

EECS 127 Lecture Notes

Dun-Ming Huang

Contents

1	Introduction and Least Squares	3
1.1	An Introductory Prompt	3
1.2	Administratives, Summary	3
1.3	Introduction to The Subject Topic: Optimization	3
1.4	Least Squares Regression	5
2	Linear Algebra Bootcamp: Norms, Gram-Schmidt, QR, FTLA	7
2.1	Vectors and Norms	7
2.2	Cauchy-Schwartz Inequality	8
2.3	Gram-Schmidt Orthonormalization and QR Decomposition	10
3	Linear Algebra: Symmetric Matrices	11
3.1	Fundamental Theorem of Linear Algebra	11
3.2	Minimum Norm Problem	12
3.3	Principal Component Analysis	13
4	Principal Component Analysis	15
4.1	Symmetric Matrix	15
4.2	PCA	16
5	SVD and Low-Rank Approximation	19
5.1	Singular Value Decomposition (SVD)	19
5.2	Low Rank Approximation	21
6	Low-Rank Approximation	23
6.1	Discussion of L-RA	23

Chapter 1

Introduction and Least Squares

For this lecture, we will traverse through the Spring 2023 versions of course logistics, and then an introduction towards optimization.

1.1 An Introductory Prompt

In this course, we discuss a problem in Computer Science called, “optimization”. Understanding optimization is not limited to just understanding mathematical details, and has more to do with “problem formulation”: for whatever technology comes, optimization is an involved technique of critical thinking.

1.2 Administratives, Summary

Logistics. For Discussion, Homework policies, please check the lecture slides. Friday sections are equivalent to the subsequent Monday sections.

Advices from Instructor. Do homeworks, collaborate with people.

New: Projects. We will have an option to do a project at the end of the semester, where we complete a project of choice. This is an effort to bridge the gap between students and research experience; where, provided a research paper, the student will extend that literature at the end of project. This will count towards grade and remove weight from exams if the project grade helps. Schemes will be announced. Projects will be in groups, released after the midterm and due during RRR week.

1.3 Introduction to The Subject Topic: Optimization

Optimization is an approach to problems.

It is the technique where we attempt to optimize some statistic resembling of a result. For example, in machine learning from DATA C100 (or EECS 16A/B), we have chosen appropriate loss functions for our learning task to characterize the performance of a model:

- Minimizing the MSE of a linear regression model.
- Minimizing the cross-entropy loss of a logistic regression model.
- Minimizing the distance traveled by an agent in a maze.

Here, we model learning tasks and attempt to optimize a metric. Depending on how we model and represent a problem, the model will perform differently on its learning task. Therefore, optimization is a study about “picking the right loss function” for some same objective. We consider these design choices, study how they affect learning process.

For another example, Air Traffic control is another optimization problem (at the point of writing, US has just experienced a flight paralysis); for another example, queueing and revenue computation are also common optimization problems. Optimization itself is omnipresent in modern applications.

In this course, we will learn about:

- Low rank approximation
- Ridge regression
- Stochastic Gradient Descent
- Dual Program (which always provides a lower bound for some optimization problem)
- Applications: LQR Control, Classification, SVM

Let's get into the Math now!!!

1.3.1 An Example of Problem Formulation

Say, we work in an oil production firm (I know, very US). We are allowed to make 10,000 barrels of crude oil, which may be produced into either *Jet Fuel* or *Gasoline*.

For every barrel of Jet Fuel produced, a revenue of 30 cents; for Gasoline, 20 cents. Meanwhile, 1 barrel of crude oil produces 0.6 barrels of Jet Fuel or 0.7 barrels of Gasoline.

Furthermore, the firm demands that we produce more than some amount of Jet Fuel and Gasoline (respectively, say, 1000 and 2000 barrels).

Finally, there are transportation capacities, where 180000 barrel miles are available. Distributing Gasoline costs 30 miles, and distributing Jet Fuel costs 10 miles.

We have now been presented a prompt: provided the above conditions, how can I maximize my revenue?

Let us first formulate this prompt mathematically:

$$\begin{cases} x_j = \text{Quantity of Jet Fuel in barrels} \\ x_g = \text{Quantity of Gasoline in barrels} \end{cases}$$

The revenue we generate would be formulated as

$$R(x_j, x_g) = 0.3x_j + 0.2x_g$$

such that we are presented the **constraints**:

$$\begin{cases} \text{Production Quantity Minimum:} & x_j \geq 1000, x_g \geq 2000 \\ \text{Total Available Resources:} & \frac{x_j}{0.7} + \frac{x_g}{0.6} \leq 10000 \\ \text{Transportation Capacity:} & 10x_j + 30x_g \leq 180000 \end{cases}$$

which are mathematically translated from the above English text.

In an optimization framework, we formulate a problem with the following frame:

Explain 1.3.1. Formulating Optimization Problems

In a general optimization problem, we attempt to minimize some function $f_0(\vec{x})$, such that some constraint exists:

$$\forall i \in \{1, \dots, m\}, f_i(\vec{x}) \leq b_i$$

and here, \vec{x} is listed as a representation of some system's state, existing within the domain of possible inputs (states).

The general objective of optimization is to find a state of system that minimizes $f_0(\vec{x})$, which we mathemati-

cally express the solution as:

$$\vec{x}^* = \operatorname{argmin}_{\forall i \in \{1, \dots, m\}, f_i(\vec{x}) \leq b_i} f_0(\vec{x})$$

See the example from above, and try matching the parts of optimization problem with the above framework!

There are some more types of optimization problems! One of them is the famous Least Squares Regression:

1.4 Least Squares Regression

The problem statement:

For some matrix A and a vector \vec{b} , solve for:

$$\min_{\vec{x}} \|A\vec{x} - \vec{b}\|_2^2$$

Such problem can be widely applied to many mathematical problems, such as regression and projection.

