TOPIC 2

Spectral and singular value decompositions

2 Matrix norms and low-rank approximations

Course: Math 535 (http://www.math.wisc.edu/~roch/mmids/) - Mathematical Methods in Data Science (MMiDS)
Author: Sebastien Roch (http://www.math.wisc.edu/~roch/), Department of Mathematics, University of Wisconsin-Madison
Updated: Sep 28, 2020

In this section, we discuss low-rank approximations of matrices. We first introduce matrix norms which allow us in particular to talk about the distance between two matrices. Throughout this section, $\|\mathbf{x}\|$ refers to the 2-norm of $\mathbf{x}$ (although the concepts we derive below can be extended to other vector norms (https://en.wikipedia.org/wiki/Matrix_norm#Matrix_norms_induced_by_vector_norms)).

2.1 Matrix norms

Recall that the Frobenius norm of an $n \times m$ matrix $A \in \mathbb{R}^{n \times m}$ is defined as

$$\|A\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^2}.$$

The Frobenius norm does not directly relate to A as a representation of a linear map (https://en.wikipedia.org/wiki/Linear_map). In particular, it is desirable in many contexts to quantify how two matrices differ in terms of how they act on vectors.

For instance, one is often interested in bounding quantities of the following form. Let $B, B' \in \mathbb{R}^{n\times m}$ and let $\mathbf{x} \in \mathbb{R}^m$ be of unit norm. What can be said about $\|B\mathbf{x} - B'\mathbf{x}\|$? Intuitively, what we would like is this: if the norm of $B - B'$ is small then B is close to B' as a linear map, that is, the vector norm $\|B\mathbf{x} - B'\mathbf{x}\|$ is small for any unit vector $\mathbf{x}$. The following definition provides us with such a notion. Define $\mathbb{S}^{m-1} = \{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\| = 1\}$.

Definition (Induced Norm): The 2-norm of a matrix $A \in \mathbb{R}^{n\times m}$ is

$$\|A\|_2 = \max_{\mathbf{0}\neq\mathbf{x}\in\mathbb{R}^m} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\mathbf{x}\in\mathbb{S}^{m-1}} \|A\mathbf{x}\|.$$

◁

The equality in the definition uses the homogeneity of the vector norm. Also the definition implicitly uses the *Extreme Value Theorem*. In this case, we use the fact that the function $f(\mathbf{x}) = \|A\mathbf{x}\|$ is continuous and the set $\mathbb{S}^{m-1}$ is closed and bounded to conclude that there exists $\mathbf{x}^* \in \mathbb{S}^{m-1}$ such that $f(\mathbf{x}^*) \geq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{S}^{m-1}$.

Exercise: Let $A \in \mathbb{R}^{n\times m}$. Use Cauchy-Schwarz to show that

$$\|A\|_2 = \max\left\{\mathbf{x}^T A\mathbf{y} : \|\mathbf{x}\| = \|\mathbf{y}\| = 1\right\}.$$

◁

Exercise: Let $A \in \mathbb{R}^{n\times m}$.

(a) Show that $\|A\|_F^2 = \sum_{j=1}^m \|A\mathbf{e}_j\|^2$.

(b) Use (a) and Cauchy-Schwarz to show that $\|A\|_2 \leq \|A\|_F$.

(c) Give an example such that $\|A\|_F = \sqrt{n}\|A\|_2$. ◁

The 2-norm of a matrix has many other useful properties.

Lemma (Properties of the Induced Norm): Let $A, B \in \mathbb{R}^{n\times m}$ and $\alpha \in \mathbb{R}$. The following hold:

(a) $\|A\mathbf{x}\| \leq \|A\|_2\|\mathbf{x}\|$, $\forall \mathbf{0} \neq \mathbf{x} \in \mathbb{R}^m$

(b) $\|A\|_2 \geq 0$

(c) $\|A\|_2 = 0$ if and only if $A = 0$

(d) $\|\alpha A\|_2 \leq |\alpha|\|A\|_2$

(e) $\|A + B\|_2 \leq \|A\|_2 + \|B\|_2$

(f) $\|AB\|_2 \leq \|A\|_2\|B\|_2$.

