mathbf{z}\|_2 = 1$.

Let us denote $\mathbf{U}_k^\perp = (\mathbf{u}_1, \ldots, \mathbf{u}_{k-1})$. Note that $\mathbf{U} = (\mathbf{U}_k^\perp \ \mathbf{U}_k)$. Hence

$$\mathbf{U}_k^\top\mathbf{U} = \mathbf{U}_k^\top(\mathbf{U}_k^\perp \ \mathbf{U}_k) = (\mathbf{0}_{n-k+1\times k-1} \ \mathbf{I}_{n-k+1}).$$

Thus, the the first part of the theorem follows since:

$$\mathbf{x}^\top\mathbf{A}\mathbf{x} = \mathbf{z}^\top\mathbf{U}_k^\top\mathbf{U}\Lambda\mathbf{U}^\top\mathbf{U}_k\mathbf{z} = \mathbf{z}^\top(\mathbf{0}, \mathbf{I})\Lambda(\mathbf{z}^\top(\mathbf{0}, \mathbf{I}))^\top = (\mathbf{0} \ \mathbf{z}^\top)\Lambda(\mathbf{0} \ \mathbf{z}^\top)^\top$$
$$= \sum_{i=k}^{n} \lambda_i(\mathbf{A})\mathbf{z}_i^2 \leq \lambda_k(\mathbf{A}) \sum_{i=k}^{n} \mathbf{z}_i^2 = \lambda_k(\mathbf{A}) = \lambda_k(\mathbf{A})\|\mathbf{z}\|_2^2 = \lambda_k(\mathbf{A}).$$

Proof of Poincaré Inequality

Fix some $\mathcal{V}$ - a k-dimensional subspace $\mathbb{R}^n$.
We have proved the first part of the theorem: $\exists \mathbf{x} \in \mathcal{V}, \|\mathbf{x}\|_2 = 1$:
$\mathbf{x}^\top \mathbf{A}\mathbf{x} \leq \lambda_k(\mathbf{A})$.

It remains to prove that $\exists \mathbf{y} \in \mathcal{V}, \|\mathbf{y}\|_2 = 1$: $\mathbf{y}^\top \mathbf{A}\mathbf{y} \geq \lambda_{n-k+1}(\mathbf{A})$

Note that the eigenvalues of $-\mathbf{A}$ are (in non-increasing order),
$-\lambda_n(\mathbf{A}) \geq \lambda_{n-1}(\mathbf{A}) \geq \cdots \geq -\lambda_1(\mathbf{A})$
(since for all $i \in [n]$, $(-\mathbf{A})\mathbf{u}_i = (-\lambda_i(\mathbf{A}))\mathbf{u}$).

Now, using the first part of the theorem with respect to the matrix $-\mathbf{A}$, we have that there is a unit vector $\mathbf{y}$ such that

$$\mathbf{y}^\top(-\mathbf{A})\mathbf{y} \leq \lambda_k(-\mathbf{A}) = -\lambda_{n-k+1}(\mathbf{A}).$$

Thus, by re-arranigng the above, the result follows.

The Minimax Principle

Poincaré inequality leads to the following corollary which is also known as *variational characterization* of the eigenvalues.

Corollary (Minimax principle)

Let $\mathbf{A} \in \mathbb{S}^n$ and let $\mathcal{V}$ denote a subspace of $\mathbb{R}^n$. Then, for $k \in \{1, \ldots, n\}$ it holds that

$$\begin{aligned} \lambda_k(\mathbf{A}) &= \max_{\mathcal{V}:\dim \mathcal{V}=k} \min_{\mathbf{x}\in\mathcal{V}, \|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x} \\ &= \min_{\mathcal{V}:\dim \mathcal{V}=n-k+1} \max_{\mathbf{x}\in\mathcal{V}, \|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x}. \end{aligned}$$

Proof of the minimax principle

First part: prove $\min_{\mathcal{V}:\dim \mathcal{V}=n-k+1} \max_{\mathbf{x}\in\mathcal{V},\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x} = \lambda_k(\mathbf{A})$.

Recall that from Poincaré inequality we know that for any subspace $\mathcal{V}$ of dim. $n-k+1$ there exists $\mathbf{x}\in\mathcal{V}, \|\mathbf{x}\|_2=1$ such that $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq \lambda_k(\mathbf{A})$.

Thus, it follows that $\min_{\mathcal{V}:\dim \mathcal{V}=n-k+1} \max_{\mathbf{x}\in\mathcal{V},\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x} \geq \lambda_k(\mathbf{A})$.

Thus, to prove the first part of the corollary it suffices to find **some subspace** $\mathcal{V}$ for which $\max_{\mathbf{x}\in\mathcal{V},\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x} = \lambda_k(\mathbf{A})$.

Let $\lambda_1,\ldots,\lambda_n$ denote the eigenvalues of $\mathbf{A}$ (in non-increasing order) and let $\mathbf{v}_1,\ldots,\mathbf{v}_n$ be the corresponding (unit) eigenvectors.

Let $\mathcal{V}$ denote the $n-k+1$ dimensional space spanned by $\mathbf{v}_k,\ldots,\mathbf{v}_n$. Now, given some unit vector $\mathbf{v}\in\mathcal{V}$ we can write it as a linear combination of $\mathbf{v}_k,\ldots,\mathbf{v}_n$: $\mathbf{v}=\sum_{i=k}^n \alpha_i \mathbf{v}_i$. Since $\mathbf{v}$ is of unit norm, we have that

$$1 = \|\mathbf{v}\|_2^2 = (\sum_{i=k}^n \alpha_i \mathbf{v}_i)^\top (\sum_{j=k}^n \alpha_j \mathbf{v}_j) = \sum_{i,j} \alpha_i \alpha_j \mathbf{v}_i^\top \mathbf{v}_j = \sum_{i=k}^n \alpha_i^2 \|\mathbf{v}_i\|_2^2 = \sum_{i=k}^n \alpha_i^2.$$

Proof of the minimax principle

Let $\mathcal{V}$ denote the $n-k+1$ dimensional space spanned by $\mathbf{v}_k,\ldots,\mathbf{v}_n$.

It suffices to show that $\max_{\mathbf{x}\in\mathcal{V},\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x} = \lambda_k(\mathbf{A})$.

Given some unit vector $\mathbf{v}\in\mathcal{V}$ we can write it as a linear combination: $\mathbf{v}=\sum_{i=k}^n \alpha_i \mathbf{v}_i$ such that $\sum_{i=k}^n \alpha_i^2 = 1$. This gives:

$$\begin{aligned}
\mathbf{v}^\top \mathbf{A}\mathbf{v} &= (\sum_{i=k}^n \alpha_i \mathbf{v}_i)^\top \mathbf{A} (\sum_{j=k}^n \alpha_j \mathbf{v}_j) = (\sum_{i=k}^n \alpha_i \mathbf{v}_i)^\top (\sum_{j=k}^n \alpha_j \mathbf{A}\mathbf{v}_j) \\
&= (\sum_{i=k}^n \alpha_i \mathbf{v}_i)^\top (\sum_{j=k}^n \alpha_j \lambda_j \mathbf{v}_j) = \sum_{i,j} \lambda_j \alpha_i \alpha_j \mathbf{v}_i^\top \mathbf{v}_j \\
&= \sum_{i=k}^n \lambda_i \alpha_i^2 \leq \sum_{i=k}^n \lambda_k \alpha_i^2 = \lambda_k.
\end{aligned}$$

On the other-hand, we know that $\mathbf{v}_k\in\mathcal{V}$ for which we have $\mathbf{v}_k^\top \mathbf{A}\mathbf{v}_k = \mathbf{v}_k^\top \lambda_k \mathbf{v}_k = \lambda_k$. Thus, indeed $\max_{\mathbf{v}\in\mathcal{V}:\|\mathbf{v}\|_2=1} \mathbf{v}^\top \mathbf{A}\mathbf{v} = \lambda_k(\mathbf{A})$.

Proof of the minimax principle

We proved the 1st equality:
$\min_{\mathcal{V}:\dim \mathcal{V}=n-k+1} \max_{\mathbf{x}\in\mathcal{V},\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x} = \lambda_k(\mathbf{A})$.

It remains to prove the 2nd equality:
$\max_{\mathcal{V}:\dim \mathcal{V}=k} \min_{\mathbf{x}\in\mathcal{V},\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x} \leq \lambda_k(\mathbf{A})$.

Using Poincaré inequality again we have that any subspace $\mathcal{V}$ of dimension k it holds that $\min_{\mathbf{x}\in\mathcal{V},\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x} \leq \lambda_k(\mathbf{A})$.

Using the same arguments as for the 1st equality, we can show that equality is obtained for $\mathcal{V} = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$.

Weyl's Inequality

An important consequence of the minimax property is the following result comparing the ordered eigenvalues of matrices $\mathbf{A}, \mathbf{B}$ with those of $\mathbf{A}+\mathbf{B}$.

Corollary (Weyl's inequality)

Let $\mathbf{A}, \mathbf{B} \in \mathbb{S}^n$. Then, for each $k = 1, \ldots, n$, we have

$$\lambda_k(\mathbf{A}) + \lambda_{\min}(\mathbf{B}) \leq \lambda_k(\mathbf{A}+\mathbf{B}) \leq \lambda_k(\mathbf{A}) + \lambda_{\max}(\mathbf{B}).$$

Proof:
From the minimax property and Rayliegh's theorem we have that

$$\begin{aligned}
\lambda_k(\mathbf{A}+\mathbf{B}) &= \min_{\mathcal{V}:\dim \mathcal{V}=n-k+1} \max_{\mathbf{x}\in\mathcal{V},\|\mathbf{x}\|_2=1} (\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{x}^\top \mathbf{B}\mathbf{x}) \\
&\geq \min_{\mathcal{V}:\dim \mathcal{V}=n-k+1} \max_{\mathbf{x}\in\mathcal{V},\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x} + \lambda_{\min}(\mathbf{B}) \\
&= \lambda_k(\mathbf{A}) + \lambda_{\min}(\mathbf{B}),
\end{aligned}$$

which proves the first inequality.

Weyl's Inequality

Corollary (Weyl's inequality)

Let $\mathbf{A}, \mathbf{B} \in \mathbb{S}^n$. Then, for each $k = 1, \ldots, n$, we have

$$\lambda_k(\mathbf{A}) + \lambda_{\min}(\mathbf{B}) \leq \lambda_k(\mathbf{A} + \mathbf{B}) \leq \lambda_k(\mathbf{A}) + \lambda_{\max}(\mathbf{B}).$$

Proof cont.:
The second inequality follows similarly since

$$\begin{aligned}
\lambda_k(\mathbf{A}+\mathbf{B}) &= \min_{\mathcal{V}:\dim\mathcal{V}=n-k+1} \max_{\mathbf{x}\in\mathcal{V},\|\mathbf{x}\|_2=1} (\mathbf{x}^\top\mathbf{A}\mathbf{x} + \mathbf{x}^\top\mathbf{B}\mathbf{x}) \\
&\leq \min_{\mathcal{V}:\dim\mathcal{V}=n-k+1} \max_{\mathbf{x}\in\mathcal{V},\|\mathbf{x}\|_2=1} \mathbf{x}^\top\mathbf{A}\mathbf{x} + \lambda_{\max}(\mathbf{B}) \\
&= \lambda_k(\mathbf{A}) + \lambda_{\max}(\mathbf{B}).
\end{aligned}$$

Example: estimating eigenvalues from noisy matrix

Suppose we observe a matrix $\mathbf{M} = \mathbf{D} + \mathbf{N}$, such that $\mathbf{M}, \mathbf{D}, \mathbf{N} \in \mathbb{S}^n$.
Here $\mathbf{D}$ is a matrix representing some data and $\mathbf{N}$ is some additive noise.
Note that only $\mathbf{M}$ (the sum) is observed.
Suppose we would like to estimate the eigenvalues of the data $\mathbf{D}$ based on the noisy observation $\mathbf{M}$.
How accurate are these estimates?

Using Weyl's inequality with $\mathbf{A} = \mathbf{M}$ and $\mathbf{B} = -\mathbf{N}$, we have for all $k \in \{1, \ldots, n\}$:

$$\lambda_k(\mathbf{D}) \leq \lambda_k(\mathbf{M}) + \lambda_{\max}(-\mathbf{N}), \quad \lambda_k(\mathbf{D}) \geq \lambda_k(\mathbf{M}) + \lambda_{\min}(-\mathbf{N}).$$

Thus,

$$\forall k \in \{1, \ldots, n\}: \quad \lambda_k(\mathbf{M}) - \lambda_{\max}(\mathbf{N}) \leq \lambda_k(\mathbf{D}) \leq \lambda_k(\mathbf{M}) - \lambda_{\min}(\mathbf{N}).$$

More such results and end of course.

Positive Semidefinite Matrices

Definition

A symmetric matrix $\mathbf{A} \in \mathbb{S}^n$ is said to be *positive semidefinite* (PSD) if the associated quadratic form is non-negative, i.e.,

$$\forall \mathbf{x} \in \mathbb{R}^n : \quad \mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0,$$

and we use the notation $\mathbf{A} \succeq 0$.
If, moreover, $\forall \mathbf{x} \neq \mathbf{0} : \quad \mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$, then $\mathbf{A}$ is said to be *positive definite*, which is denoted by $\mathbf{A} \succ 0$.
We say $\mathbf{A}$ is *negative semidefinite* and we write $\mathbf{A} \preceq 0$ if $-\mathbf{A} \succeq 0$. We say $\mathbf{A}$ is negative semidefinite, and we write $\mathbf{A} \prec 0$ if $-A \succ 0$.
Finally, given two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{S}^n$, we say $\mathbf{A} \succeq \mathbf{B}$ $(\mathbf{A} \succ \mathbf{B})$ if $\mathbf{A} - \mathbf{B} \succeq 0$ $(\mathbf{A} - \mathbf{B} \succ 0)$.

We denote the set of all positive semidefinite matrices by $\mathbb{S}^n_+$ and the set of all positive definite matrices by $\mathbb{S}^n_{++}$.

Eigenvalues of Positive Semidefinite Matrices

Lemma

Let $\mathbf{A} \in \mathbb{S}$. Then

$$\begin{aligned} \mathbf{A} \succeq 0 \quad &\Leftrightarrow \quad \lambda_i(\mathbf{A}) \geq 0, \ i = 1, \ldots, n, \\ \mathbf{A} \succ 0 \quad &\Leftrightarrow \quad \lambda_i(\mathbf{A}) > 0, \ i = 1, \ldots, n. \end{aligned}$$

Proof: To prove the first statement (the second is proven in a similar manner), let $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^\top$ be the eigen-decomposition of $\mathbf{A}$.
Given $\mathbf{x} \in \mathbb{R}^n$ and denoting $\mathbf{z} = \mathbf{U}^\top \mathbf{x}$, we have that

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \mathbf{x}^\top \mathbf{U}\Lambda\mathbf{U}^\top \mathbf{x} = \mathbf{z}^\top \Lambda \mathbf{z} = \sum_{i=1}^n \lambda_i(\mathbf{A})\mathbf{z}_i^2.$$

Now, using the relationship $\mathbf{z} = \mathbf{U}^\top \mathbf{x}$ and $\mathbf{x} = \mathbf{U}\mathbf{z}$ we have that

$$\forall \mathbf{x} \in \mathbb{R}^n : \ \mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0 \quad \Leftrightarrow \quad \forall \mathbf{z} \in \mathbb{R}^n : \ \mathbf{z}^\top \Lambda \mathbf{z} \geq 0,$$

and the latter condition is clearly equivalent to $\lambda_i(\mathbf{A}) \geq 0$, $i = 1, \ldots, n$.

Eigenvalues of Positive Semidefinite Matrices

The following lemmas are not difficult to prove.

Lemma (HW)

Let $\mathbf{A} \in \mathbb{S}^n_+$ and let $\mathbf{B} \in \mathbb{S}^n$. Then

$$\mathbf{B} \succeq 0 \;\Rightarrow\; \lambda_k(\mathbf{A}+\mathbf{B}) \geq \lambda_k(\mathbf{A}),\; k=1,\ldots,n.$$

Lemma (HW)

Let $\mathbf{A}, \mathbf{B} \in \mathbb{S}^n$ such that $\mathbf{A} \succeq \mathbf{B}$. Then

$$\forall i = 1,\ldots,n: \quad \lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B}).$$

Matrix-induced inner-products and norms

Given a matrix $\mathbf{A} \in \mathbb{S}^n$ we define the following function on $\mathbb{R}^n \times \mathbb{R}^n$ and $\mathbb{R}^n$:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} := \mathbf{x}^\top \mathbf{A} \mathbf{y}, \qquad \|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^\top \mathbf{A} \mathbf{x}}.$$

Theorem (HW)

if $\mathbf{A} \succ 0$ then $\langle \cdot, \cdot \rangle_{\mathbf{A}}$ is an inner-product over $\mathbb{R}^n$ and $\|\cdot\|_{\mathbf{A}}$ is a norm over $\mathbb{R}^n$.
If $\mathbf{A} \succeq 0$ then $\|\cdot\|_{\mathbf{A}}$ is a pseudo-norm, i.e., it satisfies all the properties of a norm except that there exists $\mathbf{x} \neq \mathbf{0}$ such that $\|\mathbf{x}\|_{\mathbf{A}} = 0$.

Algebraic Methods in Data Science: Lesson 5

Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology

Dan Garber
https://dangar.net.technion.ac.il/

Winter Semester 2020-2021

Recap: The Spectral Theorem for Symmetric Matrices

Theorem (Spectral theorem)

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be symmetric, let $\lambda_i \in \mathbb{R}$, $i = 1, \ldots, n$, be the eigenvalues of $\mathbf{A}$ (counting multiplicities). Then, there exists a set of orthonormal vectors $\mathbf{u}_i \in \mathbb{R}^n$, $i = 1, \ldots, n$, such that $\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i$. Equivalently, there exists an orthogonal matrix $\mathbf{U} = [\mathbf{u}_1 \cdots \mathbf{u}_n]$ (i.e., $\mathbf{U}\mathbf{U}^\top = \mathbf{U}^\top \mathbf{U} = \mathbf{I}_n$) such that $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^\top = \sum_{i=1}^n \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$, $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$

Positive Semidefinite Matrices

Definition

A symmetric matrix $\mathbf{A} \in \mathbb{S}^n$ is said to be *positive semidefinite* (PSD) if the associated quadratic form is non-negative, i.e.,

$$\forall \mathbf{x} \in \mathbb{R}^n : \quad \mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0,$$

and we use the notation $\mathbf{A} \succeq 0$.
If, moreover, $\forall \mathbf{x} \neq \mathbf{0} : \quad \mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$, then $\mathbf{A}$ is said to be *positive definite*, which is denoted by $\mathbf{A} \succ 0$.
We say $\mathbf{A}$ is *negative semidefinite* and we write $\mathbf{A} \preceq 0$ if $-\mathbf{A} \succeq 0$.
We say $\mathbf{A}$ is negative definite, and we write $\mathbf{A} \prec 0$ if $-\mathbf{A} \succ 0$.
Finally, given two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{S}^n$, we say $\mathbf{A} \succeq \mathbf{B}$ $(\mathbf{A} \succ \mathbf{B})$ if $\mathbf{A} - \mathbf{B} \succeq 0$ $(\mathbf{A} - \mathbf{B} \succ 0)$.

We denote the set of all positive semidefinite matrices by $\mathbb{S}^n_+$ and the set of all positive definite matrices by $\mathbb{S}^n_{++}$.

Eigenvalues of Positive Semidefinite Matrices

Lemma

Let $\mathbf{A} \in \mathbb{S}$. Then

$$\mathbf{A} \succeq 0 \quad \Leftrightarrow \quad \lambda_i(\mathbf{A}) \geq 0, \ i = 1, \ldots, n,$$
$$\mathbf{A} \succ 0 \quad \Leftrightarrow \quad \lambda_i(\mathbf{A}) > 0, \ i = 1, \ldots, n.$$

Proof: To prove the first statement (the second is proven in a similar manner), let $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^\top$ be the eigen-decomposition of $\mathbf{A}$.
Given $\mathbf{x} \in \mathbb{R}^n$ and denoting $\mathbf{z} = \mathbf{U}^\top \mathbf{x}$, we have that

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \mathbf{x}^\top \mathbf{U}\Lambda\mathbf{U}^\top \mathbf{x} = \mathbf{z}^\top \Lambda \mathbf{z} = \sum_{i=1}^n \lambda_i(\mathbf{A})\mathbf{z}_i^2.$$

Now, using the relationship $\mathbf{z} = \mathbf{U}^\top \mathbf{x}$ and $\mathbf{x} = \mathbf{U}\mathbf{z}$ we have that

$$\forall \mathbf{x} \in \mathbb{R}^n : \ \mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0 \quad \Leftrightarrow \quad \forall \mathbf{z} \in \mathbb{R}^n : \ \mathbf{z}^\top \Lambda \mathbf{z} \geq 0,$$

and the latter condition is clearly equivalent to $\lambda_i(\mathbf{A}) \geq 0$, $i = 1, \ldots, n$.

Eigenvalues of Positive Semidefinite Matrices

The following lemmas are not difficult to prove.

Lemma (HW)

Let $\mathbf{A} \in \mathbb{S}^n_+$ and let $\mathbf{B} \in \mathbb{S}^n$. Then

$$\mathbf{B} \succeq 0 \;\Rightarrow\; \lambda_k(\mathbf{A}+\mathbf{B}) \geq \lambda_k(\mathbf{A}),\; k = 1, \ldots, n.$$

Lemma (HW)

Let $\mathbf{A}, \mathbf{B} \in \mathbb{S}^n$ such that $\mathbf{A} \succeq \mathbf{B}$. Then

$$\forall i = 1, \ldots, n: \quad \lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B}).$$

Matrix-induced inner-products and norms

Given a matrix $\mathbf{A} \in \mathbb{S}^n$ we define the following functions on $\mathbb{R}^n \times \mathbb{R}^n$ and $\mathbb{R}^n$:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} := \mathbf{x}^\top \mathbf{A} \mathbf{y}, \qquad \|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^\top \mathbf{A} \mathbf{x}}.$$

Theorem (HW)

if $\mathbf{A} \succ 0$ then $\langle \cdot, \cdot \rangle_{\mathbf{A}}$ is an inner-product over $\mathbb{R}^n$ and $\|\cdot\|_{\mathbf{A}}$ is a norm over $\mathbb{R}^n$.
If $\mathbf{A} \succeq 0$ then $\|\cdot\|_{\mathbf{A}}$ is a pseudo-norm, i.e., it satisfies all the properties of a norm except that there exists $\mathbf{x} \neq \mathbf{0}$ such that $\|\mathbf{x}\|_{\mathbf{A}} = 0$.

The Singular Value Decomposition

We have seen that symmetric matrices always admit an eigen/spectral decomposition and we have already seen some examples for the great importance of this decomposition.

However, what about matrices which are not symmetric and even not rectangular? Can we obtain a similar result with similar consequences?

The answer is yes and it is called the **Singular Value Decomposition** (SVD) and it is (arguably) one of the most important algorithmic / mathematical tools that you will learn throughout your studies.

The Singular Value Decomposition

Theorem (SVD decomposition)

Any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ *can we factored as*

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$$

where $\mathbf{V} \in \mathbb{R}^{n \times n}$, $\mathbf{U} \in \mathbb{R}^{m \times m}$ *are* **orthogonal matrices**, *and* $\Sigma \in \mathbb{R}^{m \times n}$ *is a* **diagonal matri***x with the first* $r = rank(\mathbf{A})$ *diagonal entries* $(\sigma_1, \ldots, \sigma_r)$ *positive and non-increasing, and all other diagonal entries are zero.*
The values along the main diagonal of Σ *are called the singular values of* $\mathbf{A}$. *The columns of* $\mathbf{U}$ *are called the left singular vectors, and the columns of* $\mathbf{V}$ *are called the right singular vectors.*

Proof of the SVD Theorem $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$

Roadmap to proof:

1. $\mathbf{V}$ - matrix of right singular vectors is going to be simply the eignevectors (as columns) of the symmetric matrix $\mathbf{A}^\top\mathbf{A}$.

2. Σ - diagonal matrix of (non-negative) singular values will be: $\Sigma_{i,i} = \sqrt{\lambda_i(\mathbf{A}^\top\mathbf{A})}$.

   Note that using the SVD we indeed have $\mathbf{A}^\top\mathbf{A} = \mathbf{V}\Sigma\mathbf{U}^\top\mathbf{U}\Sigma\mathbf{V}^\top = \mathbf{V}\Sigma^2\mathbf{V}^\top$ - the spectral decomposition of $\mathbf{A}^\top\mathbf{A}$.

3. $\mathbf{U}$ - matrix of left singular vectors is going to be the eigenvectors (as columns) of the symmetric matrix $\mathbf{A}\mathbf{A}^\top$,**but constructed in a specialized way** so that the desired relation $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$ holds.

Proof of the SVD Theorem $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$

Consider the symmetric and positive semidefinite matrix $\mathbf{A}^\top\mathbf{A}$ and its spectral decomposition $\mathbf{A}^\top\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^\top$. This $\mathbf{V}$ is going to be the also be the $\mathbf{V}$ in the SVD of $\mathbf{A}$ - that is, the eigenvectors of $\mathbf{A}^\top\mathbf{A}$ are going to be the right singular vectors of the matrix $\mathbf{A}$.

Denote $\lambda_i = \lambda_i(\mathbf{A}^\top\mathbf{A})$ and recall that $\lambda_i \geq 0, i = 1, \ldots, n$.

We first observe that $\mathrm{rank}(\mathbf{A}^\top\mathbf{A}) = \mathrm{rank}(\mathbf{A}) = r$.
To recall why this is true, note it suffices to show that they have the same nullspace, i.e., $\mathrm{Ker}(\mathbf{A}) = \mathrm{Ker}(\mathbf{A}^\top\mathbf{A})$ (recall $\dim\mathcal{N}(\mathbf{A}) + \mathrm{rank}(\mathbf{A}) = n$).
Indeed this follows since

$$\mathbf{x} \in \mathrm{Ker}(\mathbf{A}) \Rightarrow \mathbf{A}\mathbf{x} = \mathbf{0} \Rightarrow \mathbf{A}^\top\mathbf{A}\mathbf{x} = \mathbf{0} \Rightarrow \mathbf{x} \in \mathrm{Ker}(\mathbf{A}^\top\mathbf{A}).$$

$$\begin{aligned}\mathbf{x} \in \mathrm{Ker}(\mathbf{A}^\top\mathbf{A}) \quad &\Rightarrow \mathbf{A}^\top\mathbf{A}\mathbf{x} = \mathbf{0} \Rightarrow \mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x} = 0 \\ &\Rightarrow \|\mathbf{A}\mathbf{x}\|_2^2 = 0 \Rightarrow \mathbf{x} \in \mathrm{Ker}(\mathbf{A}).\end{aligned}$$

Thus, the first r eigenvalues of $\mathbf{A}^\top\mathbf{A}$, $\lambda_1, \ldots, \lambda_r$ are strictly positive and all others are zero (recall rank = number of non-zero eigenvalues).

Proof of the SVD Theorem $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$

Recall $\mathbf{A}^\top\mathbf{A} = \mathbf{V}^\top\Lambda\mathbf{V}^\top$. By the same argument as in previous slide, we have that $\mathrm{rank}(\mathbf{A}\mathbf{A}^\top) = \mathrm{rank}(\mathbf{A}^\top) = \mathrm{rank}(\mathbf{A}) = r$. That is, $\mathbf{A}\mathbf{A}^\top$ also has exactly r positive eigenvalues.