Explain 1.4.1. Formulating Least Squares Regression

Provided a dataset of points:

$$\{(x_1, y_1), \dots, (x_n, y_n)\}$$

where we attempt to model a linear equation

$$y = mx + c$$

First of all, what is the variable we attempt to search/optimize for? It would be m, c that are the unknowns. To formulate the Least Squares Problem, we would need to formulate our linear equation for points into some matrix-vector multiplication, and knowing that we are solving for m and c , the formulation follows:

$$A \begin{bmatrix} m \\ c \end{bmatrix} = \vec{b}$$

Let us now fill in the contents of A and \vec{b} for formulation, where A and \vec{b} are of known quantities:

$$\begin{bmatrix} \vec{x} & \vec{1} \end{bmatrix} \begin{bmatrix} m \\ c \end{bmatrix} = \vec{y}$$

From prior coursework we also know that there is a closed-form solution:

$$\vec{x}^* = (A^T A)^{-1} A^T \vec{b}$$

for matrices A with nontrivial nullspaces (see EECS16A for a proof that $N(A) = N(A^T A)$).

Explain 1.4.2. Closed-Form Solution of Least Squares

To minimize the L2 norm of difference between $A\vec{x}$ and \vec{b} , we attempt to find a vector \vec{x} that minimizes the differences between them.

Note that for any choice of \vec{x} , it is by definition that $A\vec{x} \in \operatorname{Col}(A)$.

We now face two claims that we resolve to proceed on solving this problem:

1. The endpoint of $proj_{Col(A)} \vec{b}$ is the point closest to the endpoint of \vec{b} (if they share a same starting point) in $Col(A)$.
2. $\vec{OP} = A\vec{x}^*$, where \vec{x}^* has the closed-form solution as previously described.

Claim 1 can be proven by contradiction: assume another point C closer than P to B (such that \vec{OB} is closer to \vec{b}). However, Pythagorean Theorem argues against it.

Claim 2 is a series of algebraic manipulation as outlined in previous courseworks, based on that the error vector $\vec{e} = \vec{b} - A\vec{x}$ should be orthogonal to the columnspace $Col(A)$ by geometry:

$$\begin{aligned}
 A^T \vec{e} &= 0 \\
 A^T (A\vec{x}^* - \vec{b}) &= 0 \\
 A^T A\vec{x}^* &= A^T \vec{b} \\
 \vec{x}^* &= (A^T A)^{-1} A^T \vec{b}
 \end{aligned}$$

Such result is also known as the **Normal Equation**.

Interestingly, Least Squares Algorithm has a quadratic property, causing it to be some parabolic object that performs a **convex function** to optimize along; that is, the local minimum of this function is a global minimum.

Chapter 2

Linear Algebra Bootcamp: Norms, Gram-Schmidt, QR, FTLA

2.1 Vectors and Norms

In the previous lecture, we have recognized the following symbol:

Symbol 2.1.1. Euclidean Norm

The Euclidean Norm, otherwise known as a **2-Norm**, is a mathematical quantity for some vector \vec{x} :

$$\|\vec{x}\|_2 = \sqrt{\sum_i x_i^2}$$

which measures the distance between the starting and terminal point of a vector.

However, there exist more norms to that. Particularly, **norms** are functions that satisfy the following properties:

Definition 2.1.1. Norm

A norm is a function $f : \mathcal{X} \rightarrow \mathbb{R}$ such that:

1. Non-negativeness: $\forall \vec{x} \in \mathcal{X}, \|\vec{x}\| \geq 0$, and $\|\vec{x}\| = 0 \iff \vec{x} = \vec{0}$
2. Triangle Inequality: $\forall \vec{x}, \vec{y} \in \mathcal{X}, \|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|$
3. Scalar Multiplication: $\forall \alpha \in \mathbb{R}, \vec{x} \in \mathcal{X}, \|\alpha \vec{x}\| = \alpha \|\vec{x}\|$

For example, a general family of norms that satisfy the above properties would be the **LP-norms**:

Definition 2.1.2. LP-Norms

LP-Norms are norm functions defined as:

$$\|\vec{x}\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}}$$

for some natural number p .

Particularly, the Euclidean Norm, otherwise known as the *2-norm* is LP-norm with $p = 2$.

And, the interesting observation is:

$$\|\vec{x}\|_1 = \sum_{i=1}^n |x_i|$$

$$\|\vec{x}\|_\infty = \max_{i=1,\dots,n} |x_i|$$

2.2 Cauchy-Schwartz Inequality

This inequality was active in EECS 16A!

Definition 2.2.1. Cauchy-Schwartz Inequality

The inequality is phrased as:

$$|\vec{x}^T \vec{y}| \leq \|\vec{x}\|_2 \|\vec{y}\|_2$$

And using the property,

$$|\forall x \in \mathbb{R}, |\cos(x)| \leq 1$$

this inequality originates from the following algebraic work:

$$\begin{aligned} |\langle \vec{x}, \vec{y} \rangle| &= |\vec{x}^T \vec{y}| = |\vec{y}^T \vec{x}| \\ &= \|\vec{x}\|_2 \|\vec{y}\|_2 \cos(\theta_{\vec{x}, \vec{y}}) \\ &\leq \|\vec{x}\|_2 \|\vec{y}\|_2 \end{aligned}$$

The last two lines of the above derivation is justified by the mechanics along which we find the projection of \vec{x} onto \vec{y} :

$$proj_{\vec{x}} \vec{y} = \vec{y} \frac{\vec{x}^T \vec{y}}{\|\vec{y}\|_2^2}$$

Where, since the projection is a multiple of \vec{y} such that $proj_{\vec{x}} \vec{y} = t\vec{y}$, we also recognize that,

$$\cos(\theta_{\vec{x}, \vec{y}}) = \frac{\|t\vec{y}\|_2}{\|\vec{x}\|_2}$$

And upon matching the two equations, I acquire:

$$\begin{aligned} t &= \cos(\theta_{\vec{x}, \vec{y}}) \frac{\|\vec{x}\|_2}{\|\vec{y}\|_2} = \frac{\vec{x}^T \vec{y}}{\|\vec{y}\|_2^2} \\ \|\vec{x}\|_2 \|\vec{y}\|_2 &= \langle \vec{x}, \vec{y} \rangle \end{aligned}$$

2.2.1 Holder's Inequality and Norm Ball

In fact, we find that the Cauchy-Schwartz is a general case of this inequality:

Definition 2.2.2. Holder's Inequality

Provided the quantities

$$\vec{x}, \vec{y} \in \mathbb{R}^n, \text{ and some } p, q \geq 1 \text{ s.t. } \frac{1}{p} + \frac{1}{q} = 1$$

Then,

$$|\vec{x}^T \vec{y}| \leq \|\vec{x}\|_p \|\vec{y}\|_p$$

From this concept, let us consider the concept of “*norm ball*”: the geographical object containing all vectors such that their l_p -norm is 1.