Proof: These properties all follow from the definition of the induced norm and the corresponding properties for the vector norm:

- Claims (a) and (b) are immediate from the definition.
- For (c) note that $\|A\|_2 = 0$ implies $\|A\mathbf{x}\|_2 = 0, \forall \mathbf{x} \in \mathbb{S}^{m-1}$, so that $A\mathbf{x} = \mathbf{0}, \forall \mathbf{x} \in \mathbb{S}^{m-1}$. In particular, $a_{ij} = \mathbf{e}_i^T A \mathbf{e}_j = 0, \forall i, j$.
- For (d), (e), (f), observe that for all $\mathbf{x} \in \mathbb{S}^{m-1}$

$$\begin{aligned} \|\alpha A\mathbf{x}\| &= |\alpha| \|A\mathbf{x}\| \\ \|(A+B)\mathbf{x}\| = \|A\mathbf{x} + B\mathbf{x}\| &\leq \|A\mathbf{x}\| + \|B\mathbf{x}\| \leq \|A\|_2 + \|B\|_2 \\ \|(AB)\mathbf{x}\| = \|A(B\mathbf{x})\| &\leq \|A\|_2 \|B\mathbf{x}\| \leq \|A\|_2 \|B\|_2. \end{aligned}$$

□

Exercise: Use Cauchy-Schwarz to show that for any A, B it holds that

$$\|AB\|_F \leq \|A\|_F \|B\|_F.$$

◁

NUMERICAL CORNER In Julia, the Frobenius norm of a matrix can be computed using the function `norm` (https://docs.julialang.org/en/v1/stdlib/LinearAlgebra/#LinearAlgebra.norm) while the induced norm can be computed using the function `opnorm` (https://docs.julialang.org/en/v1/stdlib/LinearAlgebra/#LinearAlgebra.opnorm).

In [1]:
```
#Julia version: 1.5.1
using LinearAlgebra
```

In [2]:
```
A = [1. 0.; 0. 1.; 0. 0.]
```

Out[2]:
```
3×2 Array{Float64,2}:
 1.0  0.0
 0.0  1.0
 0.0  0.0
```

In [3]:
```
norm(A)
```

Out[3]:
```
1.4142135623730951
```

In [4]:
```
opnorm(A)
```

Out[4]:
```
1.0
```

2.2 Rank-k approximation

Now that we have defined a notion of distance between matrices, we will consider the problem of finding a good approximation to a matrix A among all matrices of rank at most k. We will start with the Frobenius norm, which is easier to work with, and we will show later on that the solution is the same under the induced norm.

From the proof of the *Row Rank Equals Column Rank Lemma*, it follows that a rank-r matrix A can be written as a sum of r rank-1 matrices

$$A = \sum_{i=1}^{r} \mathbf{b}_i \mathbf{c}_i^T.$$

We will now consider the problem of finding a "simpler" approximation to A

$$A \approx \sum_{i=1}^{k} \mathbf{b}'_i (\mathbf{c}'_i)^T$$

where $k < r$. Here we measure the quality of this approximation using a matrix norm. We will see in the next subsection that this problem has a natural interpretation in a data analysis context.

But, first, we are ready to state our key observation. In words, the best rank-k approximation to A in Frobenius norm is obtained by projecting the rows of A onto a linear subspace of dimension k. We will come back to how one finds the best such subspace below.

Lemma (Projection and Rank-k Approximation): Let $A \in \mathbb{R}^{n \times m}$. For any matrix $B \in \mathbb{R}^{n \times m}$ of rank $k \leq \min\{n, m\}$,

$$\|A - B_\perp\|_F \leq \|A - B\|_F$$

where $B_\perp \in \mathbb{R}^{n \times m}$ is the matrix of rank at most k obtained as follows. Denote row i of A, B and $B_\perp$ respectively by $\boldsymbol{\alpha}_i^T$, $\mathbf{b}_i^T$ and $\mathbf{b}_{\perp,i}^T$, $i = 1, \ldots, n$. Set $\mathbf{b}_{\perp,i}$ to be the orthogonal projection of $\boldsymbol{\alpha}_i$ onto $\mathcal{Z} = \mathrm{span}(\mathbf{b}_i, i = 1, \ldots, n)$.

Proof idea: The square of the Frobenius norm decomposes as a sum of squared row norms. Each term in the sum is minimized by the orthogonal projection.