Now, denote by $\mathbf{v}_1, \ldots, \mathbf{v}_r$ the first r columns of $\mathbf{V}$, i.e., the eigenvectors of $\mathbf{A}^\top\mathbf{A}$ associated with $\lambda_1, \ldots, \lambda_r > 0$.

By definition $\mathbf{A}^\top\mathbf{A}\mathbf{v}_i = \lambda_i\mathbf{v}_i, \quad i = 1, \ldots, r.$

Multiplying both sides by $\mathbf{A}$ we have $(\mathbf{A}\mathbf{A}^\top)\mathbf{A}\mathbf{v}_i = \lambda_i\mathbf{A}\mathbf{v}_i, i = 1, \ldots, r.$

This implies that $\mathbf{A}^\top\mathbf{A}, \mathbf{A}\mathbf{A}^\top$ have the same r positive eigenvalues, and that $\mathbf{A}\mathbf{v}_i, i = 1, \ldots, r$, are the corresponding eigenvectors of $\mathbf{A}\mathbf{A}^\top$.

In particular, for all $i \neq j$ we have

$$(\mathbf{A}\mathbf{v}_i)^\top(\mathbf{A}\mathbf{v}_j) = \mathbf{v}_i^\top\mathbf{A}^\top\mathbf{A}\mathbf{v}_j = \mathbf{v}_i^\top\lambda_j\mathbf{v}_j = 0.$$

Hence, these r eigenvectors of $\mathbf{A}\mathbf{A}^\top$ that we have just found are mutually orthogonal.

Proof of the SVD Theorem $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$

Set the r positive singular values of $\mathbf{A}$ (positive entries of matrix Σ in decomposition) as: $\sigma_i := \sqrt{\lambda_i(\mathbf{A}^\top\mathbf{A})} = \sqrt{\lambda_i(\mathbf{A}\mathbf{A}^\top)} > 0, i = 1, \ldots, r.$

Since for all $i = 1, \ldots, r$ we have $\|\mathbf{A}\mathbf{v}_i\|_2^2 = \mathbf{v}_i^\top\mathbf{A}^\top\mathbf{A}\mathbf{v}_i = \lambda_i = \sigma_i^2$, we define the corresponding unit-norm normalized eigenvectors of $\mathbf{A}\mathbf{A}^\top$ as

$$\mathbf{u}_i = \frac{\mathbf{A}\mathbf{v}_i}{\sqrt{\lambda_i}} = \frac{\mathbf{A}\mathbf{v}_i}{\sigma_i}, i = 1, \ldots, r.$$

Observe that for all $i, j = 1, \ldots, r$ we have

$$\mathbf{u}_i^\top\mathbf{A}\mathbf{v}_j = \frac{1}{\sigma_i}\mathbf{v}_i^\top\mathbf{A}^\top\mathbf{A}\mathbf{v}_j = \frac{\lambda_j}{\sigma_i}\mathbf{v}_i^\top\mathbf{v}_j = \frac{\sigma_j^2}{\sigma_i}\mathbf{v}_i^\top\mathbf{v}_j = \begin{cases} \sigma_i & \text{if } i = j \\ 0 & otherwise. \end{cases}$$

Denoting $\mathbf{U}_r = (\mathbf{u}_1, \ldots, \mathbf{u}_r)$, $\mathbf{V}_r = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$, $\Sigma_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$, this can be written in matrix form: $\mathbf{U}_r^\top\mathbf{A}\mathbf{V}_r = \Sigma_r$.

This is already the SVD in its "compact" form.

Proof of the SVD Theorem $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$

We now derive the "full"-version SVD. For this purpose we need to complete the set of vectors $\mathbf{u}_1, \ldots, \mathbf{u}_r$ into an orthonormal basis for $\mathbb{R}^m$. Recall $\mathrm{rank}(\mathbf{A}^\top\mathbf{A}) = \mathrm{rank}(\mathbf{A}) = r$. Thus,

$$\mathbf{A}^\top\mathbf{A}\mathbf{v}_i = 0, \ i = r+1, \ldots, n,$$

and since, as we have seen $\mathrm{Ker}(\mathbf{A}^\top\mathbf{A}) = \mathrm{Ker}(\mathbf{A})$, this implies that

$$\mathbf{A}\mathbf{v}_i = 0, \ i = r+1, \ldots, n.$$

Thus, we can find orthonormal vectors $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m$ such that $\mathbf{u}_1, \ldots, \mathbf{u}_m$ is an orthonormal basis for $\mathbb{R}^m$ (any completion of $\mathbf{u}_1, \ldots, \mathbf{u}_r$ to an orthonormal basis will work), and

$$\mathbf{u}_i^\top\mathbf{A}\mathbf{v}_j = \mathbf{u}_i^\top\mathbf{0} = 0, \ i = 1, \ldots, m; \ j = r+1, \ldots, n.$$

Also, recalling that for all $j \in \{1, \ldots, r\}$ $\mathbf{u}_j = \frac{\mathbf{A}\mathbf{v}_j}{\sqrt{\lambda_j}}$, we have that

$$\forall i \in \{r+1, \ldots, m\}, j \in \{1, \ldots, r\}: \quad \mathbf{u}_i^\top\mathbf{A}\mathbf{v}_j = \mathbf{u}_i^\top\sqrt{\lambda_j}\mathbf{u}_j = 0.$$

Proof of the SVD Theorem $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$

To conclude, we have found orthogonal $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n) \in \mathbb{R}^{n\times n}$, $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_m) \in \mathbb{R}^{m\times m}$ and diagonal $\Sigma \in \mathbb{R}^{m\times n}$ with non-negative entries such that

$$\mathbf{u}_i^\top\mathbf{A}\mathbf{v}_j = \begin{cases} \sigma_i > 0 & \text{if } i = j, i \le \mathrm{rank}(\mathbf{A}) \\ 0 & otherwise. \end{cases}$$

Therefore, we can write in matrix form

$$\begin{pmatrix} \mathbf{u}_1^\top \\ \vdots \\ \mathbf{u}_m^\top \end{pmatrix} \mathbf{A} \begin{pmatrix} \mathbf{v}_1 \ldots \mathbf{v}_m \end{pmatrix} = \begin{pmatrix} \Sigma_r & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix} = \Sigma,$$

were $\Sigma_r \in \mathbb{R}^{r\times r}, \Sigma_r(i,i) = \sigma_i$, $r = \mathrm{rank}(\mathbf{A})$.
Multiplying the right hand side above by $\mathbf{U}$ from the left and by $\mathbf{V}^\top$ form the right, we obtain the SVD theorem.

Compact form of the Singular Value Decomposition

Corollary (compact-form SVD)

Any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ *can be written as*

$$\mathbf{A} = \mathbf{U}_r \Sigma_r \mathbf{V}_r^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

where $r = rank(\mathbf{A})$, $\mathbf{U}_r = (\mathbf{u}_1, \ldots, \mathbf{u}_r) \in \mathbb{R}^{m \times r}$ *has orthonormal columns,* $\mathbf{V}_r = (\mathbf{v}_1, \ldots, \mathbf{v}_r) \in \mathbb{R}^{n \times r}$ *has orthonormal columns, and* $\sigma_1 \geq \sigma_2 \geq ...\sigma_r > 0$. *The positive numbers* $\sigma_1, \ldots, \sigma_r$ *are called the singular values of* $\mathbf{A}$, *the vectors* $\mathbf{u}_i$ *are called the left singular vectors of* $\mathbf{A}$ *and the vectors* $\mathbf{v}_i$ *are called the right singular vectors. These quantities satisfy*

$$\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i, \quad \mathbf{u}_i^\top \mathbf{A} = \sigma_i \mathbf{v}_i, \quad i = 1, \ldots, r.$$

Moreover, $\sigma_i^2 = \lambda_i(\mathbf{A}\mathbf{A}^\top) = \lambda_i(\mathbf{A}^\top \mathbf{A})$, $i = 1, \ldots, r$, *and* $\mathbf{u}_i, \mathbf{v}_i$ *are eigenvectors (corresponding to non-zero eignevalues) of* $\mathbf{A}\mathbf{A}^\top$ *and* $\mathbf{A}^\top \mathbf{A}$, *respectively.*

Spectral Matrix Norms

Let us take another look at the Frobenius (Euclidean) norm which is defined as $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} \mathbf{A}_{ij}^2}$. Since the norm is induced by the standard inner product for matrices we have:

$$\|\mathbf{A}\|_F^2 = \mathrm{Tr}(\mathbf{A}^\top \mathbf{A}) = \sum_{i=1}^{n} \lambda_i(\mathbf{A}^\top \mathbf{A}) = \sum_{i=1}^{n} \sigma_i^2,$$

where σ_i are the singular values of $\mathbf{A}$. Hence, the Euclidean norm of a matrix is directly related to its singular values.

Recall also that the spectral norm (2-norm) of a matrix $\mathbf{A}$ is given by

$$\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|_2 = 1} \|\mathbf{A}\mathbf{x}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}^\top \mathbf{A})} = \sigma_1(\mathbf{A}).$$

That is, the spectral norm of a given matrix is its largest (first) singular value.

Spectral Matrix Norms

Theorem (Schatten norm)

For any $p \in [1, \infty]$, the function $\|\cdot\|_{S(p)} : \mathbb{R}^{m \times n} \to \mathbb{R}$ given by

$$\|\mathbf{A}\|_{S(p)} := \|\vec{\sigma}(\mathbf{A})\|_p = \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i(\mathbf{A})^p \right)^{1/p}$$

where σ_i are the singular values of $\mathbf{A}$, is a norm.

Note in particular that $\|\mathbf{A}\|_{S(1)} = \sum_{i=1}^{\min\{m,n\}} \sigma_i(\mathbf{A})$. Hence, $\|\cdot\|_{S(1)}$ is often-called the trace-norm.

Also, $\|\mathbf{A}\|_{S(2)} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i(\mathbf{A})^2} = \|\mathbf{A}\|_F$.

Finally, $\|\mathbf{A}\|_{S(\infty)} = \max_i \sigma_i(\mathbf{A}) = \sigma_1(\mathbf{A}) = \|\mathbf{A}\|_2$.

Rank nullspace, and image via SVD

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and denote its SVD by $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$ where $\mathbf{U} = (\mathbf{U}_r, \mathbf{U}_r^\perp), \mathbf{V} = (\mathbf{V}_r, \mathbf{V}_r^\perp)$ and

$$\Sigma = \begin{pmatrix} \Sigma_r & \mathbf{0}_{r \times (n-r)} \\ \mathbf{0}_{(m-r) \times r} & \mathbf{0}_{(m-r) \times (n-r)} \end{pmatrix},$$

where $\mathbf{U}_r \Sigma_r \mathbf{V}_r^\top$ is the compact-form SVD of $\mathbf{A}$.

Corollary

The columns of the four matrices $\mathbf{U}_r, \mathbf{V}_r, \mathbf{U}_r^\perp, \mathbf{V}_r^\perp$ form an orthonormal basis to the four subspaces $Im(\mathbf{A}), Im(\mathbf{A}^\top), \mathcal{N}(\mathbf{A}^\top), \mathcal{N}(\mathbf{A})$, respectively.

The Matrix Pseudo-Inverse

Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ of rank r with SVD $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$ where

$$\Sigma = \begin{pmatrix} \Sigma_r & \mathbf{0}_{r\times(n-r)} \\ \mathbf{0}_{(m-r)\times r} & \mathbf{0}_{(m-r)\times(n-r)} \end{pmatrix},$$

the **Moore-Penrose pseduo-inverse** (or generalized inverse) of $\mathbf{A}$ is defined as

$$\mathbf{A}^\dagger = \mathbf{V}\Sigma^\dagger\mathbf{U}^\top \in \mathbb{R}^{n\times m}$$

where

$$\Sigma^\dagger = \begin{pmatrix} \Sigma_r^{-1} & \mathbf{0}_{r\times(m-r)} \\ \mathbf{0}_{(n-r)\times r} & \mathbf{0}_{(n-r)\times(m-r)} \end{pmatrix}, \quad \Sigma_r^{-1} = \operatorname{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}\right) \succ 0.$$

It can be seen that $\mathbf{A}^\dagger$ admits the following compact-form SVD:

$$\mathbf{A}^\dagger = \mathbf{V}_r\Sigma_r^{-1}\mathbf{U}_r^\top.$$

The Matrix Pseudo-Inverse

For $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$, the pseudo-inverse is $\mathbf{A}^\dagger = \mathbf{V}\Sigma^\dagger\mathbf{U}^\top$, where

$$\Sigma^\dagger = \begin{pmatrix} \Sigma_r^{-1} & \mathbf{0}_{r\times(m-r)} \\ \mathbf{0}_{(n-r)\times r} & \mathbf{0}_{(n-r)\times(m-r)} \end{pmatrix}, \quad \Sigma_r^{-1} = \operatorname{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}\right) \succ 0.$$

Observe that by these definitions we have

$$\Sigma\Sigma^\dagger = \begin{pmatrix} \mathbf{I}_r & \mathbf{0}_{r\times(m-r)} \\ \mathbf{0}_{(m-r)\times r} & \mathbf{0}_{(m-r)\times(m-r)} \end{pmatrix}, \; \Sigma^\dagger\Sigma = \begin{pmatrix} \mathbf{I}_r & \mathbf{0}_{r\times(n-r)} \\ \mathbf{0}_{(n-r)\times r} & \mathbf{0}_{(m-r)\times(n-r)} \end{pmatrix}.$$

It is straightforward to verify that the following equalities follow:

$$\begin{aligned} \mathbf{A}\mathbf{A}^\dagger &= \mathbf{U}_r\mathbf{U}_r^\top \\ \mathbf{A}^\dagger\mathbf{A} &= \mathbf{V}_r\mathbf{V}_r^\top \\ \mathbf{A}\mathbf{A}^\dagger\mathbf{A} &= \mathbf{A} \\ \mathbf{A}^\dagger\mathbf{A}\mathbf{A}^\dagger &= \mathbf{A}^\dagger. \end{aligned}$$

The Matrix Pseudo-Inverse - 3 special cases

Case 1: If $\mathbf{A}$ is square and non-singular, then $\mathbf{A}^\dagger = \mathbf{A}^{-1}$.

Case 2: If $\mathbf{A} \in \mathbb{R}^{m\times n}$ $(n \leq m)$ is full-column rank, that is $r = n \leq m$:

$$\mathbf{A}^\dagger\mathbf{A} = \mathbf{V}_r\mathbf{V}_r^\top = \mathbf{V}\mathbf{V}^\top = \mathbf{I}_n.$$

In this case $\mathbf{A}^\dagger$ is called a **left inverse** of $\mathbf{A}$. That is, for every $\mathbf{x} \in \mathbb{R}^n$ we have that $\mathbf{A}^\dagger\mathbf{A}\mathbf{x} = \mathbf{x}$.
Moreover, in this case $\mathbf{A}^\top\mathbf{A}$ is invertible and using the SVD of $\mathbf{A}$ we have that

$$\begin{aligned}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top &= (\mathbf{V}_r\Sigma_r^\top\mathbf{U}_r^\top\mathbf{U}_r\Sigma_r\mathbf{V}_r^\top)^{-1}\mathbf{A} = (\mathbf{V}_r\Sigma_r^2\mathbf{V}_r^\top)^{-1}\mathbf{V}_r\Sigma_r\mathbf{U}_r^\top \\ &= (\mathbf{V}_r\Sigma_r^{-2}\mathbf{V}_r^\top)\mathbf{V}_r\Sigma_r\mathbf{U}_r^\top = \mathbf{V}_r\Sigma_r^{-2}\mathbf{I}_r\Sigma_r\mathbf{U}_r^\top \\ &= \mathbf{V}_r\Sigma_r^{-1}\mathbf{U}_r^\top = \mathbf{A}^\dagger.\end{aligned}$$

That is, we can express $\mathbf{A}^\dagger$ explicitly in terms of $\mathbf{A}$:

$$\mathbf{A}^\dagger = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}.$$

The Matrix Pseudo-Inverse - 3 special cases

Case 3: If $\mathbf{A} \in \mathbb{R}^{m\times n}$ $(n \geq m)$ is full-row rank, that is $r = m \leq n$, then

$$\mathbf{A}\mathbf{A}^\dagger = \mathbf{U}_r\mathbf{U}_r^\top = \mathbf{U}\mathbf{U}^\top = \mathbf{I}_m.$$

In this case, $\mathbf{A}^\dagger$ is called a **right inverse** of $\mathbf{A}$. That is, for every $\mathbf{x} \in \mathbb{R}^m$ we have that $\mathbf{x}^\top\mathbf{A}\mathbf{A}^\dagger = \mathbf{x}^\top$.

In this case $\mathbf{A}\mathbf{A}^\top$ is invertible and using the SVD of $\mathbf{A}$ we have, similarly to before that

$$\begin{aligned}\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{-1} &= \mathbf{V}_r\Sigma_r\mathbf{U}_r^\top(\mathbf{U}_r\Sigma_r\mathbf{V}_r^\top\mathbf{V}_r\Sigma_r\mathbf{U}_r^\top)^{-1} \\ &= \mathbf{V}_r\Sigma_r\mathbf{U}_r^\top(\mathbf{U}_r\Sigma_r^{-2}\mathbf{U}_r^\top) = \mathbf{V}_r\Sigma_r^{-1}\mathbf{U}_r^\top = \mathbf{A}^\dagger.\end{aligned}$$

Hence, in case $\mathbf{A}$ is full-row rank we can also express $\mathbf{A}^\dagger$ explicitly in terms of $\mathbf{A}$:

$$\mathbf{A}^\dagger = \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{-1}.$$

Application of SVD to Least-Squares

Recall that for a non-singular matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and vector $\mathbf{b} \in \mathbb{R}^n$, the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ admits a unique solution which is given by $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$.

However, what if $\mathbf{A}$ is not square, i.e., $\mathbf{A} \in \mathbb{R}^{m \times n}$ or singular?
In this case we can try and solve

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2,$$

i.e., find a solution that minimizers the error in ℓ_2 norm.
This is known as the **least-squares problem**.Note that (1) is equivalent to the following problem

$$\min_{\mathbf{z} \in \mathrm{Im}(\mathbf{A})} \|\mathbf{z} - \mathbf{b}\|_2^2.$$

However, (1) is nothing but the problem of projecting the vector $\mathbf{b}$ onto the subspace $\mathrm{Im}(\mathbf{A})$.

Application of SVD to Least-Squares

The least squares problem $\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ is equivalent to the problem $\min_{\mathbf{z} \in \mathrm{Im}(\mathbf{A})} \|\mathbf{z} - \mathbf{b}\|_2^2$ - projecting the point $\mathbf{b}$ onto the subspace $\mathrm{Im}(\mathbf{A})$.

Recall we have seen that if $\mathbf{A} = \mathbf{U}_r \Sigma_r \mathbf{V}_r^\top$ is the compact-form SVD of $\mathbf{A}$ then the columns of $\mathbf{U}_r$ are an orthonormal basis for $\mathrm{Im}(\mathbf{A})$. This implies that the matrix $\mathbf{U}_r \mathbf{U}_r^\top$ is the **projection matrix** onto the subspace $\mathrm{Im}(\mathbf{A})$.

Recall that we have also seen that $\mathbf{A}\mathbf{A}^\dagger = \mathbf{U}_r \mathbf{U}_r^\top$.

Hence, the projection of $\mathbf{b}$ onto $\mathrm{Im}(\mathbf{A})$ is given by $\mathbf{z}^* = \mathbf{A}\mathbf{A}^\dagger \mathbf{b}$.

Since an optimal solution $\mathbf{x}^*$ to to the least squares problem must satisfy $\mathbf{A}\mathbf{x}^* = \mathbf{z}^* = \mathbf{A}\mathbf{A}^\dagger \mathbf{b}$, we can in particular choose

$$\mathbf{x}^* = \mathbf{A}^\dagger \mathbf{b}.$$

Note this agrees with the solution in case $\mathbf{A}$ is square and non-singular.

Algebraic Methods in Data Science: Lesson 6

Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology

Dan Garber
https://dangar.net.technion.ac.il/

Winter Semester 2020-2021

Recap: The Singular Value Decomposition

Theorem (SVD decomposition)

Any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ *can we factored as*

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$$

where $\mathbf{V} \in \mathbb{R}^{n \times n}$, $\mathbf{U} \in \mathbb{R}^{m \times m}$ *are* **orthogonal matrices**, *and* $\Sigma \in \mathbb{R}^{m \times n}$ *is a* **diagonal matri***x with the first* $r = rank(\mathbf{A})$ *diagonal entries* $(\sigma_1, \ldots, \sigma_r)$ *positive and non-increasing, and all other diagonal entries are zero.*
The values along the main diagonal of Σ *are called the singular values of* $\mathbf{A}$. *The columns of* $\mathbf{U}$ *are called the left singular vectors, and the columns of* $\mathbf{V}$ *are called the right singular vectors.*

Recap: Compact form of the Singular Value Decomposition

Corollary (compact-form SVD)

Any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ *can be written as*

$$\mathbf{A} = \mathbf{U}_r \Sigma_r \mathbf{V}_r^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

where $r = rank(\mathbf{A})$, $\mathbf{U}_r = (\mathbf{u}_1, \ldots, \mathbf{u}_r) \in \mathbb{R}^{m\times r}$ *has orthonormal columns,* $\mathbf{V}_r = (\mathbf{v}_1, \ldots, \mathbf{v}_r) \in \mathbb{R}^{n\times r}$ *has orthonormal columns, and* $\sigma_1 \geq \sigma_2 \geq ...\sigma_r > 0$. *The positive numbers* $\sigma_1, \ldots, \sigma_r$ *are called the singular values of* $\mathbf{A}$, *the vectors* $\mathbf{u}_i$ *are called the left singular vectors of* $\mathbf{A}$ *and the vectors* $\mathbf{v}_i$ *are called the right singular vectors. These quantities satisfy*

$$\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i, \quad \mathbf{u}_i^\top \mathbf{A} = \sigma_i \mathbf{v}_i, \quad i = 1, \ldots, r.$$

Moreover, $\sigma_i^2 = \lambda_i(\mathbf{A}\mathbf{A}^\top) = \lambda_i(\mathbf{A}^\top\mathbf{A})$, $i = 1, \ldots, r$, *and* $\mathbf{u}_i, \mathbf{v}_i$ *are eigenvectors (corresponding to non-zero eignevalues) of* $\mathbf{A}\mathbf{A}^\top$ *and* $\mathbf{A}^\top\mathbf{A}$, *respectively.*

Low-rank Matrix Approximation via SVD

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ of rank $\mathrm{rank}(\mathbf{A}) = r > 0$. We consider the problem of finding a matrix $\mathbf{A}_k$ of rank-k, $k < r$, that is the best rank-k approximation of $\mathbf{A}$.
That is, we are interested in solving the following rank-constrained optimization problem:

$$\min_{\mathbf{A}_k \in \mathbb{R}^{m\times n}} \|\mathbf{A} - \mathbf{A}_k\|_F^2 \quad \text{s.t.} \quad \mathrm{rank}(\mathbf{A}_k) = k. \tag{1}$$

Let us begin by writing the compact-form SVD of $\mathbf{A}$:

$$\mathbf{A} = \mathbf{U}_r \Sigma_r \mathbf{V}_r^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top.$$

Theorem

An optimal solution to (1) *is given by* $\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$.
That is, we take the rank-k *truncation of the SVD of* $\mathbf{A}$.

Low-rank Matrix Approximation via SVD - Motivation

Recall that if $\mathbf{A}_k \in \mathbb{R}^{m\times n}$ is of rank-k and given by its compact-form SVD $\mathbf{A} = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ then

1. Storing $\mathbf{A}_k$ requires $O(k(m+n))$ space (as opposed to mn required for storing $\mathbf{A}$ explicitly.
2. Computing the matrix-vector product $\mathbf{A}_k\mathbf{x}$ for some $\mathbf{x}$ requires $k(m+n)$ time (instead of mn). Similarly for computing $\mathbf{y}^\top\mathbf{A}_t$ for some $\mathbf{y}$.
3. While $\mathbf{A}_k$ can be quite different than $\mathbf{A}$, as we shall see, for certain tasks $\mathbf{A}_k$ might provide a suffiecinately good approximation of $\mathbf{A}$ while enjoying the above computational benefits.

Proof of Low-rank Approximation

In order to prove the theorem, we first observe that the Frobenius norm is unitarily invariant, meaning $\|\mathbf{Y}\|_F = \|\mathbf{QYR}\|_F$ for every $\mathbf{Y} \in \mathbb{R}^{m\times n}$ and orthogonal matrices $\mathbf{Q} \in \mathbb{R}^{m\times m}$, $\mathbf{R} \in \mathbb{R}^{n\times n}$.
To see why this is true, note that

$$\begin{aligned}\|\mathbf{QYR}\|_F^2 &= \mathrm{Tr}((\mathbf{QYR})(\mathbf{QYR})^\top) = \mathrm{Tr}(\mathbf{QYRR}^\top\mathbf{Y}^\top\mathbf{Q}^\top) \\ &= \mathrm{Tr}(\mathbf{QYY}^\top\mathbf{Q}^\top) = \mathrm{Tr}(\mathbf{YY}^\top\mathbf{Q}^\top\mathbf{Q}) = \mathrm{Tr}(\mathbf{YY}^\top) = \|\mathbf{Y}\|_F^2.\end{aligned}$$

Therefore, considering some candidate solution to our problem - $\mathbf{A}_k$ and using the SVD of $\mathbf{A}$ ($\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$) we have

$$\|\mathbf{A} - \mathbf{A}_k\|_F^2 = \|\mathbf{U}^\top(\mathbf{A} - \mathbf{A}_k)\mathbf{V}\|_F^2 = \|\Sigma - \mathbf{Z}\|_F^2, \quad \mathbf{Z} = \mathbf{U}^\top\mathbf{A}_k\mathbf{V}.$$

Note $\mathbf{ZZ}^\top = \mathbf{U}^\top\mathbf{A}_k\mathbf{A}_k^\top\mathbf{U}$. Since $\mathbf{U}^\top = \mathbf{U}^{-1}$, we have that $\mathbf{ZZ}^\top$ and $\mathbf{A}_k\mathbf{A}_k^\top$ are similar matrices and thus $\mathrm{rank}(\mathbf{A}_k\mathbf{A}_k^\top) = \mathrm{rank}(\mathbf{ZZ}^\top)$. During the proof of the SVD theorem we have seen that for any $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathrm{rank}(\mathbf{AA}^\top) = \mathrm{rank}(\mathbf{A})$. Thus, we have that $\mathrm{rank}(\mathbf{Z}) = \mathrm{rank}(\mathbf{A}_k) = k$.

Proof of Low-rank Approximation

We have showed that our problem

$$\min_{\mathbf{A}_k \in \mathbb{R}^{m\times n}} \|\mathbf{A} - \mathbf{A}_k\|_F^2 \quad \text{s.t.} \quad \text{rank}(\mathbf{A}_k) = k$$

is equivalent to the following problem:

$$\min_{\mathbf{Z}\in\mathbb{R}^{m\times n}} \left\| \begin{pmatrix} \text{diag}(\sigma_1,\ldots,\sigma_r) & \mathbf{0}_{r,n-r} \\ \mathbf{0}_{m-r,r} & \mathbf{0}_{m-r,n-r} \end{pmatrix} - \mathbf{Z} \right\|_F^2 \quad \text{s.t.} \quad \text{rank}(\mathbf{Z}) = k.$$

Claim (without proof): An optimal rank-k approximation to a diagonal matrix $\mathbf{D} = \text{diag}(d_1,\ldots,d_{\min\{m,n\}})$, where $d_1 \geq d_2 \geq \ldots$ is the **diagonal matrix** $\mathbf{D}_k = \text{diag}(d_1, d_2, \ldots, d_k, 0, \ldots, 0)$.

