Singular value decomposition

(Singular value decomposition)
Every $m \times n$ matrix A can be decomposed as $A = U\Sigma V^T$, where

- Σ is a (rectangular) diagonal matrix with nonnegative entries, $(m \times n)$

The diagonal entries σ_i are called the **singular values** of A.

- U is orthogonal, $(m \times m)$
- V is orthogonal. $(n \times n)$

Comment. If A is symmetric, then the singular value decomposition is already provided by the spectral theorem (the diagonalization of A). Moreover, in that case, $V = U$.
Important observations. If $A = U\Sigma V^T$, then $A^TA = V\Sigma^T\Sigma V^T$.

- Note that $\Sigma^T\Sigma$ is an $n \times n$ diagonal matrix. Its entries are σ_i^2 (the squares of the entries in Σ).
- A^TA is a symmetric matrix! (Why?!) Hence, by the spectral theorem, we are able to find V and $\Sigma^T\Sigma$.

In other words, V is obtained from the (orthonormally chosen) eigenvectors of A^TA. Likewise, the entries of $\Sigma^T\Sigma$ are the eigenvalues of A^TA; their square roots are the entries of Σ, the singular values.
Finally, the equation $AV = U\Sigma$ allows us to determine U. How?! (Hint: $A\boldsymbol{v}_i = \sigma_i \boldsymbol{u}_i$)

This results in the following **recipe** to determine the SVD $A = U\Sigma V^T$ for any matrix A.
Find an orthonormal basis of eigenvectors $\boldsymbol{v}_i$ of A^TA. Let λ_i be the eigenvalue of $\boldsymbol{v}_i$.

- V is the matrix with columns $\boldsymbol{v}_i$.
- Σ is the diagonal matrix with entries $\sigma_i = \sqrt{\lambda_i}$.
- U is the matrix with columns $\boldsymbol{u}_i = \frac{1}{\sigma_i} A\boldsymbol{v}_i$. If needed, fill in additional columns to make U orthogonal.

Example 154. Determine the SVD of $A = \begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix}$.

Solution. $A^TA = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$ has 8-eigenvector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and 2-eigenvector $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$.

Since $A^TA = V\Sigma^2V^T$ (here, $\Sigma^T\Sigma = \Sigma^2$), we conclude that $V = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \sqrt{8} & \\ & \sqrt{2} \end{bmatrix}$.

From $A\boldsymbol{v}_i = \sigma_i \boldsymbol{u}_i$, we find $\boldsymbol{u}_1 = \frac{1}{\sigma_1} A\boldsymbol{v}_1 = \frac{1}{\sqrt{8}}\begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

Likewise, $\boldsymbol{u}_2 = \frac{1}{\sigma_2} A\boldsymbol{v}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Hence, $U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Check that, indeed, $A = U\Sigma V^T$!

Comment. For applications, it is common to arrange the singular values in decreasing order like we did.

Comment. If we had chosen $V = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 & -1 \\ -1 & 1 \end{bmatrix}$ instead, then $U = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \sqrt{8} & \\ & \sqrt{2} \end{bmatrix}$.

As with diagonalization, there are choices! (A lot fewer choices though.) This is another perfectly fine SVD. In fact, it's what Sage computes below.

Armin Straub
straub@southalabama.edu

Sage. Let's have Sage do the work for us. In Sage, the SVD is currently only implemented for floating point numbers. (RDF is the real numbers as floating point numbers with double precision)

```
Sage] A = matrix(RDF, [[2,2],[-1,1]])
```

```
Sage] U,S,V = A.SVD()
```

```
Sage] U
```

$$\begin{bmatrix} -1.0 & 1.11022302463 \times 10^{-16} \\ 8.64109131471 \times 10^{-17} & 1.0 \end{bmatrix}$$

```
Sage] S
```

$$\begin{bmatrix} 2.82842712475 & 0.0 \\ 0.0 & 1.41421356237 \end{bmatrix}$$

```
Sage] V
```

$$\begin{bmatrix} -0.707106781187 & -0.707106781187 \\ -0.707106781187 & 0.707106781187 \end{bmatrix}$$

Review. SVD

Example 155. Determine the SVD of $A=\begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix}$.

Comment. In contrast to our previous example, $\operatorname{rank}(A)=1$. It follows that A^TA has eigenvalue 0, so that 0 is a singular value of A.

Solution. $A^TA=\begin{bmatrix} 5 & 5 \\ 5 & 5 \end{bmatrix}$ has 10-eigenvector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and 0-eigenvector $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$.

We conclude that $V=\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ and $\Sigma=\begin{bmatrix} \sqrt{10} & \\ & 0 \end{bmatrix}$.

$\boldsymbol{u}_1=\frac{1}{\sigma_1}A\boldsymbol{v}_1=\frac{1}{\sqrt{10}}\begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}=\frac{1}{\sqrt{20}}\begin{bmatrix} 4 \\ 2 \end{bmatrix}=\frac{1}{\sqrt{5}}\begin{bmatrix} 2 \\ 1 \end{bmatrix}$

We cannot obtain $\boldsymbol{u}_2$ in the same way because $\sigma_2=0$. Since for every vector $\boldsymbol{u}_2$, $A\boldsymbol{v}_2=\sigma_2\boldsymbol{u}_2$, we can choose $\boldsymbol{u}_2$ as we wish, as long as the columns of U are orthonormal in the end.

$\boldsymbol{u}_2=\frac{1}{\sqrt{5}}\begin{bmatrix} -1 \\ 2 \end{bmatrix}$ (but $\boldsymbol{u}_2=\frac{1}{\sqrt{5}}\begin{bmatrix} 1 \\ -2 \end{bmatrix}$ works just as well)

Hence, $U=\frac{1}{\sqrt{5}}\begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}$.

In summary, $A=U\Sigma V^T$ with $U=\frac{1}{\sqrt{5}}\begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}$, $\Sigma=\begin{bmatrix} \sqrt{10} & \\ & 0 \end{bmatrix}$, $V=\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$.

Check. Do check that, indeed, $A=U\Sigma V^T$.

Example 156. Determine the SVD of $A=\begin{bmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$.

Solution. $A^TA=\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ has 3-eigenvector $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$ and 1-eigenvector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

Since $A^TA=V\Sigma^T\Sigma V^T$, we conclude that $V=\frac{1}{\sqrt{2}}\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}$ and $\Sigma=\begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$.

$\boldsymbol{u}_1=\frac{1}{\sigma_1}A\boldsymbol{v}_1=\frac{1}{\sqrt{3}}\begin{bmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 1 \end{bmatrix}=\frac{1}{\sqrt{6}}\begin{bmatrix} -2 \\ 1 \\ -1 \end{bmatrix}$

$\boldsymbol{u}_2=\frac{1}{\sigma_2}A\boldsymbol{v}_2=\frac{1}{1}\begin{bmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}=\frac{1}{\sqrt{2}}\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$

$\boldsymbol{u}_3$ is chosen so that the matrix U is orthogonal. Hence, $\boldsymbol{u}_3=\frac{1}{\sqrt{3}}\begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}$ (or $\boldsymbol{u}_3=\frac{1}{\sqrt{3}}\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$).

Hence, $U=\begin{bmatrix} -2/\sqrt{6} & 0 & -1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{2} & -1/\sqrt{3} \\ -1/\sqrt{6} & 1/\sqrt{2} & 1/\sqrt{3} \end{bmatrix}$.

In summary, $A=U\Sigma V^T$ with $U=\begin{bmatrix} -2/\sqrt{6} & 0 & -1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{2} & -1/\sqrt{3} \\ -1/\sqrt{6} & 1/\sqrt{2} & 1/\sqrt{3} \end{bmatrix}$, $\Sigma=\begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$, $V=\frac{1}{\sqrt{2}}\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}$.

How did we find $\boldsymbol{u}_3$? We already have the vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, and need a vector orthogonal to both.

That is, we need to find the vector spanning $\operatorname{span}\left\{\begin{bmatrix} -2 \\ 1 \\ -1 \end{bmatrix},\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}\right\}^\perp=\operatorname{col}\left(\begin{bmatrix} -2 & 0 \\ 1 & 1 \\ -1 & 1 \end{bmatrix}\right)^\perp=\operatorname{null}\left(\begin{bmatrix} -2 & 1 & -1 \\ 0 & 1 & 1 \end{bmatrix}\right)$.

[Without the intermediate steps, can you see why the null space consists of precisely the vectors orthogonal to both $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$?]

More generally, proceeding like this, we can always fill in "missing" vectors $\boldsymbol{u}_i$ to obtain an orthonormal basis $\boldsymbol{u}_1,\boldsymbol{u}_2,\ldots,\boldsymbol{u}_m$ that we can use as the columns of U.

Example 157. Determine the SVD of $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$.

Solution. $A^TA = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$ has characteristic polynomial $(1-\lambda)(2-\lambda)-1 = \lambda^2 - 3\lambda + 1$.

The eigenvalues of A^TA are $\lambda_{1,2} = \frac{3 \pm \sqrt{5}}{2}$.

$\frac{3+\sqrt{5}}{2}$-eigenvector $\begin{bmatrix} 2 \\ 1+\sqrt{5} \end{bmatrix}$ and $\frac{3-\sqrt{5}}{2}$-eigenvector $\begin{bmatrix} 2 \\ 1-\sqrt{5} \end{bmatrix}$.