Proof: By definition of the Frobenius norm, we note that

$$\|A - B\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} (a_{i,j} - b_{i,j})^2 = \sum_{i=1}^{n} \|\boldsymbol{\alpha}_i - \mathbf{b}_i\|^2$$

and similarly for $\|A - B_\perp\|_F$. We make two key observations:

(1) Because the orthogonal projection of $\boldsymbol{\alpha}_i$ onto $\mathcal{Z}$ minimizes the distance to $\mathcal{Z}$, it follows that term by term $\|\boldsymbol{\alpha}_i - \mathbf{b}_{\perp,i}\| \leq \|\boldsymbol{\alpha}_i - \mathbf{b}_i\|$ so that

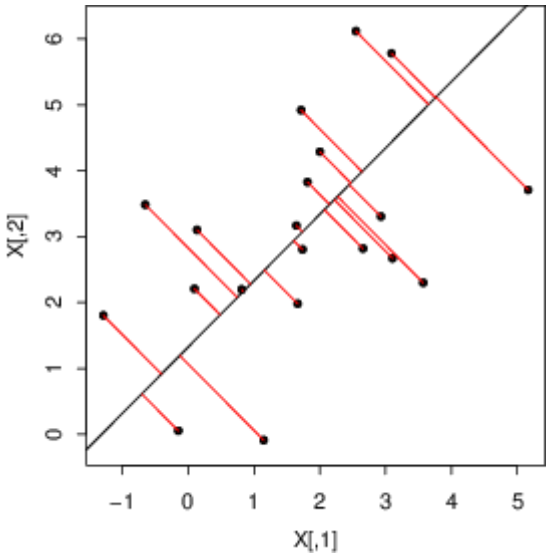
$$\|A - B_\perp\|_F^2 = \sum_{i=1}^{n} \|\boldsymbol{\alpha}_i - \mathbf{b}_{\perp,i}\|^2 \leq \sum_{i=1}^{n} \|\boldsymbol{\alpha}_i - \mathbf{b}_i\|^2 = \|A - B\|_F^2.$$

(2) Moreover, because the projections satisfy $\mathbf{b}_{\perp,i} \in \mathcal{Z}$ for all i, $\text{row}(B_\perp) \subseteq \text{row}(B)$ and, hence, the rank of $B_\perp$ is at most the rank of B.

That concludes the proof. □

2.3 Approximating subspaces

Think of the rows $\boldsymbol{\alpha}_i^T$ of $A \in \mathbb{R}^{n \times m}$ as a collection of n data points in $\mathbb{R}^m$. A natural way to identify a low-dimensional structure in this dataset is to find a low-dimensional linear subspace $\mathcal{Z}$ of $\mathbb{R}^m$ such that the $\boldsymbol{\alpha}_i$'s are "close to it."



(Source (https://i.stack.imgur.com/DAX3R.png))

Again the squared 2-norm turns out to be convenient computationally. So we are looking for a linear subspace $\mathcal{Z}$ that minimizes

$$\sum_{i=1}^{n} \|\boldsymbol{\alpha}_i - \text{proj}_{\mathcal{Z}}(\boldsymbol{\alpha}_i)\|^2$$

over all linear subspaces of $\mathbb{R}^m$ of dimension at most k. By the *Projection and Rank-k Approximation Lemma*, this problem is equivalent to finding a matrix $\boldsymbol{B}$ that minimizes

$$\|A - \boldsymbol{B}\|_F$$

among all matrices in $\mathbb{R}^{n \times m}$ of rank at most k.

The following observation gives a related, useful characterization.

Lemma: Let $\boldsymbol{\alpha}_i$, $i = 1 \ldots, n$, be vectors in $\mathbb{R}^m$. A linear subspace $\mathcal{Z}$ of $\mathbb{R}^m$ that minimizes