Thus, the optimal solution to our new problem is
$\mathbf{Z}^* = \text{diag}(\sigma_1,\ldots,\sigma_k,0,\ldots,0)$.

Recall we have used the transformation $\mathbf{Z} = \mathbf{U}^\top \mathbf{A}_k \mathbf{V}$. Thus, the optimal solution to our original problem is:
$\mathbf{A}_k = \mathbf{U}\mathbf{Z}^*\mathbf{V}^\top = \mathbf{U}\text{diag}(\sigma_1,\ldots,\sigma_k,0,\ldots,0)\mathbf{V}^\top = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$.

Low-rank Matrix Approximation via SVD

Theorem

The best rank-k approximation to a matrix $\mathbf{A}$ with $rank(\mathbf{A}) \geq k$ (in the sense $\min_{\mathbf{A}_k : rank(\mathbf{A}_k)=k} \|\mathbf{A}_k - \mathbf{A}\|_F^2$) is given by $\mathbf{A}_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, where $\sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ is the compact SVD of $\mathbf{A}$.

We can also readily quantify the relative approximation error of $\mathbf{A}_k$:

$$\frac{\|\mathbf{A} - \mathbf{A}_k\|_F^2}{\|\mathbf{A}\|_F^2} = \frac{\|\sum_{i=k+1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top\|_F^2}{\|\sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top\|_F^2} = \frac{\sum_{i=k+1}^r \sigma_i^2}{\sum_{i=1}^r \sigma_i^2} \in [0,1).$$

Low-rank Matrix Approximation with the Spectral Norm

It is important to note that one of our main uses of the fact that we measured the quality of the low-rank approximation using the Frobenius norm was that it is **unitary invariant** (recall the change of variables from $\mathbf{A}_k$ to $\mathbf{Z} = \mathbf{U}\mathbf{A}_k\mathbf{V}^\top$).

It can be shown (HW) that the spectral norm is also unitary invariant. Indeed, for the following variant of the low-rank approximation problem:

$$\min_{\mathbf{A}_k \in \mathbb{R}^{m\times n}} \|\mathbf{A} - \mathbf{A}_k\|_2^2 \quad \text{s.t.} \quad \text{rank}(\mathbf{A}_k) = k,$$

that is, finding the best rank-k approximation w.r.t. the **spectral norm**, the optimal solution is also given by $\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$.

Here also we can quantify the relative approximation error of $\mathbf{A}_k$ given by

$$\frac{\|\mathbf{A} - \mathbf{A}_k\|_2^2}{\|\mathbf{A}\|_2^2} = \frac{\|\sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top\|_2^2}{\|\sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top\|_2^2} = \frac{\sigma_{k+1}^2}{\sigma_1^2} \in [0, 1].$$

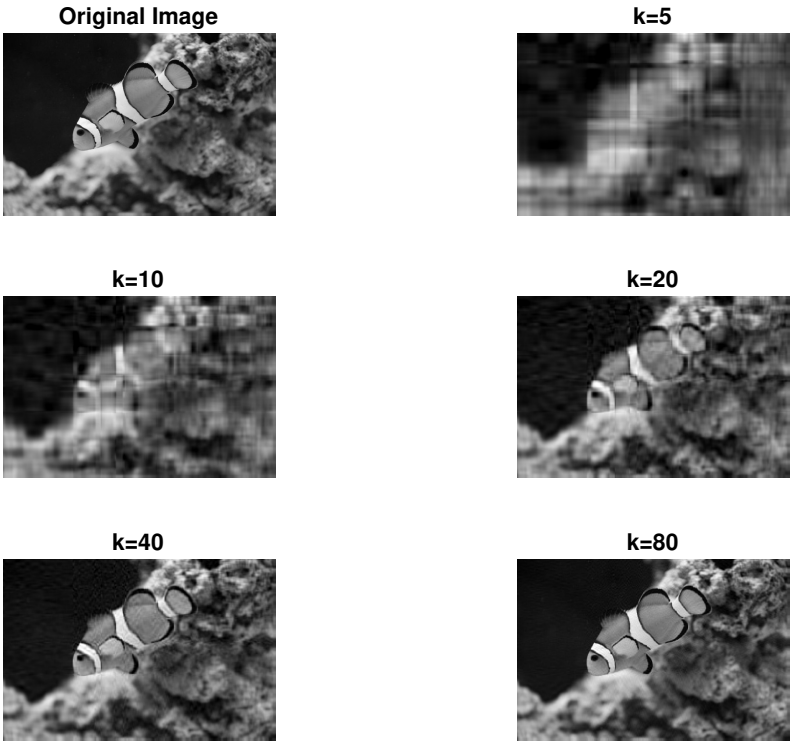
Applications to Image Compression

Suppose $\mathbf{A} \in \mathbb{R}^{m\times n}$ represents an $m \times n$ grayscale image, i.e., every entry $\mathbf{A}_{i,j}$ is the light-intensity of the corresponding pixel.

Storing the complete image matrix (which is usually full-rank) requires $O(mn)$ space. However, storing it's best rank-k approximation requires only $O(k(m+n))$ space.

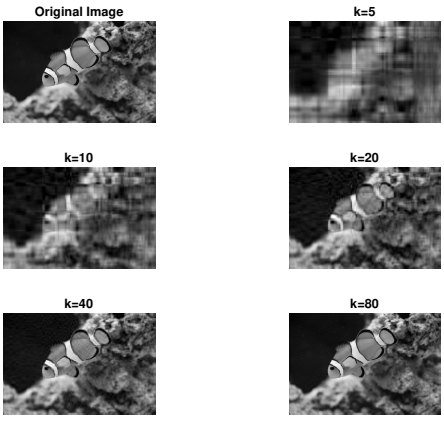
Applications to Image Compression

Example with $m = 475, n = 712$ (rank $= 475$).



Applications to Image Compression

Example with $m = 475, n = 712$ (rank $= 475$).



For $k = 80$ the image already looks very much like the original. What is the saving in space?

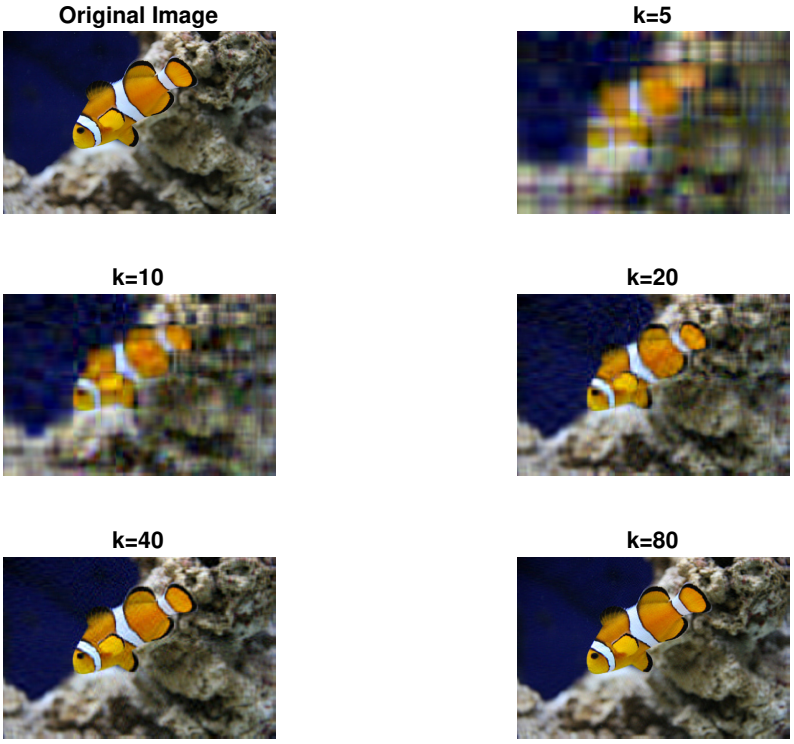
$$\frac{k(m+n)}{mn} = \frac{80 \cdot (475 + 712)}{475 \cdot 712} \approx 0.28.$$

Applications to Image Compression

Color images are represented via the RGB format. That is each image is represented by three $m \times n$ matrices, each representing the red, green and blue intensities, respectively.
Note that now saving an image requires $3mn$ space.

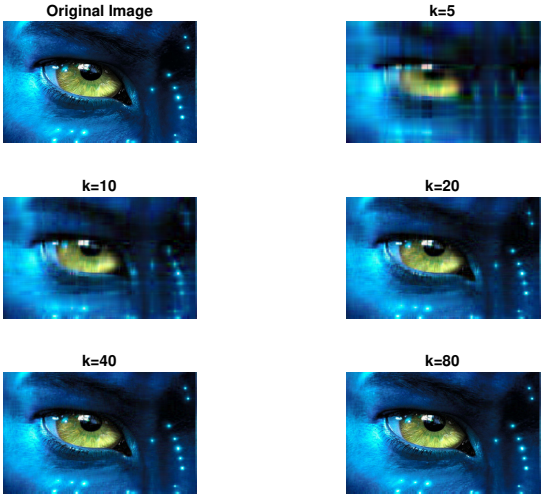
We can apply the same compression idea as before, by approximating each color matrix via a low-rank matrix independently.

Applications to Image Compression



Applications to Image Compression

Here is another example with a 707x1131 RGB image (here also the RGB matrices are full-rank). We get the following compression results:



Note that already for $k = 5$ (only 0.7% of space!!) we can identify that the image is a face.

Applications to Image Compression

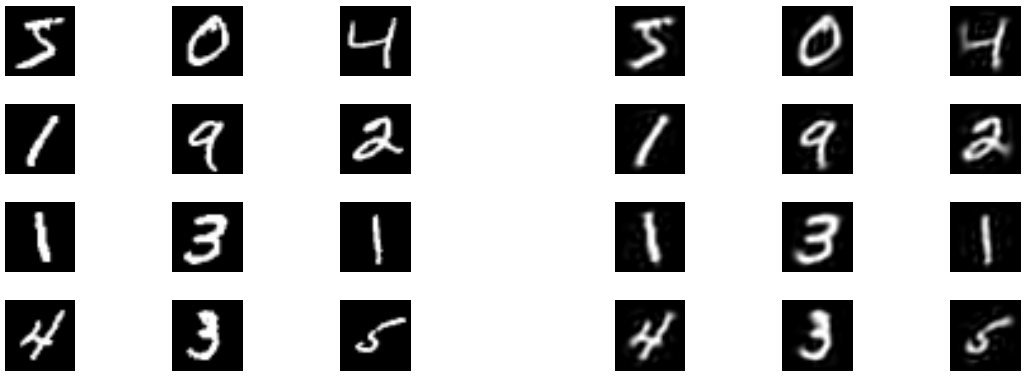
We can also use low-rank approximation to compress many related images at once.

The MNIST dataset contains $m \approx 60,000$ 28x28 grayscale images of 0-9 digits used to train machine learning algorithms (very successfully!!).

To compress the entire dataset, we construct a matrix in which each row corresponds to a single image, that is, we save each 28x28 image as a 1x784 vector. Hence we get a matrix with $m \approx 60000$ and $n = 784$ (here also the matrix is almost full rank: $r = 712$).

Applications to Image Compression

Below we plot the first 16 images of the original dataset and the first 16 images of the compressed dataset when using $k = 70$ (that is, using only 10% of original space!): Original on left, compressed on right.

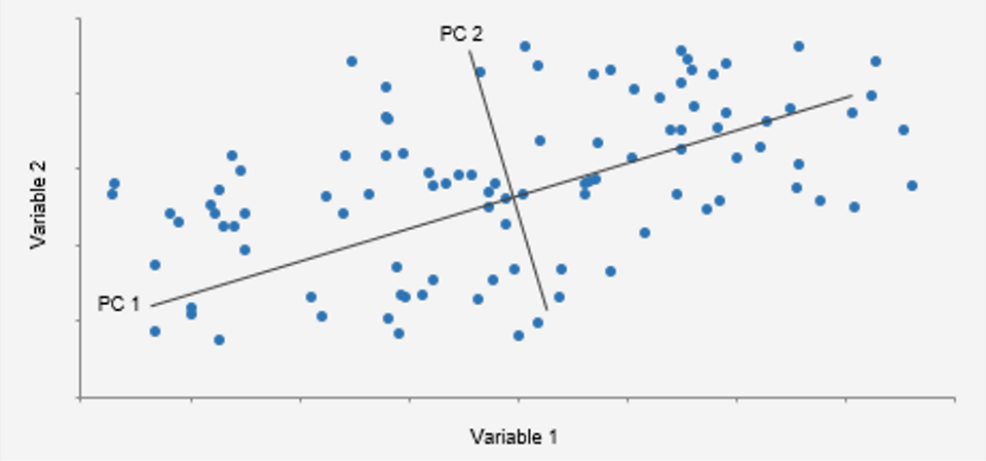


Why not compress each image by itslef? can you do the calculation what to see is preferred in terms for space?

More Applications of SVD: Principal Component Analysis

Motivation: given many data-points, each is a collection of many features, we would like to find a "simple" explanation to the data, or more precisely to the variation in the data.

In our case, "simple" will correspond to a low-dimensional subspace that "nearly" contains the data points. This could be thought of also as finding a small number of "new features" that provide a "good" approximation of the data (which often contains many many features).



Principal Component Analysis

Let $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ be given data points. Let us denote their average by $\bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i$.
The mean square variation of the data (or simply variance) is given by

$$\frac{1}{m}\sum_{i=1}^m \|\mathbf{x}_i - \bar{\mathbf{x}}\|_2^2 = \frac{1}{m}\sum_{i=1}^m \|\tilde{\mathbf{x}}_i\|_2^2, \qquad \tilde{\mathbf{x}}_i := \mathbf{x}_i - \bar{\mathbf{x}}.$$

Similarly, the variance along some normalized direction in space, represented by some unit vector $\mathbf{z} \in \mathbb{R}^n, \|\mathbf{z}\|_2 = 1$ is given by

$$\frac{1}{m}\sum_{i=1}^m \left((\mathbf{x}_i - \bar{\mathbf{x}})^\top \mathbf{z}\right)^2 = \frac{1}{m}\sum_{i=1}^m (\tilde{\mathbf{x}}_i^\top \mathbf{z})^2 = \frac{1}{m}\sum_{i=1}^m \|\mathbf{z}\mathbf{z}^\top \tilde{\mathbf{x}}_i\|_2^2.$$

Thus, the variance along the direction $\mathbf{z}$ is equal to the average of squared norms of the projections of the data onto the subspace $\mathrm{span}(\mathbf{z})$.

Principal Component Analysis

We define the $n \times m$ matrix which holds the centered data points as columns:

$$\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_m), \quad \tilde{\mathbf{x}}_i := \mathbf{x}_i - \bar{\mathbf{x}}, \ i = 1, \ldots, m.$$

We are looking for a (normalized) direction in space, i.e, a vector $\mathbf{z} \in \mathbb{R}^n$, $\|\mathbf{z}\|_2 = 1$, such that the variance along this direction is maximized.

The direction $\mathbf{z}$ along which the data has the largest variation can be found as the solution to the following optimization problem:

$$\max_{\mathbf{z}:\|\mathbf{z}\|_2=1} \frac{1}{m}\sum_{i=1}^m \left((\mathbf{x}_i - \bar{\mathbf{x}})^\top \mathbf{z}\right)^2 = \max_{\mathbf{z}:\|\mathbf{z}\|_2=1} \frac{1}{m}\sum_{i=1}^m (\tilde{\mathbf{x}}_i^\top \mathbf{z})^2$$
$$= \max_{\mathbf{z}:\|\mathbf{z}\|_2=1} \frac{1}{m}\sum_{i=1}^m \mathbf{z}^\top \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top \mathbf{z} = \max_{\mathbf{z}:\|\mathbf{z}\|_2=1} \mathbf{z}^\top \left(\frac{1}{m}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top\right)\mathbf{z}.$$

We now show how to solve this problem via SVD.

Principal Component Analysis

We want to solve $\max_{\mathbf{z}:\|\mathbf{z}\|_2=1} \mathbf{z}^\top \left(\frac{1}{m}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top\right)\mathbf{z}$.
Write the compact-form SVD of $\tilde{\mathbf{X}}$: $\tilde{\mathbf{X}} = \mathbf{U}_r\Sigma_r\mathbf{V}_r^\top = \sum_{i=1}^r \sigma_i\mathbf{u}_i\mathbf{v}_i^\top$.
Then,

$$\mathbf{H} = \frac{1}{m}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top = \frac{1}{m}\mathbf{U}_r\Sigma_r^2\mathbf{U}_r^\top.$$

Note that $\mathbf{H}$ is symmetric and $\mathbf{U}_r\frac{1}{m}\Sigma_r^2\mathbf{u}_r^\top$ is its spectral decomposition.

Recall that by Rayleigh's Theorem, we have that the optimal solution to our optimization problem is given by the leading eigenvector of $\mathbf{H}$. That is, $\mathbf{z}^* = \mathbf{u}_1$ - the first column of $\mathbf{U}_r$, or in other words, the left singular vector corresponding to the first singular value of $\tilde{\mathbf{X}}$ - σ_1.

Note further that the variance along this optimal direction is given by

$$\mathbf{u}_1^\top\left(\frac{1}{m}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top\right)\mathbf{u}_1 = \mathbf{u}_1^\top\left(\sum_{i=1}^r \frac{\sigma_i^2}{m}\mathbf{u}_i\mathbf{u}_i^\top\right)\mathbf{u}_1 = \sigma_1^2/m.$$

Subsequent Principal Components

After determining the direction which maximizes the variance, we can proceed to determine a second-largest variation direction.
To this end, we first remove from data the component along the already found largest variation direction $\mathbf{u}_1$:

$$\tilde{\mathbf{x}}_i^{(1)} = \tilde{\mathbf{x}}_i - \mathbf{u}_1\mathbf{u}_1^\top\tilde{\mathbf{x}}_i = \tilde{\mathbf{x}}_i - (\mathbf{u}_1^\top\tilde{\mathbf{x}}_i)\mathbf{u}_1, \quad i = 1,\ldots,m,$$

and the corresponding deflated matrix

$$\tilde{\mathbf{X}}^{(1)} = (\tilde{\mathbf{x}}_1^{(1)},\ldots,\tilde{\mathbf{x}}_m^{(1)}) = (\mathbf{I}_n - \mathbf{u}_1\mathbf{u}_1^\top)\tilde{\mathbf{X}}.$$

Note, the we can readily obtain the SVD of $\tilde{\mathbf{X}}^{(1)}$ from that of $\tilde{\mathbf{X}}$ by observing that

$$\tilde{\mathbf{X}}^{(1)} = (\mathbf{I}_n - \mathbf{u}_1\mathbf{u}_1^\top)\sum_{i=1}^r \sigma_i\mathbf{u}_i\mathbf{v}_i^\top = \sum_{i=1}^r \sigma_i\mathbf{u}_i\mathbf{v}_i^\top - \sigma_1\mathbf{u}_1\mathbf{v}_1^\top = \sum_{i=2}^r \sigma_i\mathbf{u}_i\mathbf{v}_i.$$

The second-largest direction of variation $\mathbf{z}$ is thus the solution to the optimization problem: $\max_{\mathbf{z}:\|\mathbf{z}\|_2=1} \mathbf{z}^\top\left(\frac{1}{m}\tilde{\mathbf{X}}^{(1)}\tilde{\mathbf{X}}^{(1)\top}\right)\mathbf{z}$.

Subsequent Principal Components

The second-largest direction of variation $\mathbf{z}$ is thus the solution to the optimization problem: $\max_{\mathbf{z}:\|\mathbf{z}\|_2=1} \mathbf{z}^\top \left(\frac{1}{m}\tilde{\mathbf{X}}^{(1)}\tilde{\mathbf{X}}^{(1)\top}\right)\mathbf{z}$ for $\tilde{\mathbf{X}}^{(1)} = (\mathbf{I}_n - \mathbf{u}_1\mathbf{u}_1^\top)\tilde{\mathbf{X}} = \sum_{i=2}^r \sigma_i \mathbf{u}_i \mathbf{v}_i$.

As before, the solution is given by the left singular vector associated with the largest singular value of $\tilde{\mathbf{X}}^{(1)}$, that is, the vector $\mathbf{u}_2$ - the left singular value corresponding to the second largest singular value of $\tilde{\mathbf{X}}$ - σ_2.

Note that again, the variance along this second optimal direction is given by

$$\mathbf{u}_2^\top \left(\frac{1}{m}\tilde{\mathbf{X}}^{(1)}\tilde{\mathbf{X}}^{(1)\top}\right)\mathbf{u}_2 = \mathbf{u}_2^\top \left(\sum_{i=2}^r \frac{\sigma_i^2}{m}\mathbf{u}_i\mathbf{u}_i^\top\right)\mathbf{u}_2 = \sigma_2^2/m.$$

We can iterate this process to find subsequent principal directions.
We see that these principal directions are nothing more than the left singular vectors of $\tilde{\mathbf{X}}$.
Moreover, the corresponding mean-square data variation along the top k directions are given by $\sigma_1^2/m, \ldots, \sigma_k^2/m$.

Principal Component Analysis

After finding the first k principal components, i.e., the top left singular vectors of $\tilde{\mathbf{X}}$ - $\mathbf{u}_1, \ldots, \mathbf{u}_k$, we have a k-dimensional subspace which is $\text{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$.
We can also readily quantify the quality of using the first k principal components by the "explained" variance ratio:

$$\frac{\sum_{i=1}^k \mathbf{u}_i^\top \left(\frac{1}{m}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top\right)\mathbf{u}_i}{\sum_{i=1}^n \mathbf{u}_i^\top \left(\frac{1}{m}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top\right)\mathbf{u}_i} = \frac{\sum_{i=1}^k \sigma_i^2}{\sum_{i=1}^n \sigma_i^2} \in (0, 1].$$

When this ratio is close to 1 it means that the majority of variance in the data is indeed explained by the top k principal components.

PCA and Low-rank Approximation

Recall that if the compact-form SVD of $\tilde{\mathbf{X}}$ is given by $\tilde{\mathbf{X}} = \mathbf{U}_r \Sigma_r \mathbf{V}_r^\top = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, then the best rank-$k$ ($k \leq r$) matrix approximation to $\tilde{\mathbf{X}}$ (in Frobenius norm) is given by

$$\tilde{\mathbf{X}}_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^\top = \mathbf{U}_k \Sigma_k \mathbf{V}_k^\top.$$

Recall that the projection of each centered data-point $\tilde{\mathbf{x}}_i$ onto the span of top k principal components is given by $\mathbf{U}_k \mathbf{U}_k^\top \tilde{\mathbf{x}}_i$, or in matrix form:

$$(\mathbf{U}_k \mathbf{U}_k^\top \tilde{\mathbf{x}}_1 \ \mathbf{U}_k \mathbf{U}_k^\top \tilde{\mathbf{x}}_2 \ \ldots \mathbf{U}_k \mathbf{U}_k^\top \tilde{\mathbf{x}}_m) = \mathbf{U}_k \mathbf{U}_k^\top \tilde{\mathbf{X}}.$$

Plugging the SVD of $\tilde{\mathbf{X}}$ we have

$$\mathbf{U}_k \mathbf{U}_k^\top \mathbf{U}_r \Sigma_r \mathbf{V}_r^\top = \mathbf{U}_k \left(\mathbf{I}_k \ \mathbf{0}_{k \times r-k}\right) \Sigma_r \mathbf{V}_r^\top = \mathbf{U}_k \left(\mathbf{I}_k \ \mathbf{0}_{k \times r-k}\right) \begin{pmatrix} \Sigma_k & \mathbf{0} \\ \mathbf{0} & \Sigma_{k+1:r} \end{pmatrix} \mathbf{V}_r^\top$$

$$= \mathbf{U}_k \left(\Sigma_k \ \mathbf{0}_{k \times r-k}\right) \mathbf{V}_r^\top = \mathbf{U}_k \left(\Sigma_k \ \mathbf{0}\right) \left(\mathbf{V}_k \ \mathbf{V}_{r+1:k}\right)^\top = \mathbf{U}_k \Sigma_k \mathbf{V}_k^\top.$$

Thus, the projection of the centered data points onto the optimal rank-k subspace coincides with the optimal rank-k approximation of $\tilde{\mathbf{X}}$.

PCA as a Procedure for Dimension Reduction

PCA is often used a a technique to reduce the dimensionality of data and by that reduce the runtime of algorithms that operate on the data and even the memory required to store it.

Let $\mathbf{U}_k = (\mathbf{u}_1, \ldots, \mathbf{u}_k)$ be the matrix whose columns are the k top principal components of the (centered) data matrix $\tilde{\mathbf{X}}$.

The projection of the data onto the $\mathrm{span}(\mathbf{u}_1, \ldots, \mathbf{u}_k)$ is given by

$$\mathbf{y}_i = \mathbf{U}_k \mathbf{U}_k^\top \tilde{\mathbf{x}}_i \in \mathbb{R}^n, \ i = 1, \ldots, m.$$

Consider now the following points in $\mathbb{R}^k$:

$$\mathbf{w}_i = \mathbf{U}_k^\top \tilde{\mathbf{x}}_i \in \mathbb{R}^k, \ i = 1, \ldots, m.$$

Note that $\mathbf{w}_i$ is nothing but the vector of coefficients when we write $\mathbf{y}_i$ as a linear combination of $\mathbf{u}_1, \ldots, \mathbf{u}_k$.

PCA as a Procedure for Dimension Reduction

Many algorithms, given the data $\tilde{\mathbf{x}}_i, i = 1, \ldots, m$ access the data only by comparing pairs of points $\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j$ (e.g., Nearest Neighbours, cosine similarity etc.). Recall $\mathbf{y}_i = \mathbf{U}_k \mathbf{U}_k^\top \tilde{\mathbf{x}}_i = \mathbf{U}_k \mathbf{w}_i$, $\mathbf{y}_i, \tilde{\mathbf{x}}_i \in \mathbb{R}^n, \mathbf{w}_i \in \mathbb{R}^k$.

We now observe the following:

$$\begin{aligned} \forall i, j: \quad \|\mathbf{y}_i - \mathbf{y}_j\|_2^2 &= \|\mathbf{U}_k \mathbf{U}_k^\top (\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j)\|_2^2 \\ &= (\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j)^\top \mathbf{U}_k \mathbf{U}_k^\top \mathbf{U}_k \mathbf{U}_k^\top (\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j) \\ &= (\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j)^\top \mathbf{U}_k \mathbf{U}_k^\top (\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j) \\ &= \|\mathbf{U}_k^\top (\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j)\|_2^2 = \|\mathbf{w}_i - \mathbf{w}_j\|_2^2. \end{aligned}$$

Moreover,

$$\forall i, j: \quad \mathbf{y}_i^\top \mathbf{y}_j = \tilde{\mathbf{x}}_i^\top \mathbf{U}_k \mathbf{U}_k^\top \mathbf{U}_k \mathbf{U}_k^\top \tilde{\mathbf{x}}_j = \tilde{\mathbf{x}}_i^\top \mathbf{U}_k \mathbf{U}_k^\top \tilde{\mathbf{x}}_j = \mathbf{w}_i^\top \mathbf{w}_j.$$

It thus follows, that if our use of the projected vectors $\mathbf{y}_1, \ldots, \mathbf{y}_m$ is to compute distances, inner-products or angles, then we can readily use the low-dimensional vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m$ instead.