It would be rather painful to continue with exact expressions, and that is not how applications typically proceed. Numerically:

- 2.618-eigenvector $\begin{bmatrix} 0.526 \\ 0.851 \end{bmatrix}$ and 0.382-eigenvector $\begin{bmatrix} -0.851 \\ 0.526 \end{bmatrix}$. These eigenvectors are normalized, and it is now actually immediately obvious that they are orthogonal. (Of course, they had to be!)
- Hence, $\Sigma = \begin{bmatrix} \sqrt{2.618} & \\ & \sqrt{0.382} \end{bmatrix} = \begin{bmatrix} 1.618 & \\ & 0.618 \end{bmatrix}$ and $V = \begin{bmatrix} 0.526 & -0.851 \\ 0.851 & 0.526 \end{bmatrix}$.

  [We chose $\begin{bmatrix} -0.851 \\ 0.526 \end{bmatrix}$ instead of $\begin{bmatrix} 0.851 \\ -0.526 \end{bmatrix}$, so that, for the resulting V, $\det V = +1$.]
- $\boldsymbol{u}_1 = \frac{1}{\sigma_1} A\boldsymbol{v}_1 = \frac{1}{1.618} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.526 \\ 0.851 \end{bmatrix} = \begin{bmatrix} 0.851 \\ 0.526 \end{bmatrix}$

  $\boldsymbol{u}_2 = \frac{1}{\sigma_2} A\boldsymbol{v}_1 = \frac{1}{0.618} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -0.851 \\ 0.526 \end{bmatrix} = \begin{bmatrix} -0.526 \\ 0.851 \end{bmatrix}$.

  Hence, $U = \begin{bmatrix} 0.851 & -0.526 \\ 0.526 & 0.851 \end{bmatrix}$. (Again, notice the obvious orthogonality!)

Comment. The matrix A itself has eigenvalues $1, 1$, but the 1-eigenspace is only 1-dimensional. We are missing an eigenvector, which renders A not diagonalizable.

Comment. If we had continued symbolically, there are some magical simplifications like $\sqrt{\frac{3+\sqrt{5}}{2}} = \frac{1+\sqrt{5}}{2}$ going on. By the way, this is the golden ratio!

Sage. In Sage, the SVD is currently only implemented for floating point numbers (`RDF` is the real numbers as floating point numbers with double precision). Here's our computation:

```
Sage] A = matrix(RDF, [[1,1],[0,1]])

Sage] U,S,V = A.SVD()

Sage] U
```

$$\begin{bmatrix} 0.850650808352 & -0.525731112119 \\ 0.525731112119 & 0.850650808352 \end{bmatrix}$$

```
Sage] S
```

$$\begin{bmatrix} 1.61803398875 & 0.0 \\ 0.0 & 0.61803398875 \end{bmatrix}$$

```
Sage] V
```

$$\begin{bmatrix} 0.525731112119 & -0.850650808352 \\ 0.850650808352 & 0.525731112119 \end{bmatrix}$$

Armin Straub
straub@southalabama.edu

Example 158. (continued) The matrices U and V are rotation matrices. By what angle?

Why rotations? Recall that orthogonal matrices have determinant $+1$ or -1.
Since $\det U = +1$ and $\det V = +1$, the orthogonal matrices U, V are rotations.

Solution. Being rotation matrices, each of them equals $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ for some angle θ.

To find the angle θ_V for V, we compute $\arccos(0.526) = 1.017$. This means that $\theta_V = 1.017$ or $\theta_V = 2\pi - 1.017$ (make a sketch of $\cos(\theta)$ if that's unclear!). Since $\sin(1.017) = 0.851$ (whereas $\sin(2\pi - 1.017) = -0.851$), we conclude that V is a rotation by $\theta_V = 1.017 = 58.3°$. Keep that angle in mind for the next example!
Likewise, U is a rotation by $\theta_U = 0.554 = 31.7°$.

Comment. The two angles add up to $90°$. That's a consequence of the (atypical) fact that the matrices U and V have essentially the same entries.

Example 159. Explain the geometric meaning of the SVD in the previous example.

- The map $\boldsymbol{x} \mapsto A\boldsymbol{x}$ with $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ sends the (orthogonal) grid spanned by $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ to the (nonorthogonal) grid spanned by $A\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $A\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.
  Make a sketch! The two grids are overlayed in the first plot on the next page.
- Likewise, for instance, the (orthogonal) grid spanned by $\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 1 \end{bmatrix}$ (the $45°$ degree rotated version of the previous grid) is sent to the (again, nonorthogonal) grid spanned by $\frac{1}{\sqrt{2}}\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $\frac{1}{\sqrt{2}}\begin{bmatrix} 0 \\ 1 \end{bmatrix}$.
  Make a sketch! The two grids are overlayed in the second plot on the next page.
- Can we find an orthogonal grid which is sent to another orthogonal grid by A?
  Solution. Yes! The SVD $A = U\Sigma V^T$ is equivalent to $AV = U\Sigma$. That is, $A\boldsymbol{v}_i = \sigma_i \boldsymbol{u}_i$.
  In other words, the orthogonal grid spanned by $\boldsymbol{v}_1, \boldsymbol{v}_2$ is sent to the orthogonal grid spanned by $\sigma_1\boldsymbol{u}_1, \sigma_2\boldsymbol{u}_2$.
  As we observed earlier, the grid spanned by $\boldsymbol{v}_1, \boldsymbol{v}_2$ is the $58.3°$ degree rotated version of the standard grid)
  While the input grid consists of little squares, the output grid consists of rectangles with sides σ_1, σ_2.
  Make a sketch! The two grids are overlayed in the third plot on the next page.

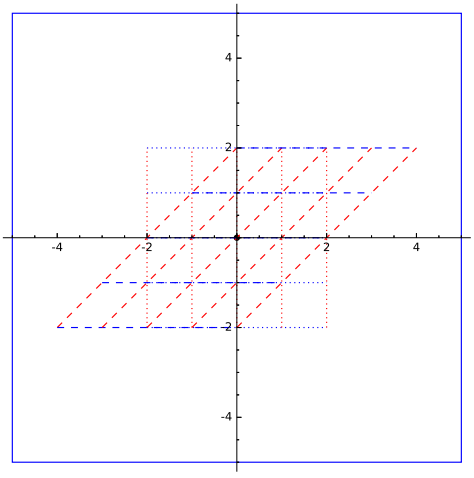
The following Sage code prepares the plots on the next page. Even if you have no coding background, can you see, roughly, what is happening?

```
Sage] def grid_lines(v1, v2, n, args={}):
          lines = Graphics()
          for i in [-n..n]:
              lines += line([i*v1-n*v2, i*v1+n*v2], color='red', **args)
              lines += line([i*v2-n*v1, i*v2+n*v1], color='blue', **args)
          return lines

Sage] def svd_rotate(angle):
          A = matrix([[1,1],[0,1]])
          t = angle*2*pi/360
          R = matrix([[cos(t),-sin(t)],[sin(t),cos(t)]])
          G1 = grid_lines(R*vector([1,0]), R*vector([0,1]), 2, {'linestyle':':'})
          G2 = grid_lines(A*R*vector([1,0]), A*R*vector([0,1]), 2, {'linestyle':'--'})
          B = polygon([(-5,-5), (-5,5), (5,5), (5,-5)], fill=False)
          O = point((0,0), pointsize=30,color='black')
          return B+O+G1+G2
```

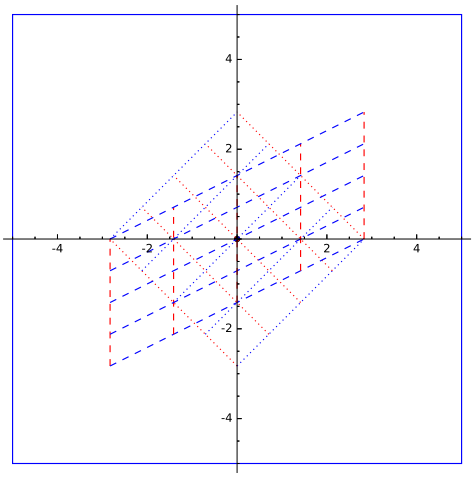
Grid spanned by $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ (dotted), and grid spanned by $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ (dashed):

```
Sage] svd_rotate(angle = 0)
```



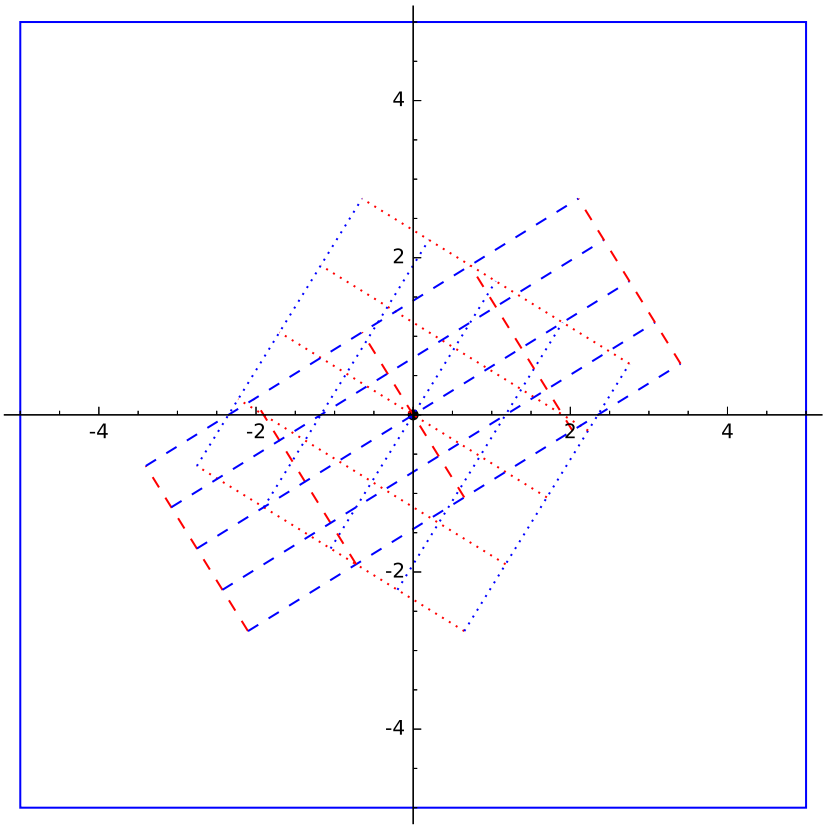
Grid spanned by $\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 1 \end{bmatrix}$ (dotted), and grid spanned by $\frac{1}{\sqrt{2}}\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $\frac{1}{\sqrt{2}}\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ (dashed):

```
Sage] svd_rotate(angle = 45)
```



Finally, here is the special situation (given by the SVD!) which shows an orthogonal grid (rotated by $58.3°$ degree) that is sent to another orthogonal grid (rotated by $31.7°$ degree):

```
Sage] svd_rotate(angle = 58.3)
```



For more pictures and detailed comments see the beautiful article:
http://www.ams.org/samplings/feature-column/fcarc-svd

Example 160. Show that the eigenvalues of A^TA are all nonnegative.