- For a 2 – *norm*, for example, the norm ball looks like a unit circle (containing all 2D vectors with length 1, hence a circle of radius 1).
- For a 1 – *norm*, similar logic guides us to a diagonally placed circle (rotated 45°) centered at the origin with side lengths 2.
- For the ∞ – *norm*, has a norm ball of a square centered at origin with side length 2. This embeds the unit ball from 2 – *norm*, which embeds the unit ball from 1 – *norm*.

In this sense, the area of norm ball is larger for any increase in p such that it embeds the prior norm balls.

Explain 2.2.1. Optimization Problem Regarding Norm Ball

Problem Formulation: Solve for

$$\max_{\|\vec{x}\|_2 \leq 1} \vec{x}^T \vec{y}$$

p = 2: The solution would be, by the Cauchy Schwartz’s implication,

$$\vec{x}^* = \frac{\vec{y}}{\|\vec{y}\|_2}$$

p = 1: The expression of dot product is equivalently

$$x_1 y_1 + \cdots + x_n y_n$$

Where the constraint is

$$|x_1| + \cdots + |x_n| \leq 1$$

For each value the components of \vec{x} , which is finite and upper bounded by 1, we should allocate maximum contribution to the maximum element of \vec{y} to maximize this dot product (a weighted sum of \vec{y} ’s component, essentially).

Therefore, let i be the index at which \vec{y} has the component of largest absolute value,

There is an achievable solution for \vec{x}^* , being the unit vector \vec{e}_i multiplied by $\text{sgn}(y_i)$, so to counter for cases where \vec{y} is negative.

In turn, we see that

$$\vec{x}^T \vec{y} = \max_i |y_i| = \|\vec{y}\|_\infty$$

Meanwhile, let us use the Holder’s Inequality to achieve a more rigorous proof. Holder’s Inequality states that,

$$|\vec{x}^T \vec{y}| \leq \|\vec{x}\|_1 \|\vec{y}\|_\infty = \max_i |y_i|$$

Where,

The formulation in above section shows achievability, and Holder’s Inequality expresses an upper bound to prove the approach correct.

Alternative proof of $p = 1$ via Triangle Inequality: Via the triangle inequality, we acquire:

$$\begin{aligned}
 |\vec{x}^T \vec{y}| &= \left| \sum_i x_i y_i \right| \\
 &\stackrel{\text{Triangle Inequality}}{\leq} \sum_i |x_i y_i| \\
 &= \sum_i |x_i| |y_i| \\
 &\leq \sum_i |x_i| \max_i |y_i| = \max_i |y_i| = \|\vec{y}\|_\infty
 \end{aligned}$$

We see that the upperbound of $|\vec{x}^T \vec{y}|$ **has an upperbound** that is **achievable**, assembling all necessary aspects of a complete proof.

$p = \infty$: We can find upper bound via Holder's Inequality,

$$|\vec{x}^T \vec{y}| \leq \|\vec{x}\|_1 \|\vec{y}\|_\infty = \sum_i |y_i|$$

The infinity norm of \vec{x} shows the constraint that:

$$\max_i |x_i| = 1$$

and we may simply compute:

$$x_i = \text{sgn}(y_i)$$

to find and certify the achievability of this upper bound.

2.3 Gram-Schmidt Orthonormalization and QR Decomposition

Gram-Schmidt Orthonormalization is an algorithmic technique to find an orthonormal basis for a set of vectors. The procedure of such algorithm is portrayed as defined in the following cell:

Definition 2.3.1. The Procedure of Gram-Schmidt Orthonormalization

Let us have a set of vectors $\{\vec{a}_1, \dots, \vec{a}_n\}$.

- 1 $\vec{q}_1 = \frac{\vec{a}_1}{\|\vec{a}_1\|_2}$
- 2 for i in $\{2, \dots, n\}$:
- 3 $\vec{z}_i = \text{proj}_{\{\vec{q}_1, \dots, \vec{q}_{i-1}\}} \vec{a}_i$
- 4 $\vec{s}_i = \vec{a}_i - \vec{z}_i$
- 5 $\vec{q}_i = \frac{\vec{s}_i}{\|\vec{s}_i\|_2}$
- 6 return $\{\vec{q}_1, \dots, \vec{q}_n\}$

And note that,

$$\vec{z}_i = \sum_{j=1}^{i-1} \vec{q}_j \langle \vec{a}_i, \vec{q}_j \rangle$$

Chapter 3

Linear Algebra: Symmetric Matrices

3.1 Fundamental Theorem of Linear Algebra

Theorem 3.1.1. Orthogonal Decomposition of Space

Let us consider some vector space \mathcal{X} and some subspace $\mathcal{S} \subseteq \mathcal{X}$.

Then, we may state that,

$$\forall \vec{x} \in \mathcal{X}, \vec{x} = \vec{s} + \vec{r} \text{ uniquely, s.t. } \vec{s} \in \mathcal{S} \wedge \vec{r} \in \mathcal{S}^\perp$$

such that,

$$\mathcal{S}^\perp = \{\vec{y} : \langle \vec{x}, \vec{y} \rangle = 0, \vec{x} \in \mathcal{S}\}$$

Consequently, we are stating that for the space \mathcal{X} ,

$$\mathcal{X} = \mathcal{S} \oplus \mathcal{S}^\perp$$

it is a **direct sum** of the vector spaces \mathcal{S} and \mathcal{S}^\perp (we may write any vector in \mathcal{X} as the sum of two components where each component belongs to some specific subspace of the direct sum.)

This leads to a theorem about decomposing a vector space from some given matrix:

Theorem 3.1.2. Fundamental Theorem of Linear Algebra

Theorem. Consider a matrix $A \in \mathbb{R}^{m \times n}$, then,

$$\mathbb{R}^n = N(A) \oplus R(A^T)$$

where, similarly,

$$\mathbb{R}^m = R(A) \oplus N(A^T)$$

Proof. We may use Orthogonal Decomposition of Space (Theorem 3.1.1) to aid our proof. If so, we would only need to show the two following facts:

1. $N(A) \perp R(A^T)$, or equivalently, $N(A) = R(A^T)^\perp$
2. $N(A^T) \perp R(A)$

Let us perform a proof on the first fact, which would be to prove the equivalence of two sets. First, we may prove that $N(A) \subseteq R(A^T)^\perp$:

Suppose that there is an arbitrary vector $\vec{u} \in N(A)$, where $A\vec{u} = \vec{0}$.