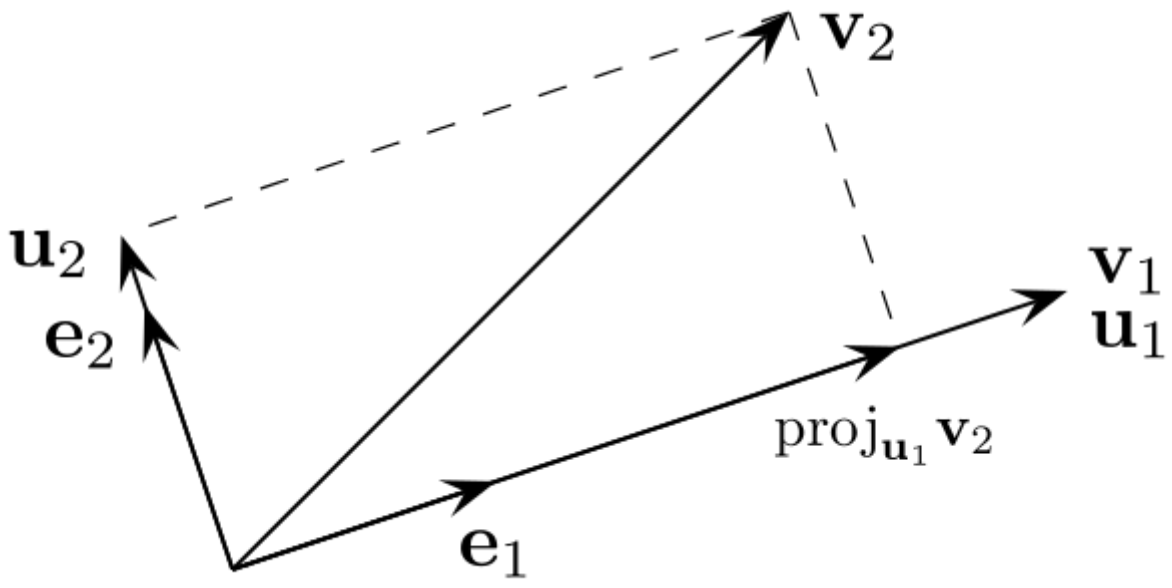
$$\sum_{i=1}^{n} \|\boldsymbol{\alpha}_i - \text{proj}_{\mathcal{Z}}(\boldsymbol{\alpha}_i)\|^2$$

over all subspaces of dimension at most k also maximizes

$$\sum_{i=1}^{n} \|\text{proj}_{\mathcal{Z}}(\boldsymbol{\alpha}_i)\|^2$$

over the same subspaces.

Proof idea: This is a straightforward application of the triangle inequality.



(Source (https://commons.wikimedia.org/wiki/File:Gram–Schmidt_process.svg))

Proof: By Pythagoras,

$$\|\boldsymbol{\alpha}_i - \mathrm{proj}_{\mathcal{Z}}(\boldsymbol{\alpha}_i)\|^2 + \|\mathrm{proj}_{\mathcal{Z}}(\boldsymbol{\alpha}_i)\|^2 = \|\boldsymbol{\alpha}_i\|^2$$

since, by the *Orthogonal Decomposition Lemma*, $\mathrm{proj}_{\mathcal{Z}}(\boldsymbol{\alpha}_i)$ is orthogonal to $\boldsymbol{\alpha}_i - \mathrm{proj}_{\mathcal{Z}}(\boldsymbol{\alpha}_i)$. Rearranging,

$$\|\boldsymbol{\alpha}_i - \mathrm{proj}_{\mathcal{Z}}(\boldsymbol{\alpha}_i)\|^2 = \|\boldsymbol{\alpha}_i\|^2 - \|\mathrm{proj}_{\mathcal{Z}}(\boldsymbol{\alpha}_i)\|^2.$$

The result follows from the fact that the first term on the right-hand side does not depend on the choice of $\mathcal{Z}$. □

The lemma immediately gives a characterization of the solution to the simplest version of the problem, the rank-1 case. Indeed a one-dimensional space $\mathcal{Z}$ is determined by a unit vector $\mathbf{v}$. The projection $\boldsymbol{\alpha}_i$ onto $\mathbf{v}$ is given by the inner product formula $\langle \boldsymbol{\alpha}_i, \mathbf{v} \rangle \mathbf{v}$. So

$$\sum_{i=1}^{n} \|\mathrm{proj}_{\mathcal{Z}}(\boldsymbol{\alpha}_i)\|^2 = \sum_{i=1}^{n} \langle \boldsymbol{\alpha}_i, \mathbf{v} \rangle^2 = \|A\mathbf{v}\|^2$$

where, again, A is the matrix with rows $\boldsymbol{\alpha}_i^T$, $i = 1, \ldots, n$. So the solution is

$$\mathbf{v}_1 \in \arg\max\{\|A\mathbf{v}\| : \|\mathbf{v}\| = 1\}.$$

Here $\arg\max$ means that $\mathbf{v}_1$ is a $\mathbf{v}$ that achieves the maximum. Note that there could be more than one such $\mathbf{v}$, in which case we pick an arbitrary one.