Proof. Suppose that λ is an eigenvalue of A^TA. Then $A^TA\boldsymbol{v} = \lambda\boldsymbol{v}$ (where $\boldsymbol{v}$ is a λ-eigenvector).
It follows that $\underbrace{\boldsymbol{v}^TA^TA\boldsymbol{v}}_{=\|A\boldsymbol{v}\|^2 \geqslant 0} = \lambda\boldsymbol{v}^T\boldsymbol{v} = \lambda\|\boldsymbol{v}\|^2$. Finally, $\lambda\|\boldsymbol{v}\|^2 \geqslant 0$ implies that $\lambda \geqslant 0$. □

The **pseudoinverse** of an $m \times n$ matrix A is the matrix A^+ such that the system $A\boldsymbol{x} = \boldsymbol{b}$ has "optimal" solution $\boldsymbol{x} = A^+\boldsymbol{b}$.

Here, "optimal" means that $\boldsymbol{x}$ is the smallest least squares solution.
In particular:

- If $A\boldsymbol{x} = \boldsymbol{b}$ has a unique solution, then $\boldsymbol{x} = A^+\boldsymbol{b}$ is that solution.
- If $A\boldsymbol{x} = \boldsymbol{b}$ has many solutions, then $\boldsymbol{x} = A^+\boldsymbol{b}$ is the one of smallest norm (the "optimal" one; and there is indeed only one such optimal solution).
- If $A\boldsymbol{x} = \boldsymbol{b}$ is inconsistent but has a unique least squares solution, then $\boldsymbol{x} = A^+\boldsymbol{b}$ is that least squares solution.
- If $A\boldsymbol{x} = \boldsymbol{b}$ has many least squares solutions, then $\boldsymbol{x} = A^+\boldsymbol{b}$ is the one with smallest norm.

When there is a unique (least squares) solution, we know how to find the pseudoinverse:

- If A is invertible, then $A^+ = A^{-1}$.
- If A has full column rank, then $A^+ = (A^TA)^{-1}A^T$.
  Recall. If $A\boldsymbol{x} = \boldsymbol{b}$ is inconsistent, a least squares solution can be determined by solving $A^TA\boldsymbol{x} = A^T\boldsymbol{b}$. If A has full column rank (i.e. the columns of A are independent; in this context, the typical case), then $\boldsymbol{x} = (A^TA)^{-1}A^T\boldsymbol{b}$ is the **unique** least squares solution to $A\boldsymbol{x} = \boldsymbol{b}$.

Example 161.

(a) What is the pseudoinverse of $\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 3 \\ 0 & 0 \end{bmatrix}$?

(b) What is the pseudoinverse of $\Sigma = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix}$?

(c) What is the pseudoinverse of $\Sigma = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$?

(d) In each case, compute $\Sigma^+\Sigma$ and $\Sigma\Sigma^+$.

Armin Straub
straub@southalabama.edu

Solution.

(a) Recall that, if A has full column rank, then $A^+ = (A^TA)^{-1}A^T$.

Here, $\Sigma^T\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 9 \end{bmatrix}$, so that $\Sigma^+ = (\Sigma^T\Sigma)^{-1}\Sigma^T = \begin{bmatrix} 1/4 & \\ & 1/9 \end{bmatrix}\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \end{bmatrix}$.

Alternative. Let us think about the optimal solution to $\Sigma\boldsymbol{x} = \boldsymbol{b}$, that is, $\begin{bmatrix} 2 & 0 \\ 0 & 3 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$.

The (unique) least squares solution is $\boldsymbol{x} = \begin{bmatrix} b_1/2 \\ b_2/3 \end{bmatrix}$. (Review if this is not obvious!)

Since $\begin{bmatrix} b_1/2 \\ b_2/3 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \end{bmatrix}\boldsymbol{b}$, we conclude that $\Sigma^+ = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \end{bmatrix}$.

(b) Let us think about the smallest norm ("optimal") solution to $\Sigma\boldsymbol{x} = \boldsymbol{b}$, that is, $\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$.

The general solution is $\boldsymbol{x} = \begin{bmatrix} b_1/2 \\ b_2/3 \\ t \end{bmatrix}$, where t is a free parameter.

Clearly, the smallest norm solution is $\begin{bmatrix} b_1/2 \\ b_2/3 \\ 0 \end{bmatrix}$.

Since $\begin{bmatrix} b_1/2 \\ b_2/3 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/3 \\ 0 & 0 \end{bmatrix}\boldsymbol{b}$, we conclude that $\Sigma^+ = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/3 \\ 0 & 0 \end{bmatrix}$.

(c) Now, $\Sigma\boldsymbol{x} = \boldsymbol{b}$, that is, $\begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ has no solution (unless $b_2 = 0$).

We therefore need to think about least squares solutions.

The general least squares solution (why?!) is $\boldsymbol{x} = \begin{bmatrix} b_1/2 \\ s \\ t \end{bmatrix}$, where s, t are free parameters.

Clearly, the smallest norm least squares solution is $\begin{bmatrix} b_1/2 \\ 0 \\ 0 \end{bmatrix}$.

Since $\begin{bmatrix} b_1/2 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}\boldsymbol{b}$, we conclude that $\Sigma^+ = \begin{bmatrix} 1/2 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$.

(d) Firstly, $\Sigma^+\Sigma = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 0 & 3 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $\Sigma\Sigma^+ = \begin{bmatrix} 2 & 0 \\ 0 & 3 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$.

Secondly, $\Sigma^+\Sigma = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/3 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ and $\Sigma\Sigma^+ = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix}\begin{bmatrix} 1/2 & 0 \\ 0 & 1/3 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

[Note how the pseudoinverse tries to behave like the regular inverse. But since Σ has only 2 columns, $\Sigma^+\Sigma$ and $\Sigma\Sigma^+$ can have rank at most 2 (so cannot be the full 3×3 identity).]

Thirdly, $\Sigma^+\Sigma = \begin{bmatrix} 1/2 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ and $\Sigma\Sigma^+ = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} 1/2 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$.

[Here, Σ has rank 1, so that $\Sigma^+\Sigma$ and $\Sigma\Sigma^+$ can have rank at most 1.]

In general. Proceeding, as in this example, we find that the pseudoinverse of any $m \times n$ diagonal matrix Σ is the $n \times m$ (transposed dimensions!) diagonal matrix whose nonzero entries are the inverses of the entries of Σ.

Comment. Observe that, in all three cases, $\Sigma^{++} = \Sigma$.

Comment. Note that $\begin{bmatrix} 1 & 0 \\ 0 & \varepsilon \end{bmatrix}^+ = \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon^{-1} \end{bmatrix}$ for small $\varepsilon \neq 0$, while $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}^+ = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$. This shows that the pseudoinverse is not a continuous operation.

It turns out that the pseudoinverse A^+ can be easily obtained from the SVD of A:

Theorem 162. The **pseudoinverse** of an $m \times n$ matrix A with SVD $A = U\Sigma V^T$ is

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

where Σ^+, the pseudoinverse of Σ, is the $n \times m$ diagonal matrix, whose nonzero entries are the inverses of the entries of Σ.

Proof. The equation $A\boldsymbol{x} = \boldsymbol{b}$ is equivalent to $U\Sigma V^T\boldsymbol{x} = \boldsymbol{b}$ and, thus, $\Sigma V^T\boldsymbol{x} = U^T\boldsymbol{b}$.
Write $\boldsymbol{y} = V^T\boldsymbol{x}$ and note that $\boldsymbol{y}$ and $\boldsymbol{x}$ have the same norm (why?!).
We already know that the equation $\Sigma\boldsymbol{y} = U^T\boldsymbol{b}$ has optimal solution $\boldsymbol{y} = \Sigma^+ U^T\boldsymbol{b}$.
Since $\boldsymbol{y}$ and $\boldsymbol{x}$ have the same norm, it follows that $\boldsymbol{x} = V\boldsymbol{y} = V\Sigma^+ U^T\boldsymbol{b}$ is the optimal solution to $A\boldsymbol{x} = \boldsymbol{b}$.
Hence, $A^+ = V\Sigma^+ U^T$. □

Lemma 163. The pseudoinverse of A^+ is $A^{++} = A$.

Proof. Starting with the SVD $A = U\Sigma V^T$, we have $A^+ = V\Sigma^+ U^T$, which is the SVD of A^+.
Therefore, $A^{++} = U\Sigma^{++} V^T$. The claim thus follows from $\Sigma^{++} = \Sigma$. □

Example 164. Determine the pseudoinverse of $A = \begin{bmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$ in two ways.

First, using the SVD and, second, using the fact that A has full column rank.