Then, we realize that,

$$(A\vec{u}) = \vec{0}, \vec{0}^T = \vec{u}^T A^T$$

then, for any arbitrary vector \vec{u}' , we may find that

$$\vec{u}'^T A^T \vec{u}' = \vec{0}^T \vec{u}' = 0$$

Therefore, the vector \vec{u} is orthogonal to any vector belonging to $\mathcal{R}(A^T)$, which would mathematically state,

$$\vec{u} \in \mathcal{R}(A^T)^\perp$$

Then, we may prove the opposite direction: $\mathcal{R}(A^T)^\perp \subseteq N(A)$

Suppose that there is an arbitrary vector $\vec{w} \in \mathcal{R}(A^T)$, and $\vec{x} \in \mathcal{R}(A^T)^\perp$; by which, we would state for the arbitrary pair (\vec{w}, \vec{x}) that

$$\forall \vec{w} \in \mathcal{R}(A^T), \vec{x} \in \mathcal{R}(A^T)^\perp (\vec{x} \cdot \vec{w} = 0)$$

Thus, $\forall \vec{w} \in \mathcal{R}(A^T)$:

$$\begin{aligned} \langle \vec{x}, \vec{w} \rangle &= 0 \\ \langle \vec{x}, A^T \vec{w} \rangle &= 0 \\ \vec{x}^T A^T \vec{w} &= 0 \end{aligned}$$

Which qualifies us to state that $A\vec{x} = \vec{0}$.

Therefore, $\vec{x} \in N(A)$.

Applying a symmetrical proof to the second fact allows us to provide a complete proof for this theorem.

3.2 Minimum Norm Problem

This is some sister problem to the Least Squares Problem:

Least Squares	Minimum Norm
Solves an overdetermined system	Solves an underdetermined system
Formulation: $\min_{\vec{x}} \ A\vec{x} - \vec{b}\ _2^2$	Formulation: $\min_{A\vec{x}=\vec{b}} \ \vec{x}\ _2^2$

Explain 3.2.1. Solution to Minimum Norm Problem

Let us provide some \vec{x} that is a solution to $A\vec{x} = \vec{b} \neq \vec{0}$.

Some component of \vec{x} might belong to the nullspace of A . Meanwhile, we also acknowledge that \vec{x} is some arbitrary vector in the real space. Therefore, the vector \vec{x} may be decomposed into $\vec{x} = \vec{n} + \vec{r}$, where

$$\vec{n} \in N(A), \vec{r} \in \mathcal{R}(A^T)$$

In that case, we now minimize for $\vec{r} \in \mathcal{R}(A^T)$ under the setup that,

$$A\vec{x} = A(\vec{n} + \vec{r}) = A\vec{r} = \vec{b}$$

We may recognize that $\vec{n} \cdot \vec{r} = 0$ due to the orthogonal decomposition's nature. Therefore,

$$\|\vec{x}\|_2^2 = \|\vec{n}\|_2^2 + \|\vec{r}\|_2^2$$

Therefore, $\vec{x}^* = A^T \vec{w}$

$$\begin{aligned} A\vec{x} &= \vec{b} \\ (AA^T)\vec{w} &= \vec{b} \\ \vec{w} &= (AA^T)^{-1}\vec{b} \\ \vec{x}^* &= A^T(AA^T)^{-1}\vec{b} \end{aligned}$$

3.3 Principal Component Analysis

Principal Component Analysis is a technique to reduce dimensionality in high-dimension datapoints, so to find underlying feature patterns and enable plotting.

This can be achieved via projecting data onto a lower dimension structure to recover lower dimensional structure from the original high dimensional data.

Mathematically, it would be to project $\vec{x}_1, \dots, \vec{x}_n$ onto $\vec{w} \in \mathbb{R}^p$ such that the projected vectors are all as close to the original vectors as possible.

Explain 3.3.1. Projections

For projecting some vector \vec{x} onto \vec{w} , we would obtain result

$$(\vec{w}^T \vec{w})^{-1} \vec{w}^T \vec{x}$$

Where, projection vector \vec{x} onto a column space of matrix A then follows the least squares formula,

$$(A^T A)^{-1} A^T \vec{x}$$

Now, let's discuss how would PCA work:

Explain 3.3.2. PCA as an Optimization Problem

Formulation. Let the individual projection errors be phrased as

$$\|\vec{x}_i - \langle \vec{w}, \vec{x}_i \rangle \vec{w}\|_2^2$$

Where, the average projection error can thus be quantified as,

$$R(\vec{w}) = \frac{1}{n} \sum_{i=1}^n \|\vec{x}_i - \langle \vec{w}, \vec{x}_i \rangle \vec{w}\|_2^2 = MSE(\vec{w})$$

Therefore, the formulation of optimization problem is:

$$\min_{\vec{w}} R(\vec{w}) \text{ subject to } \|\vec{w}\|_2^2 = 1$$

And the assumption of PCA dataset is that it is zero-meaned, because de-meaned data will allow less bias when deciding the weight vector for PCA.

Solution. Let us consider the projection error,

$$\|\vec{x}_i - \langle \vec{w}, \vec{x}_i \rangle \vec{w}\|_2^2 = \|\vec{e}_i\|_2^2$$

The error can be algebraically simplified:

$$\begin{aligned}\|\vec{e}_i\|_2^2 &= (\vec{x}_i - \langle \vec{w}, \vec{x}_i \rangle \vec{w})^T \cdot (\vec{x}_i - \langle \vec{w}, \vec{x}_i \rangle \vec{w}) \\ &= \|\vec{x}_i\|_2^2 - 2\langle \vec{x}_i, \vec{w} \rangle^2 + \langle \vec{x}_i, \vec{w} \rangle^2 \|\vec{w}\|_2^2 \\ &= \|\vec{x}_i\|_2^2 - \langle \vec{x}_i, \vec{w} \rangle^2\end{aligned}$$

Therefore, removing the fixed costs in the error, we are essentially facing the following optimization problem:

$$\max_{\vec{w}} \frac{1}{n} \sum_{i=1}^n \langle \vec{x}_i, \vec{w} \rangle^2$$

Chapter 4

Principal Component Analysis

4.1 Symmetric Matrix

Let us begin with a formal definition of Symmetric Matrix:

Definition 4.1.1. Symmetric Matrix

A matrix $A \in \mathbb{R}^{n \times n}$ is a **symmetric matrix** if

$$A = A^T$$

or equivalently,

$$\forall i, j \in \{1, \dots, n\}, a_{i,j} = a_{j,i}$$

and in such case we express,

$$A \in \mathbb{S}^n$$

Symmetric Matrices have such properties summarized by the Spectral Theorem:

Theorem 4.1.1. Spectral Theorem

If $A \in \mathbb{S}^n$, then:

1. All of its eigenvalues are real.
2. Eigenspaces corresponding to distinct eigenvalues are orthogonal.
3. The algebraic multiplicity of an eigenvalue is equal to its geometric multiplicity
(alternatively, the matrix is diagonalizable, such that there exists an orthonormal U and diagonal Σ to compose $A = U\Sigma U^T$).

An interlude here:

Definition 4.1.2. Multiplicity of Eigenvalues

A review from EECS 16B:

- The algebraic multiplicity of an eigenvalue (m_A) is the number of times such eigenvalue appears in its matrix's characteristic polynomial.
- The geometric multiplicity of an eigenvalue (m_G) is the number of linearly independent eigenvectors that such eigenvalue corresponds to.

Meanwhile, symmetric matrices also enjoy special properties in their eigenvalues:

Theorem 4.1.2. Variational Characteristics of Rayleigh Coefficient

The Rayleigh Coefficient of some symmetric matrix $A \in \mathbb{S}^n$ and vector $\vec{x} \in \mathbb{R}^n$ is defined as:

$$R_{A,\vec{x}} = \frac{\vec{x}^T A \vec{x}}{\vec{x}^T \vec{x}}$$

Theorem. Then, we may acknowledge via proof that:

$$\forall \vec{x} \neq \vec{0}, \lambda_{\min}(A) \leq R_{A,\vec{x}} \leq \lambda_{\max}(A)$$

and alternatively we may formulate the above inequality such that,

$$\begin{cases} \lambda_{\max}(A) = \max_{\|\vec{x}\|_2=1} \vec{x}^T A \vec{x} \\ \lambda_{\min}(A) = \min_{\|\vec{x}\|_2=1} \vec{x}^T A \vec{x} \end{cases}$$

Proof. Let us define that $\vec{y} = U^T \vec{x}$, and we will proceed onto the algebraic work:

$$\begin{aligned} \vec{x}^T A \vec{x} &= \vec{x} U \Lambda U^T \vec{x} \\ &= \vec{y}^T \Lambda \vec{y} \\ \|\vec{y}\|_2^2 &= 1 \end{aligned}$$

And, furthermore,

$$\begin{aligned} \vec{y}^T \Lambda \vec{y} &= \sum \lambda_i y_i^2 \\ &< \sum \lambda_{\max} y_i^2 = \lambda_{\max} \\ \vec{y}^T \Lambda \vec{y} &= \sum \lambda_i y_i^2 \\ &> \sum \lambda_{\min} y_i^2 = \lambda_{\min} \end{aligned}$$

So we have now found the upper bound and lower bound of the optimization problem, justifying the boundaries of inequality.

And, we may also acknowledge that providing \vec{x} as the normalized eigenvector of maximum and minimum eigenvalues provide the solution for achieving the lower and upper bound:

$$\begin{aligned} \vec{v}_{\lambda_{\max}}^T A \vec{v}_{\lambda_{\max}} &= \lambda_{\max} \|\vec{v}_{\lambda_{\max}}\|_2^2 \\ &= \lambda_{\max} \\ \vec{v}_{\lambda_{\min}}^T A \vec{v}_{\lambda_{\min}} &= \lambda_{\min} \|\vec{v}_{\lambda_{\min}}\|_2^2 \\ &= \lambda_{\min} \end{aligned}$$

Below, let us discuss how we may use the properties of symmetric matrices to solve Principal Component Analysis, which is a problem we were studying in the previous lecture.

4.2 PCA

Principal Component Analysis is a technique to reduce dimensionality in high-dimension datapoints, so to find underlying feature patterns and enable plotting. This can be achieved via projecting data onto a lower dimension structure to recover lower dimensional structure from the original high dimensional data.

We will pick up from last lecture with a new formulation for optimization problem:

Explain 4.2.1. Introduction of Principal Component Analysis as Problem

Suppose we have a set of vectors:

$$\{\vec{x}_1, \dots, \vec{x}_n\}, \forall i (\vec{v}_i \in \mathbb{R}^p)$$

And we would like to project our data into a lower dimensional subspace, such that we may reduce computational cost, and projected vectors are close to the original data.

Let us find the first principal component, \vec{w} , then, such that the dataset projected on this direction is as close to the original data as possible. Then, we will repeat the process of:

- Find principal component \vec{w}_i
- ↓ Remove all projected components onto principal component i
- Repeat process to find principal component \vec{w}_{i+1}

Let us now formulate the mathematical aspects of PCA:

Explain 4.2.2. Mathematical Formulation of Principal Component Analysis as Problem

Metric of Optimization. Suppose we have a set of vectors:

$$\{\vec{x}_1, \dots, \vec{x}_n\}, \forall i (\vec{v}_i \in \mathbb{R}^p)$$

Let us find a vector \vec{w} such that:

$$\begin{cases} \|\vec{w}\|_2 = 1 \\ MSE(\vec{w}) = \frac{1}{n} \sum_{i=1}^n e_i^2 \end{cases}$$

where, the individual error terms were defined and derived into:

$$\begin{aligned} \|\vec{e}_i\|_2^2 &= (\vec{x}_i - \langle \vec{w}, \vec{x}_i \rangle \vec{w})^T \cdot (\vec{x}_i - \langle \vec{w}, \vec{x}_i \rangle \vec{w}) \\ &= \|\vec{x}_i\|_2^2 - 2\langle \vec{x}_i, \vec{w} \rangle^2 + \langle \vec{x}_i, \vec{w} \rangle^2 \|\vec{w}\|_2^2 \\ &= \|\vec{x}_i\|_2^2 - \langle \vec{x}_i, \vec{w} \rangle^2 \end{aligned}$$

Therefore,

$$MSE(\vec{w}) = \frac{1}{n} \sum_{i=1}^n (\|\vec{x}_i\|_2^2 - \langle \vec{w}, \vec{x}_i \rangle^2)$$

Formulation of Optimization. To consider the optimization problem holistically, we should formulate that,

$$\underset{\vec{w}}{\operatorname{argmin}} MSE(\vec{w}) = \underset{\vec{w}}{\operatorname{argmin}} - \frac{1}{n} \sum_{i=1}^n \langle \vec{w}, \vec{x}_i \rangle^2$$

which is equivalently solving for

$$\underset{\vec{w}}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^n \langle \vec{w}, \vec{x}_i \rangle^2$$

and we would recognize so because the aspect of MSE, $\frac{1}{n} \sum_{i=1}^n \|\vec{x}_i\|_2^2$, is a fixed cost.