Exercise: Construct a matrix $A \in \mathbb{R}^{n \times n}$ for which there exist multiple solutions to the maximization problem above. ◁

For general k, we are looking for an orthonormal set $\mathbf{v}_1, \ldots, \mathbf{v}_k$ of size k that maximizes

$$\begin{aligned}\sum_{i=1}^{n} \|\text{proj}_{\mathcal{Z}}(\boldsymbol{\alpha}_i)\|^2 &= \sum_{i=1}^{n} \left\| \sum_{j=1}^{k} \langle \boldsymbol{\alpha}_i, \mathbf{v}_j \rangle \, \mathbf{v}_j \right\|^2 \\ &= \sum_{i=1}^{n} \sum_{j=1}^{k} \langle \boldsymbol{\alpha}_i, \mathbf{v}_j \rangle^2 \\ &= \sum_{j=1}^{k} \left(\sum_{i=1}^{n} \langle \boldsymbol{\alpha}_i, \mathbf{v}_j \rangle^2 \right) \\ &= \sum_{j=1}^{k} \|A\mathbf{v}_j\|^2\end{aligned}$$

by orthonormality, where $\mathcal{Z}$ is the subspace spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_k$. We show next that a simple algorithm solves this problem.

2.4 A greedy algorithm

Remarkably, the following sequence of subproblems leads to the solution. We first compute

$$\mathbf{v}_1 \in \arg\max\{\|A\mathbf{v}\| : \|\mathbf{v}\| = 1\}.$$

By the *Extreme Value Theorem*, we know such a $\mathbf{v}_1$ exists (but may not be unique).

Then we consider all unit vectors orthogonal to $\mathbf{v}_1$ and compute

$$\mathbf{v}_2 \in \arg\max\{\|A\mathbf{v}\| : \|\mathbf{v}\| = 1, \ \langle \mathbf{v}, \mathbf{v}_1 \rangle = 0\}.$$

Again, such a $\mathbf{v}_2$ exists by the *Extreme Value Theorem*. Then proceeding by induction

$$\mathbf{v}_i \in \arg\max\{\|A\mathbf{v}\| : \|\mathbf{v}\| = 1, \ \langle \mathbf{v}, \mathbf{v}_j \rangle = 0, \forall j \leq i-1\}.$$

In words, the claim -- which requires a proof -- is that the best k-dimensional approximating subspace is obtained by finding the best 1-dimensional subspace, then the best 1-dimensional subspace orthogonal to the first one, and so on. That is, we can proceed greedily (https://en.wikipedia.org/wiki/Greedy_algorithm). While it is clear that, after k steps, this procedure constructs an orthonormal set of size k, it is far from obvious that it maximizes $\sum_{j=1}^{k} \|A\mathbf{v}_j\|^2$ over all such sets. We establish this claim next.

First, we will need the following basic linear algebra fact, which is left as an exercise.

Exercise: Establish the following facts:

(a) Let $\mathcal{Z} \subseteq \mathcal{W}$ be linear subspaces such that $\dim(\mathcal{Z}) < \dim(\mathcal{W})$. Show that there exists a unit vector $\mathbf{w} \in \mathcal{W}$ that is orthogonal to $\mathcal{Z}$.

(b) Let $\mathcal{W} = \text{span}(\mathbf{w}_1, \ldots, \mathbf{w}_\ell)$ and $\mathbf{z} \in \mathcal{W}$ of unit norm. Use the *Linear Dependence Lemma* and Gram-Schmidt to show that there exists an orthonormal basis of $\mathcal{W}$ that includes $\mathbf{z}$. $\lhd$

Theorem (Greedy Finds Best Fit): Let $A \in \mathbb{R}^{n \times m}$ be a matrix with rows $\boldsymbol{\alpha}_i^T$, $i = 1, \ldots, n$. For any $k \leq \text{rk}(A)$, let $\mathbf{v}_1, \ldots, \mathbf{v}_k$ be the greedy sequence constructed above. Then $\mathcal{Z}^* = \text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$ is a solution to the minimization problem

$$\min \left\{ \sum_{i=1}^{n} \|\boldsymbol{\alpha}_i - \text{proj}_{\mathcal{Z}}(\boldsymbol{\alpha}_i)\|^2 \ : \ \mathcal{Z} \text{ is a linear subspace of dimension} \leq k \right\}.$$

Proof idea: We proceed by induction. For an arbitrary orthonormal set $\mathbf{w}_1, \ldots, \mathbf{w}_k$, we use the exercise above to find an orthonormal basis of their span containing an element orthogonal to $\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}$. Then we use the defintion of $\mathbf{v}_k$ to conclude.