Solution. (SVD) We have computed the SVD of this matrix before.
Since $A = U\Sigma V^T$ with $U = \begin{bmatrix} -2/\sqrt{6} & 0 & -1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{2} & -1/\sqrt{3} \\ -1/\sqrt{6} & 1/\sqrt{2} & 1/\sqrt{3} \end{bmatrix}$, $\Sigma = \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$, $V = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}$,
the pseudoinverse is $A^+ = V\Sigma^+ U^T$ where $\Sigma^+ = \begin{bmatrix} 1/\sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$.
Multiplying these matrices, $A^+ = \frac{1}{3}\begin{bmatrix} 1 & 1 & 2 \\ -1 & 2 & 1 \end{bmatrix}$.

Comment. For many applications, it may be neither necessary nor helpful to multiply V, Σ^+, U^T.

Solution. (full column rank) Since A clearly has full column rank, we also have $A^+ = (A^TA)^{-1}A^T$.
Indeed, $A^+ = (A^TA)^{-1}A^T = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}^{-1}\begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 1 & 1 & 2 \\ -1 & 2 & 1 \end{bmatrix}$.

Example 165. What is the pseudoinverse of $A = \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix}$?

Solution. Recall (or compute) that $A = U\Sigma V^T$ with $U = \frac{1}{\sqrt{5}}\begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} \sqrt{10} & \\ & 0 \end{bmatrix}$, $V = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$.
Hence, $A^+ = V\Sigma^+ U^T$ where $\Sigma^+ = \begin{bmatrix} 1/\sqrt{10} & 0 \\ 0 & 0 \end{bmatrix}$.
Multiplying these matrices (which may not be necessary or helpful for applications), $A^+ = \frac{1}{10}\begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix}$.

Note. Since A does not have full column rank, $A^+ = (A^TA)^{-1}A^T$ cannot be used. That's because A^TA is not invertible.
Comment. Here, $A^+A = \boldsymbol{v}_1\boldsymbol{v}_1^T = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and $AA^+ = \boldsymbol{u}_1\boldsymbol{u}_1^T = \frac{1}{5}\begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix}$ are not visually like the identity. However, note that these are the (orthogonal) projections onto $\boldsymbol{v}_1$ and $\boldsymbol{u}_1$ respectively (in particular, the eigenvalues are $1, 0$).

Review.

- If the $m \times n$ matrix A has SVD $A = U\Sigma V^T$, then its pseudoinverse is $A^+ = V\Sigma^+ U^T$.

  Here, Σ^+, the pseudoinverse of Σ, is the $n \times m$ diagonal matrix, whose nonzero entries are the inverses of the entries of Σ.

- The system $A\boldsymbol{x} = \boldsymbol{b}$ has "optimal" solution $\boldsymbol{x} = A^+\boldsymbol{b}$.

  Here, "optimal" means that $\boldsymbol{x}$ is the smallest least squares solution.

Example 166.

(a) Find the pseudoinverse of $A = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$.

(b) Find the smallest solution to $x_1 + 2x_2 + 3x_3 = 6$.

As before, smallest solutions means the solution $\boldsymbol{x}$ such that $\|\boldsymbol{x}\|$ is as small as possible. One obvious solution is $\begin{bmatrix} 1,1,1 \end{bmatrix}^T$, but is it the smallest?

Solution.

(a) $A^TA = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}\begin{bmatrix} 1 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}$ has 14-eigenvector $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ and 0-eigenvectors $\begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -3 \\ 0 \\ 1 \end{bmatrix}$.

$\boldsymbol{u}_1 = \frac{1}{\sigma_1} A\boldsymbol{v}_1 = \frac{1}{\sqrt{14}}\begin{bmatrix} 1 & 2 & 3 \end{bmatrix}\frac{1}{\sqrt{14}}\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = 1$

Hence, $A = U\Sigma V^T$ with $U = \begin{bmatrix} 1 \end{bmatrix}$, $\Sigma = \begin{bmatrix} \sqrt{14} & 0 & 0 \end{bmatrix}$, $V = \begin{bmatrix} 1/\sqrt{14} & * & * \\ 2/\sqrt{14} & * & * \\ 3/\sqrt{14} & * & * \end{bmatrix}$.

$A^+ = V\Sigma^+U^T = \begin{bmatrix} 1/\sqrt{14} & * & * \\ 2/\sqrt{14} & * & * \\ 3/\sqrt{14} & * & * \end{bmatrix}\begin{bmatrix} 1/\sqrt{14} \\ 0 \\ 0 \end{bmatrix}\begin{bmatrix} 1 \end{bmatrix} = \frac{1}{14}\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$

Comment. No surprise on U. The only options for U are $U = \begin{bmatrix} 1 \end{bmatrix}$ and $U = \begin{bmatrix} -1 \end{bmatrix}$.

Comment. Realizing what we did here allows us to write down A^+ immediately for all $1 \times n$ matrices A. See Example 167.

Homework. Complete the SVD of A. That is, find an option for the two missing columns of V, so that V is an orthogonal matrix. In other words, find an orthonormal basis for the 0-eigenspace.

Comment. An even better approach would be to compute AA^T first (instead of A^TA) which would allow us to compute U first (rather than V first). Can you fill in the blanks?

(b) We are solving $A\boldsymbol{x} = \begin{bmatrix} 6 \end{bmatrix}$ with $A = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$ as in the previous example.

We conclude that the smallest solution is $\boldsymbol{x} = A^+\begin{bmatrix} 6 \end{bmatrix} = \frac{3}{7}\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$.

Compare. $\left\| \frac{3}{7}\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \right\| = \frac{3}{7}\sqrt{14} \approx 1.604$ is indeed smaller than, say, $\left\| \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\| = \sqrt{3} \approx 1.732$.

Geometric picture. The equation $x_1 + 2x_2 + 3x_3 = 6$ describes a plane (not through the origin), and we are asking for the point on that plane which is closest to the origin. That's a typical question in Calculus III. Note that $\begin{bmatrix} 1 & 2 & 3 \end{bmatrix}^T$ is the normal vector of the plane. Explain why the answer had to be a multiple of that normal vector!

Example 167. More generally, find the pseudoinverse of $A = \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}$.

Solution. As in the previous example, we see that the answer will be $A^+ = \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|^2}$ with $\boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$.

Comment. Likewise for $A = \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix}$.

Example 168. How is the rank of A reflected in its singular value decomposition $A=U\Sigma V^T$?

Solution. The rank of A is equal to the number of nonzero singular values.

Theorem 169. (matrix approximation lemma) Suppose A is a $m\times n$ matrix, and we want to approximate A using a matrix B of rank s (smaller than the rank of A).

Let $A=U\Sigma V^T$ be the SVD of A (with singular values in decreasing order).

Then, the best such approximation is $B=U_s\Sigma_s V_s^T$, where Σ_s is the $s\times s$ diagonal matrix with entries $\sigma_1,\sigma_2,\ldots,\sigma_s$ and U_s, V_s are obtained from U, V by only taking the first s columns.

Comment. Note that, by choosing s small compared to r, we can store an approximation of A using much less data. This approximation will be good if the omitted singular values $\sigma_{s+1},\sigma_{s+2},\ldots,\sigma_r$ are all "small".

Comment. Equivalently, $B=U\Sigma_s V^T$, where Σ_s is now obtained from Σ by setting all but the largest s singular values to 0. In other words, Σ_s has the values $\sigma_1,\sigma_2,\ldots,\sigma_s$ on its diagonal, followed by zeros.

In other words. Here is another common way to say the same thing:

- Observe that $A=U\Sigma V^T$ is equivalent to $A=\sum_{i=1}^{r}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^T$.
- Each matrix $\boldsymbol{u}_i\boldsymbol{v}_i^T$ has rank 1.
- The best rank s approximation to A is $B=\sum_{i=1}^{s}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^T$.

Advanced comment. Here, "best" approximation is measured using the Frobenius norm of a matrix A (which is the same as the norm of a vector with all the entries of A).

Example 170. Determine the best rank 1 approximation of $A=\begin{bmatrix}1&1&1\\1&0&-1\end{bmatrix}$.

Solution. We determine (do it!) that A has the SVD

$$A=\begin{bmatrix}1&0\\0&-1\end{bmatrix}\begin{bmatrix}\sqrt{3}&0&0\\0&\sqrt{2}&0\end{bmatrix}\begin{bmatrix}1/\sqrt{3}&-1/\sqrt{2}&1/\sqrt{6}\\1/\sqrt{3}&0&-2/\sqrt{6}\\1/\sqrt{3}&1/\sqrt{2}&1/\sqrt{6}\end{bmatrix}^T=\begin{bmatrix}1&1&1\\1&0&-1\end{bmatrix}.$$

Hence, the best rank 1 approximation of A is (that is, we keep 1 singular value only) is

$$\begin{bmatrix}1\\0\end{bmatrix}\begin{bmatrix}\sqrt{3}\end{bmatrix}\begin{bmatrix}1/\sqrt{3}\\1/\sqrt{3}\\1/\sqrt{3}\end{bmatrix}^T=\begin{bmatrix}1&1&1\\0&0&0\end{bmatrix}.$$

Comment. Equivalently, $\begin{bmatrix}1&0\\0&-1\end{bmatrix}\begin{bmatrix}\sqrt{3}&0&0\\0&0&0\end{bmatrix}\begin{bmatrix}1/\sqrt{3}&-1/\sqrt{2}&1/\sqrt{6}\\1/\sqrt{3}&0&-2/\sqrt{6}\\1/\sqrt{3}&1/\sqrt{2}&1/\sqrt{6}\end{bmatrix}^T=\begin{bmatrix}1&1&1\\0&0&0\end{bmatrix}$.