Now, let us attempt to solve for the above formulation:

Explain 4.2.3. Solving PCA Optimization Problem

Keep in mind that our data matrix would appear as:

$$X = \begin{bmatrix} \vec{x}_1^T \\ \vdots \\ \vec{x}_n^T \end{bmatrix}$$

We may begin with the algebraic work then.

$$X\vec{w} = \begin{bmatrix} \vec{x}_1^T \vec{w} \\ \vdots \\ \vec{x}_n^T \vec{w} \end{bmatrix}$$

Let us capture the information and make some algebraic derivation:

$$\begin{aligned} \frac{1}{n} \sum \langle \vec{w}, \vec{x}_i \rangle^2 &= \frac{1}{n} \sum (\vec{x}_i^T \vec{w})^2 \\ &= \frac{1}{n} \|X\vec{w}\|_2^2 \\ &= \vec{w}^T \frac{X^T X}{n} \vec{w} \end{aligned}$$

We call the matrix in above intermediate form the covariance matrix:

$$C = \frac{X^T X}{n} = C^T$$

And, using the knowledge from Section 4.1, we obtain that the solution of:

$$\operatorname{argmax}_{\vec{w}} \vec{w}^T C \vec{w} \text{ subject to } \|\vec{w}\|_2^2$$

would be the eigenvector of $\lambda_{max}(C)$.

Chapter 5

SVD and Low-Rank Approximation

5.1 Singular Value Decomposition (SVD)

Let us suppose that there is a matrix $A \in \mathbb{R}^{m \times n}$ whose rank is $rk(A) = r$, then there exists a decomposition of A in the shape of:

$$A = \sum_{i=1}^n \sigma_i \vec{u}_i \vec{v}_i^T$$

where,

- The set $\{\vec{u}_1, \dots, \vec{u}_n\}$ is orthonormal
- The set $\{\vec{v}_1, \dots, \vec{v}_n\}$ is orthonormal
- $\forall i \in [1, r], \sigma_i > 0$
 $\forall i \in [1, n], \sigma_i \geq 0$

Symbol 5.1.1. Compact SVD

The above dyadic decomposition of SVD written in terms of matrix transformation would be:

$$\begin{aligned} A &= [\vec{u}_1 \quad \dots \quad \vec{u}_r] \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \begin{bmatrix} \vec{v}_1^T \\ \vdots \\ \vec{v}_r^T \end{bmatrix} \\ &= U_r \Sigma_r V_r^T \end{aligned}$$

Meanwhile, the full SVD that still considers zero-valued σ_i is written as follows:

Symbol 5.1.2. Full SVD

The formulation follows:

$$\begin{aligned} A &= [\vec{u}_1 \quad \dots \quad \vec{u}_m] \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \vec{v}_1^T \\ \vdots \\ \vec{v}_n^T \end{bmatrix} \\ &= U \Sigma V^T \end{aligned}$$

from which, we may observe that the shape of Σ is highly dependent on the shape of A :

- If A is tall, so is Σ
- If A is wide, so is Σ

The squareness of U and V in full SVD allows us to also decompose $A^T A$ as:

$$A^T A = (U \Sigma V^T)^T (U \Sigma V^T) = V \Sigma^T U^T U \Sigma V^T = V \Sigma^2 V^T$$

which is more convenient than the compact form SVD that instead uses rectangular U and V .

5.1.1 Formation of SVD

As singular vectors are eigenvectors of symmetric matrices $A^T A$, by applying Spectral Theorem, we find that singular vectors of A would all be orthonormal. Then, let's use this point as an opportunity to praise $A^T A$'s symmetricity:

Theorem 5.1.1. The Properties and Formation of SVD

Let the nonzero eigenvalues of $A^T A$ be

$$\lambda_1 \geq \dots \geq \lambda_r > 0$$

where, upon dimming that $rk(A^T A) = r$, we may obtain $rk(A) = rk(A^T A) = r$ (using rank-nullity theorem, and the fact that $N(A^T A) = N(A)$).

Let us also suppose that orthonormal vectors in V form eigenpairs with the above eigenvalues:

$$\forall i \in [1, r], A^T A \vec{v}_i = \lambda_i \vec{v}_i$$

Now, define that,

$$\sigma_i := \sqrt{\lambda_i}$$

such that:

$$\vec{u}_i := \frac{A \vec{v}_i}{\sigma_i}$$

Let us now check, whether the above formation grants an orthonormal U :

Explain 5.1.1. Orthonormality of Set of \vec{u}_i

Let us follow the definition of \vec{u}_i from the last block.

Then,

$$A \vec{v}_i = \sigma_i \vec{u}_i$$

Suppose now we have some vectors \vec{u}_i and \vec{u}_j , then

$$\begin{aligned} \vec{u}_i \cdot \vec{u}_j &= \frac{\vec{v}_i^T A^T}{\sigma_i} \frac{A \vec{v}_j}{\sigma_j} \\ &= \frac{\vec{v}_i^T \lambda_j \vec{v}_j}{\sigma_i \sigma_j} = 0 \end{aligned}$$

Now that we have proven the orthonormality and note the definitions of \vec{u}_i and \vec{v}_i , we arrive at the conclusion:

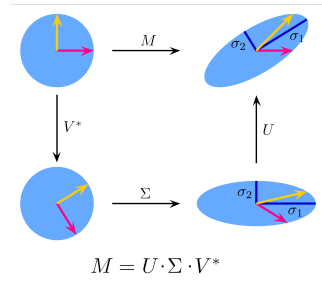
$$AV = U\Sigma, A = U\Sigma V^T$$

such arithmetic operation is only possible under the full SVD form (such that V can be noted as orthonormal at all). This is once again a demonstration of significance of full SVD.