Proof: As we explained in the previous subsection, we reformulate the problem as the maximization

$$\max \left\{ \sum_{j=1}^{k} \|A\mathbf{w}_j\|^2 \ : \ \{\mathbf{w}_1, \ldots, \mathbf{w}_k\} \text{ is an orthonormal set} \right\}.$$

We proceed by induction. For $k = 1$, we define $\mathbf{v}_1$ precisely as a solution of the above maximization. For $\ell < k$, assume that for any orthonormal set $\{\mathbf{w}'_1, \ldots, \mathbf{w}'_\ell\}$, we have

$$\sum_{j=1}^{\ell} \|A\mathbf{w}'_j\|^2 \leq \sum_{j=1}^{\ell} \|A\mathbf{v}_j\|^2.$$

Now consider an orthonormal set $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ and let its span be $\mathcal{W} = \text{span}(\mathbf{w}_1, \ldots, \mathbf{w}_k)$.

Step 1. For $j = 1, \dots, k-1$, let $\mathbf{v}'_j$ be the orthogonal projection of $\mathbf{v}_j$ onto $\mathcal{W}$ and let $\mathcal{V}' = \operatorname{span}(\mathbf{v}'_1, \dots, \mathbf{v}'_{k-1})$. Because $\mathcal{V}' \subseteq \mathcal{W}$ has dimension at most $k-1$ while $\mathcal{W}$ itself has dimension k, by the exercise above we can find an orthonormal basis $\mathbf{w}'_1, \dots, \mathbf{w}'_k$ of $\mathcal{W}$ such that $\mathbf{w}'_k$ is orthogonal to $\mathcal{V}'$. Then, for any $j = 1, \dots, k-1$, we have the decomposition $\mathbf{v}_j = \mathbf{v}'_j + (\mathbf{v}_j - \mathbf{v}'_j)$ where $\mathbf{v}'_j \in \mathcal{V}'$ is orthogonal to $\mathbf{w}'_k$ and $\mathbf{v}_j - \mathbf{v}'_j$ is also orthogonal to $\mathbf{w}'_k \in \mathcal{W}$ by the properties of the orthogonal projection. Hence

$$\left\langle \sum_{j=1}^{k-1} \beta_j \mathbf{v}_j, \mathbf{w}'_k \right\rangle = 0$$

for any β_j's. That is, $\mathbf{w}'_k$ is orthogonal to $\operatorname{span}(\mathbf{v}_1, \dots, \mathbf{v}_{k-1})$.

Step 2. By the induction hypothesis

$$\sum_{j=1}^{k-1} \|A\mathbf{w}'_j\|^2 \leq \sum_{j=1}^{k-1} \|A\mathbf{v}_j\|^2.$$

Moreover, recalling that the $\boldsymbol{\alpha}_i^T$'s are the rows of A,

$$\sum_{j=1}^{k} \|A\mathbf{w}_j\|^2 = \sum_{i=1}^{n} \|\operatorname{proj}_{\mathcal{W}}(\boldsymbol{\alpha}_i)\|^2 = \sum_{j=1}^{k} \|A\mathbf{w}'_j\|^2$$

since the $\mathbf{w}_j$'s and $\mathbf{w}'_j$'s form an orthonormal basis of the same subspace $\mathcal{W}$. Since $\mathbf{w}'_k$ is orthogonal to $\operatorname{span}(\mathbf{v}_1, \dots, \mathbf{v}_{k-1})$, by definition of $\mathbf{v}_k$

$$\|A\mathbf{w}'_k\|^2 \leq \|A\mathbf{v}_k\|^2.$$

Step 3. Putting everything together

$$\sum_{j=1}^{k} \|A\mathbf{w}_j\|^2 = \sum_{j=1}^{k-1} \|A\mathbf{w}'_j\|^2 + \|A\mathbf{w}'_k\|^2 \leq \sum_{j=1}^{k} \|A\mathbf{v}_j\|^2$$

which proves the claim. □

Note that we have not entirely solved the best approximating subspace problem from a computational point of view, as we have not given an efficient computational procedure to find a solution to the 1-dimensional subproblems. We've only shown that the solutions exist and have the right properties. We will take care of this in the next section.