Indeed, storing $\{\mathbf{w}_i\}_{i=1}^m$ requires only mk memory (instead of mn) and each of the above operations takes only $O(k)$ time instead of $O(n)$!

© Dan Garber (Technion) Lesson 6 Winter 2020-21 27 / 27

Algebraic Methods in Data Science: Lesson 7

Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology

Dan Garber
https://dangar.net.technion.ac.il/

Winter Semester 2020-2021

Recap: The Singular Value Decomposition

Theorem (SVD decomposition)

Any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ *can we factored as*

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$$

where $\mathbf{V} \in \mathbb{R}^{n \times n}$, $\mathbf{U} \in \mathbb{R}^{m \times m}$ *are* **orthogonal matrices**, *and* $\Sigma \in \mathbb{R}^{m \times n}$ *is a* **diagonal matri***x with the first* $r = rank(\mathbf{A})$ *diagonal entries* $(\sigma_1, \ldots, \sigma_r)$ *positive and non-increasing, and all other diagonal entries are zero.*
The values along the main diagonal of Σ *are called the singular values of* $\mathbf{A}$. *The columns of* $\mathbf{U}$ *are called the left singular vectors, and the columns of* $\mathbf{V}$ *are called the right singular vectors.*

Recap: Compact form of the Singular Value Decomposition

Corollary (compact-form SVD)

Any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ *can be written as*

$$\mathbf{A} = \mathbf{U}_r \Sigma_r \mathbf{V}_r^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

where $r = rank(\mathbf{A})$, $\mathbf{U}_r = (\mathbf{u}_1, \ldots, \mathbf{u}_r) \in \mathbb{R}^{m \times r}$ *has orthonormal columns,* $\mathbf{V}_r = (\mathbf{v}_1, \ldots, \mathbf{v}_r) \in \mathbb{R}^{n \times r}$ *has orthonormal columns, and* $\sigma_1 \geq \sigma_2 \geq ...\sigma_r > 0$. *The positive numbers* $\sigma_1, \ldots, \sigma_r$ *are called the singular values of* $\mathbf{A}$, *the vectors* $\mathbf{u}_i$ *are called the left singular vectors of* $\mathbf{A}$ *and the vectors* $\mathbf{v}_i$ *are called the right singular vectors. These quantities satisfy*

$$\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i, \quad \mathbf{u}_i^\top \mathbf{A} = \sigma_i \mathbf{v}_i, \quad i = 1, \ldots, r.$$

Moreover, $\sigma_i^2 = \lambda_i(\mathbf{A}\mathbf{A}^\top) = \lambda_i(\mathbf{A}^\top \mathbf{A})$, $i = 1, \ldots, r$, *and* $\mathbf{u}_i, \mathbf{v}_i$ *are eigenvectors (corresponding to non-zero eignevalues) of* $\mathbf{A}\mathbf{A}^\top$ *and* $\mathbf{A}^\top \mathbf{A}$, *respectively.*

Computing the SVD

Theorem

Given any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$*, the* **compact-form** *singular value decomposition of* $\mathbf{A}$ *can be computed in* $O(\min\{m,n\}^2 \max\{m,n\})$ *time.*

Unfortunately, the proof of this theorem is highly technical and involved, and as such is beyond to scope of this course.

Computing the SVD

Theorem

Given any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, the **compact-form** *singular value decomposition of $\mathbf{A}$ can be computed in $O(\min\{m,n\}^2 \max\{m,n\})$ time.*

Some computational consequences:

1. if $\mathbf{A}$ is $n \times n$ and full-rank (non-singular), then computing $\mathbf{A}^{-1}$ can be carried out in $O(n^3)$ time.
2. if $\mathbf{A}$ is $n \times n$ and full-rank (non-singular), then for any vector $\mathbf{b} \in \mathbb{R}^n$, a unique solution to the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ can be computed in $O(n^3)$ time.
3. if $\mathbf{A}$ is $n \times n$ and symmetric, then its **compact-form spectral decomposition** can be computed in $O(n^3)$ time (HW).
4. if $\mathbf{A}$ is $n \times n$ and symmetric, then $\det(\mathbf{A})$ can be computed in $O(n^3)$ time (HW).

Computing the SVD

Theorem

Given any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, the **compact-form** *singular value decomposition of $\mathbf{A}$ can be computed in $O(\min\{m,n\}^2 \max\{m,n\})$ time.*

Some computational consequences:

5. if $\mathbf{A}$ is $n \times n$ and symmetric, then for any integer $k \geq 0$, the matrix $\mathbf{A}^k$ can be computed in $O(n^3)$ time (HW).
6. for any k, the best rank-k approximation of $\mathbf{A}$ can be computed in $O(mn^2)$ time (assuming $n \leq m$)
7. Given m vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ and for any k, the k principal components of the data can be computed in $O(mn^2)$ (assuming $n \leq m$).

Computational disadvantages of the full SVD computation

As we have seen, many times we are only interested in several top components of the SVD and not the complete decomposition (low rank approximation, PCA).

In such cases it is better to use algorithms that are much faster in case only part of the SVD is needed.

Moreover, the runtime of the full-SVD algorithm does not reduce even when the input matrix $\mathbf{A}$ is very sparse (i.e., many of the entries are zero).

As we shall see, there are fast algorithms which can readily exploit the matrix sparsity to reduce the runtime. This property is highly important in applications, since data matrices are very often sparse.

Computing the Top Eigenvector of a PSD Matrix

We begin our discussion of fast and efficient methods for computing the top components of the SVD with the most basic building block - an algorithm for approximating the top eigenvector and eigenvalue of a real positive semidefinite matrix.

Recall that the SVD of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is constructed from the eigenvectors and eigenvalues of $\mathbf{A}^\top \mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$ which are positive semidefinite).

Thus, from the leading eigenvecor and eigenvalue of either $\mathbf{A}^\top \mathbf{A}$ or $\mathbf{A}\mathbf{A}^\top$ we can extract the leading singular value σ_1 and the leading left and right singular vectors $\mathbf{u}_1, \mathbf{v}_1$.

Computing the Top Eigenvector of a PSD Matrix

Theorem (Power Method)

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be symmetric, positive semidefinite and of rank r. Let $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_r > 0$ be its non-zero eigenvalues, and assume that $\lambda_1 > \lambda_2$. Let $\mathbf{u}_1$ be the eigenvector corresponding to eigenvalue λ_1. Consider the sequence of vectors $\{\mathbf{x}_t\}_{t \geq 1}$ produced by the following updates:

$$\forall t \geq 0 : \qquad \mathbf{x}_{t+1} \leftarrow \frac{\mathbf{A}\mathbf{x}_t}{\|\mathbf{A}\mathbf{x}_t\|_2},$$

where $\mathbf{x}_0$ is some unit-norm vector such that $(\mathbf{u}_1^\top \mathbf{x}_0)^2 > 0$. Then, for any $\epsilon > 0$ it holds that

$$\forall t \geq \frac{1}{2} \frac{\lambda_1}{\lambda_1 - \lambda_2} \log \frac{1}{(\mathbf{u}_1^\top \mathbf{x}_0)^2 \epsilon} : \qquad i. \quad \|\mathbf{u}_1 \mathbf{u}_1^\top \mathbf{x}_{t+1}\|_2^2 = (\mathbf{u}_1^\top \mathbf{x}_{t+1})^2 \geq 1 - \epsilon$$

$$ii. \quad \lambda_1 \geq \mathbf{x}_{t+1}^\top \mathbf{A} \mathbf{x}_{t+1} \geq (1 - \epsilon)\lambda_1.$$

Moreover, each iteration takes $O(N + n)$ time, where N is the total number of non-zero entries in the matrix $\mathbf{A}$.

Proof of the Power Method

The update step of the power method is $\mathbf{x}_{t+1} \leftarrow \frac{\mathbf{A}\mathbf{x}_t}{\|\mathbf{A}\mathbf{x}_t\|_2}$.

First, observe that for all $t \geq 0$:

$$\mathbf{x}_{t+1} = \frac{\mathbf{A}\mathbf{x}_t}{\|\mathbf{A}\mathbf{x}_t\|_2} = \frac{\mathbf{A}\frac{\mathbf{A}\mathbf{x}_{t-1}}{\|\mathbf{A}\mathbf{x}_{t-1}\|_2}}{\left\|\mathbf{A}\frac{\mathbf{A}\mathbf{x}_{t-1}}{\|\mathbf{A}\mathbf{x}_{t-1}\|_2}\right\|_2} = \frac{\mathbf{A}^2\mathbf{x}_{t-1}}{\|\mathbf{A}^2\mathbf{x}_{t-1}\|_2} = \ldots = \frac{\mathbf{A}^{t+1}\mathbf{x}_0}{\|\mathbf{A}^{t+1}\mathbf{x}_0\|_2}.$$

Recall that the eigendecomposition of $\mathbf{A}^{t+1}$ is given by $\sum_{i=1}^r \lambda_i^{t+1}\mathbf{u}_i\mathbf{u}_i^\top$.
We thus have that for all $i = 1, \ldots, r$:

$$\begin{aligned}(\mathbf{u}_i^\top \mathbf{x}_{t+1})^2 &= \frac{(\mathbf{u}_i^\top \mathbf{A}^{t+1}\mathbf{x}_0)^2}{\|\mathbf{A}^{t+1}\mathbf{x}_0\|_2^2} = \frac{(\mathbf{u}_i^\top \sum_{j=1}^r \lambda_j^{t+1}\mathbf{u}_j\mathbf{u}_j^\top \mathbf{x}_0)^2}{\|\sum_{j=1}^r \lambda_j^{t+1}\mathbf{u}_j\mathbf{u}_j^\top \mathbf{x}_0\|_2^2} \\ &= \frac{\lambda_i^{2(t+1)}(\mathbf{u}_i^\top \mathbf{x}_0)^2}{\|\sum_{j=1}^r \lambda_j^{t+1}\mathbf{u}_j\mathbf{u}_j^\top \mathbf{x}_0\|_2^2}.\end{aligned}$$

Proof of the Power Method

We have showed that $(\mathbf{u}_i^\top \mathbf{x}_{t+1})^2 = \frac{\lambda_i^{2(t+1)}(\mathbf{u}_i^\top \mathbf{x}_0)^2}{\|\sum_{j=1}^r \lambda_j^{t+1}\mathbf{u}_j\mathbf{u}_j^\top \mathbf{x}_0\|_2^2}$. Note that

$$\|\sum_{j=1}^r \lambda_j^{t+1}\mathbf{u}_j\mathbf{u}_j^\top \mathbf{x}_0\|_2^2 = \sum_{i,j} \lambda_i^{t+1}\lambda_j^{t+1}(\mathbf{u}_i^\top \mathbf{x}_0)(\mathbf{u}_j^\top \mathbf{x}_0)\mathbf{u}_i^\top \mathbf{u}_j = \sum_{j=1}^r \lambda_j^{2(t+1)}(\mathbf{u}_j^\top \mathbf{x}_0)^2.$$

Thus, we have that

$$\begin{aligned}(\mathbf{u}_i^\top \mathbf{x}_{t+1})^2 &= \frac{\lambda_i^{2(t+1)}(\mathbf{u}_i^\top \mathbf{x}_0)^2}{\sum_{j=1}^r \lambda_j^{2(t+1)}(\mathbf{u}_j^\top \mathbf{x}_0)^2} = \frac{(\mathbf{u}_i^\top \mathbf{x}_0)^2}{\sum_{j=1}^r \left(\frac{\lambda_j}{\lambda_i}\right)^{2(t+1)}(\mathbf{u}_j^\top \mathbf{x}_0)^2} \\ &\leq \frac{(\mathbf{u}_i^\top \mathbf{x}_0)^2}{\left(\frac{\lambda_1}{\lambda_i}\right)^{2(t+1)}(\mathbf{u}_1^\top \mathbf{x}_0)^2}.\end{aligned}$$

Recall we assumed $\lambda_1 > \lambda_2$. Thus, for all $i > 1$, $(\lambda_1/\lambda_i)^2 > 1$, and we have that$(\mathbf{u}_i^\top \mathbf{x}_{t+1})^2$ decays to zero exponentially fast with t.

Proof of the Power Method

Observe that

$$\begin{aligned}
(\mathbf{u}_1^\top \mathbf{x}_{t+1})^2 &= \mathbf{x}_{t+1}^\top \mathbf{u}_1 \mathbf{u}_1^\top \mathbf{x}_{t+1} = \mathbf{x}_{t+1}^\top (\sum_{i=1}^{n} \mathbf{u}_i \mathbf{u}_i^\top - \sum_{j=2}^{n} \mathbf{u}_j \mathbf{u}_j^\top) \mathbf{x}_{t+1} \\
&= \mathbf{x}_{t+1}^\top (\mathbf{U}\mathbf{U}^\top - \sum_{j=2}^{n} \mathbf{u}_j \mathbf{u}_j^\top) \mathbf{x}_{t+1} \\
&= \mathbf{x}_{t+1}^\top \mathbf{I} \mathbf{x}_{t+1} - \mathbf{x}_{t+1}^\top \Big(\sum_{j=2}^{n} \mathbf{u}_j \mathbf{u}_j^\top \Big) \mathbf{x}_{t+1} \\
&= \|\mathbf{x}_{t+1}\|_2^2 - \sum_{j=2}^{n} (\mathbf{u}_j^\top \mathbf{x}_{t+1})^2 = 1 - \sum_{j=2}^{r} (\mathbf{u}_j^\top \mathbf{x}_{t+1})^2.
\end{aligned}$$

The last equality follows since for all $j > r$: $\mathbf{u}_j^\top \mathbf{x}_{t+1} = \frac{\mathbf{u}_j^\top \mathbf{A}\mathbf{x}_t}{\|\mathbf{A}\mathbf{x}_t\|} = 0$, since $\mathbf{u}_j$ is eigenvector corresponding to eigenvalue $\lambda_j = 0$.

By the above argument, we also have that

$$\sum_{j=2}^{r} (\mathbf{u}_j^\top \mathbf{x}_0)^2 \leq \sum_{j=1}^{n} (\mathbf{u}_j^\top \mathbf{x}_0)^2 = \mathbf{x}_0^\top \mathbf{U}\mathbf{U}^\top \mathbf{x}_0 = \mathbf{x}_0^\top \mathbf{I} \mathbf{x}_0 = \|\mathbf{x}_0\|_2^2 = 1.$$

Proof of the Power Method

We have seen that

1. for all i: $(\mathbf{u}_i^\top \mathbf{x}_{t+1})^2 \leq \frac{(\mathbf{u}_i^\top \mathbf{x}_0)^2}{(\lambda_1/\lambda_i)^{2(t+1)}(\mathbf{u}_1^\top \mathbf{x}_0)^2}$
2. $(\mathbf{u}_1^\top \mathbf{x}_{t+1})^2 = 1 - \sum_{j=2}^r (\mathbf{u}_j^\top \mathbf{x}_{t+1})^2$
3. $\sum_{j=2}^r (\mathbf{u}_j^\top \mathbf{x}_0)^2 \leq 1$.

Combining these ingredients we get:

$$
\begin{aligned}
(\mathbf{u}_1^\top \mathbf{x}_{t+1})^2 &\geq 1 - \sum_{j=2}^r \frac{(\mathbf{u}_j^\top \mathbf{x}_0)^2}{\left(\frac{\lambda_1}{\lambda_j}\right)^{2(t+1)} (\mathbf{u}_1^\top \mathbf{x}_0)^2} \geq 1 - \frac{\sum_{j=2}^r (\mathbf{u}_j^\top \mathbf{x}_0)^2}{\left(\frac{\lambda_1}{\lambda_2}\right)^{2(t+1)} (\mathbf{u}_1^\top \mathbf{x}_0)^2} \\
&\geq 1 - \frac{1}{\left(\frac{\lambda_1}{\lambda_2}\right)^{2(t+1)} (\mathbf{u}_1^\top \mathbf{x}_0)^2} = 1 - \frac{1}{(\mathbf{u}_1^\top \mathbf{x}_0)^2} \left(\frac{\lambda_2}{\lambda_1}\right)^{2(t+1)} \\
&= 1 - \frac{1}{(\mathbf{u}_1^\top \mathbf{x}_0)^2} \left(1 - \frac{\lambda_1 - \lambda_2}{\lambda_1}\right)^{2(t+1)} \\
&\geq 1 - \frac{1}{(\mathbf{u}_1^\top \mathbf{x}_0)^2} \exp\left(-\frac{\lambda_1 - \lambda_2}{\lambda_1} 2(t+1)\right) \qquad \{1 - x \leq e^{-x}\}
\end{aligned}
$$

Proof of the Power Method

We got $(\mathbf{u}_1^\top \mathbf{x}_{t+1})^2 \geq 1 - \frac{1}{(\mathbf{u}_1^\top \mathbf{x}_0)^2} \exp\left(-\frac{\lambda_1 - \lambda_2}{\lambda_1} 2(t+1)\right)$.

Thus, we conclude that

$$\forall t \geq \frac{1}{2} \frac{\lambda_1}{\lambda_1 - \lambda_2} \log \frac{1}{(\mathbf{u}_1^\top \mathbf{x}_0)^2 \epsilon} : \quad (\mathbf{u}_1^\top \mathbf{x}_{t+1})^2 \geq 1 - \epsilon.$$

Moreover, if $(\mathbf{u}_1^\top \mathbf{x}_{t+1})^2 \geq 1 - \epsilon$, then

$$\lambda_1 \geq \mathbf{x}_{t+1}^\top \mathbf{A} \mathbf{x}_{t+1} = \mathbf{x}_{t+1}^\top \left(\sum_{i=1}^{r} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top \right) \mathbf{x}_{t+1} \geq \lambda_1 (\mathbf{u}_1^\top \mathbf{x}_{t+1})^2 \geq \lambda_1 (1 - \epsilon),$$

where the first inequality follows from Rayleigh's theorem.

Runtime analysis: Each step t requires computing $\mathbf{A}\mathbf{x}_t$ and computing $\|\mathbf{A}\mathbf{x}_t\|_2$. In the TA you shall see that $\mathbf{A}\mathbf{x}_t$ can be computed in $O(N + n)$ time. Given $\mathbf{A}\mathbf{x}_t$, computing $\|\mathbf{A}\mathbf{x}_t\|_2$ requires $O(n)$ time.

The Power Method

Theorem (Power Method)

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be symmetric, positive semidefinite and of rank r. Let $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_r > 0$ be its non-zero eigenvalues, and assume that $\lambda_1 > \lambda_2$. Let $\mathbf{u}_1$ be the eigenvector corresponding to eigenvalue λ_1. Consider the sequence of vectors $\{\mathbf{x}_t\}_{t \geq 1}$ produced by the following updates:

$$\forall t \geq 0: \qquad \mathbf{x}_{t+1} \leftarrow \frac{\mathbf{A}\mathbf{x}_t}{\|\mathbf{A}\mathbf{x}_t\|_2},$$

where $\mathbf{x}_0$ is some unit-norm vector such that $(\mathbf{u}_1^\top \mathbf{x}_0)^2 > 0$. Then, for any $\epsilon > 0$ it holds that

$$\forall t \geq \frac{1}{2}\frac{\lambda_1}{\lambda_1 - \lambda_2}\log\frac{1}{(\mathbf{u}_1^\top \mathbf{x}_0)^2 \epsilon}: \qquad i. \quad \|\mathbf{u}_1\mathbf{u}_1^\top \mathbf{x}_{t+1}\|_2^2 = (\mathbf{u}_1^\top \mathbf{x}_{t+1})^2 \geq 1 - \epsilon$$

$$ii. \quad \lambda_1 \geq \mathbf{x}_{t+1}^\top \mathbf{A}\mathbf{x}_{t+1} \geq (1-\epsilon)\lambda_1.$$

Moreover, each iteration takes $O(N + n)$ time, where N is the total number of non-zero entries in the matrix $\mathbf{A}$.

Initializing the Power Method

Lemma (Choosing initial vector $\mathbf{x}_0$ for the Power Method (without proof))

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a positive semidefinite matrix and let $\mathbf{u}_1$ be its leading eigenvector. Let $\mathbf{x}_0 \in \mathbb{R}^n$ be a unit vector chosen at random (uniformly from all unit vectors). Then, for any $\delta > 0$, it holds with probability at least $1 - \delta$ that $(\mathbf{u}_1^\top \mathbf{x}_0)^2 \geq \frac{\delta^2}{9n}$.

Question: what if the matrix $\mathbf{A}$ satisfies: $\lambda_1 \approx \lambda_2$? In this case the analysis we have seen for the power method is not effective since $\frac{\lambda_1}{\lambda_1 - \lambda_2} \to \infty$ as $\lambda_2 \to \lambda_1$.

Gap-free Analysis of the Power Method

Theorem (Gap-free analysis of the Power Method)

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be symmetric, positive semidefinite and of rank r. and let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ be its non-zero eigenvalues. Consider the sequence of vectors $\{\mathbf{x}_t\}_{t \geq 1}$ produced by the following rule:

$$\forall t \geq 0: \qquad \mathbf{x}_{t+1} \leftarrow \frac{\mathbf{A}\mathbf{x}_t}{\|\mathbf{A}\mathbf{x}_t\|_2},$$

where $\mathbf{x}_0$ is some unit-norm vector such that $(\mathbf{u}_1^\top \mathbf{x}_0)^2 > 0$.
Then, for any $\epsilon > 0$ it holds that

$$\forall t \geq \frac{1}{\epsilon} \log \frac{2}{(\mathbf{u}_1^\top \mathbf{x}_0)^2 \epsilon} - 1: \qquad \lambda_1 \geq \mathbf{x}_{t+1}^\top \mathbf{A} \mathbf{x}_{t+1} \geq (1-\epsilon)\lambda_1.$$

Moreover, each iteration of the above algorithm, i.e., computing $\mathbf{x}_{t+1}$ given $\mathbf{x}_t$, takes $O(N+n)$ time, where N is the total number of non-zero entries in the matrix $\mathbf{A}$.

Proof of Gap-free Convergence

Fix some integer $k \in \{1, 2, \ldots, r = \text{rank}(\mathbf{A})$.

Let $\mathbf{U}_k$ denote the $n \times k$ matrix whose columns are the corresponding eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$. Similarly, let $\mathbf{U}_k^{\perp}$ denote the $n \times (n-k)$ matrix whose columns are the rest of the $n-k$ eigenvectors. Recall that $\mathbf{U}_k\mathbf{U}_k^{\top} + \mathbf{U}_k^{\perp}\mathbf{U}_k^{\perp\top} = \mathbf{U}\mathbf{U}^{\top} = \mathbf{I}$. Observe that

$$\mathbf{x}_{t+1}^{\top}\mathbf{A}_t\mathbf{x}_{t+1} \quad = \quad \mathbf{x}_{t+1}^{\top}(\mathbf{U}_k\mathbf{U}_k^{\top} + \mathbf{U}_k^{\perp}\mathbf{U}_k^{\perp\top})\mathbf{A}(\mathbf{U}_k\mathbf{U}_k^{\top} + \mathbf{U}_k^{\perp}\mathbf{U}_k^{\perp\top})\mathbf{x}_{t+1}.$$

Note that

$$\mathbf{x}_{t+1}^{\top}\mathbf{U}_k\mathbf{U}_k^{\top}\mathbf{A}\mathbf{U}_k^{\perp}\mathbf{U}_k^{\perp\top}\mathbf{x}_{t+1} = \mathbf{x}_{t+1}^{\top}\mathbf{U}_k\mathbf{U}_k^{\top}(\sum_{i=1}^{n}\lambda_i\mathbf{u}_i\mathbf{u}_i^{\top})\mathbf{U}_k^{\perp}\mathbf{U}_k^{\perp\top}\mathbf{x}_{t+1} = 0.$$

Thus,

$$\begin{aligned}\mathbf{x}_{t+1}^{\top}\mathbf{A}\mathbf{x}_{t+1} &\geq \mathbf{x}_{t+1}^{\top}\mathbf{U}_k\mathbf{U}_k^{\top}\mathbf{A}\mathbf{U}_k\mathbf{U}_k^{\top}\mathbf{x}_{t+1} + \mathbf{x}_{t+1}^{\top}\mathbf{U}_k^{\perp}\mathbf{U}_k^{\perp\top}\mathbf{A}\mathbf{U}_k^{\perp}\mathbf{U}_k^{\perp\top}\mathbf{x}_{t+1} \\ &\geq \mathbf{x}_{t+1}^{\top}\mathbf{U}_k\mathbf{U}_k^{\top}\mathbf{A}\mathbf{U}_k\mathbf{U}_k^{\top}\mathbf{x}_{t+1}.\end{aligned}$$

Proof of Gap-free Convergence

Recall : $\mathbf{x}_{t+1}^\top \mathbf{A}_t \mathbf{x}_{t+1} \geq \mathbf{x}_{t+1}^\top \mathbf{U}_k \mathbf{U}_k^\top \mathbf{A} \mathbf{U}_k \mathbf{U}_k^\top \mathbf{x}_{t+1}$. Thus,

$$
\begin{aligned}
\mathbf{x}_{t+1}^\top \mathbf{A} \mathbf{x}_{t+1} &\geq \mathbf{x}_{t+1}^\top \mathbf{U}_k \mathbf{U}_k^\top \mathbf{A} \mathbf{U}_k \mathbf{U}_k^\top \mathbf{x}_{t+1} \\
&= \mathbf{x}_{t+1}^\top \Big(\sum_{i=1}^{k} \mathbf{u}_i \mathbf{u}_i^\top \Big) \Big(\sum_{j=1}^{r} \lambda_j \mathbf{u}_j \mathbf{u}_j^\top \Big) \Big(\sum_{l=1}^{k} \mathbf{u}_l \mathbf{u}_l \Big) \mathbf{x}_{t+1} \\
&= \mathbf{x}_{t+1}^\top \Big(\sum_{i=1}^{k} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top \Big) \mathbf{x}_{t+1} \\
&\geq \lambda_k \sum_{i=1}^{k} (\mathbf{u}_i^\top \mathbf{x}_{t+1})^2 \underset{(a)}{=} \lambda_k \Big(1 - \sum_{i=k+1}^{n} (\mathbf{u}_i^\top \mathbf{x}_{t+1})^2 \Big) \\
&\underset{(b)}{=} \lambda_k \Big(1 - \sum_{i=k+1}^{r} (\mathbf{u}_i^\top \mathbf{x}_{t+1})^2 \Big),
\end{aligned}
$$

were (a) follows since $1 = \|\mathbf{x}_{t+1}\|_2^2 = \sum_{i=1}^{n} (\mathbf{u}_i^\top \mathbf{x}_{t+1})^2$, and (b) follows since $\mathbf{u}_i^\top \mathbf{x}_{t+1} = 0$ for all $i = r+1, \ldots, n$.

Proof of Gap-free Convergence

We have showed that for $k \in \{1, \ldots, r\}$:
$\mathbf{x}_{t+1}^\top \mathbf{A}\mathbf{x}_{t+1} \geq \lambda_k \left(1 - \sum_{i=k+1}^r (\mathbf{u}_i^\top \mathbf{x}_{t+1})^2\right)$.

Recall that in proof of previous theorem we have showed:
for all $i = 2, \ldots, r$: $(\mathbf{u}_i^\top \mathbf{x}_{t+1})^2 \leq \frac{(\mathbf{u}_i^\top \mathbf{x}_0)^2}{(\lambda_1/\lambda_i)^{2(t+1)}(\mathbf{u}_1^\top \mathbf{x}_0)^2}$.