5.1.2 Geometry of SVD

Figure 5.1.1. Geometric Property of SVD as Transformation

We may see from the below image's operations,



that SVD is, as priorly mentioned, a combination of three transformations. Where,

- The orthonormal matrices U , V are reflections or rotations.
- The matrix Σ is a dilation.

Furthermore, let us observe the vectors drawn on the unit circle.

Since V is orthonormal, we may conclude that $\vec{v}_1 \perp \vec{v}_2$. However, bizzarely, we may also find that

$$A\vec{v}_1 \perp A\vec{v}_2$$

5.2 Low Rank Approximation

Low Rank Approximation is the task of approximating a matrix A in a lower rank than $rk(A)$, to conserve computational cost and perform efficient computations.

5.2.1 Matrix Norms

Let us first discuss the optimization function of this task: matrix norm.

The norm of matrix also comes in diverse form, because matrices can be interpreted in many forms.

For example:

- Chunk of data (like an array)
- Operator of transformation

Therefore, along these interpretations, we may present respective definitions of matrix norms.

Definition 5.2.1. Frobenius Norm

insert most recent COMPSCI 189 weeder homework PTSD.

The Frobenius Norm is defined as:

$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n A_{i,j}^2} = \sqrt{\text{Trace}(A^T A)}$$

This is because the Frobenius inner product may be defined as:

$$\langle A, B \rangle_F = \text{Trace}(A^T B) = \sum_{i=1}^n \vec{A}_i \cdot \vec{B}_i$$

and, as we recognize the relation between norm and inner product,

$$\|A\|_F = \sqrt{\langle A, A \rangle_F} = \sqrt{\sum_{i=1}^n \vec{A}_i \cdot \vec{A}_i}$$

The Frobenius Norm has some interesting properties:

Theorem 5.2.1. Orthonormal Matrices and Frobenius Norm

Let $U_1 \in \mathbb{R}^{m \times m}$, $A \in \mathbb{R}^{m \times n}$, and $U_2 \in \mathbb{R}^{n \times n}$, where U_1, U_2 are orthonormal. Then,

$$\begin{aligned} \|U_1 A\|_F &= \sqrt{\text{Trace}(A^T U_1^T U_1 A)} \\ &= \sqrt{\text{Trace}(A^T A)} = \|A\|_F \\ \|AU_2\|_F &= \sqrt{\text{Trace}(U_2^T A^T A U_2)} \\ &\stackrel{\text{Commutivity in Trace}}{=} \sqrt{\text{Trace}(A^T A)} = \|A\|_F \end{aligned}$$

Now, let's discuss another definition of matrix norm:

Definition 5.2.2. Operator Norm (aka. Spectral Norm, L2-Norm)

Such norm is defined as:

$$\|A\|_2 = \max_{\|\vec{x}\|_2=1} \|A\vec{x}\|_2 = \sigma_{\max}(A)$$

Solution to Above Optimization Problem.

$$\max_{\|\vec{x}\|_2=1} \|A\vec{x}\|_2 = \max_{\|\vec{x}\|_2=1} \sqrt{\vec{x}^T A^T A \vec{x}} = \sqrt{\lambda_{\max}(A^T A)} = \sigma_{\max}(A)$$

Suppose we attach another orthonormal matrix U to the original matrix A , then the optimization problem develops a solution fairly similarly:

$$\begin{aligned} \max_{\|\vec{x}\|_2=1} \|AU\vec{x}\|_2 &= \max_{\|\vec{x}\|_2=1} \sqrt{\vec{x}^T U^T A^T A U \vec{x}} \\ &= \max_{\|\vec{x}\|_2=1} \sqrt{\vec{y}^T A^T A \vec{y}} \\ &= \sqrt{\lambda_{\max}(A^T A)} = \sigma_{\max}(A) \end{aligned}$$

where since we still find that \vec{y} is normal, the attachment of U providing no change to the optimization problem. Meanwhile, we find this phenomenon to persist even if U is multiplied from an opposite direction:

$$\begin{aligned} \max_{\|\vec{x}\|_2=1} \|UA\vec{x}\|_2 &= \max_{\|\vec{x}\|_2=1} \sqrt{\vec{x}^T A^T U^T U A \vec{x}} \\ &= \max_{\|\vec{x}\|_2=1} \sqrt{\vec{x}^T A^T A \vec{x}} \\ &= \sqrt{\lambda_{\max}(A^T A)} = \sigma_{\max}(A) \end{aligned}$$

Therefore, we may also observe similar properties in norms, where:

$$\|A\|_2 = \|U_1 A\|_2 = \|AU_2\|_2 \text{ for orthonormal matrices } U_1, U_2$$

Chapter 6

Low-Rank Approximation

6.1 Discussion of L-RA

Suppose we have some matrix $A \in \mathbb{R}^{m \times n}$, where storing this matrix would take significant computational resources. Is there a method to store an approximation of this matrix, such that the approximation is more efficient to store for, and we would thus reduce computational cost of storage?

We have discussed the possibility of such technique within the previous lecture, where we have solved the optimization problem:

$$\max_{\|\vec{x}\|_2=1} \|A\vec{x}\|_2$$

Today, we will discuss the extension of such solution: Low-Rank Approximation. It would be characterized from the optimization problem:

$$\operatorname{argmin}_{B \in \mathbb{R}^{m \times n}, rk(B)=k} \|A - B\|_F$$

where we may see a contextually similar optimization problem:

$$\operatorname{argmin}_{B \in \mathbb{R}^{m \times n}, rk(B)=k} \|A - B\|_2$$

Let us perform a proof on this:

Explain 6.1.1. The LRA Optimization Problem on Spectral Norm

Problem.

$$\operatorname{argmin}_{B \in \mathbb{R}^{m \times n}, rk(B)=k} \|A - B\|_2$$

Solution.

Keep in mind that our proof proceeds in the direction of:

- Finding a possible upper bound
- Finding the achievability of such upper bound
- Finding the legitimacy of such upper bound

Now, let us move on with the subparts of a proof.

Finding a Possible Upper Bound.