Example 171. Determine the best rank 1 approximation of $A=\begin{bmatrix}1&-1\\0&1\\1&0\end{bmatrix}$.

Solution. Recall that $A=U\Sigma V^T$ with $U=\begin{bmatrix}-2/\sqrt{6}&0&-1/\sqrt{3}\\1/\sqrt{6}&1/\sqrt{2}&-1/\sqrt{3}\\-1/\sqrt{6}&1/\sqrt{2}&1/\sqrt{3}\end{bmatrix}$, $\Sigma=\begin{bmatrix}\sqrt{3}&0\\0&1\\0&0\end{bmatrix}$, $V=\frac{1}{\sqrt{2}}\begin{bmatrix}-1&1\\1&1\end{bmatrix}$.

Hence, the best rank 1 approximation of A is $\frac{1}{\sqrt{6}}\begin{bmatrix}-2\\1\\-1\end{bmatrix}\begin{bmatrix}\sqrt{3}\end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix}-1\\1\end{bmatrix}^T=\frac{1}{2}\begin{bmatrix}2&-2\\-1&1\\1&-1\end{bmatrix}$.

Example 172. (image compression) Let us load a 341x512 grayscale photo and store it as a matrix A. Each entry of the matrix is a value between 0 (black) and 1 (white).

The beautiful picture is taken from: http://www.southalabama.edu/departments/publicrelations/brand/photography.html

[The same approach works with color pictures. These are often represented by three matrices: one for the red component of the pixel, one for the green and for the blue component (RGB color scheme).]

```
Sage] import pylab
```

```
Sage] A = matrix(pylab.imread('/home/armin/photo.png'))
```

```
Sage] A.dimensions()
```

$(341, 512)$

```
Sage] A[0,0]
```

0.137254908681

```
Sage] matrix_plot(A, cmap='gray')
```

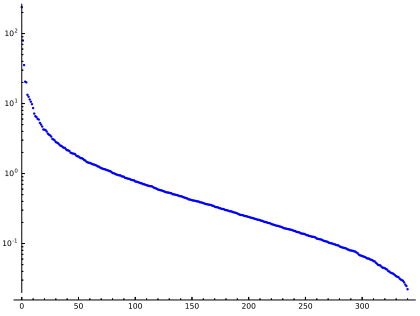
Next, we compute the SVD of A. Despite the size of A that takes the computer only a fraction of a second:

```
Sage] U,S,V = A.SVD()
```

```
Sage] S.diagonal()[:6]
```

$[238.443435709, 79.4429775448, 35.4540786319, 20.5662302846, 20.0697710337, 13.3421216529]$

```
Sage] list_plot(S.diagonal(), scale='semilogy')
```



As we can see, the magnitude of the singular values drops off quickly. We get a good approximation to A (our original photo) by computing a best rank s approximation to A by computing $U_s\Sigma_s V_s$ where Σ_s is the $s \times s$ diagonal matrix with entries $\sigma_1, \sigma_2, ..., \sigma_s$ and U_s, V_s are obtained from the corresponding matrices in the SVD $A = U\Sigma V^T$ by only taking the first s columns.

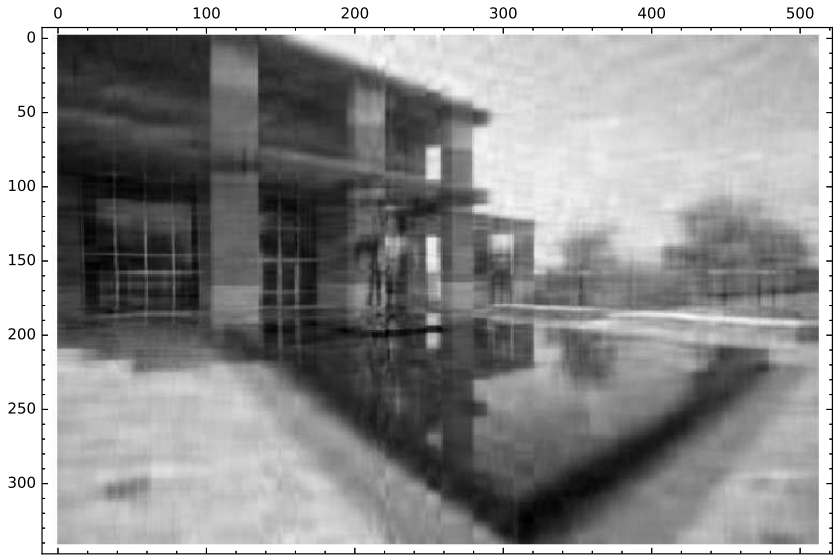
```
Sage] def A_approx(s):
          U0 = U.matrix_from_columns([0..s-1])
          S0 = diagonal_matrix(S.diagonal()[:s])
          V0 = V.matrix_from_columns([0..s-1])
          return U0*S0*V0.transpose()
```

Taking only 100 of the 341 singular values, we get an approximation, which is almost as good as the original:

```
Sage] matrix_plot(A_approx(100), cmap='gray')
```

But notice the development of artifacts. Taking only 20 of the 341 singular values, a lot is lost:

```
Sage] matrix_plot(A_approx(20), cmap='gray')
```



Comment. Image compression is just one (nice visual) example of the power of SVD. A variation of this approach can, for instance, also be used for image denoising. Much more generally, the SVD is able to extract the most important features of any sort of data!

Review. matrix approximation and compression

Function spaces

Recall the following:

- We call objects **vectors** if they can be added and scaled (subject to the usual laws).
- A set of vectors is a **vector space** if it is closed under addition and scaling.

  In other words, vector spaces are spans.

We will now discuss spaces of vectors, where the vectors are functions.

Why? Just one example why it is super useful to apply our linear algebra machinery to functions: we discussed the **distance** between vectors and how to find vectors closest to interesting subspaces (i.e. orthogonal projections). These notions are important for functions, too. For instance, given a (complicated) function, we want to find the closest function in a subspace of (simple) functions. In other words, we want to approximate functions using other (typically, simpler) functions.

Comment. Functions $f(x)$ and $g(x)$ can also be multiplied. This is an extra structure (it makes appropriate sets of functions an **algebra**, which is something more special than a **vector space**), which we ignore during our discussion of vector spaces.

An inner product on function spaces

On the space of, say, (piecewise) continuous functions $f\colon [a,b] \to \mathbb{R}$, it is natural to consider the dot product

$$\langle f, g \rangle = \int_a^b f(t)g(t)\mathrm{d}t.$$

Why? A (sensible) dot product provides a (sensible) notion of distance between functions. The dot product above is the continuous analog of the usual dot product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{t=1}^n x_t y_t$ for vectors in $\mathbb{R}^n$. Do you see it?!

As a consequence, once we have the dot product, we can orthogonally project functions onto spaces of simple functions. In other words, we can compute best approximations of functions by simple functions (for instance, best quadratic approximations).

Why continuous? We need that any product $f(x)g(x)$ is integrable. That means we cannot work with all functions. Continuity is certainly sufficient. In fact, the right condition is that $f(x)^2$ should be integrable on $[a,b]$ (i.e. $f(x)$ is square-integrable). Such a function is said to be in $\mathcal{L}^2[a,b]$.

Example 173. What is the orthogonal projection of $f\colon [a,b] \to \mathbb{R}$ onto the space of constant functions (that is, $\operatorname{span}\{1\}$)?

Solution. The orthogonal projection of $f\colon [a,b] \to \mathbb{R}$ onto $\operatorname{span}\{1\}$ is

$$\frac{\langle f, 1 \rangle}{\langle 1, 1 \rangle} 1 = \frac{\int_a^b f(t)1\mathrm{d}t}{\int_a^b 1^2 \mathrm{d}t} = \frac{1}{b-a}\int_a^b f(t)\mathrm{d}t.$$

This is the average of $f(x)$ on $[a,b]$.

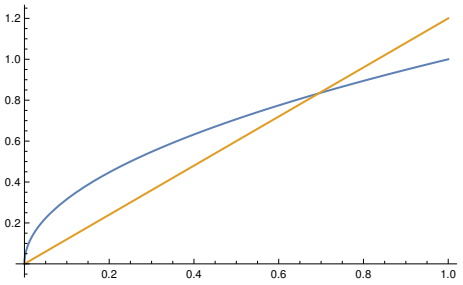
Comment. Makes perfect sense, doesn't it? Intuitively, the best approximation of a function by a constant should indeed be the one where the constant is the average.

Armin Straub
straub@southalabama.edu

Example 174. Find the best approximation of $f(x)=\sqrt{x}$ on the interval $[0,1]$ using a function of the form $y=ax$.

Solution. The orthogonal projection of $f\colon[0,1]\to\mathbb{R}$ onto $\operatorname{span}\{x\}$ is

$$\frac{\langle f,x\rangle}{\langle x,x\rangle}x=\frac{\int_0^1 f(t)t\mathrm{d}t}{\int_0^1 t^2\mathrm{d}t}x=3x\int_0^1 t\,f(t)\mathrm{d}t.$$



In our case, the best approximation is

$$3x\int_0^1 t\sqrt{t}\mathrm{d}t=3x\int_0^1 t^{3/2}\mathrm{d}t=3x\Big[\frac{1}{5/2}t^{5/2}\Big]_0^1=\frac{6}{5}x.$$

Example 175. Find the best approximation of $f(x)=\sqrt{x}$ on the interval $[0,1]$ using a function of the form $y=a+bx$.