Let us set k such that for all $i \leq k$: $\lambda_i \geq (1 - \epsilon/2)\lambda_1$. Then,

$$\sum_{i=k+1}^r (\mathbf{u}_i^\top \mathbf{x}_{t+1})^2 \leq \sum_{i=k+1}^r \frac{(\mathbf{u}_i^\top \mathbf{x}_0)^2}{\left(\frac{\lambda_1}{\lambda_i}\right)^{2(t+1)} (\mathbf{u}_1^\top \mathbf{x}_0)^2} \leq \sum_{i=k+1}^r \frac{(\mathbf{u}_i^\top \mathbf{x}_0)^2}{\left(\frac{\lambda_1}{(1-\epsilon/2)\lambda_1}\right)^{2(t+1)} (\mathbf{u}_1^\top \mathbf{x}_0)^2}$$

$$\leq \frac{1}{(\mathbf{u}_1^\top \mathbf{x}_0)^2}(1 - \epsilon/2)^{2(t+1)} \leq \frac{1}{(\mathbf{u}_1^\top \mathbf{x}_0)^2} \exp\left(-\frac{t+1}{\epsilon}\right).$$

Thus, for any $t \geq \frac{1}{\epsilon} \log \frac{2}{(\mathbf{u}_1^\top \mathbf{x}_0)^2 \epsilon} - 1$ we have $\sum_{i=k+1}^r (\mathbf{u}_i^\top \mathbf{x}_{t+1})^2 \leq \epsilon/2$ and

$$\mathbf{x}_{t+1}^\top \mathbf{A}\mathbf{x}_{t+1} \geq (1 - \epsilon/2)\lambda_1 \left(1 - \sum_{i=k+1}^r (\mathbf{u}_i^\top \mathbf{x}_{t+1})^2\right) \geq (1 - \epsilon/2)\lambda_1 (1 - \epsilon/2)$$

$$\geq (1 - \epsilon/2)^2 = 1 - \epsilon + \epsilon^2/4 > 1 - \epsilon,$$

as required.

Computation of subsequent components

Note that given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ with SVD $\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. Given the top components $\sigma_1, \mathbf{u}_1, \mathbf{v}_1$ we can continue to compute $\sigma_2 \mathbf{u}_2, \mathbf{v}_2$ by computing the leading SVD component of the matrix:

$$\mathbf{A}^{(1)} = \mathbf{A} - \sigma_1 \mathbf{u}_1 \mathbf{v}_1^\top = \sum_{i=2}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top.$$

Computing the top components of $\mathbf{A}^{(1)}$ will thus give us the second component of $\mathbf{A}$: $\sigma_2, \mathbf{u}_2, \mathbf{v}_2$.
Note however, that in practice we never compute $\sigma_1, \mathbf{u}_1, \mathbf{v}_1$ precicely, but only compute approximations $\hat{\sigma}_1, \hat{\mathbf{u}}_1, \hat{\mathbf{v}}_1$, in which case we only have that

$$\mathbf{A}^{(1)} := \mathbf{A} - \hat{\sigma}_1 \hat{\mathbf{u}}_1 \hat{\mathbf{v}}_1^\top \approx \sum_{i=2}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top.$$

Hence, as we continue to compute subsequent components we accumulate errors.

Extracting the Top-k Subspace QR Iterations

Theorem (QR iterations algorithm)

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be positive semidefinite and with rank r. Let $k \leq r$ be a positive integer and suppose that $\lambda_k > \lambda_{k+1}$. Let us write eigendecomposition of $\mathbf{A}$ as

$$\mathbf{A} = (\mathbf{V}_1 \ \mathbf{V}_2) \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix} (\mathbf{V}_1 \ \mathbf{V}_2)^\top,$$

where $\mathbf{V}_1$ contains the top k eigenvectors of $\mathbf{A}$ and Λ_1 is diagonal with $\lambda_1, \ldots, \lambda_k$ along the main diagonal.
Consider an algorithm that performs the following iterations:

$$\begin{aligned} \forall t \geq 0: \qquad & \mathbf{Z}_{t+1} = \mathbf{A}\mathbf{Q}_t \\ & \mathbf{Q}_{t+1}\mathbf{R}_{t+1} = \mathbf{Z}_{t+1} \quad (QR \text{ decomposition of } \mathbf{Z}_{t+1}), \end{aligned}$$

where $\mathbf{Q}_0 \in \mathbb{R}^{n \times k}$ has orthonormal columns and suppose that $\sigma_{\min}(\mathbf{V}_1^\top \mathbf{Q}_0) > 0$. Then it holds that

$$\forall t \geq 1: \quad \|\mathbf{V}_2 \mathbf{V}_2^\top \mathbf{Q}_t\|_2 = \|\mathbf{V}_2^\top \mathbf{Q}_t\|_2 \leq \frac{1}{\sigma_{\min}(\mathbf{V}_1^\top \mathbf{Q}_0)} e^{-\frac{\lambda_k - \lambda_{k+1}}{\lambda_k}(t+1)}.$$

Moreover, each iteration of the algorithm requires $O(kN + k^2 n)$ runtime, where N is the total number of non-zero entries in $\mathbf{A}$.

Extracting the Top-k Subspace QR Iterations

Note that the QR iterations algorithm does not extract exactly the k leading eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$!!

Instead, it finds an orthonormal basis (the k columns of $\mathbf{Q}_t$) to the subspace $\mathrm{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$. This is often enough as we shall see.

Extracting the Top-k Subspace QR Iterations

Let $\mathbf{Q} = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k) \in \mathbb{R}^{n \times k}$ be such that the columns of $\mathbf{Q}$ are orthonormal and $\mathrm{span}(\mathbf{q}_1, \ldots, \mathbf{q}_k) = \mathrm{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$. It follows that $\mathbf{Q}\mathbf{Q}^\top = \mathbf{V}_1\mathbf{V}_1^\top$. The equality holds since both matrices are projection matrices onto the same subspace (and the projection is unique).

Now consider the problem of low rank matrix approximation. Let $\mathbf{A} = \mathbf{U}_r\Sigma_r\mathbf{V}_r^\top \in \mathbb{R}^{m \times n}$ (the compact SVD form) and let $\mathbf{U}_k\Sigma_k\mathbf{V}_k^\top$ be its best rank-k approximation (top k components of SVD).

Now, consider applying the QR iterations algorithm to the matrix $\mathbf{A}\mathbf{A}^\top = \mathbf{U}_r\Sigma_r^2\mathbf{U}_r^\top$ to find $\mathbf{Q} \in \mathbb{R}^{m \times k}$ such that $\mathbf{Q}\mathbf{Q}^\top \approx \mathbf{U}_k\mathbf{U}_k^\top$. Then, we have that

$$\mathbf{Q}\mathbf{Q}^\top\mathbf{A} \approx \mathbf{U}_k\mathbf{U}_k^\top\mathbf{U}_r\Sigma_r\mathbf{V}_r^\top = \mathbf{U}_k(\mathbf{I}_k, \mathbf{0}_{k \times r-k})\Sigma_r\mathbf{V}_r^\top = \cdots = \mathbf{U}_k\Sigma_k\mathbf{V}_k^\top.$$

Thus, from a basis to the top-k left singular vectors we can readily obtain the best rank-k approximation of the matrix $\mathbf{A}$.

Extracting the Top-k Subspace QR Iterations

Similarly, recall that for the PCA problem, given the centred-data matrix $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_m) \in \mathbb{R}^{n \times n}$ (data points as columns), the projection of the data onto the top k principal components is given by

$$\mathbf{U}_k \mathbf{U}_k^\top \tilde{\mathbf{X}}$$

where $\mathbf{U}_k \in \mathbb{R}^{n \times k}$ contain the k top eigenvectors of $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top$ as columns. Thus, by applying QR iterations to the matrix $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top$, we can readily compute (or at least approximate) this projection.

Algebraic Methods in Data Science: Lesson 8

Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology

Dan Garber
https://dangar.net.technion.ac.il/

Winter Semester 2020-2021

Recap: Computing the SVD

Theorem

Given any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$*, the* **compact-form** *singular value decomposition of* $\mathbf{A}$ *can be computed in* $O(\min\{m,n\}^2 \max\{m,n\})$ *time.*

Unfortunately, the proof of this theorem is highly technical and involved, and as such is beyond to scope of this course.

Recap: Computing the Top Eigenvector of a PSD Matrix

Theorem (Power Method)

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be symmetric, positive semidefinite and of rank r. Let $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_r > 0$ be its non-zero eigenvalues, and assume that $\lambda_1 > \lambda_2$. Let $\mathbf{u}_1$ be the eigenvector corresponding to eigenvalue λ_1. Consider the sequence of vectors $\{\mathbf{x}_t\}_{t \geq 1}$ produced by the following updates:

$$\forall t \geq 0: \qquad \mathbf{x}_{t+1} \leftarrow \frac{\mathbf{A}\mathbf{x}_t}{\|\mathbf{A}\mathbf{x}_t\|_2},$$

where $\mathbf{x}_0$ is some unit-norm vector such that $(\mathbf{u}_1^\top \mathbf{x}_0)^2 > 0$.
Then, for any $\epsilon > 0$ it holds that

$$\forall t \geq \frac{1}{2}\frac{\lambda_1}{\lambda_1 - \lambda_2}\log\frac{1}{(\mathbf{u}_1^\top \mathbf{x}_0)^2 \epsilon}: \quad i. \quad \|\mathbf{u}_1\mathbf{u}_1^\top \mathbf{x}_{t+1}\|_2^2 = (\mathbf{u}_1^\top \mathbf{x}_{t+1})^2 \geq 1 - \epsilon$$

$$ii. \quad \lambda_1 \geq \mathbf{x}_{t+1}^\top \mathbf{A}\mathbf{x}_{t+1} \geq (1-\epsilon)\lambda_1.$$

Moreover, each iteration takes $O(N+n)$ time, where N is the total number of non-zero entries in the matrix $\mathbf{A}$.

Recap: Gap-free Analysis of the Power Method

Theorem (Gap-free analysis of the Power Method)

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be symmetric, positive semidefinite and of rank r. and let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ be its non-zero eigenvalues. Consider the sequence of vectors $\{\mathbf{x}_t\}_{t \geq 1}$ produced by the following rule:

$$\forall t \geq 0: \qquad \mathbf{x}_{t+1} \leftarrow \frac{\mathbf{A}\mathbf{x}_t}{\|\mathbf{A}\mathbf{x}_t\|_2},$$

where $\mathbf{x}_0$ is some unit-norm vector such that $(\mathbf{u}_1^\top \mathbf{x}_0)^2 > 0$.
Then, for any $\epsilon > 0$ it holds that

$$\forall t \geq \frac{1}{\epsilon}\log\frac{2}{(\mathbf{u}_1^\top \mathbf{x}_0)^2 \epsilon} - 1: \qquad \lambda_1 \geq \mathbf{x}_{t+1}^\top \mathbf{A}\mathbf{x}_{t+1} \geq (1-\epsilon)\lambda_1.$$

Moreover, each iteration of the above algorithm, i.e., computing $\mathbf{x}_{t+1}$ given $\mathbf{x}_t$, takes $O(N+n)$ time, where N is the total number of non-zero entries in the matrix $\mathbf{A}$.

Recap: Computation of subsequent components

Note that given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ with SVD $\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, given the top components $\sigma_1, \mathbf{u}_1, \mathbf{v}_1$ we can continue to compute $\sigma_2 \mathbf{u}_2, \mathbf{v}_2$ by computing the leading SVD component of the matrix:

$$\mathbf{A}^{(1)} = \mathbf{A} - \sigma_1 \mathbf{u}_1 \mathbf{v}_1^\top = \sum_{i=2}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top.$$

Computing the top components of $\mathbf{A}^{(1)}$ will thus give us the second component of $\mathbf{A}$: $\sigma_2, \mathbf{u}_2, \mathbf{v}_2$.
Note however, that in practice we never compute $\sigma_1, \mathbf{u}_1, \mathbf{v}_1$ precisely, but only compute approximations $\hat{\sigma}_1, \hat{\mathbf{u}}_1, \hat{\mathbf{v}}_1$, in which case we only have that

$$\mathbf{A}^{(1)} := \mathbf{A} - \hat{\sigma}_1 \hat{\mathbf{u}}_1 \hat{\mathbf{v}}_1^\top \approx \sum_{i=2}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top.$$

Hence, as we continue to compute subsequent components we accumulate errors.

Extracting the Top-k Subspace QR Iterations

Theorem (QR iterations algorithm)

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be positive semidefinite and with rank r. Let $k \leq r$ be a positive integer and suppose that $\lambda_k > \lambda_{k+1}$. Let us write eigendecomposition of $\mathbf{A}$ as

$$\mathbf{A} = (\mathbf{V}_1 \ \mathbf{V}_2) \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix} (\mathbf{V}_1 \ \mathbf{V}_2)^\top,$$

where $\mathbf{V}_1$ contains the top k eigenvectors of $\mathbf{A}$ and Λ_1 is diagonal with $\lambda_1, \ldots, \lambda_k$ along the main diagonal.
Consider an algorithm that performs the following iterations:

$$\forall t \geq 0: \qquad \mathbf{Z}_{t+1} = \mathbf{A}\mathbf{Q}_t$$
$$\mathbf{Q}_{t+1}\mathbf{R}_{t+1} = \mathbf{Z}_{t+1} \quad (QR \text{ decomposition of } \mathbf{Z}_{t+1}),$$

where $\mathbf{Q}_0 \in \mathbb{R}^{n \times k}$ has orthonormal columns and suppose that $\sigma_{\min}(\mathbf{V}_1^\top \mathbf{Q}_0) > 0$. Then it holds that

$$\forall t \geq 1: \quad \|\mathbf{V}_2 \mathbf{V}_2^\top \mathbf{Q}_t\|_2 = \|\mathbf{V}_2^\top \mathbf{Q}_t\|_2 \leq \frac{1}{\sigma_{\min}(\mathbf{V}_1^\top \mathbf{Q}_0)} e^{-\frac{\lambda_k - \lambda_{k+1}}{\lambda_k}(t+1)}.$$

Moreover, each iteration of the algorithm requires $O(kN + k^2 n)$ runtime, where N is the total number of non-zero entries in $\mathbf{A}$.

Extracting the Top-k Subspace QR Iterations

Note that the QR iterations algorithm does not extract exactly the k leading eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$!!

Instead, it finds an orthonormal basis (the k columns of $\mathbf{Q}_t$) to the subspace $\mathrm{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$. This is often enough as we shall see.

Extracting the Top-k Subspace QR Iterations

Let $\mathbf{Q} = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k) \in \mathbb{R}^{n \times k}$ be such that the columns of $\mathbf{Q}$ are orthonormal and $\mathrm{span}(\mathbf{q}_1, \ldots, \mathbf{q}_k) = \mathrm{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$. It follows that $\mathbf{Q}\mathbf{Q}^\top = \mathbf{V}_1\mathbf{V}_1^\top$. The equality holds since both matrices are projection matrices onto the same subspace (and the projection is unique).

Now consider the problem of low rank matrix approximation. Let $\mathbf{A} = \mathbf{U}_r\Sigma_r\mathbf{V}_r^\top \in \mathbb{R}^{m \times n}$ (the compact SVD form) and let $\mathbf{U}_k\Sigma_k\mathbf{V}_k^\top$ be its best rank-k approximation (top k components of SVD).

Now, consider applying the QR iterations algorithm to the matrix $\mathbf{A}\mathbf{A}^\top = \mathbf{U}_r\Sigma_r^2\mathbf{U}_r^\top$ to find $\mathbf{Q} \in \mathbb{R}^{m \times k}$ such that $\mathbf{Q}\mathbf{Q}^\top \approx \mathbf{U}_k\mathbf{U}_k^\top$. Then, we have that

$$\mathbf{Q}\mathbf{Q}^\top\mathbf{A} \approx \mathbf{U}_k\mathbf{U}_k^\top\mathbf{U}_r\Sigma_r\mathbf{V}_r^\top = \mathbf{U}_k(\mathbf{I}_k, \mathbf{0}_{k \times r-k})\Sigma_r\mathbf{V}_r^\top = \cdots = \mathbf{U}_k\Sigma_k\mathbf{V}_k^\top.$$

Thus, from a basis to the top-k left singular vectors we can readily obtain the best rank-k approximation of the matrix $\mathbf{A}$.

Extracting the Top-k Subspace QR Iterations

Similarly, recall that for the PCA problem, given the centred-data matrix $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_m) \in \mathbb{R}^{n \times n}$ (data points as columns), the projection of the data onto the top k principal components is given by

$$\mathbf{U}_k \mathbf{U}_k^\top \tilde{\mathbf{X}}$$

where $\mathbf{U}_k \in \mathbb{R}^{n \times k}$ contain the k top eigenvectors of $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top$ as columns. Thus, by applying QR iterations to the matrix $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top$, we can readily compute (or at least approximate) this projection.

Part 3: Linear Systems and Least Squares Problems

We are moving to a new chapter: an in-depth look at linear systems and the highly related least squares problem (which we already met).

These are arguably two of the most important mathematical problems in applied sciences and engineering.

Recap: Linear Systems

Consider the system of linear equations

$$\mathbf{A}\mathbf{x} = \mathbf{y}, \qquad \mathbf{A} \in \mathbb{R}^{m \times n},\ \mathbf{y} \in \mathbb{R}^{m}. \tag{1}$$

Clearly, there exists a solution if and only if $\mathbf{y}$ can be expressed as a linear combination of the columns of $\mathbf{A}$. Note this is equivalent to the condition

$$\mathrm{rank}((\mathbf{A},\ \mathbf{y})) = \mathrm{rank}(\mathbf{A}). \tag{2}$$

Suppose condition (2) holds and let $\mathbf{x}^*$ be some solution.
There exists another solution $\mathbf{x} \neq \mathbf{x}^*$ to the system (1) if and only if

$$\mathbf{A}(\mathbf{x} - \mathbf{x}^*) = \mathbf{0},$$

hence $\mathbf{x} - \mathbf{x}^*$ must lie in the nullspace of $\mathbf{A}$. It thus follows that all possible solution for the system are of the form

$$\mathbf{x} = \mathbf{x}^* + \mathbf{z}, \quad \mathbf{z} \in \mathcal{N}(\mathbf{A}).$$

It further follows that $\mathbf{x}^*$ is a unique solution if and only if $\mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}$.

Recap: Linear Systems

Theorem (The solution set of a linear system)

The linear system

$$\mathbf{A}\mathbf{x} = \mathbf{y}, \qquad \mathbf{A} \in \mathbb{R}^{m \times n}$$

admits a solution if and only if $rank((\mathbf{A}\ \mathbf{y})) = rank(\mathbf{A})$. *When this condition is satisfied, the set of all solutions is the set*

$$S = \{\mathbf{x}^* + \mathbf{z}\ :\ \mathbf{z} \in \mathcal{N}(\mathbf{A})\},$$

where $\mathbf{x}^*$ *is any vector such that* $\mathbf{A}\mathbf{x}^* = \mathbf{y}$. *In particular, the system has a unique solution if* $rank((\mathbf{A},\ \mathbf{y})) = rank(\mathbf{A})$ *and* $\mathcal{N}(\mathbf{A}) = \mathbf{0}$.

Types of Linear Systems: Overdetermined Systems

The system $\mathbf{Ax} = \mathbf{y}$ is called overdetermined when it has more equations than unknowns, that is $m > n$.

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \\ A_{41} & A_{42} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

Now, suppose that $\mathbf{A}$ is full-column rank, that is $\mathrm{rank}(\mathbf{A}) = n$. Since

$$\mathrm{rank}(\mathbf{A}) + \mathrm{dim}\mathcal{N}(\mathbf{A}) = n$$

it follows that $\mathrm{dim}\mathcal{N}(\mathbf{A}) = 0$ and thus, the system has either one solution or no solution at all.

The common case is that indeed $\mathbf{y} \notin \mathrm{Im}(\mathbf{A})$ and thus the system has no solution.

Types of Linear Systems: Underdetermined Systems

The system $\mathbf{Ax} = \mathbf{y}$ is called underdetermined if it has more unknowns than equations, i.e., $m < n$.

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

Suppose now that $\mathbf{A}$ is full row-rank, that is $\mathrm{rank}(\mathbf{A}) = m$, and then $\mathrm{Im}(\mathbf{A}) = \mathbb{R}^m$. Recall that since

$$\mathrm{rank}(\mathbf{A}) + \mathrm{dim}\,\mathcal{N} = n,$$

and thus, $\mathrm{dim}\,\mathcal{N}(\mathbf{A}) = n - m > 0$. The system of linear equations is therefore solvable with infinite possible selections, and the set of solutions has "dimension" $n - m$ (this is not a subspace but an affine subspace).

Types of Linear Systems: Square Systems

In case $m = n$, the linear system $\mathbf{Ax} = \mathbf{y}$ is called square.

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

Clearly, if $m = n$ and $\mathbf{A}$ is full rank, then it is invertible and the unique solution is given by

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}.$$

This solution is unique since, as we have seen above, in this case $\dim \mathcal{N}(\mathbf{A}) = 0$.

The Least-Squares Problem

When $\mathbf{y} \notin \mathrm{Im}(\mathbf{A})$, the linear system has no solution.

This situation happens frequently in the case of overdetermined systems.

In such a case it may, however, make sense to determine an "approximate solution" to the system, that is a solution that renders the *residual* vector $\mathbf{r} := \mathbf{Ax} - \mathbf{y}$ as "small" as possible.

A natural way to measure the size of this residual vector is by a norm. The most common case is when the norm used is the Euclidean norm, in which case the problem becomes:

$$\min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{y}\|_2^2.$$

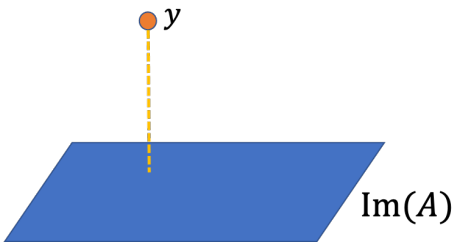
This is known as the Least-Squares (LS) problem.

The Least-Squares Problem

The Least Squares problem:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2.$$

Geometric interpretation: since the point $\mathbf{A}\mathbf{x}$ lies in $\mathrm{Im}(\mathbf{A})$, the LS problem amounts to finding a point $\tilde{\mathbf{y}} = \mathbf{A}\mathbf{x} \in \mathrm{Im}(\mathbf{A})$ at a minimum distance from $\mathbf{y}$. That is, the point $\tilde{\mathbf{y}} = \mathbf{A}\mathbf{x}$ is the orthogonal projection of $\mathbf{y}$ onto the subspace $\mathrm{Im}(\mathbf{A})$.



Thus, we can use the projection theorem to characterize the solutions to the LS problem. This leads us to the following theorem.

Least Squares and the Normal Equations

Theorem (Normal Equation)

Let $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{y} \in \mathbb{R}^m$. The LS problem

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 \tag{3}$$

always admits a solution. Moreover, $\mathbf{x}^* \in \mathbb{R}^n$ is a solution of (3) if and only if it is a solution of the following system of linear equations (the **normal** *equations)*

$$\mathbf{A}^\top \mathbf{A}\mathbf{x} = \mathbf{A}^\top \mathbf{y}.$$

Furthermore, if $\mathbf{A}$ is full column rank (i.e., $rank(\mathbf{A}) = n$), then the solution to (3) is unique, and it is given by

$$\mathbf{x}^* = (\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top \mathbf{y}.$$

Proof of theorem

Recall $\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 = \min_{\mathbf{w} \in \mathrm{Im}(\mathbf{A})} \|\mathbf{w} - \mathbf{y}\|_2^2$.

Given any $\mathbf{y} \in \mathbb{R}^m$, by the **projection theorem**, there exists a unique point $\tilde{\mathbf{y}} \in \mathrm{Im}(\mathbf{A})$ at minimal distance from $\mathbf{y}$ (the projection of $\mathbf{y}$ onto the subspace $\mathrm{Im}(\mathbf{A})$), and this point satisfies:

$$(\mathbf{y} - \tilde{\mathbf{y}}) \perp \mathrm{Im}(\mathbf{A}). \tag{4}$$

Since $\tilde{\mathbf{y}} \in \mathrm{Im}(\mathbf{A})$ there exists some $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{A}\mathbf{x} = \tilde{\mathbf{y}}$. This prove that the LS problem (3) indeed has a solution.

In particular, it follows from (5) that for all $\mathbf{z} \in \mathbb{R}^n$, $(\mathbf{y} - \tilde{\mathbf{y}})^\top \mathbf{A}\mathbf{z} = 0$. Thus, for any standard basis vector $\mathbf{e}_i$, we have that

$$(\mathbf{y} - \tilde{\mathbf{y}})^\top \mathbf{A}\mathbf{e}_i = (\mathbf{y} - \tilde{\mathbf{y}})^\top \mathbf{A}_i = 0,$$

where $\mathbf{A}_i$ denotes the ith column of $\mathbf{A}$. As a result it follows that

$$(\mathbf{y} - \tilde{\mathbf{y}})^\top \mathbf{A} = \left((\mathbf{y} - \tilde{\mathbf{y}})^\top \mathbf{A}\mathbf{e}_1, (\mathbf{y} - \tilde{\mathbf{y}})^\top \mathbf{A}\mathbf{e}_2, \ldots, (\mathbf{y} - \tilde{\mathbf{y}})^\top \mathbf{A}\mathbf{e}_n\right) = (0, \ldots, 0).$$

Rearranging we have $\mathbf{A}^\top(\mathbf{y} - \tilde{\mathbf{y}}) = \mathbf{0}$.

Proof of theorem

We showed that $\mathbf{A}^\top(\mathbf{y} - \tilde{\mathbf{y}}) = \mathbf{0}$, where $\tilde{\mathbf{y}} = \Pi_{\mathrm{Im}(\mathbf{A})}[\mathbf{y}]$.
Furthermore, plugging-in $\mathbf{A}\mathbf{x} = \tilde{\mathbf{y}}$ (recall such $\mathbf{x}$ is an optimal solution of the LS) and rearranging, we have

$$\mathbf{A}^\top \mathbf{A}\mathbf{x} = \mathbf{A}^\top \mathbf{y}. \tag{5}$$

Thus, a solution to the LS problem is indeed a solution to the normal equation (6).

Note that the other direction (solution to normal Eq. is solution to LS problem) also holds from the same arguments:

$$\begin{aligned} \mathbf{A}^\top \mathbf{A}\mathbf{x} = \mathbf{A}^\top \mathbf{y} \quad &\Rightarrow \quad \mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{y}) = \mathbf{0} \Rightarrow \forall \mathbf{z}: \ (\mathbf{A}\mathbf{x} - \mathbf{y})^\top \mathbf{A}\mathbf{z} = 0 \\ &\Rightarrow \quad (\mathbf{A}\mathbf{x} - \mathbf{y}) \perp \mathrm{Im}(\mathbf{A}) \Rightarrow \mathbf{A}\mathbf{x} = \Pi_{\mathrm{Im}(\mathbf{A})}[\mathbf{y}]. \end{aligned}$$

Thus, we have established the equivalence between the LS problem and the normal equation (6).

Proof of theorem

It remains to show that if $\mathbf{A}$ is full column rank ($\text{rank}(\mathbf{A}) = n$), then hen the solution to (3) is unique, and it is given by $\mathbf{x}^* = (\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top \mathbf{y}$.

Recall we have seen in the tutorials that if $\mathbf{A}$ is full column rank, then $\mathbf{A}^\top \mathbf{A} \succ 0$, which means $\mathbf{A}^\top \mathbf{A}$ is invertible.