Let us define

$$A_k = \sum_{i=1}^k \sigma_i \vec{u}_i \vec{v}_i^T$$

Then, the difference between such approximation and A would be:

$$\|A - A_k\|_2 = \|A_n - A_k\|_2 = \left\| \sum_{i=k+1}^n \sigma_i \vec{u}_i \vec{v}_i^T \right\|_2$$

which, following along the optimization problem phrased last lecture, we would obtain the solution to be:

$$\left\| \sum_{i=k+1}^n \sigma_i \vec{u}_i \vec{v}_i^T \right\|_2 = \sigma_{k+1}(A)$$

Finding the Legitimacy (and Achievability) of Upper Bound.

Then, let us show that for every B where $rk(B) \leq k$, the difference $\|A - B\|_2$ is larger than our current possible minimum, σ_{k+1} . By proving so we secure the result of maximization is as illustrated above.

Let us first define \vec{w} that,

$$\|A - B\|_2 = \max_{\|\vec{w}\|=1} \|(A - B)\vec{w}\|_2$$

Where, if $\vec{w} \in N(B)$, then we would be able to simplify the above optimization problem into something easier to solve: along such above constraint, let us note that,

$$\|A - B\|_2^2 \geq \|(A - B)\vec{w}\|_2^2 = \|A\vec{w}\|_2^2$$

And decomposing A into SVD form, which contains orthonormal matrices,

$$\|A - B\|_2^2 \geq \|U\Sigma V^T \vec{w}\|_2^2 = \|\Sigma V^T \vec{w}\|_2^2$$

Provided that \vec{w} is a unit vector, and V is invertible, we can guarantee that

$$\exists \vec{a}, V\vec{a} = \vec{w}$$

Therefore, let us now substitute such value in, and we will obtain that

$$\begin{aligned} \|A - B\|_2^2 &\geq \|\Sigma V^T \vec{w}\|_2^2 \\ &= \|\Sigma \vec{a}\|_2^2 \\ &= \vec{a}^T \Sigma^T \Sigma \vec{a} = \vec{a}^T \Lambda \vec{a} \end{aligned}$$

Let us come back to our assumption that $\vec{w} \in N(B)$, and extract more insights from it.

Now, we may define along the V of $A = U\Sigma V^T$ that:

$$V_{k+1} = [\vec{v}_1 \quad \cdots \quad \vec{v}_{k+1}]$$

where along the property of SVD we understand that $rk(V_{k+1}) = k + 1$.

To find some \vec{w} such that $\vec{w} \in N(B) \wedge \vec{w} \in \mathcal{R}(V_{k+1})$, we should notice that:

1. Along rank-nullity theorem, $\dim(N(B)) \geq n - k$.
2. Along above description, $\dim(\mathcal{R}(V_{k+1})) = k + 1$

At this point, we must note that the achievability in our proposed upper bound exists in the possibility of having such \vec{w} . Therefore, **the achievability of upper bound is protected by the existence of \vec{w} , which we must guarantee.**

These two subspaces of \mathbb{R}^n must have overlap. Therefore, there must exist some vector that belongs to both subspaces. This guarantees the existence of our \vec{w} .

Back to the Achievability of Upper Bound, Final Step.

Let us now follow along, and revisit the derivation that $\vec{a} = V^T \vec{w}$ on the change of basis.

Then, here we perform the final computations:

$$\begin{aligned}
 \vec{a} &= V^T \vec{w} \\
 &= V^T \left(\sum_{i=1}^{k+1} a_i \vec{V}_i \right) \\
 &= [a_1 \quad \cdots \quad a_{k+1} \quad 0 \quad \cdots \quad 0]^T \\
 \vec{a}^T \Lambda \vec{a} &= \sum_{i=1}^{k+1} a_i^2 \lambda_i \\
 &\geq \sum_{i=1}^{k+1} a_i^2 \lambda_{k+1} = \lambda_{k+1} \sum_{i=1}^{k+1} a_i^2 \\
 &= \lambda_{k+1} = \sigma_{k+1}^2 (\text{Note, } \|\tilde{a}\|_2 = \|\tilde{w}\|_2 = 1)
 \end{aligned}$$

Explain 6.1.2. The LRA Optimization Problem on Frobenius Norm

Problem.

$$\operatorname{argmin}_{B \in \mathbb{R}^{m \times n}, \operatorname{rk}(B)=k} \|A - B\|_F$$

Solution.

Let us consider A_k as priorly defined to be the solution of optimization again.

$$\begin{aligned}
 \|A\|_F^2 &= \|U \Sigma V^T\|_F^2 \\
 &= \|\Sigma\|_F^2 = \sum_{i=1}^n \sigma_i^2 \\
 \|A - A_k\|_F^2 &= \sum_{i=k+1}^n \sigma_i^2 \\
 \|A - B\|_F^2 &= \sum_{i=1}^n \sigma_i^2 (A - B)
 \end{aligned}$$

Now, following the logic before, we want to prove this upperbound:

$$\text{Prove that } \forall B \in \mathbb{R}^{m \times n} \text{ s.t. } \operatorname{rk}(B) = k, \|A - B\|_F \geq \|A - A_k\|_F$$

Using the aligned equations above, let us find an equivalent proof prompt,

$$\text{Prove that } \forall i, \sigma_i(A - B) \geq \sigma_{k+i}(A)$$

Where, along spectral norm's definition, we can also obtain that,

$$\sigma_{k+1}(A) = \|A - A_{k+i-1}\|_2$$

translated in text stating that A_{k+i-1} is the best rank $(k + i - 1)$ approximation to A in spectral norm sense.

Now, let us return to the above equivalent prompt and define such that $C := A - B$.

$$\begin{aligned}
 \sigma_i(A - B) &= \sigma_i(C) \\
 &= \|C - C_{i-1}\|_2 = \|A - B - C_{i-1}\|_2
 \end{aligned}$$

Pay attention to the rank of these matrices:

$$\begin{aligned}
 B &= B_k, \text{rk}(B_k) \leq k \\
 \text{rk}(C_{i-1}) &\leq i-1 \\
 \text{rk}(B_k + C_{i-1}) &\leq k+i-1 \\
 \sigma_i(A-B) &= \|A-D\|_2 \\
 &\text{, where } \text{rk}(D) \leq k+i-1
 \end{aligned}$$

Using the solution from spectral norm side proof of LRA,

$$\|A-D\|_2 \geq \|A-A_{k+i-1}\|_2 = \sigma_{k+1}(A)$$