Important observation. The orthogonal projection of $f\colon[0,1]\to\mathbb{R}$ onto $\operatorname{span}\{1,x\}$ is not simply the projection onto 1 plus the projection onto x. That's because 1 and x are not orthogonal:

$$\langle 1,x\rangle=\int_0^1 t\,\mathrm{d}t=\frac{1}{2}\neq 0.$$

Solution. To find an orthogonal basis for $\operatorname{span}\{1,x\}$, following Gram–Schmidt, we compute

$$x-\begin{pmatrix}\text{projection of}\\ x\text{ onto }1\end{pmatrix}=x-\frac{\langle x,1\rangle}{\langle 1,1\rangle}1=x-\frac{1}{2}.$$

Hence, $1, x-\frac{1}{2}$ is an orthogonal basis for $\operatorname{span}\{1,x\}$.
The orthogonal projection of $\sqrt{x}$ on $[0,1]$ onto $\operatorname{span}\{1,x\}=\operatorname{span}\left\{1,x-\frac{1}{2}\right\}$ therefore is

$$\frac{\langle\sqrt{x},1\rangle}{\langle 1,1\rangle}1+\frac{\left\langle\sqrt{x},x-\frac{1}{2}\right\rangle}{\left\langle x-\frac{1}{2},x-\frac{1}{2}\right\rangle}\left(x-\frac{1}{2}\right) \;=\; \frac{\int_0^1\sqrt{t}\,\mathrm{d}t}{\int_0^1 1\,\mathrm{d}t}+\frac{\int_0^1\sqrt{t}\left(t-\frac{1}{2}\right)\mathrm{d}t}{\int_0^1\left(t-\frac{1}{2}\right)^2\mathrm{d}t}\left(x-\frac{1}{2}\right).$$

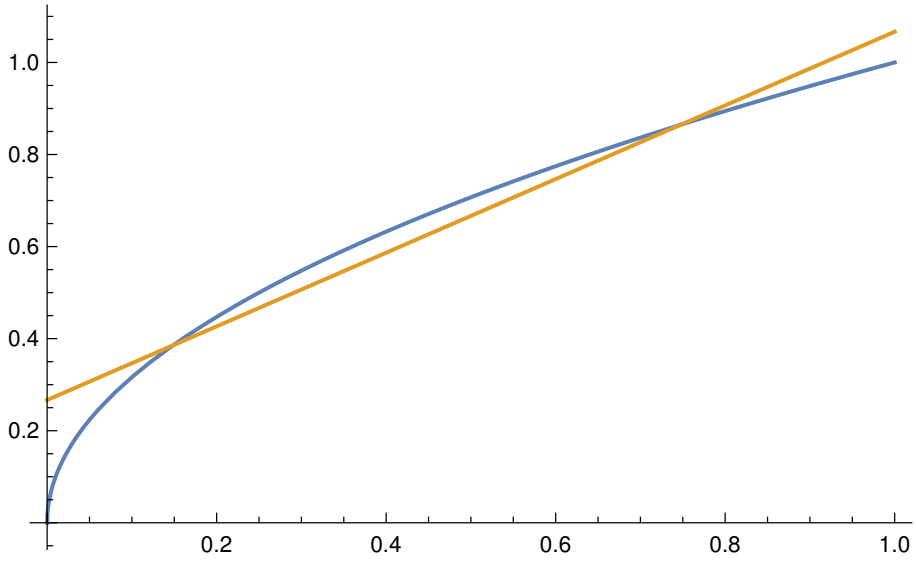
We compute the three new integrals:

$$\begin{aligned}\int_0^1\sqrt{t}\,\mathrm{d}t &= \left[\frac{2}{3}t^{3/2}\right]_0^1=\frac{2}{3}\\ \int_0^1\sqrt{t}\left(t-\frac{1}{2}\right)\mathrm{d}t &= \int_0^1\left(t^{3/2}-\frac{1}{2}t^{1/2}\right)\mathrm{d}t=\left[\frac{2}{5}t^{5/2}-\frac{1}{3}t^{3/2}\right]_0^1=\frac{2}{5}-\frac{1}{3}=\frac{1}{15}\\ \int_0^1\left(t-\frac{1}{2}\right)^2\mathrm{d}t &= \int_0^1\left(t^2-t+\frac{1}{4}\right)\mathrm{d}t=\left[\frac{1}{3}t^3-\frac{1}{2}t^2+\frac{1}{4}t\right]_0^1=\frac{1}{3}-\frac{1}{2}+\frac{1}{4}=\frac{1}{12}\end{aligned}$$

Using these values, the best approximation is

$$\frac{\int_0^1\sqrt{t}\,\mathrm{d}t}{\int_0^1 1\,\mathrm{d}t}+\frac{\int_0^1\sqrt{t}\left(t-\frac{1}{2}\right)\mathrm{d}t}{\int_0^1\left(t-\frac{1}{2}\right)^2\mathrm{d}t}\left(x-\frac{1}{2}\right) \;=\; \frac{2}{3}+\frac{12}{15}\left(x-\frac{1}{2}\right)=\frac{4}{5}x+\frac{4}{15}$$

The plot below confirms how good this linear approximation is (compare with the previous example):



Armin Straub
straub@southalabama.edu

Example 176. Give a basis for the space of all polynomials.

Solution. $1, x, x^2, x^3, \ldots$

Indeed, every polynomial $p(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$ can be written uniquely as a sum of these basis elements. ("can be" = span; "uniquely" = independent)

Comment. The dimension is ∞. But we can make a list of basis elements, which is the "smallest kind of ∞" and is referred to as **countably infinite**. For the space of all functions, no such list can be made.

Just for fun. Let us indicate this difference in infiniteness in a slightly simpler situation: first, the natural numbers $0, 1, 2, 3\ldots$ are infinite but they are countable, because we can make a (infinite but complete) list starting with a first, then a second element and so on (hence, the name "countable"). On the other hand, consider the real numbers between 0 and 1. Clearly, there are infinitely many such numbers. The somewhat shocking fact (first realized by Georg Cantor in 1874) is that every attempt of making a complete list of these numbers must fail because every list will inevitably miss some numbers. Here's a brief indication of how the famous diagonal argument goes: suppose you can make a list, say:

$$\begin{array}{ll} \#1 & 0.111111\ldots \\ \#2 & 0.123456\ldots \\ \#3 & 0.750000\ldots \\ \vdots & \end{array}$$

Now, we are going to construct a new number $x = 0.x_1x_2x_3\ldots$ with decimal digits x_i in such a way that the digit x_i differs (by more than 1) from the ith digit of number $\#i$ on our list. For instance, $0.352\ldots$ in our case (for instance, $x_3 = 2$ differs from 0, the 3rd digit of sequence $\#3$). By construction, the number x is missing from the list.

Comment on fun. The statement "some infinities are bigger than others" nicely captures our observation. It appears in the book *The Fault in Our Stars* by John Green, where it is said by a cranky old author who attributes it to Cantor. Hazel, the main character, later reflects on that statement and compares $[0, 1]$ to $[0, 2]$. Can you explain why that is actually not what Cantor meant...?

Orthogonal polynomials

Let us think about the space of all polynomials (with real coefficients). On that space, we consider the dot product

$$\langle p_1, p_2 \rangle = \int_{-1}^{1} p_1(t) p_2(t) \mathrm{d}t. \tag{1}$$

Comment. That dot product is useful if we are thinking about the polynomials as functions on $[-1, 1]$ only. You can, of course, consider any other interval and you will obtain a shifted version of what we get here.

Example 177. Are $1, x, x^2, \ldots$ orthogonal (with respect to the inner product (1))?

Solution. Since $\langle x^r, x^s \rangle = \int_{-1}^{1} t^r t^s \mathrm{d}t = \int_{-1}^{1} t^{r+s} \mathrm{d}t$, we find that $\langle x^r, x^s \rangle = \begin{cases} \frac{2}{r+s+1}, & \text{if } r+s \text{ is even,} \\ 0, & \text{otherwise.} \end{cases}$

Hence, if $r + s$ is odd, then the monomials x^r and x^s are orthogonal. On the other hand, if $r + s$ is even, then x^r and x^s are not orthogonal.

Armin Straub
straub@southalabama.edu

Example 178. Use Gram-Schmidt to produce an orthogonal basis $\boldsymbol{p}_0, \boldsymbol{p}_1, \boldsymbol{p}_2, \ldots$ for the space of polynomials with the dot product (1). Compute $\boldsymbol{p}_0, \boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3, \boldsymbol{p}_4$.

Instead of normalizing these polynomials, **standardize** them so that $\boldsymbol{p}_n(1) = 1$.

Solution. We construct an orthogonal basis $\boldsymbol{p}_0, \boldsymbol{p}_1, \boldsymbol{p}_2, \ldots$ from $1, x, x^2, \ldots$ as follows:

- Starting with 1, we find $\boldsymbol{p}_0(x) = 1$.

  For future reference, let us note that $\|\boldsymbol{p}_0\|^2 = \int_{-1}^{1} 1 \mathrm{d}x = 2$.

- Starting with x, Gram–Schmidt produces $x - \begin{pmatrix}\text{projection of}\\ x \text{ onto } \boldsymbol{p}_0\end{pmatrix} = x - \frac{\langle x, \boldsymbol{p}_0\rangle}{\langle \boldsymbol{p}_0, \boldsymbol{p}_0\rangle}\boldsymbol{p}_0 = x - \int_{-1}^{1} t \mathrm{d}t = x$.

  Again, that's already standardized, so that $\boldsymbol{p}_1(x) = x$.

  Comment. The previous problem already told us that x is orthogonal to 1.

  For future reference, let us note that $\|\boldsymbol{p}_1\|^2 = \int_{-1}^{1} t^2 \mathrm{d}t = \frac{2}{3}$.