Plugging this into the normal equation $\mathbf{A}^\top \mathbf{A}\mathbf{x} = \mathbf{A}^\top \mathbf{y}$, we get that in this case, the **unique solution** of both the normal equation and the LS problem is given by

$$\mathbf{x}^* = (\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top \mathbf{y}.$$

Recall we have seen this result when we talked about the pseudo-inverse matrix, since we have seen that $\mathbf{A}^\dagger \mathbf{y}$ is always a solution to the LS problem and we have seen that when $\mathbf{A}$ is column full rank then $\mathbf{A}^\dagger = (\mathbf{A}^\top \mathbf{A})^{-1}$. Now we see it from a different angel.

Set of solutions of LS problem

Theorem (set of solutions of LS problem)

The set of optimal solutions of the LS problem $\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$ *can be expressed as*

$$\mathcal{X}^* = \mathbf{A}^\dagger \mathbf{y} + \mathcal{N}(\mathbf{A}),$$

where $\mathbf{A}^\dagger \mathbf{y}$ *is the optimal solution of* **minimum Euclidean norm**.

If $\mathbf{A}$ *is full column rank, then the solution is unique and equal to*

$$\mathbf{x}^* = \mathbf{A}^\dagger \mathbf{y} = (\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top \mathbf{y}.$$

Proof: We begin with the last part of the theorem. We have already seen in previous theorem that if $\mathbf{A}$ is full column rank then $(\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top \mathbf{y}$ is the unique optimal solution to the LS problem. We have also previously seen that $\mathbf{A}^\dagger \mathbf{y}$ is always a solution to the LS. Thus, it follows that $\mathbf{x}^* = \mathbf{A}^\dagger \mathbf{y} = (\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top \mathbf{y}$.

Proof of theorem cont.

First, recall that according to the previous theorem, the LS problem always admits a solution and hence $\mathcal{X}^*$ is not empty.

Furthermore, since according to same theorem:

$$\mathbf{x}^* \in \mathcal{X}^* \quad \Leftrightarrow \quad \mathbf{A}^\top \mathbf{A}\mathbf{x}^* = \mathbf{A}^\top \mathbf{y}$$

it follows that $\mathcal{X}^*$ contains multiple solutions if and only if $\mathcal{N}(\mathbf{A}^\top \mathbf{A}) = \mathcal{N}(\mathbf{A})$ is not trivial, that is $\dim \mathcal{N}(\mathbf{A}) > 0$.
(recall we have seen in proof of SVD theorem that $\mathcal{N}(\mathbf{A}^\top \mathbf{A}) = \mathcal{N}(\mathbf{A})$).

In this case, it follows that any optimal solution $\mathbf{x} \in \mathcal{X}^*$ can be written as $\mathbf{x} = \mathbf{x}^* + \mathbf{z}$, where $\mathbf{z} \in \mathcal{N}(\mathbf{A})$ and $\mathbf{x}^*$ is some optimal solution.

Or equivalently:

$$\mathcal{X}^* = \mathbf{x}^* + \mathcal{N}(\mathbf{A}).$$

Proof of theorem cont.

Thus, the optimal solution to the LS problem of minimum norm is given by

$$\min_{\mathbf{x} \in \mathcal{X}^*} \|\mathbf{x}\|_2^2 = \min_{\mathbf{x} = \mathbf{x}^* + \mathbf{z},\ \mathbf{z} \in \mathcal{N}(\mathbf{A})} \|\mathbf{x}\|_2^2.$$

Consider now the change of variables: $\mathbf{z} = \mathbf{x} - \mathbf{x}^* \in \mathcal{N}(\mathbf{A})$. We have that

$$\min_{\mathbf{x} \in \mathcal{X}^*} \|\mathbf{x}\|_2^2 = \min_{\mathbf{z} \in \mathcal{N}(\mathbf{A})} \|\mathbf{z} - (-\mathbf{x}^*)\|_2^2.$$

The optimal solution $\mathbf{z}^*$ is given by the projection of the point $-\mathbf{x}^*$ onto the subspace $\mathcal{N}(\mathbf{A})$.
Thus, according to the **projection theorem**, the unique optimal solution $\mathbf{z}^*$ is the one which satisfies:

$$i.\ \mathbf{z}^* \in \mathcal{N}(\mathbf{A}), \qquad ii.\ -\mathbf{x}^* - \mathbf{z}^* \perp \mathcal{N}(\mathbf{A})$$

However, using the change of variables, these conditions are equivalent to

$$i.\ \mathbf{x} \in \mathbf{x}^* + \mathcal{N}(\mathbf{A}), \qquad ii.\ \mathbf{x} \perp \mathcal{N}(\mathbf{A})$$

Proof of theorem cont.

the optimal solution to the LS problem of minimum norm is the point $\mathbf{x}$ which satisfies

$$i.\ \mathbf{x} \in \mathbf{x}^* + \mathcal{N}(\mathbf{A}), \qquad ii.\ \mathbf{x} \perp \mathcal{N}(\mathbf{A}).$$

We now show that $\mathbf{x} = \mathbf{A}^\dagger \mathbf{y} = \mathbf{V}_r \Sigma_r^{-1} \mathbf{U}_r^\top \mathbf{y}$ indeed satisfies both requirements.

First note, that since $\mathbf{V}_r \perp \mathcal{N}(\mathbf{A})$ (i.e., columns of $\mathbf{V}_r$ orthogonal to $\mathcal{N}(\mathbf{A})$) it clearly follows that $\mathbf{A}^\dagger \mathbf{y} \perp \mathcal{N}(\mathbf{A})$.

Second, it holds that

$$\mathbf{A}^\top \mathbf{A}\mathbf{x} = \mathbf{V}_r \Sigma_r \mathbf{U}_r^\top \mathbf{U}_r \Sigma_r \mathbf{V}_r^\top (\mathbf{V}_r \Sigma_r^{-1} \mathbf{U}_r^\top \mathbf{y}) = \mathbf{V}_r \Sigma_r \mathbf{U}_r^\top \mathbf{y} = \mathbf{A}^\top \mathbf{y}.$$

Thus, $\mathbf{x} = \mathbf{A}^\dagger \mathbf{y}$ is a solution to the normal equation and thus an optimal solution, which means that indeed $\mathbf{x} \in \mathbf{x}^* + \mathcal{N}(\mathbf{A})$, as required.

Algebraic Methods in Data Science: Lesson 9

Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology

Dan Garber
https://dangar.net.technion.ac.il/

Winter Semester 2020-2021

Recap: Least Squares

Recall the least-squares problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x}-\mathbf{y}\|^2, \quad \mathbf{A}\in\mathbb{R}^{m\times n}$$

which generalizes the problem of solving a linear system:

- If $\mathbf{A}\mathbf{x}=\mathbf{y}$ has a solution then solving the LS will give such a solution.
- If $\mathbf{A}\mathbf{x}=\mathbf{y}$ does not have a solution then solving the LS will find a point $\mathbf{x}$ that is closest (in ℓ_2 norm) to solving the linear system.

Recall that $\mathbf{x}^*=\mathbf{A}^\dagger\mathbf{y}$ is always a solution to the LS ($\mathbf{A}^\dagger$ is the pseudo-inverse of $\mathbf{A}$).

Solving Least Squares Problems and Linear System

Since both problems can be solved by computing $\mathbf{x}^* = \mathbf{A}^\dagger \mathbf{y}$, from a computational point of view, the most expensive step in computing $\mathbf{x}^*$ is to compute the (compact) SVD of $\mathbf{A}$: $\mathbf{A} = \mathbf{U}_r \Sigma_r \mathbf{V}_r^\top$ (recall $\mathbf{A}^\dagger = \mathbf{V}_r \Sigma_r^{-1} \mathbf{U}_r^\top$). This leads to the following conclusion.

Theorem

The least squares problem $\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2$, $\mathbf{A} \in \mathbb{R}^{m \times n}$ *can be solved in* $O(\min\{m,n\}^2 \max\{m,n\})$.

In the tutorial you will see additional approaches to solving linear systems and least squares problems (QR decomposition, Cholesky decomposition), but these all have a similar runtime - $O(\min\{m,n\}^2 \max\{m,n\})$.

Fast Iterative Methods for Least Squares Problems

Theorem

The least squares problem $\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2$, $\mathbf{A} \in \mathbb{R}^{m \times n}$ *can be solved in* $O(\min\{m,n\}^2 \max\{m,n\})$.

Similarly, to the **leading eigenvector problem**, when the dimension m, n are very large, computing the SVD of $\mathbf{A}$ is not practical. Moreover, the runtime in the theorem does not benefit from any sparsity of $\mathbf{A}$ (small number of non-zero entries), which is often the case.

Thus, similarly to the leading eigenvector problem, we would like to develop an **iterative method** which can approximate the optimal solution of the LS problem to arbitrarily small error and such that each iteration is very efficient (e.g. $O(m+n+N)$ runtime, where N is number of non-zero entries in $\mathbf{A}$).

Towards an Iterative Method for LS

We will attempt to develop an iterative method from first principles. Be patient (:

Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{y} \in \mathbb{R}^m$, we will denote the LS objective by $f(\mathbf{x}) := \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$.

Now let $\mathbf{x} \in \mathbb{R}^n$ be some candidate solution to the LS problem. Let us try to improve $\mathbf{x}$. That is, let us try to find a new point $\mathbf{x}' \in \mathbb{R}^n$ such that $f(\mathbf{x}') < f(\mathbf{x})$.

Towards this, suppose $\mathbf{x}'$ is produced by a small change to $\mathbf{x}$. Concretely, consider $\mathbf{x}'$ of the form:

$$\mathbf{x}' = \mathbf{x} + \eta \cdot \mathbf{w},$$

where $\mathbf{w}$ is a unit-vector - the direction in which we want to move, and $\eta > 0$ is a (small) scalar (step-size).

Towards an Iterative Method for LS

Denote the LS objective by $f(\mathbf{x}) := \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$.

Given current point $\mathbf{x} \in \mathbb{R}^n$ consider a new point $\mathbf{x}'$ of the form:

$$\mathbf{x}' = \mathbf{x} + \eta \cdot \mathbf{w},$$

where $\mathbf{w}$ is a unit-vector - the direction in which we want to move, and $\eta > 0$ is a (small) scalar (step-size).
We have

$$\begin{aligned} f(\mathbf{x}') &= \|\mathbf{A}(\mathbf{x} + \eta\mathbf{w}) - \mathbf{y}\|_2^2 = \|\mathbf{A}\mathbf{x} - \mathbf{y} + \eta\mathbf{A}\mathbf{w}\|_2^2 \\ &= \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + 2\eta\mathbf{w}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{y}) + \|\eta\mathbf{A}\mathbf{w}\|_2^2 \\ &= f(\mathbf{x}) + 2\eta\mathbf{w}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{y}) + \eta^2\|\mathbf{A}\mathbf{w}\|_2^2. \end{aligned}$$

We would like now like to choose $\mathbf{w}, \eta$ in such a way that indeed $f(\mathbf{x}') < f(\mathbf{x})$.

Towards an Iterative Method for LS

Recall $f(\mathbf{x}) := \|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2$, $\mathbf{x}' = \mathbf{x} + \eta \cdot \mathbf{w}$.
We saw that

$$f(\mathbf{x}') = f(\mathbf{x}) + 2\eta \mathbf{w}^\top \mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y}) + \eta^2\|\mathbf{A}\mathbf{w}\|_2^2. \tag{1}$$

We would like now like to choose $\mathbf{w}, \eta$ so that $f(\mathbf{x}') < f(\mathbf{x})$.

Since η is (typically) small, we have $\eta^2\|\mathbf{A}\mathbf{w}\|_2^2 << 2\eta\mathbf{w}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})$.

Thus, when choosing $\mathbf{w}$ we will focus only on the term $2\eta\mathbf{w}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})$ (which is only linear in $\mathbf{w}$).

Since our goal is to make (1) as small as possible, we will naturally choose $\mathbf{w}$ that minimizes the product $2\eta\mathbf{w}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})$. Since $\|\mathbf{w}\|_2 = 1$, we take

$$\mathbf{w} = \arg\min_{\|\mathbf{v}\|_2=1} \mathbf{v}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y}) = -\frac{\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})}{\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2}$$

Recall from Cauchy-Swartz: $\forall \mathbf{z}$: $|\mathbf{z}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})| \le \|\mathbf{z}\|_2 \cdot \|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2$.

Towards an Iterative Method for LS

$$f(\mathbf{x}') = f(\mathbf{x}) + 2\eta \mathbf{w}^\top \mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y}) + \eta^2\|\mathbf{A}\mathbf{w}\|_2^2.$$

Plugging-in our choice $\mathbf{w} = -\frac{\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})}{\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2}$ we get

$$\begin{aligned} f(\mathbf{x}') &= f(\mathbf{x}) - 2\eta\frac{\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^2}{\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2} + \eta^2\|\mathbf{A}\frac{\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})}{\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2}\|_2^2 \\ &= f(\mathbf{x}) - 2\eta\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2 + \eta^2\frac{\|\mathbf{A}\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^2}{\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2}. \end{aligned}$$

Now, let us turn to choose the step-size η. Note we got a "smiling" parabola in η. Thus, the η that achieves the minimum is

$$\eta = 2\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2/2\frac{\|\mathbf{A}\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^2}{\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2} = \frac{\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^3}{\|\mathbf{A}\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^2}.$$

Plugging back the value of η we get

$$f(\mathbf{x}') = f(\mathbf{x}) - \frac{\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^4}{\|\mathbf{A}\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^2}.$$

Towards an Iterative Method for LS

We got:

$$\begin{aligned} f(\mathbf{x}') &= f(\mathbf{x}) - \frac{\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^4}{\|\mathbf{A}\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^2} \leq f(\mathbf{x}) - \frac{1}{\sigma_{\max}(\mathbf{A})^2}\frac{\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^4}{\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^2} \\ &= f(\mathbf{x}) - \frac{1}{\sigma_{\max}(\mathbf{A})^2}\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^2. \end{aligned}$$

Let us now make an observation: note that $\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^2 = 0$ if and only if $\mathbf{A}^\top\mathbf{A}\mathbf{x} = \mathbf{A}^\top\mathbf{y}$. That is, if and only if $\mathbf{x}$ is a solution to the normal equation of the LS which is true if and only if $\mathbf{x}$ is an optimal solution to the LS.
Thus, if $\mathbf{x}$ is not an optimal solution to the LS we have that $\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^2 > 0$ and thus indeed $f(\mathbf{x}') < f(\mathbf{x})$ and our method indeed reduces the function value on each iteration.

Nevertheless, as in the leading eigenvalue problem, we would like to get an explicit convergence rate, that is bound on number of iterations required to get desired approximation error.

Towards an Iterative Method for LS

We got:

$$f(\mathbf{x}') \leq f(\mathbf{x}) - \frac{1}{\sigma_{\max}(\mathbf{A})^2}\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^2.$$

As a preliminary result, let us assume **for now** that $\mathbf{A}$ has full row-rank. In this case, the $m \times m$ matrix $\mathbf{A}\mathbf{A}^\top$ is **positive definite**. In this case using **Rayleigh's Theorem** we have

$$\begin{aligned} \|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^2 &= (\mathbf{A}\mathbf{x}-\mathbf{y})^\top\mathbf{A}\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y}) \geq \lambda_{\min}(\mathbf{A}\mathbf{A}^\top)\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 \\ &= \sigma_{\min}(\mathbf{A})^2\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2. \end{aligned}$$

Here we denote by $\sigma_{\min}(\mathbf{A})$ - the smallest non-zero singular value of $\mathbf{A}$.

Thus, when $\mathbf{A}$ is full row-rank we get

$$\|\mathbf{A}\mathbf{x}'-\mathbf{y}\|_2^2 \leq \|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 - \frac{\sigma_{\min}(\mathbf{A})^2}{\sigma_{\max}(\mathbf{A})^2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 = \left(1 - \frac{\sigma_{\min}(\mathbf{A})^2}{\sigma_{\max}(\mathbf{A})^2}\right)\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2.$$

Thus, if $\mathbf{x}_t$ is the point we get after t such updates, starting from $\mathbf{x}_0$ we will have

$$\|\mathbf{A}\mathbf{x}_t-\mathbf{y}\|_2^2 \leq \left(1 - \frac{\sigma_{\min}(\mathbf{A})^2}{\sigma_{\max}(\mathbf{A})^2}\right)^t\|\mathbf{A}\mathbf{x}_0-\mathbf{y}\|_2^2 \leq \|\mathbf{A}\mathbf{x}_0-\mathbf{y}\|_2^2 \cdot \exp\left(-\frac{\sigma_{\min}(\mathbf{A})^2}{\sigma_{\max}(\mathbf{A})^2}t\right).$$

Towards an Iterative Method for LS

Under the assumption that $\mathbf{A}$ is full row-rank, we got exponential convergence: after t updates of the form we discussed we get to $\mathbf{x}_t$ satisfying:

$$\|\mathbf{A}\mathbf{x}_t - \mathbf{y}\|_2^2 \leq \|\mathbf{A}\mathbf{x}_0 - \mathbf{y}\|_2^2 \cdot \exp\left(-\frac{\sigma_{\min}(\mathbf{A})^2}{\sigma_{\max}(\mathbf{A})^2} t\right).$$

Note that if $\mathbf{A}$ is full row-rank then $\mathrm{rank}(\mathbf{A}) = m$ meaning $\text{column-span}(\mathbf{A}) = \mathbb{R}^m$ and thus in this case $\mathbf{y} \in \mathrm{Im}(\mathbf{A})$ and the optimal value of the LS problem is 0 (i.e., the linear system $\mathbf{A}\mathbf{x} = \mathbf{y}$ has a solution).

We now continue to discuss the general case in which $\mathbf{A}$ is not necessarily full row-rank. That is, we can't use the inequality:

$$\|\mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{y})\|_2^2 \geq \lambda_{\min}(\mathbf{A}\mathbf{A}^\top)\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2,$$

since $\lambda_{\min}(\mathbf{A}\mathbf{A}^\top)$ might be zero, and thus the above will not give reduction in function value.

Towards an Iterative Method for LS

We got:

$$f(\mathbf{x}') \leq f(\mathbf{x}) - \frac{1}{\sigma_{\max}(\mathbf{A})^2}\|\mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{y})\|_2^2.$$

Recall we no longer assume $\mathbf{A}$ is full row-rank.

Let $\mathbf{x}^*$ be an optimal solution to the LS. Thus, $\mathbf{x}^*$ satisfies the normal equation: $\mathbf{A}^\top\mathbf{A}\mathbf{x}^* = \mathbf{A}^\top\mathbf{y}$. We thus have that

$$\|\mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{y})\|_2^2 = \|\mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*)\|_2^2 = \|\mathbf{A}^\top\mathbf{A}(\mathbf{x} - \mathbf{x}^*)\|_2^2$$

We are going to show the following two claims hold:

$$i.\ \|\mathbf{A}^\top\mathbf{A}(\mathbf{x} - \mathbf{x}^*)\|_2^2 \geq \sigma_{\min}(\mathbf{A})^2 \cdot \|\mathbf{A}(\mathbf{x} - \mathbf{x}^*)\|_2^2$$
$$ii.\ \|\mathbf{A}(\mathbf{x} - \mathbf{x}^*)\|_2^2 = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|_2^2.$$

Combined they give:

$$\|\mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{y})\|_2^2 \geq \sigma_{\min}(\mathbf{A})^2 \cdot \left(\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|_2^2\right)$$

Thus, from the equation above we will again get $f(\mathbf{x}') < f(\mathbf{x})$.

Towards an Iterative Method for LS

We are going to show the following two claims hold:

$$i.\ \|\mathbf{A}^\top\mathbf{A}(\mathbf{x}-\mathbf{x}^*)\|_2^2 \geq \sigma_{\min}(\mathbf{A})^2 \cdot \|\mathbf{A}(\mathbf{x}-\mathbf{x}^*)\|_2^2$$
$$ii.\ \|\mathbf{A}(\mathbf{x}-\mathbf{x}^*)\|_2^2 = \|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}^*-\mathbf{y}\|_2^2.$$

For the first claim, you are going to prove (HW) that:

$$\forall \mathbf{z}: \quad \|\mathbf{A}^\top\mathbf{A}\mathbf{z}\|_2^2 \geq \sigma_{\min}(\mathbf{A})^2\|\mathbf{A}\mathbf{z}\|_2^2.$$

Towards an Iterative Method for LS

We are going to show the following two claims hold:

$$i.\ \|\mathbf{A}^\top\mathbf{A}(\mathbf{x}-\mathbf{x}^*)\|_2^2 \geq \sigma_{\min}(\mathbf{A})^2 \cdot \|\mathbf{A}(\mathbf{x}-\mathbf{x}^*)\|_2^2$$
$$ii.\ \|\mathbf{A}(\mathbf{x}-\mathbf{x}^*)\|_2^2 = \|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}^*-\mathbf{y}\|_2^2.$$

For the second claim we have

$$\begin{aligned}\|\mathbf{A}(\mathbf{x}-\mathbf{x}^*)\|_2^2 &= \|(\mathbf{A}\mathbf{x}-\mathbf{y})-(\mathbf{A}\mathbf{x}^*-\mathbf{y})\|_2^2 = \|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 + \|\mathbf{A}\mathbf{x}^*-\mathbf{y}\|_2^2 \\ &\quad -2(\mathbf{A}\mathbf{x}-\mathbf{y})^\top(\mathbf{A}\mathbf{x}^*-\mathbf{y}).\end{aligned}$$

Also

$$\begin{aligned}(\mathbf{A}\mathbf{x}-\mathbf{y})^\top(\mathbf{A}\mathbf{x}^*-\mathbf{y}) &= (\mathbf{A}\mathbf{x}^*-\mathbf{y})^\top(\mathbf{A}\mathbf{x}^*-\mathbf{y}) + (\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*)^\top(\mathbf{A}\mathbf{x}^*-\mathbf{y}) \\ &= \|\mathbf{A}\mathbf{x}^*-\mathbf{y}\|_2^2 + (\mathbf{x}-\mathbf{x}^*)^\top(\mathbf{A}^\top\mathbf{A}\mathbf{x}^*-\mathbf{A}^\top\mathbf{y}) = \|\mathbf{A}\mathbf{x}^*-\mathbf{y}\|_2^2,\end{aligned}$$

where last equality follows since $\mathbf{x}^*$ is a solution to the normal equation. Combining two last equations we indeed get

$$\|\mathbf{A}(\mathbf{x}-\mathbf{x}^*)\|_2^2 = \|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}^*-\mathbf{y}\|_2^2.$$

Towards an Iterative Method for LS

We have showed:

$$i. f(\mathbf{x}') \leq f(\mathbf{x}) - \frac{1}{\sigma_{\max}(\mathbf{A})^2}\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^2$$
$$ii. \|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^2 \geq \sigma_{\min}(\mathbf{A})^2 \cdot \left(\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}^*-\mathbf{y}\|_2^2\right).$$

Combining we get

$$\|\mathbf{A}\mathbf{x}'-\mathbf{y}\|_2^2 \leq \|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 - \frac{\sigma_{\min}(\mathbf{A})^2}{\sigma_{\max}(\mathbf{A})^2}\left(\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}^*-\mathbf{y}\|_2^2\right).$$

Subtracting $\|\mathbf{A}\mathbf{x}^*-\mathbf{y}\|_2^2$ from both sides we have

$$\|\mathbf{A}\mathbf{x}'-\mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}^*-\mathbf{y}\|_2^2 \leq \left(1 - \frac{\sigma_{\min}(\mathbf{A})^2}{\sigma_{\max}(\mathbf{A})^2}\right)\left(\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}^*-\mathbf{y}\|_2^2\right).$$

Thus, as before, after t such updates we will reach a point $\mathbf{x}_t$ such that

$$\|\mathbf{A}\mathbf{x}_t-\mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}^*-\mathbf{y}\|_2^2 \leq \left(\|\mathbf{A}\mathbf{x}_0-\mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}^*-\mathbf{y}\|_2^2\right)\exp\left(-\frac{\sigma_{\min}(\mathbf{A})^2}{\sigma_{\max}(\mathbf{A})^2}t\right)$$

Iterative Method for LS

Recall our update to a point $\mathbf{x}$ was $\mathbf{x}' = \mathbf{x} + \eta\mathbf{w}$ for

$$\mathbf{w} = -\frac{\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})}{\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2}, \quad \eta = \frac{\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^3}{\|\mathbf{A}\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^2}.$$

This results in the update:

$$\mathbf{x}' = \mathbf{x} - \frac{\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^2}{\|\mathbf{A}\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})\|_2^2}\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y}).$$

Note that given $\mathbf{x}$, computing $\mathbf{x}'$ requires a finite amount of operation that either multiply a vector with $\mathbf{A}$, or add vectors in $\mathbb{R}^m, \mathbb{R}^n$, or basic arithmetics with scalars. Thus, obtaining $\mathbf{x}'$ from $\mathbf{x}$ can be carried out in overall $O(N+n+m)$ time.

Iterative Method for LS

Our efforts have accumulated to the following theorem.

Theorem

Let $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{y} \in \mathbb{R}^m$. Consider a sequence of points in $\mathbb{R}^n$, $\{\mathbf{x}_t\}_{t \geq 1}$ produced by the updates:

$$\forall t \geq 1: \quad \mathbf{x}_t = \mathbf{x}_{t-1} - \frac{\|\mathbf{A}^\top(\mathbf{A}\mathbf{x}_{t-1} - \mathbf{y})\|_2^2}{\|\mathbf{A}\mathbf{A}^\top(\mathbf{A}\mathbf{x}_{t-1} - \mathbf{y})\|_2^2} \mathbf{A}^\top(\mathbf{A}\mathbf{x}_{t-1} - \mathbf{y}),$$

where $\mathbf{x}_0$ is some arbitrary point in $\mathbb{R}^n$. Then, it holds for all $t \geq 1$ that

$$\|\mathbf{A}\mathbf{x}_t - \mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|_2^2 \leq \left(\|\mathbf{A}\mathbf{x}_0 - \mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|_2^2\right) \exp\left(-\frac{\sigma_{\min}(\mathbf{A})^2}{\sigma_{\max}(\mathbf{A})^2} t\right),$$

where $\mathbf{x}^*$ is an optimal solution to $\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$.

Moreover, computing $\mathbf{x}_t$ from $\mathbf{x}_{t-1}$ can be carried-out in $O(N + n + m)$ time, where N is the number of non-zero entries in $\mathbf{A}$.

Iterative Method for LS

The previous theorem establishes the speed of convergence of the sequence $\{\mathbf{x}_t\}_{t \geq 1}$ to the optimal value of the LS problem. It is often of interest to establish the speed of convergence to the optimal solution itself, i.e., to $\mathbf{x}^*$ (if possible).

Theorem

Suppose the matrix $\mathbf{A}$ is full column-rank. Then, it holds that the solution to the LS problem is unique and given by $\mathbf{x}^* = \mathbf{A}^\dagger \mathbf{y}$. Moreover, for any $\mathbf{x} \in \mathbb{R}^n$ it holds that

$$\|\mathbf{x} - \mathbf{x}^*\|_2^2 \leq \frac{1}{\sigma_{\min}(\mathbf{A})^2} \left(\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|_2^2\right).$$

In particular, if $\mathbf{A}$ is full-column rank, then the sequence $\{\mathbf{x}_t\}_{t \geq 1}$ generated by the iterative algorithm above satisfies:

$$\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 \leq \frac{1}{\sigma_{\min}(\mathbf{A})^2} \left(1 - \left(\frac{\sigma_{\min}(\mathbf{A})}{\sigma_{\max}(\mathbf{A})}\right)^2\right)^t \left(\|\mathbf{A}\mathbf{x}_0 - \mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|_2^2\right)$$

Proving the theorem

We have already seen that when $\mathbf{A}$ has full-column rank that $\mathbf{x}^* = \mathbf{A}^\dagger \mathbf{y}$ is the unique optimal solution.