- Starting with x, Gram–Schmidt produces $x^2 - \begin{pmatrix}\text{projection of } x^2\\ \text{onto } \operatorname{span}\{\boldsymbol{p}_0, \boldsymbol{p}_1\}\end{pmatrix} = x^2 - \frac{\langle x^2, \boldsymbol{p}_0\rangle}{\langle \boldsymbol{p}_0, \boldsymbol{p}_0\rangle}\boldsymbol{p}_0 - \frac{\langle x^2, \boldsymbol{p}_1\rangle}{\langle \boldsymbol{p}_1, \boldsymbol{p}_1\rangle}\boldsymbol{p}_1$
  $= x^2 - \frac{1}{2}\int_{-1}^{1} t^2 \mathrm{d}t - \frac{x}{2/3}\int_{-1}^{1} t^3 \mathrm{d}t = x^2 - \frac{1}{3}$.

  Hence, standardizing, $\boldsymbol{p}_2(x) = \frac{1}{2}(3x^2 - 1)$.

  Comment. The previous problem told us that x^2 is orthogonal to x (but not to 1).

- Continuing, we find $\boldsymbol{p}_3(x) = \frac{1}{2}(5x^3 - 3x)$ and $\boldsymbol{p}_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3)$.

Comment. These famous polynomials are known as the **Legendre polynomials**. The Legendre polynomial $\boldsymbol{p}_n$ is an even function if n is even, and an odd function if n is odd (can you explain why?!).

An explicit formula is $\boldsymbol{p}_n(x) = 2^{-n}\sum_{k=0}^{n}\binom{n}{k}^2 (x+1)^k (x-1)^{n-k}$.

For instance, $\boldsymbol{p}_2(x) = \frac{1}{4}((x-1)^2 + 2^2(x-1)(x+1) + (x+1)^2) = \frac{1}{2}(3x^2 - 1)$.

https://en.wikipedia.org/wiki/Legendre_polynomials

Comment. Legendre polynomials are an example of **orthogonal polynomials**. Each choice of dot product gives rise to a family of such orthogonal polynomials.

https://en.wikipedia.org/wiki/Orthogonal_polynomials

Comment. It is also particularly natural to consider the dot product (1), where the integral is from 0 to 1. In that case, we obtain what's known as the shifted Legendre polynomials $\tilde{\boldsymbol{p}}_n(x) = \boldsymbol{p}_n(2x-1)$.

Comment on other norms. Our choice of inner product

$$\langle f, g\rangle = \int_a^b f(t)g(t)\mathrm{d}t$$

for (square-integrable) functions on $[a, b]$ gives rise to the norm $\|f\| = \left(\int_a^b f(t)^2 \mathrm{d}t\right)^{1/2}$. This is known as the L^2-norm (and often written as $\|f\|_2$).

It is the continuous analog of the usual Euclidean norm $\|\boldsymbol{v}\| = (v_1^2 + v_2^2 + \ldots)^{1/2}$ (known as ℓ^2-norm).

There do exist other norms to measure the magnitude of vectors, such as the ℓ_1-norm $\|\boldsymbol{v}\|_1 = |v_1| + |v_2| + \ldots$ or, more generally, for $p \geqslant 1$, the ℓ_p-norms $\|\boldsymbol{v}\|_p = (|v_1|^p + |v_2|^p + \ldots)^{1/p}$.

Likewise, for functions, we have the L^p-norms $\|f\|_p = \left(\int_a^b f(t)^p \mathrm{d}t\right)^{1/p}$.

Only in the case $p = 2$ do these norms come from an inner product. That's a mathematical (as opposed to geometric) reason why we especially care about that case.

Linear transformations

Throughout, V and W are vector spaces.

Just like we went from column vectors to abstract vectors (such as polynomials), the concept of a matrix leads to abstract linear transformations.

In the other direction, picking a basis, abstract vectors can be represented as column vectors (see Lecture 35). Correspondingly, linear transformations can then be represented as matrices.

Definition 179. A map $T: V \to W$ is a **linear transformation** if

$$T(c\boldsymbol{x}+d\boldsymbol{y}) = cT(\boldsymbol{x}) + dT(\boldsymbol{y}) \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \text{ in } V \text{ and all } c, d \text{ in } \mathbb{R}.$$

In other words, a linear transformation respects addition and scaling:

- $T(\boldsymbol{x}+\boldsymbol{y}) = T(\boldsymbol{x}) + T(\boldsymbol{y})$
- $T(c\boldsymbol{x}) = cT(\boldsymbol{x})$

It necessarily sends the zero vector in V to the zero vector in W:

- $T(\mathbf{0}) = \mathbf{0}$ [because $T(\mathbf{0}) = T(0 \cdot \mathbf{0}) = 0 \cdot T(\mathbf{0}) = \mathbf{0}$]

Comment. Linear transformations are special functions and, hence, can be composed. For instance, if $T: V \to W$ and $S: U \to V$ are linear transformations, then $T \circ S$ is a linear transformation $U \to W$ (sending $\boldsymbol{x}$ to $T(S(\boldsymbol{x}))$). If S, T are represented by matrices A, B, then $T \circ S$ is represented by the matrix BA. In other words, matrix multiplication arises as the composition of (linear) functions.

Example 180. The **derivative** you know from Calculus I is linear.

Indeed, the map $D: \left\{\begin{array}{c}\text{space of all}\\ \text{differentiable}\\ \text{functions}\end{array}\right\} \to \left\{\begin{array}{c}\text{space of all}\\ \text{functions}\end{array}\right\}$ defined by $f(x) \mapsto f'(x)$ is a linear transformation:

- $\underbrace{D(f(x)+g(x))}_{(f(x)+g(x))'} = \underbrace{D(f(x))}_{f'(x)} + \underbrace{D(g(x))}_{g'(x)}$
- $\underbrace{D(cf(x))}_{(cf(x))'} = \underbrace{cD(f(x))}_{cf'(x)}$

These are among the first properties you learned about the derivative.

Similarly, the **integral** you love from Calculus II is linear:

$$\int_a^b (f(x)+g(x))\mathrm{d}x = \int_a^b f(x)\mathrm{d}x + \int_a^b g(x)\mathrm{d}x, \quad \int_a^b cf(x)\mathrm{d}x = c\int_a^b f(x)\mathrm{d}x$$

In this form, we are looking at a map $T: \left\{\begin{array}{c}\text{space of all}\\ \text{continuous}\\ \text{functions}\end{array}\right\} \to \mathbb{R}$ defined by $T(f(x)) = \displaystyle\int_a^b f(x)\mathrm{d}x$.

Example 181. Consider the space V of all polynomials $p(x)$ of degree 3 or less. The map $D: V \to V$ given by $p(x) \mapsto p'(x)$ is a linear. Write down the matrix M for this linear map with respect to the basis $1, x, x^2, x^3$.

Armin Straub
straub@southalabama.edu

Solution. $M = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

For instance, the 3rd column says that x^2 (the 3rd basis element) gets sent to $0 \cdot 1 + 2 \cdot x + 0 \cdot x^2 + 0 \cdot x^3 = 2x$.

Example 182. Consider the map

$$D: \begin{Bmatrix} \text{space of poly's} \\ \text{of degree } \leqslant 3 \end{Bmatrix} \to \begin{Bmatrix} \text{space of poly's} \\ \text{of degree } \leqslant 2 \end{Bmatrix}, \quad p(x) \mapsto p'(x).$$

Write down the matrix M for this linear map with respect to the bases $1, x, x^2, x^3$ and $1, x, x^2$.

Solution. $M = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$

For instance, the 3rd column says that x^2 (the 3rd basis element) gets sent to $0 \cdot 1 + 2 \cdot x + 0 \cdot x^2 = 2x$.

Example 183. What is the pseudo-inverse of the matrix M from the previous example. Interpret your finding.

Solution. (final answer only) The pseudo-inverse of $\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$ is $\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/3 \end{bmatrix}$.

The corresponding linear map sends 1 to x, x to $\frac{1}{2}x^2$ and x^2 to $\frac{1}{3}x^3$. That is, the pseudo-inverse computes the antiderivative of each monomial.

Comment. This is not surprising, since we are familiar from Calculus with the concepts of derivatives and antiderivatives (or integrals), and that these are "pseudo" inverse to each other.

Comment. Similarly, the pseudo-inverse of $\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ is $\begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/3 & 0 \end{bmatrix}$.

Now, the corresponding linear map sends 1 to x, x to $\frac{1}{2}x^2$, x^2 to $\frac{1}{3}x^3$, and x^3 to 0. That is, the pseudo-inverse computes the antiderivative of each monomial, with the exception of x^3 which gets send to 0 (its antiderivative does not live in the space of polynomials of degree 3).

Example 184. (The April Fools' Day "proof" that $\pi = 4$, cont'd)

In that "proof", we are constructing curves c_n with the property that $c_n \to c$ where c is the circle. This convergence can be understood, for instance, in the same sense $\|c_n - c\| \to 0$ with the norm introduced as we did for functions.

Since $c_n \to c$ we then wanted to conclude that $\text{perimeter}(c_n) \to \text{perimeter}(c)$, leading to $4 \to \pi$.

However, in order to conclude from $x_n \to x$ that $f(x_n) \to f(x)$ we need that f is continuous (at x)!!

The "function" perimeter, however, is not continuous. In words, this means that (as we see in this example) curves can be arbitrarily close, yet have very different arc length.

We can dig a little deeper: as you learned in Calculus II, the arc length of a function $y = f(x)$ for $x \in [a, b]$ is

$$\int_a^b \sqrt{(\mathrm{d}x)^2 + (\mathrm{d}y)^2} = \int_a^b \sqrt{1 + f'(x)^2}\mathrm{d}x.$$

Observe that this involves f'. Try to see why the operator D that sends f to f' is not continuous with respect to the distance induced by the norm

$$\|f\| = \left(\int_a^b f(x)^2 \mathrm{d}x \right)^{1/2}.$$

In words, two functions f and g can be arbitrarily close, yet have very different derivatives f' and g'.