Recall we have seen during the proof of the previous theorem that

$$\|\mathbf{A}(\mathbf{x} - \mathbf{x}^*)\|_2^2 = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|_2^2.$$

Recall that via Rayleigh's theorem, it holds that

$$\begin{aligned} \|\mathbf{A}(\mathbf{x} - \mathbf{x}^*)\|_2^2 &= (\mathbf{x} - \mathbf{x}^*)^\top \mathbf{A}^\top \mathbf{A}(\mathbf{x} - \mathbf{x}^*) \geq \lambda_{\min}(\mathbf{A}^\top \mathbf{A}) \cdot \|\mathbf{x} - \mathbf{x}^*\|_2^2 \\ &= \sigma_{\min}(\mathbf{A})^2 \cdot \|\mathbf{x} - \mathbf{x}^*\|_2^2. \end{aligned}$$

Finally, since $\mathbf{A}$ is full column-rank, it follows that $\mathbf{A}^\top \mathbf{A}$ is positive definite, and hence $\lambda_{\min}(\mathbf{A}^\top \mathbf{A}) = \sigma_{\min}(\mathbf{A})^2 > 0$.

AIgebraic Methods in Data Science: Lesson 10

Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology

Dan Garber
https://dangar.net.technion.ac.il/

Winter Semester 2020-2021

Application of Least Squares: Linear Regression

Suppose we want to predict a scalar quantity $y \in \mathbb{R}$ from some related scalar features $a_1, \ldots a_n \in \mathbb{R}$. For instance:

1. y is the price of a home on the real estate market and the features are a_1 - size of house in sqrm, a_2 - floor, a_3 - social economic status of neighbourhood, and so on.
2. y is life expectancy of an individual, and the features are a_1 - average income, a_2 - age, a_3 - number of degrees from Technion, and so on.
3. y is the price of the Apple stock on day t, and $a_1, \ldots, a_n$ are the stock prices of leading tech companies on day $t-1$.

Application of Least Squares: Linear Regression

Suppose we want to predict a scalar quantity $y \in \mathbb{R}$ from some related scalar features $a_1, \ldots a_n \in \mathbb{R}$. For instance:

1. y is the price of a home on the real estate market and the features are a_1 - size of house in sqrm, a_2 - floor, a_3 - social economic status of neighbourhood, and so on.
2. y is life expectancy of an individual, and the features are a_1 - average income, a_2 - age, a_3 - number of degrees from Technion, and so on.
3. y is the price of the Apple stock on day t, and $a_1, \ldots, a_n$ are the stock prices of leading tech companies on day $t-1$.

We can try to predict y by a **linear** predictor (i.e., linear combination of features):

$$\hat{y} = \sum_{i=1}^{n} w_i \cdot a_i + w_0.$$

$w_1, \ldots, w_n$ are the weights given to the different features, and w_0 is and additional bias term.

Application of Least Squares: Linear Regression

Suppose we want to predict a scalar quantity $y \in \mathbb{R}$ from some related scalar features $a_1, \ldots a_n \in \mathbb{R}$.

We can try to predict y by a **linear** predictor:

$$\hat{y} = \sum_{i=1}^{n} w_i \cdot a_i + w_0.$$

$w_1, \ldots, w_n$ are the weights given to the different features, and w_0 is and additional bias term.

Why linear predictor? simple, intuitive and interpretable, and easy to solve (as we shall see).

Main challenge: how do we find the "correct" weights $w_0, \ldots, w_n$?
Solution: we shall assume we are given data with correct labels: $(y^{(1)}, a_1^{(1)}, \ldots, a_n^{(1)}), \ldots, (y^{(m)}, a_1^{(m)}, \ldots, a_n^{(m)})$, and we will find weights $w_0, \ldots, w_n$ that "explain" these data best.

Application of Least Squares: Linear Regression

Given data with correct labels: $(y^{(1)}, a_1^{(1)}, \ldots, a_n^{(1)}), \ldots, (y^{(m)}, a_1^{(m)}, \ldots, a_n^{(m)})$, we would like to find the weights $w_0, \ldots, w_n$ that "explain" these data best.
These weights can then be used to predict the labels of new unlabeled instances (given only by the features $a_1, \ldots, a_n$).

First attempt: Since we are trying to find a linear predictor of the form $\hat{y} = \sum_{i=1}^{n} w_i \cdot a_i + w_0$, we can try and find weights $w_0, \ldots, w_n$ such that:

$$\forall j \in \{1, \ldots, m\}: \quad y^{(j)} = \sum_{i=1}^{n} w_i \cdot a_i^{(j)} + w_0.$$

This is actually a system of m linear equation in $n+1$ unknowns.

What is wrong with this approach?
We are trying to predict the labels (y) with a linear predictor, but who said that REALITY is indeed (exactly) linear?

Application of Least Squares: Linear Regression

Given data with correct labels: $(y^{(1)}, a_1^{(1)}, \ldots, a_n^{(1)})$, $\ldots$, $(y^{(m)}, a_1^{(m)}, \ldots, a_n^{(m)})$, we would like to find the weights $w_0, \ldots, w_n$ that "explain" these data best.
These weights can then be used to predict the labels of new unlabeled instances (given only by the features $a_1, \ldots, a_n$).

Second attempt: try to find weights $w_0, \ldots, w_n$, that **best approximate** the data:

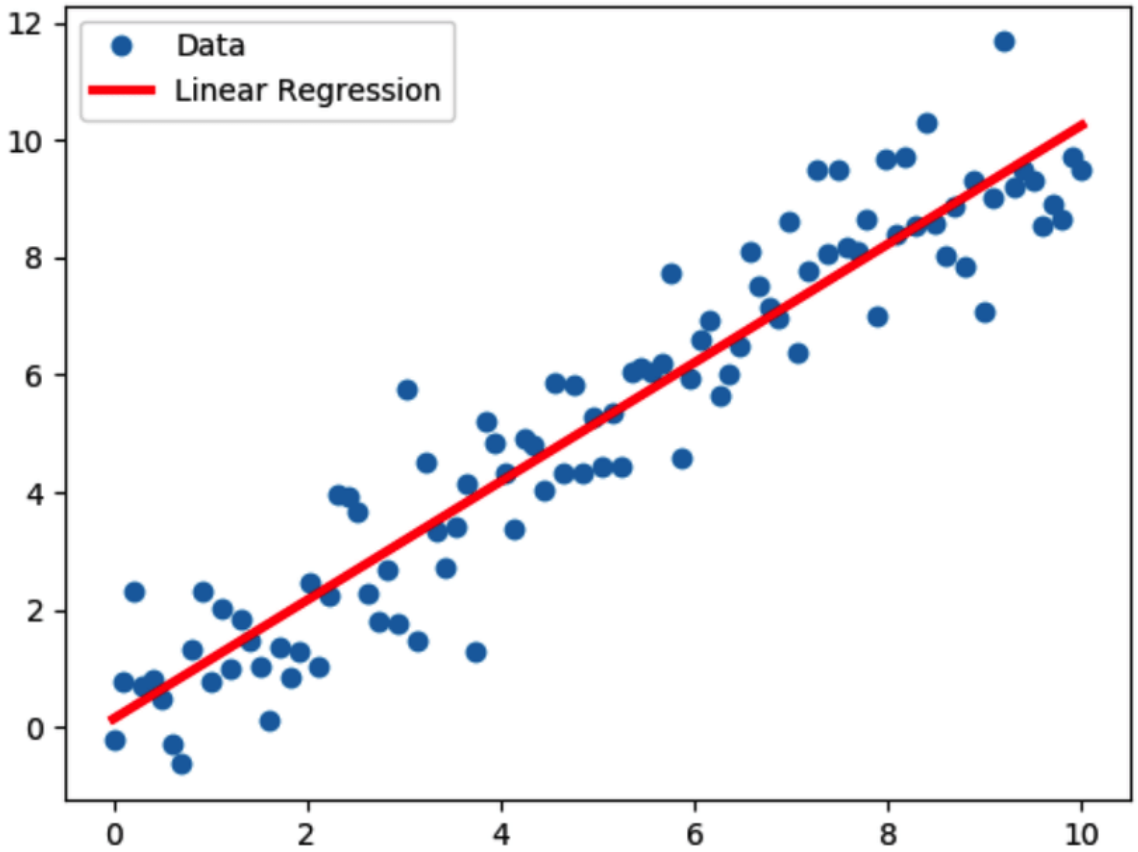
$$\min_{w_0,\ldots,w_n \in \mathbb{R}} \sum_{j=1}^{m} \left(y^{(j)} - \left[\sum_{i=1}^{n} w_i \cdot a_i^{(j)} + w_0 \right] \right)^2 .$$

That is, find $w_0, \ldots, w_n$ that minimize the square error.

In particular, no assumption that the data $\{(y^{(j)}, a_1^{(j}, \ldots, a_n^{(j)})\}_{j=1}^{m}\}$ exactly matches some linear predictor.

Application of Least Squares: Linear Regression

2D illustration: consider the case that we only have a single feature a. In this case the linear predictor takes the form $\hat{y} = w_1 \cdot a + w_0$.
That is, the mapping from feature to label is a line in the plane.
We want to find such a line that approximates the data best.



Application of Least Squares: Linear Regression

We can find weights $w_0, \ldots, w_n$ by solving the problem:

$$\min_{w_0,\ldots,w_n \in \mathbb{R}} \sum_{j=1}^{m} \left(y^{(j)} - \left[\sum_{i=1}^{n} w_i \cdot a_i^{(j)} + w_0 \right] \right)^2 . \tag{1}$$

Consider the $m \times (n+1)$ matrix $\mathbf{A}$, vector $\mathbf{w}$ of length $(n+1)$ and vector $\mathbf{y}$ of length m given by:

$$\mathbf{A} = \begin{pmatrix} 1 & a_1^{(1)} & a_2^{(1)} & \ldots & a_n^{(1)} \\ 1 & a_1^{(2)} & a_2^{(2)} & \ldots & a_n^{(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & a_1^{(m)} & a_2^{(m)} & \ldots & a_n^{(m)} \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_m \end{pmatrix}.$$

Letting $\mathbf{A}_i$ denote the ith row of $\mathbf{A}$, we now see that (1) is the same as:

$$\min_{\mathbf{w} \in \mathbb{R}^{n+1}} \sum_{j=1}^{m} \left(y^{(j)} - \mathbf{A}_i^\top \mathbf{w} \right)^2 = \min_{\mathbf{w} \in \mathbb{R}^{n+1}} \|\mathbf{A}\mathbf{w} - \mathbf{y}\|_2^2.$$

Thus, we have arrived at a Least Squares formulation.

Application of Least Squares: Linear Regression

We have seen that we can find a linear predictor of the form $\hat{\mathbf{y}} = \sum_{i=1}^{n} w_i \cdot a_i + w_0$ that best matches given data (in square error) by solving the corresponding least squares problem $\min_{\mathbf{w}} \|\mathbf{A}\mathbf{w} - \mathbf{y}\|_2^2$, where $\mathbf{A}, \mathbf{y}$ encode the available labeled data.

This is called Linear Regression and it is the single most important data analysis tool in science and engineering.

Often the case is that m - number of data points available, satisfies $m \geq n + 1$, since otherwise we will usually have infinite solutions to the LS problem and it won't make sense to pick one over the other for making future predictions.

In case $\mathbf{A}$ has linearly independent columns (which is usually the case if $m \geq n + 1$) then we know the LS has a unique solution and it is given by $\mathbf{w}^* = \mathbf{A}^\dagger \mathbf{y} = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{y}$.

Recursive Least Squares

Suppose now that we want to build some system that uses linear regression for prediction. For instance, a system to predict the values of homes on the real estate market.

As before, given some data (list of homes with their features and "correct" price), we can use least squares to find a linear predictor for predicting the prices of new homes on the market based on their features.

Suppose now that every once in a while we get new labeled data. For instance, houses are being sold on the market and we can see the selling price (the "correct" label). We can use these new data to come up with a (potentially) more accurate predictor, because it is based on more data.

However, such an approach will require to solve a LS problem every time new data arrives. This can be expensive if data arrives frequently and we need to frequently solve a LS. Also, the time to solve the LS grows with m - number of data points, which keeps increasing!

Recursive Least Squares

Formal model for sequential least squares:
We begin with some matrix $\mathbf{A}_0 \in \mathbb{R}^{m_0 \times n}$ and vector $\mathbf{y}_0 \in \mathbb{R}^n$.

Each (integer) time step $t = 1, 2, \ldots,$ we get a new vector $\mathbf{a}_t \in \mathbb{R}^n$ and new scalar y_t.

Our goal is on each time t to produce a vector $\mathbf{x}_t \in \mathbb{R}^n$ which is a solution to the new least squares problem:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{A}_t \mathbf{x} - \mathbf{y}_t\|_2^2, \quad \text{for} \quad \mathbf{A}_t := \begin{pmatrix} \mathbf{A}_{t-1} \\ \mathbf{a}_t^\top \end{pmatrix}, \mathbf{y}_t := \begin{pmatrix} \mathbf{y}_{t-1} \\ y_t \end{pmatrix}.$$

Question: Suppose that on some time t we have some $\mathbf{x}_t$ which is an exact solution to the the LS $\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{A}_t \mathbf{x} - \mathbf{y}_t\|_2^2$. Is there a way to quickly update $\mathbf{x}_t$ to $\mathbf{x}_{t+1}$ - a solution to the next LS $\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{A}_{t+1} \mathbf{x} - \mathbf{y}_{t+1}\|_2^2$?

Hope: the LS on time $t+1$ differs from that of time t only in the last row in $\mathbf{A}_{t+1}, \mathbf{y}_{t+1}$.
Answer: Yes we can! We will show an algorithm with fast updates.

The Recursive Least Squares Algorithm

The following lemma will be a main technical ingredient in our Recursive LS (RLS) algorithm.

Lemma (Sherman-Morrison formula)

Let $\mathbf{M} \in \mathbb{R}^{n \times n}$ *be an invertible matrix and let* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. *Then* $\mathbf{M} + \mathbf{u}\mathbf{v}^\top$ *is invertible if and only if* $1 + \mathbf{v}^\top \mathbf{M}^{-1}\mathbf{u} \neq 0$, *and if* $\mathbf{M} + \mathbf{u}\mathbf{v}^\top$ *is invertible, its inverse is given by*

$$(\mathbf{M} + \mathbf{u}\mathbf{v}^\top)^{-1} = \mathbf{M}^{-1} - \frac{\mathbf{M}^{-1}\mathbf{u}\mathbf{v}^\top\mathbf{M}^{-1}}{1 + \mathbf{v}^\top\mathbf{M}^{-1}\mathbf{u}}. \tag{2}$$

In particular, note that if $\mathbf{M}^{-1}$ *is given explicitly, one can compute* $(\mathbf{M} + \mathbf{u}\mathbf{v}^\top)^{-1}$ *in only* $O(n^2)$ *time (instead of* $O(n^3)$*).*

Proof: First note that if $\mathbf{M}^{-1}$ is given, then by computing $\mathbf{M}^{-1}\mathbf{u}$, $\mathbf{v}^\top\mathbf{M}^{-1}$, each in $O(n^2)$ time, we can indeed compute the RHS of (2) in $O(n^2)$ time.

The Recursive Least Squares Algorithm

Lemma (Sherman-Morrison formula)

Let $\mathbf{M} \in \mathbb{R}^{n \times n}$ be an invertible matrix and let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. Then $\mathbf{M} + \mathbf{u}\mathbf{v}^\top$ is invertible if and only if $1 + \mathbf{v}^\top \mathbf{M}^{-1} \mathbf{u} \neq 0$, and if $\mathbf{M} + \mathbf{u}\mathbf{v}^\top$ is invertible, its inverse is given by

$$(\mathbf{M} + \mathbf{u}\mathbf{v}^\top)^{-1} = \mathbf{M}^{-1} - \frac{\mathbf{M}^{-1}\mathbf{u}\mathbf{v}^\top \mathbf{M}^{-1}}{1 + \mathbf{v}^\top \mathbf{M}^{-1}\mathbf{u}}. \tag{3}$$

In particular, note that if $\mathbf{M}^{-1}$ is given explicitly, one can compute $(\mathbf{M} + \mathbf{u}\mathbf{v}^\top)^{-1}$ in only $O(n^2)$ time (instead of $O(n^3)$).

Proof cont.: If $1 + \mathbf{v}^\top \mathbf{M}^{-1} \mathbf{u} \neq 0$ then it is straightforward to check that the RHS of (3) is indeed the inverse of $\mathbf{M} + \mathbf{u}\mathbf{v}^\top$.

If $1 + \mathbf{v}^\top \mathbf{M}^{-1} \mathbf{u} = 0$, then denoting $\mathbf{z} = \mathbf{M}^{-1}\mathbf{u}$ we have that $(\mathbf{M} + \mathbf{u}\mathbf{v}^\top)\mathbf{z} = \mathbf{u} + \mathbf{u}\mathbf{v}^\top \mathbf{M}^{-1}\mathbf{u} = \mathbf{u}(1 + \mathbf{v}^\top \mathbf{M}^{-1}\mathbf{u}) = \mathbf{0}$. Thus, the Kernel of $\mathbf{M} + \mathbf{u}\mathbf{v}^\top$ is not trivial and thus it is not invertible.

The Recursive Least Squares Algorithm

Suppose that the initial matrix in the RLS problem, $\mathbf{A}_0 \in \mathbb{R}^{m_0 \times n}$ has **linearly independent columns**.
Note this implies that for all t, $\mathbf{A}_t$ has linearly independent columns.

In such a case, we know that $\forall t$: $\mathbf{x}_t = (\mathbf{A}_t^\top \mathbf{A}_t)^{-1} \mathbf{A}_t^\top \mathbf{y}_t$.

Recall: $\mathbf{A}_t = (\mathbf{A}_{t-1}^\top \ \mathbf{a}_t)^\top$. Thus,

$$\mathbf{A}_t^\top \mathbf{A}_t = (\mathbf{A}_{t-1}^\top \ \mathbf{a}_t) \begin{pmatrix} \mathbf{A}_{t-1} \\ \mathbf{a}_t^\top \end{pmatrix} = \mathbf{A}_{t-1}^\top \mathbf{A}_{t-1} + \mathbf{a}_t \mathbf{a}_t^\top.$$

Thus, using Sherman-Morison (with $\mathbf{M} = \mathbf{A}_{t-1}^\top \mathbf{A}_{t-1}$, $\mathbf{u} = \mathbf{v} = \mathbf{a}_t$), we have that

$$(\mathbf{A}_t^\top \mathbf{A}_t)^{-1} = (\mathbf{A}_{t-1}^\top \mathbf{A}_{t-1})^{-1} - \frac{(\mathbf{A}_{t-1}^\top \mathbf{A}_{t-1})^{-1} \mathbf{a}_t \mathbf{a}_t^\top (\mathbf{A}_{t-1}^\top \mathbf{A}_{t-1})^{-1}}{1 + \mathbf{a}_t^\top (\mathbf{A}_{t-1}^\top \mathbf{A}_{t-1})^{-1} \mathbf{a}_t}.$$

Thus, if we have $(\mathbf{A}_{t-1}^\top \mathbf{A}_{t-1})^{-1}$ we can compute $(\mathbf{A}_t^\top \mathbf{A}_t)^{-1}$ in $O(n^2)$ time (instead of the standard $O(n^3)$ time)!!

The Recursive Least Squares Algorithm

Suppose that the initial matrix in the RLS problem, $\mathbf{A}_0 \in \mathbb{R}^{m_0 \times n}$ has **linearly independent columns**.
Note this implies that for all t, $\mathbf{A}_t$ has linearly independent columns.

In such a case, we know that $\forall t$: $\mathbf{x}_t = (\mathbf{A}_t^\top \mathbf{A}_t)^{-1} \mathbf{A}_t^\top \mathbf{y}_t$.

We have seen that if we keep $(\mathbf{A}_{t-1}^\top \mathbf{A}_{t-1})^{-1}$ in memory we can compute $(\mathbf{A}_t^\top \mathbf{A}_t)^{-1}$ in $O(n^2)$ time.

Note also that

$$\mathbf{A}_t^\top \mathbf{y}_t = (\mathbf{A}_{t-1}^\top \ \mathbf{a}_t) \begin{pmatrix} \mathbf{y}_{t-1} \\ y_t \end{pmatrix} = \mathbf{A}_{t-1}^\top \mathbf{y}_{t-1} + y_t \mathbf{a}_t$$

Thus, if we store $\mathbf{A}_{t-1}^\top \mathbf{y}_{t-1} \in \mathbb{R}^n$ in memory, we can compute $\mathbf{A}_t^\top \mathbf{y}_t$ in $O(n)$ time!

Conclusion: if on any time $t-1$ we store $(\mathbf{A}_{t-1}^\top \mathbf{A}_{t-1})^{-1}, \mathbf{A}_{t-1}^\top \mathbf{y}_{t-1}$ in memory, we can compute $\mathbf{x}_t$ and $(\mathbf{A}_t^\top \mathbf{A}_t)^{-1}, \mathbf{A}_t^\top \mathbf{y}_t$ in $O(n^2)$ time!!!

The Recursive Least Squares Algorithm

Here is the complete algorithm:

1. Given initalization $\mathbf{A}_0, \mathbf{y}_0$, compute $(\mathbf{A}_0^\top \mathbf{A}_0)^{-1}$, $\mathbf{A}_0^\top \mathbf{y}_0$ and output LS solution $\mathbf{x}_0 = (\mathbf{A}_0^\top \mathbf{A}_0)^{-1}\mathbf{A}_0^\top \mathbf{y}_0$
2. for all $t \geq 1$:
   1. observe new data $\mathbf{a}_t \in \mathbb{R}^n$, $y_t \in \mathbb{R}$
   2. compute $(\mathbf{A}_t^\top \mathbf{A}_t)^{-1}$ in $O(n^2)$ time via

$$(\mathbf{A}_t^\top \mathbf{A}_t)^{-1} = (\mathbf{A}_{t-1}^\top \mathbf{A}_{t-1})^{-1} - \frac{(\mathbf{A}_{t-1}^\top \mathbf{A}_{t-1})^{-1}\mathbf{a}_t\mathbf{a}_t^\top(\mathbf{A}_{t-1}^\top \mathbf{A}_{t-1})^{-1}}{1 + \mathbf{a}_t^\top(\mathbf{A}_{t-1}^\top \mathbf{A}_{t-1})^{-1}\mathbf{a}_t}.$$

   3. compute $\mathbf{A}_t^\top \mathbf{y}_t = \mathbf{A}_{t-1}^\top \mathbf{y}_{t-1} + y_t\mathbf{a}_t$ in $O(n)$ time
   4. output solution to $\min_{\mathbf{x}} \|\mathbf{A}_t\mathbf{x} - \mathbf{y}_t\|_2^2$ by computing $\mathbf{x}_t = (\mathbf{A}_t^\top \mathbf{A}_t)^{-1}(\mathbf{A}_t\mathbf{y}_t)$ in $O(n^2)$ time.

Note that on anytime t the algorithm only requires to store in memory the matrix $(\mathbf{A}_{t-1}^\top \mathbf{A}_{t-1})^{-1} \in \mathbb{R}^{n\times n}$ and the vector $\mathbf{A}_{t-1}^\top \mathbf{y}_{t-1} \in \mathbb{R}^n$, and thus requires only $O(n^2)$ memory - independent of the overall size of the data!!

Algebraic Methods in Data Science: Lesson 11

Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology

Dan Garber
https://dangar.net.technion.ac.il/

Winter Semester 2020-2021

Perturbation Bounds

In this part of the course we study the following two related questions:

1. For two given real matrices $\mathbf{X}, \mathbf{Y}$, what can be said about the distance between their decompositions (either spectral or SVD) as a function of the distance $\mathbf{X} - \mathbf{Y}$?
2. For a linear system (or least squares problem), how much does the solution changes if we perturb a little bit the coefficients matrix and the output vector?

Both of these questions are motivated both from an information perspective and a computational perspective.

Perturbation Bounds

Take for instance the task of computing the top k principal components of a data matrix $\mathbf{X} \in \mathbb{R}^{n \times N}$ (data points are columns).

As we have seen, computing these k vectors amounts to computing the top k eigenvectors of the covariance matrix $\mathbf{X}\mathbf{X}^\top$.

Now, suppose that $\mathbf{X}$ contains the medical data of all people of Israel (each person's medical data is a column). Clearly, no-one has $\mathbf{X}$.

However, it is reasonable to obtain a matrix $\mathbf{Y} \in \mathbb{R}^{n \times m}$, where $m << N$ (for instance, by asking for the records of maybe only 50,000-500,000 people).

It is also many times reasonable (and can even be made rigorous) to assume that $\mathbf{X}\mathbf{X}^\top \approx \mathbf{Y}\mathbf{Y}^\top$ (note that $\mathbf{X}, \mathbf{Y}$ themselves are of different dimension).

The question now is: if we compute the k leading eigenvectors of $\mathbf{Y}\mathbf{Y}^\top$, how close is the corresponding subspace going to be to the subspace of the k leading eigenvectors of $\mathbf{X}\mathbf{X}^\top$?

Perturbation Bounds

Take for instance the task of computing the top k principal components of a data matrix $\mathbf{X} \in \mathbb{R}^{n \times N}$ (data points are columns).

From a computational perspective, even if we had $\mathbf{X}$, computing the top k eigenvectors of $\mathbf{X}\mathbf{X}^\top = \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\top$ requires as a first step to compute the matrix $\mathbf{X}\mathbf{X}^\top$.

This operation alone takes $O(Nn^2)$ time which is very expensive when $N >> n$.

However, computing the matrix $\mathbf{Y}\mathbf{Y}^\top$ takes only $O(mn^2)$ time (where m is number of columns in $\mathbf{Y}$).

Thus, for $m << N$, working with $\mathbf{Y}\mathbf{Y}^\top$ can be much faster.

Thus, when $\mathbf{Y}\mathbf{Y}^\top$ is sufficiently close to $\mathbf{X}\mathbf{X}^\top$, this allows us to compute the principal components much faster in favor of loosing some accuracy.

Perturbation Bounds for the Spectral Decomposition

We begin by asking how do the eigenvalues and eigenvectors of two symmetric matrices $\mathbf{X}, \mathbf{Y}$ change as a function of the difference $\mathbf{X} - \mathbf{Y}$.

We have already seen the following result which followed from our study of the eigenvalues as solutions to optimization problems.

Theorem (Weyl's eigenvalue inequality)

Let $\mathbf{X}, \mathbf{Y} \in \mathbb{S}^n$. It holds for any $i = 1, \ldots, n$ that

$$\lambda_i(\mathbf{X}) + \lambda_n(\mathbf{Y} - \mathbf{X}) \leq \lambda_i(\mathbf{Y}) \leq \lambda_i(\mathbf{X}) + \lambda_1(\mathbf{Y} - \mathbf{X}).$$

Note in particular that since

$$-\|\mathbf{X} - \mathbf{Y}\|_2 \leq \lambda_n(\mathbf{Y} - \mathbf{X}) \leq \lambda_1(\mathbf{Y} - \mathbf{X}) \leq \|\mathbf{Y} - \mathbf{X}\|_2,$$

we have $|\lambda_i(\mathbf{X}) - \lambda_i(\mathbf{Y})| \leq \|\mathbf{X} - \mathbf{Y}\|_2$.

Question: Is this bound tight (in worst case)?

Perturbation Bounds for the Spectral Decomposition

The following theorem study the change in the leading eigenvector. Next we will generalize it to the k leading eigenvectors.