That's a huge issue in **functional analysis**, which is the generalization of linear algebra to infinite dimensional spaces (like the space of all differentiable functions). The linear operators ("matrices") on these spaces frequently fail to be continuous.

Fourier series

A **Fourier series** for a function $f(x)$ is a series of the form

$$f(x) = a_0 + a_1\cos(x) + b_1\sin(x) + a_2\cos(2x) + b_2\sin(2x) + \cdots$$

You may have seen Fourier series in other classes before. Our goal here is to tie them in with what we have learned about orthogonality.

In these other classes, you would have seen formulas for the coefficients a_k and b_k. We will see where those come from.

Observe that the right-hand side combination of cosines and sines is 2π-periodic.

Let us consider (nice) functions on $[0, 2\pi]$.

Or, equivalently, functions that are 2π-periodic.

We know that a natural inner product for that space of functions is

$$\langle f, g\rangle = \int_0^{2\pi} f(t)g(t)\mathrm{d}t.$$

Example 185. Show that $\cos(x)$ and $\sin(x)$ are orthogonal (in that sense).

Solution. $\langle\cos(x), \sin(x)\rangle = \int_0^{2\pi} \cos(t)\sin(t)\mathrm{d}t = \left[\frac{1}{2}(\sin(t))^2\right]_0^{2\pi} = 0$

In fact:

All the functions $1, \cos(x), \sin(x), \cos(2x), \sin(2x), \ldots$ are orthogonal to each other!

Moreover, they form a basis in the sense that every other (nice) function can be written as a (infinite) linear combination of these basis functions.

Example 186. What is the norm of $\cos(x)$?

Solution. $\langle\cos(x), \cos(x)\rangle = \int_0^{2\pi} \cos(t)\cos(t)\mathrm{d}t = \pi$

Why? There's many ways to evaluate this integral. For instance:

- integration by parts
- using a trig identity
- here's a simple way:
  - $\int_0^{2\pi} \cos^2(t)\mathrm{d}t = \int_0^{2\pi} \sin^2(t)\mathrm{d}t$ ($\cos$ and $\sin$ are just a shift apart)
  - $\cos^2(t) + \sin^2(t) = 1$
  - So: $\int_0^{2\pi} \cos^2(t)\mathrm{d}t = \frac{1}{2}\int_0^{2\pi} 1\,\mathrm{d}x = \pi$

Hence, $\cos(x)$ is not normalized. It has norm $\|\cos(x)\| = \sqrt{\pi}$.

Similarly. The same calculation shows that $\cos(kx)$ and $\sin(kx)$ have norm $\sqrt{\pi}$ as well.

Example 187. How do we find, say, b_2?

Solution. Since the functions 1, $\cos(x)$, $\sin(x)$, $\cos(2x)$, $\sin(2x)$, ..., the term $b_2\sin(2x)$ is the orthogonal projection of $f(x)$ onto $\sin(2x)$.

In particular, $b_2 = \dfrac{\langle f(x), \sin(2x)\rangle}{\langle \sin(2x), \sin(2x)\rangle} = \dfrac{1}{\pi}\displaystyle\int_0^{2\pi} f(t)\sin(2t)\mathrm{d}x.$

In conclusion:

A (nice) $f(x)$ on $[0, 2\pi]$ has the Fourier series

$$f(x) = a_0 + a_1\cos(x) + b_1\sin(x) + a_2\cos(2x) + b_2\sin(2x) + \cdots$$

where

$$\begin{aligned} a_k = \frac{\langle f(x), \cos(kx)\rangle}{\langle \cos(kx), \cos(kx)\rangle} &= \frac{1}{\pi}\int_0^{2\pi} f(t)\cos(kt)\mathrm{d}t, \\ b_k = \frac{\langle f(x), \sin(kx)\rangle}{\langle \sin(kx), \sin(kx)\rangle} &= \frac{1}{\pi}\int_0^{2\pi} f(t)\sin(kt)\mathrm{d}t, \\ a_0 = \frac{\langle f(x), 1\rangle}{\langle 1, 1\rangle} &= \frac{1}{2\pi}\int_0^{2\pi} f(t)\mathrm{d}t. \end{aligned}$$

How little we actually know!

Q: *How fast can we solve N linear equations in N unknowns?*

Estimated cost of Gaussian elimination:

$$\begin{bmatrix} \blacksquare & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & & & \vdots \\ 0 & * & * & \cdots & * \end{bmatrix}$$

- to create the zeros below the first pivot:
  $\Longrightarrow$ on the order of N^2 operations
- if there are N pivots total:
  $\Longrightarrow$ on the order of $N \cdot N^2 = N^3$ operations

- A more careful count places the cost at $\sim\frac{1}{3}N^3$ operations.
- For large N, it is only the N^3 that matters.

  It says that if $N \to 10N$ then we have to work 1000 times as hard.

That's not optimal! We can do better than Gaussian elimination:

- Strassen algorithm (1969): $N^{\log_2 7} = N^{2.807}$
- Coppersmith–Winograd algorithm (1990): $N^{2.375}$
- ... Stothers–Williams–Le Gall (2014): $N^{2.373}$ (If $N \to 10N$ then we have to work 229 times as hard.)

Is $N^{2+(\text{a tiny bit})}$ possible? **We don't know!** (People increasingly suspect so.) (Better than N^2 is impossible; why?)

Comment. The above algorithms actually are for computing matrix products. It can be shown that, if $M(N)$ is the cost for multiplying two $N \times N$ matrices, then $N \times N$ systems can also be solved for cost on the order of $M(N)$. In other words, we don't even know how costly it is to multiply two matrices.

Good news for applications:

- Matrices typically have lots of structure and zeros

  which makes solving so much faster.

Armin Straub
straub@southalabama.edu

Just for fun and curiousity!

Recall that we introduced the **dimension** of a vector space as the number of vectors in a/any basis. In Calculus, on the other hand, you learn about curves (1-dimensional), surfaces (2-dimensional) and solids (3-dimensional).

The reason that Linear Algebra is relevant for curved objects like surfaces is that locally these (typically) do look flat (like a plane), so that our tools apply at least locally.

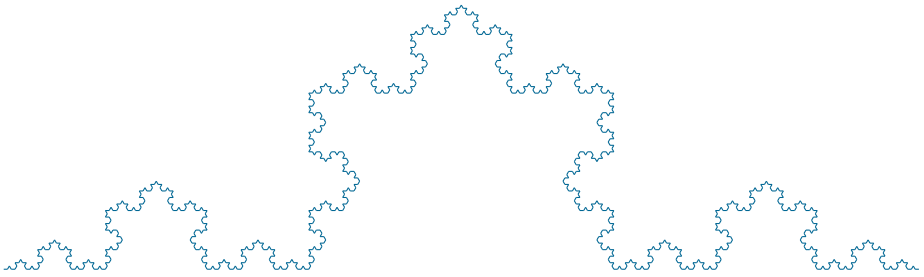
What should a 1.5 dimensional thing look like?

Something between a curve and a surface...

(Note that our linear algebra approach to dimension is not helpful.)

Here is a candidate.

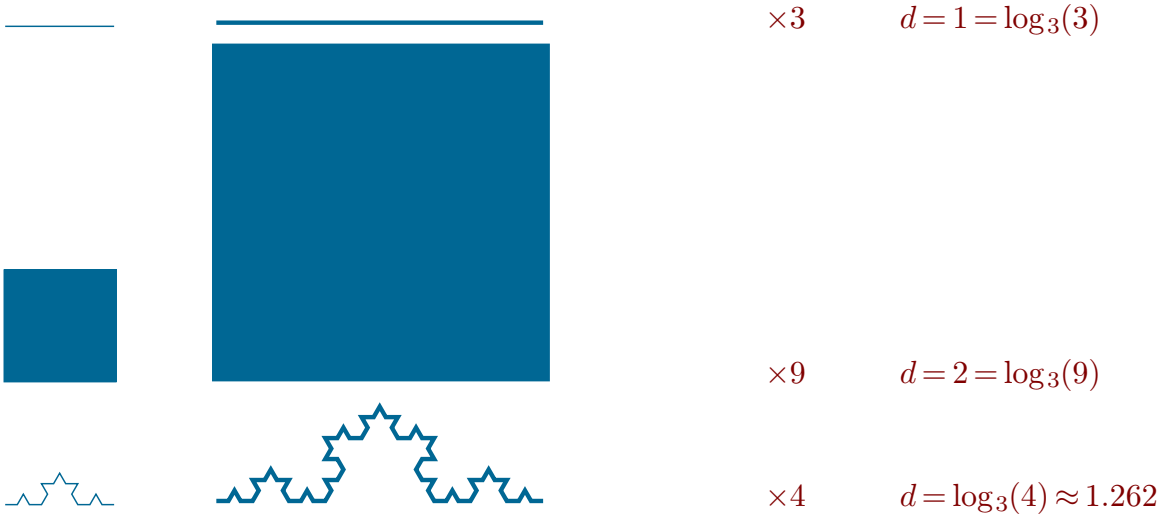
Continuing this process, results in the **Koch snowflake**, a **fractal**:



- Its perimeter is infinite!

  Why? At each iteration, the perimeter gets multiplied by $4/3$.

- The table below indicates that its boundary has dimension $\log_3(4) \approx 1.262$!!



Does this have any practical relevance? Surprisingly, yes!

Have you ever wondered why perimeters of countries are missing from wikipedia? Or, why the coastline of the UK is listed as 11,000 miles by the UK mapping authority but 7,700 miles by the CIA Factbook?

Some of the fun can be found at: https://en.wikipedia.org/wiki/Coastline_paradox