Theorem (Davis-Kahan $\sin\theta$ theorem)

Let $\mathbf{X}, \mathbf{Y} \in \mathbb{S}^n$. Let $\mathbf{u}$ be the leading eigenvector of $\mathbf{X}$ (corresponding to the largest eigenvalue $\lambda_1(\mathbf{X})$) and let $\mathbf{v}$ be the leading eigenvector of $\mathbf{Y}$. Suppose further $\lambda_1(\mathbf{X}) - \lambda_2(\mathbf{X}) > 0$. Then, it holds that

$$\begin{aligned} \|\mathbf{u}\mathbf{u}^\top - \mathbf{v}\mathbf{v}^\top\|_F &= \sqrt{2}|\sin\angle\mathbf{u}, \mathbf{v}| \\ &\leq \min\left\{2\frac{\|\mathbf{X} - \mathbf{Y}\|_F}{\lambda_1(\mathbf{X}) - \lambda_2(\mathbf{X})}, \; 2\sqrt{2}\frac{\|\mathbf{X} - \mathbf{Y}\|_2}{\lambda_1(\mathbf{X}) - \lambda_2(\mathbf{X})}\right\}. \end{aligned}$$

Note that since often
$\|\mathbf{X} - \mathbf{Y}\|_2 = \sigma_1(\mathbf{X}) << \sqrt{\sum_i \sigma_i(\mathbf{X})^2} = \|\mathbf{X} - \mathbf{Y}\|_F$, the bound which depends on $\|\mathbf{X} - \mathbf{Y}\|_2$ is much more preferable.

Proof of the Davis-Kahan $\sin\theta$ theorem

Let $\mathbf{u}$ be top eigenvector of $\mathbf{X}$ and $\mathbf{v}$ be top eigenvector of $\mathbf{y}$.
We want to upper-bound $\|\mathbf{u}\mathbf{u}^\top - \mathbf{v}\mathbf{v}^\top\|_F$.

The projection of $\mathbf{v}$ onto $\mathrm{span}(\mathbf{u})$ is given by $\mathbf{u}\mathbf{u}^\top\mathbf{v} = (\mathbf{u}^\top\mathbf{v})\mathbf{u}$.

Recall that by the projection theorem we have that $\mathbf{v} - (\mathbf{u}^\top\mathbf{v})\mathbf{u} \perp \mathbf{u}$.
Thus, we can write $\mathbf{v}$ as

$$\mathbf{v} = (\mathbf{u}^\top\mathbf{v})\mathbf{u} + \mathbf{w} \quad \text{s.t.} \quad \mathbf{w} \perp \mathbf{u}.$$

Note that it follows that

$$\|\mathbf{w}\|_2^2 = \|\mathbf{v} - (\mathbf{u}^\top\mathbf{v})\mathbf{u}\|_2^2 = \|\mathbf{v}\|_2^2 - 2(\mathbf{u}^\top\mathbf{v})^2 + (\mathbf{u}^\top\mathbf{v})^2\|\mathbf{v}\|_2^2 = 1 - (\mathbf{u}^\top\mathbf{v})^2.$$

Let $\bar{\mathbf{w}}$ be a unit vector in the direction of $\mathbf{w}$. Since $\mathbf{u}$ is the leading eigenvector of $\mathbf{X}$, that is $\mathbf{u}^\top\mathbf{X}\mathbf{u} = \lambda_1(\mathbf{X})$ and $\bar{\mathbf{w}} \perp \mathbf{u}$, from the **minimax theorem** it follows that

$$\bar{\mathbf{w}}^\top\mathbf{X}\bar{\mathbf{w}} \leq \max_{\mathbf{z}:\|\mathbf{z}\|_2=1, \mathbf{z}\perp\mathbf{u}} \mathbf{z}^\top\mathbf{X}\mathbf{z} = \lambda_2(\mathbf{X}).$$

Thus, we have that

$$\mathbf{w}^\top\mathbf{X}\mathbf{w} \leq \|\mathbf{w}\|_2^2 \cdot \lambda_2(\mathbf{X}) = ((1 - (\mathbf{u}^\top\mathbf{v})^2)\lambda_2(\mathbf{X}).$$

Proof of the Davis-Kahan $\sin\theta$ theorem

Let $\mathbf{u}$ be top eigenvector of $\mathbf{X}$ and $\mathbf{v}$ be top eigenvector of $\mathbf{y}$.
We want to upper-bound $\|\mathbf{u}\mathbf{u}^\top - \mathbf{v}\mathbf{v}^\top\|_F$.

We have seen: $\mathbf{v} = (\mathbf{u}^\top\mathbf{v})\mathbf{u} + \mathbf{w}$ for some $\mathbf{w} \perp \mathbf{u}$ and $\mathbf{w}^\top\mathbf{X}\mathbf{w} \leq ((1 - (\mathbf{u}^\top\mathbf{v})^2)\lambda_2(\mathbf{X})$.

Now we can write,

$$\begin{aligned}
\mathbf{v}^\top\mathbf{X}\mathbf{v} &= ((\mathbf{u}^\top\mathbf{v})\mathbf{u} + \mathbf{w})^\top\mathbf{X}((\mathbf{u}^\top\mathbf{v})\mathbf{u} + \mathbf{w}) \underset{(a)}{=} (\mathbf{u}^\top\mathbf{v})^2\mathbf{u}^\top\mathbf{X}\mathbf{u} + \mathbf{w}^\top\mathbf{X}\mathbf{w} \\
&\leq (\mathbf{u}^\top\mathbf{v})^2 \cdot \lambda_1(\mathbf{X}) + (1 - (\mathbf{u}^\top\mathbf{v})^2) \cdot \lambda_2(\mathbf{X}) \\
&= (\mathbf{u}^\top\mathbf{v})^2 \cdot \lambda_1(\mathbf{X}) + (1 - (\mathbf{u}^\top\mathbf{v})^2) \cdot (\lambda_1(\mathbf{X}) - (\lambda_1(\mathbf{X}) - \lambda_2(\mathbf{X}))) \\
&= \lambda_1(\mathbf{X}) - (1 - (\mathbf{u}^\top\mathbf{v})^2) \cdot (\lambda_1(\mathbf{X}) - \lambda_2(\mathbf{X})),
\end{aligned}$$

where (a) follows since $\mathbf{w}^\top\mathbf{X}\mathbf{u} = \lambda_1(\mathbf{X})\mathbf{w}^\top\mathbf{u} = 0$.
On the other hand,

$$\begin{aligned}
\mathbf{v}^\top\mathbf{X}\mathbf{v} &= \mathbf{v}^\top\mathbf{Y}\mathbf{v} - \mathbf{v}^\top(\mathbf{Y} - \mathbf{X})\mathbf{v} \geq \mathbf{u}^\top\mathbf{Y}\mathbf{u} - \mathbf{v}^\top(\mathbf{Y} - \mathbf{X})\mathbf{v} \\
&= \mathbf{u}^\top\mathbf{X}\mathbf{u} - \mathbf{u}^\top(\mathbf{X} - \mathbf{Y})\mathbf{u} - \mathbf{v}^\top(\mathbf{Y} - \mathbf{X})\mathbf{v} \\
&= \lambda_1(\mathbf{X}) - \langle \mathbf{X} - \mathbf{Y}, \mathbf{v}\mathbf{v}^\top - \mathbf{u}\mathbf{u}^\top \rangle.
\end{aligned}$$

Proof of the Davis-Kahan $\sin\theta$ theorem

Let $\mathbf{u}$ be top eigenvector of $\mathbf{X}$ and $\mathbf{v}$ be top eigenvector of $\mathbf{y}$.
We want to upper-bound $\|\mathbf{u}\mathbf{u}^\top - \mathbf{v}\mathbf{v}^\top\|_F$.

We have seen: $\mathbf{v} = (\mathbf{u}^\top\mathbf{v})\mathbf{u} + \mathbf{w}$ for some $\mathbf{w} \perp \mathbf{u}$. Also,

$$\mathbf{v}^\top\mathbf{X}\mathbf{v} \geq \lambda_1(\mathbf{X}) - \langle \mathbf{X} - \mathbf{Y}, \mathbf{v}\mathbf{v}^\top - \mathbf{u}\mathbf{u}^\top\rangle,$$
$$\mathbf{v}^\top\mathbf{X}\mathbf{v} \leq \lambda_1(\mathbf{X}) - (1 - (\mathbf{u}^\top\mathbf{v})^2) \cdot (\lambda_1(\mathbf{X}) - \lambda_2(\mathbf{X}))\,.$$

Combining both bounds and using Cauchy-Schwartz inequality, we obtain

$$(1 - (\mathbf{u}^\top\mathbf{v})^2) \leq \frac{\langle \mathbf{X} - \mathbf{Y}, \mathbf{v}\mathbf{v}^\top - \mathbf{u}\mathbf{u}^\top\rangle}{\lambda_1(\mathbf{X}) - \lambda_2(\mathbf{X})} \leq \frac{\|\mathbf{X} - \mathbf{Y}\|_F\|\mathbf{u}\mathbf{u}^\top - \mathbf{v}\mathbf{v}^\top\|_F}{\lambda_1(\mathbf{X}) - \lambda_2(\mathbf{X})}.$$

Finally, noticing that $(1 - (\mathbf{u}^\top\mathbf{v})^2) = \frac{1}{2}\|\mathbf{u}\mathbf{u}^\top - \mathbf{v}\mathbf{v}^\top\|_F^2$, we obtain

$$\|\mathbf{u}\mathbf{u}^\top - \mathbf{v}\mathbf{v}^\top\|_F \leq 2\frac{\|\mathbf{X} - \mathbf{Y}\|_F}{\lambda_1(\mathbf{X}) - \lambda_2(\mathbf{X})}.$$

Proof of the Davis-Kahan $\sin\theta$ theorem

Another way to go towards upper-bounding $|\langle \mathbf{X} - \mathbf{Y}, \mathbf{v}\mathbf{v}^\top - \mathbf{u}\mathbf{u}^\top\rangle|$ is as follows. Note that the matrix $\mathbf{v}\mathbf{v}^\top - \mathbf{u}\mathbf{u}^\top$ has rank at most two, and thus, we can write its spectral decomposition as
$\mathbf{v}\mathbf{v}^\top - \mathbf{u}\mathbf{u}^\top = \lambda_1\mathbf{w}_1\mathbf{w}_1^\top + \lambda_2\mathbf{w}_2\mathbf{w}_2^\top$. Now,

$$\begin{aligned}
|\langle \mathbf{X} - \mathbf{Y}, \mathbf{v}\mathbf{v}^\top - \mathbf{u}\mathbf{u}^\top\rangle| &= |\lambda_1\mathbf{w}_1^\top(\mathbf{X} - \mathbf{Y})\mathbf{w}_1 + \lambda_2\mathbf{w}_2^\top(\mathbf{X} - \mathbf{Y})\mathbf{w}_2| \\
&\leq |\lambda_1||\mathbf{w}_1^\top(\mathbf{X} - \mathbf{Y})\mathbf{w}_1| + |\lambda_2||\mathbf{w}_2^\top(\mathbf{X} - \mathbf{Y})\mathbf{w}_2| \\
&\underset{(a)}{=} (|\lambda_1| + |\lambda_2|)\|\mathbf{X} - \mathbf{Y}\|_2 \underset{(b)}{\leq} \sqrt{2}\sqrt{\lambda_1^2 + \lambda_2^2} \cdot \|\mathbf{X} - \mathbf{Y}\|_2 \\
&= \sqrt{2}\|\mathbf{u}\mathbf{u}^\top - \mathbf{v}\mathbf{v}^\top\|_F \cdot \|\mathbf{X} - \mathbf{Y}\|_2,
\end{aligned}$$

where (a) follows from Rayleigh's theorem and (b) follows since $(a+b)^2 \leq 2(a^2 + b^2)$.

Continuing as before, we will now get

$$\|\mathbf{u}\mathbf{u}^\top - \mathbf{v}\mathbf{v}^\top\|_F \leq 2\sqrt{2}\frac{\|\mathbf{X} - \mathbf{Y}\|_2}{\lambda_1(\mathbf{X}) - \lambda_2(\mathbf{X})}.$$

Perturbation Bounds for the Spectral Decomposition

Theorem (Davis-Kahan $\sin\theta$ theorem)

Let $\mathbf{X}, \mathbf{Y} \in \mathbb{S}^n$. Let $\mathbf{u}$ be the leading eigenvector of $\mathbf{X}$ (corresponding to the largest eigenvalue $\lambda_1(\mathbf{X})$) and let $\mathbf{v}$ be the leading eigenvector of $\mathbf{Y}$. Suppose further $\lambda_1(\mathbf{X}) - \lambda_2(\mathbf{X}) > 0$. Then, it holds that

$$\begin{aligned} \|\mathbf{u}\mathbf{u}^\top - \mathbf{v}\mathbf{v}^\top\|_F &= \sqrt{2}|\sin\angle\mathbf{u},\mathbf{v}| \\ &\leq \min\left\{2\frac{\|\mathbf{X}-\mathbf{Y}\|_F}{\lambda_1(\mathbf{X})-\lambda_2(\mathbf{X})},\ 2\sqrt{2}\frac{\|\mathbf{X}-\mathbf{Y}\|_2}{\lambda_1(\mathbf{X})-\lambda_2(\mathbf{X})}\right\}. \end{aligned}$$

Question: Is this bound tight (in worst case)?

Generalization of the Davis-Kahan $\sin\theta$ theorem

Theorem

Let $\mathbf{X}, \mathbf{Y} \in \mathbb{S}^n$. Fix some positive integer k and let $\mathbf{U}_k \in \mathbb{R}^{n\times k}$ a matrix whose columns are the k leading eigenvectors of $\mathbf{X}$. Similarly, let $\mathbf{V}_k \in \mathbb{R}^{n\times k}$ be a matrix whose columns are the k leading eigenvectors of $\mathbf{Y}$. Suppose further $\lambda_k(\mathbf{X}) - \lambda_{k+1}(\mathbf{X}) > 0$. Then, it holds that

$$\|\mathbf{U}_k\mathbf{U}_k^\top - \mathbf{V}_k\mathbf{V}_k^\top\|_F \leq \min\left\{2\frac{\|\mathbf{X}-\mathbf{Y}\|_F}{\lambda_k(\mathbf{X})-\lambda_{k+1}(\mathbf{X})},\ 2\sqrt{2k}\frac{\|\mathbf{X}-\mathbf{Y}\|_2}{\lambda_k(\mathbf{X})-\lambda_{k+1}(\mathbf{X})}\right\}.$$

The proof relies on similar arguments to those we used for the $k = 1$ case, but naturally, it is more complex.

Note this theorem does not say anything about the difference between the individual eigenvectors $\mathbf{u}_i, \mathbf{v}_i, i = 1, \ldots, k$, but considers only the difference between the k-dimensional subspaces spanned by the top k eigenvectors.

Perturbation Bounds for the SVD

So far we have discussed perturbations bounds for the eigenvalues and eigenvectors of symmetric matrices. What about general real matrices?

Clearly, since for a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ we know that the singular values are the square root of eigenvalues of $\mathbf{A}^\top\mathbf{A}$ and singular vectors are eigenvector of $\mathbf{A}^\top\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$, we can try and bound the perturbations using the tools we have for symmetric matrices.

The result however is not always optimal. Attempting to apply Weyl's inequality for the eigenvalues will give:

$$|\sigma_i(\mathbf{X})^2 - \sigma_i(\mathbf{Y})^2| = |\lambda_i(\mathbf{X}^\top\mathbf{X}) - \lambda_i(\mathbf{Y}^\top\mathbf{Y})| \leq \|\mathbf{X}^\top\mathbf{X} - \mathbf{Y}^\top\mathbf{Y}\|_2.$$

This is not identical to the version we had for symmetric matrices: $|\lambda_i(\mathbf{X}) - \lambda_i(\mathbf{Y})| \leq \|\mathbf{X} - \mathbf{Y}\|_2$.

Perturbation Bounds for the SVD

Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times n}$ and suppose we want to upper-bound the difference $|\sigma_i(\mathbf{X}) - \sigma_i(\mathbf{Y})|$.

In HW3 you developed a version of Rayleigh's theorem for non-symmetric matrices and used it to show that for any $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$ we have that $\sigma_1(\mathbf{A}+\mathbf{B}) \leq \sigma_1(\mathbf{A}) + \sigma_1(\mathbf{B})$.

Applying this result with $\mathbf{A} = \mathbf{X}$, $\mathbf{B} = \mathbf{Y} - \mathbf{X}$ we get

$$\sigma_1(\mathbf{Y}) \leq \sigma_1(\mathbf{X}) + \sigma_1(\mathbf{Y} - \mathbf{X}) \implies \sigma_1(\mathbf{Y}) - \sigma_1(\mathbf{X}) \leq \|\mathbf{X} - \mathbf{Y}\|_2.$$

Similarly we can obtain the second bound $\sigma_1(\mathbf{X}) - \sigma_1(\mathbf{Y}) \leq \|\mathbf{X} - \mathbf{Y}\|_2$.

More generally, we have the following theorem (without proof).

Theorem (Weyl's singular value inequality)

Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times n}$. It holds for any $i = 1, \ldots, \min\{m, n\}$ that

$$|\sigma_i(\mathbf{X}) - \sigma_i(\mathbf{Y})| \leq \|\mathbf{X} - \mathbf{Y}\|_2.$$

Perturbation Bounds for Linear Systems

Suppose we have some linear system of equations $\mathbf{A}\mathbf{x} = \mathbf{y}$. In the following we study how much does a solution $\mathbf{x}$ to the system changes if we consider instead the coefficients matrix $\mathbf{A} + \delta_{\mathbf{A}}$ and the output vector $\mathbf{y} + \delta_{\mathbf{y}}$, where $\delta_{\mathbf{A}}, \delta_{\mathbf{y}}$ are small perturbations.

As in the case of matrix decompositions, the motivation here is also that often we need to solve a linear system that is only some approximation to the "real" system of interest, where, as before, the approximation can be either due to noise in data/partial information or due to computational constraints (assembling the full system is too expensive).

Perturbation Bounds for Linear Systems

Theorem

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be non-singular (invertible) and let $\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^n, \delta_{\mathbf{y}} \in \mathbb{R}^n, \delta_{\mathbf{A}} \in \mathbb{R}^{n \times n}, \delta_{\mathbf{x}} \in \mathbb{R}^n$ be such that

$$\mathbf{A}\mathbf{x} = \mathbf{y}, \quad (\mathbf{A} + \delta_{\mathbf{A}})(\mathbf{x} + \delta_{\mathbf{x}}) = \mathbf{y} + \delta_{\mathbf{y}},$$

that is $\mathbf{x}$ is a solution to original system, and $\mathbf{x} + \delta_{\mathbf{x}}$ is a solution to the perturbed system.
Suppose that $\kappa(\mathbf{A}) := \frac{\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})}$ satisfies $\kappa(\mathbf{A})\frac{\|\delta_{\mathbf{y}}\|_2}{\|\mathbf{y}\|_2} < 1$. Then,

$$\frac{\|\delta_{\mathbf{x}}\|_2}{\|\mathbf{x} + \delta_{\mathbf{x}}\|_2} \leq \frac{\kappa(\mathbf{A})}{1 - \kappa(\mathbf{A})\frac{\|\delta_{\mathbf{y}}\|_2}{\|\mathbf{y}\|_2}} \left(\frac{\|\delta_{\mathbf{y}}\|_2}{\|\mathbf{y}\|_2} + \frac{\|\delta_{\mathbf{A}}\|_2}{\|\mathbf{A}\|_2} \right).$$

Proof of perturbation bound for linear systems

Since $\mathbf{A}\mathbf{x} = \mathbf{y}$, $(\mathbf{A} + \delta_{\mathbf{A}})(\mathbf{x} + \delta_{\mathbf{x}}) = \mathbf{y} + \delta_{\mathbf{y}}$, it follows that

$$\begin{aligned} \mathbf{A}\delta_{\mathbf{x}} + \delta_{\mathbf{A}}(\mathbf{x} + \delta_{\mathbf{x}}) = \delta_{\mathbf{y}} \quad &\Longrightarrow \quad \mathbf{A}\delta_{\mathbf{x}} = \delta_{\mathbf{y}} - \delta_{\mathbf{A}}(\mathbf{x} + \delta_{\mathbf{x}}) \\ &\Longrightarrow \quad \delta\mathbf{x} = \mathbf{A}^{-1}\delta_{\mathbf{y}} - \mathbf{A}^{-1}\delta_{\mathbf{A}}(\mathbf{x} + \delta_{\mathbf{x}}). \end{aligned}$$

Therefore,

$$\begin{aligned} \|\delta_{\mathbf{x}}\|_2 &= \|\mathbf{A}^{-1}\delta_{\mathbf{y}} - \mathbf{A}^{-1}\delta_{\mathbf{A}}(\mathbf{x} + \delta_{\mathbf{x}})\|_2 \\ &\leq \|\mathbf{A}^{-1}\delta_{\mathbf{y}}\|_2 + \|\mathbf{A}^{-1}\delta_{\mathbf{A}}(\mathbf{x} + \delta_{\mathbf{x}})\|_2 \\ &\leq \|\mathbf{A}^{-1}\|_2\|\delta_{\mathbf{y}}\|_2 + \|\mathbf{A}^{-1}\|_2\|\delta_{\mathbf{A}}\|_2\|\mathbf{x} + \delta_{\mathbf{x}}\|_2. \end{aligned}$$

Dividing by $\|\mathbf{x} + \delta_{\mathbf{x}}\|_2$ we obtain

$$\frac{\|\delta_{\mathbf{x}}\|_2}{\|\mathbf{x} + \delta_{\mathbf{x}}\|_2} \leq \|\mathbf{A}^{-1}\|_2 \frac{\|\delta_{\mathbf{y}}\|_2}{\|\mathbf{y}\|_2} \frac{\|\mathbf{y}\|_2}{\|\mathbf{x} + \delta_{\mathbf{x}}\|_2} + \kappa(\mathbf{A}) \frac{\|\delta_{\mathbf{A}}\|_2}{\|\mathbf{A}\|_2}.$$

Using the fact that $\|\mathbf{y}\|_2 = \|\mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{A}\|_2\|\mathbf{x}\|_2$ we have

$$\frac{\|\delta\mathbf{x}\|_2}{\|\mathbf{x} + \delta\mathbf{x}\|_2} \leq \kappa(\mathbf{A}) \frac{\|\delta\mathbf{y}\|_2}{\|\mathbf{y}\|_2} \frac{\|\mathbf{x}\|_2}{\|\mathbf{x} + \delta\mathbf{x}\|_2} + \kappa(\mathbf{A}) \frac{\|\delta_{\mathbf{A}}\|_2}{\|\mathbf{A}\|_2}.$$

Proof of perturbation bound for linear systems

We have seen that

$$\frac{\|\delta\mathbf{x}\|_2}{\|\mathbf{x} + \delta\mathbf{x}\|_2} \leq \kappa(\mathbf{A}) \frac{\|\delta\mathbf{y}\|_2}{\|\mathbf{y}\|_2} \frac{\|\mathbf{x}\|_2}{\|\mathbf{x} + \delta\mathbf{x}\|_2} + \kappa(\mathbf{A}) \frac{\|\delta_{\mathbf{A}}\|_2}{\|\mathbf{A}\|_2}.$$

Now we use the fact that $\|\mathbf{x}\|_2 = \|\mathbf{x} + \delta_{\mathbf{x}} - \delta_{\mathbf{x}}\|_2 \leq \|\mathbf{x} + \delta_{\mathbf{x}}\|_2 + \|\delta_{\mathbf{x}}\|_2$ and get

$$\frac{\|\delta_{\mathbf{x}}\|_2}{\|\mathbf{x} + \delta_{\mathbf{x}}\|_2} \leq \kappa(\mathbf{A}) \frac{\|\delta_{\mathbf{y}}\|_2}{\|\mathbf{y}\|_2} \left(1 + \frac{\|\delta_{\mathbf{x}}\|_2}{\|\mathbf{x} + \delta_{\mathbf{x}}\|_2}\right) + \kappa(\mathbf{A}) \frac{\|\delta_{\mathbf{A}}\|_2}{\|\mathbf{A}\|_2}.$$

Rearranging, we obtain

$$\frac{\|\delta_{\mathbf{x}}\|_2}{\|\mathbf{x} + \delta_{\mathbf{x}}\|_2} \left(1 - \kappa(\mathbf{A}) \frac{\|\delta_{\mathbf{y}}\|_2}{\|\mathbf{y}\|_2}\right) \leq \kappa(\mathbf{A}) \left(\frac{\|\delta_{\mathbf{y}}\|_2}{\|\mathbf{y}\|_2} + \frac{\|\delta_{\mathbf{A}}\|_2}{\|\mathbf{A}\|_2}\right),$$

which finally gives us

$$\frac{\|\delta_{\mathbf{x}}\|_2}{\|\mathbf{x} + \delta_{\mathbf{x}}\|_2} \leq \frac{\kappa(\mathbf{A})}{1 - \kappa(\mathbf{A}) \frac{\|\delta_{\mathbf{y}}\|_2}{\|\mathbf{y}\|_2}} \left(\frac{\|\delta_{\mathbf{y}}\|_2}{\|\mathbf{y}\|_2} + \frac{\|\delta_{\mathbf{A}}\|_2}{\|\mathbf{A}\|_2}\right).$$

Perturbation Bounds for Linear Systems

Theorem

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ *be non-singular (invertible) and let* $\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^n, \delta_{\mathbf{y}} \in \mathbb{R}^n, \delta_{\mathbf{A}} \in \mathbb{R}^{n\times n}, \delta_{\mathbf{x}} \in \mathbb{R}^n$ *be such that*

$$\mathbf{A}\mathbf{x} = \mathbf{y}, \quad (\mathbf{A} + \delta_{\mathbf{A}})(\mathbf{x} + \delta_{\mathbf{x}}) = \mathbf{y} + \delta_{\mathbf{y}},$$

that is $\mathbf{x}$ *is a solution to original system, and* $\mathbf{x} + \delta_{\mathbf{x}}$ *is a solution to the perturbed system.*
Suppose that $\kappa(\mathbf{A}) := \frac{\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})}$ *satisfies* $\kappa(\mathbf{A})\frac{\|\delta_{\mathbf{y}}\|_2}{\|\mathbf{y}\|_2} < 1$. *Then,*

$$\frac{\|\delta_{\mathbf{x}}\|_2}{\|\mathbf{x} + \delta_{\mathbf{x}}\|_2} \leq \frac{\kappa(\mathbf{A})}{1 - \kappa(\mathbf{A})\frac{\|\delta_{\mathbf{y}}\|_2}{\|\mathbf{y}\|_2}} \left(\frac{\|\delta_{\mathbf{y}}\|_2}{\|\mathbf{y}\|_2} + \frac{\|\delta_{\mathbf{A}}\|_2}{\|\mathbf{A}\|_2} \right).$$

Perturbation Bounds for Least Squares

Perturbation bounds for the solutions of a least squares problem (with both perturbation in the matrix $\mathbf{A}$ and both in the vector $\mathbf{y}$) can be handled by applying the result for linear systems to the corresponding normal equation:

$$\mathbf{A}^\top \mathbf{A}\mathbf{x} = \mathbf{A}^\top \mathbf{y}.$$

In this case, the associated matrix is $\mathbf{A}^\top \mathbf{A}$ which should be **invertible**. Hence, to apply the result from linear systems, it suffices to require that $\mathbf{A}$ is full column rank, which in turn implies that $\mathbf{A}^\top \mathbf{A}$ is indeed invertible. More details in the tutorial.