COMP 3206: Machine Learning
Linear Algebra Background

Srinandan Dasmahapatra

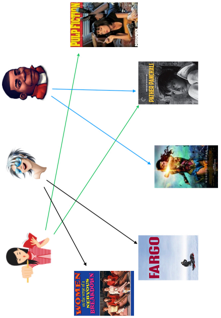
2017/8

From sets to vectors

- A vector space is a set with additional structure
- The structure allows you to
  - multiply elements by scalars and
  - add elements together

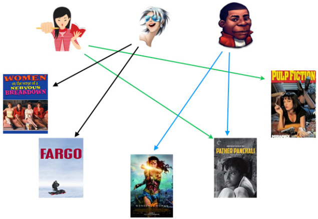
  to get other elements of the set
- Impose transformation rules on all elements – linear transformations
- Discover hidden parts/patterns in collections

Matrix representation of associations: storage and access



- Information storage matrix is **A** may be viewed as map from space of **users** to space of **movies**
- $$\begin{aligned} \mathbf{A} : \mathcal{U} &\rightarrow \mathcal{V} \\ A_{uv} &= \begin{cases} 1 & \text{if } (u, v) \text{ connected} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$
- Retrieval of information is by matrix-vector operations

Matrix representation of associations: storage and access



	x			x
		x		x
x			x	
?	?	?	?	?

- Data as matrix

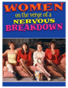
$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

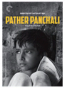
- Information retrieval by matrix-vector ops
- Given: preferences of 3 subscribers to Netflix, predict: what movies would this new user rent?

Represent elements of sets as vectors

 $\hat{\mathbf{u}}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$  $\hat{\mathbf{u}}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$  $\hat{\mathbf{u}}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$

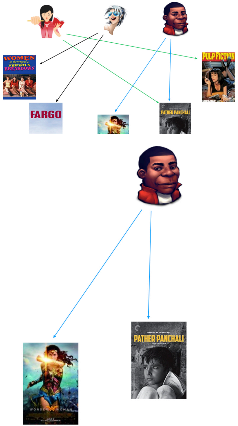
$$\hat{\mathbf{v}}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \hat{\mathbf{v}}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \hat{\mathbf{v}}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \quad \hat{\mathbf{v}}_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \quad \hat{\mathbf{v}}_5 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$



- Retrieval: movie preferences of k-th user $\hat{\mathbf{u}}_k$ obtained by performing $\mathbf{A}^T\hat{\mathbf{e}}_k$

Retrieval of information by matrix-vector multiplication



- For user 2,

$$\mathbf{A}^T \hat{\mathbf{u}}_2 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

- $\mathbf{A}^T \hat{\mathbf{u}}_2 = \hat{\mathbf{v}}_3 + \hat{\mathbf{v}}_5$

Information Retrieval, Lexical Semantics: meaning of word determined by company it keeps

- $\mathcal{D}$, a collection of documents $d_j \in \mathcal{D}$, $j = 1, \ldots, n = |\mathcal{D}|$.
- $\mathcal{W}$, the vocabulary, *i.e.* all the words $w_i \in \mathcal{W}$ contained in $\mathcal{D}$, $i = 1, \ldots, m = |\mathcal{W}|$
- Construct $(m \times n)$ matrix $\mathcal{T}$ with entries $(\mathcal{T})_{ij} = t_{ij}$, where t_{ij} counts the number of times word w_i appears in document d_j:

$$\mathcal{T} = \begin{pmatrix} | & & | \\ \mathbf{d}_1 & \cdots & \mathbf{d}_n \\ | & & | \end{pmatrix} = \begin{pmatrix} — & \mathbf{w}_1 & — \\ & \vdots & \\ — & \mathbf{w}_m & — \end{pmatrix}$$

- Column view – document retrieval (library catalogue)
- Row view – lexical semantics (*distributional similarity*): if 2 words a, b appear in same documents, they are similar: $\mathbf{w}_a$ close to $\mathbf{w}_b$.

Representing images as vectors

$$\mapsto \begin{array}{|c|c|c|c|} \hline x_{11} & x_{12} & \cdots & x_{1L} \\ \hline x_{21} & x_{22} & \cdots & x_{2L} \\ \hline x_{31} & x_{32} & \cdots & x_{3L} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline x_{L1} & x_{L2} & \cdots & x_{LL} \\ \hline \end{array} \mapsto \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1L} \\ x_{21} \\ \vdots \\ x_{LL} \end{pmatrix}$$

- $\mathbf{x} = (x_{11}, x_{12}, \ldots, x_{LL})^T = (x(1), x(2), \ldots, x(D))^T$.

Matrices

You should all know the following

- Matrix notation
- Matrix transpose
- Scalar multiplication
- Matrix addition & multiplication
- Matrix inverse
- System of linear equations in matrix form
- Matrix determinant

Reminder: Solving Linear Equations – Geometrical Picture

- Solve set of equations:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r \\ s \end{pmatrix}$$

- Geometrically viewed as intersection of linear linear combination of vectors:

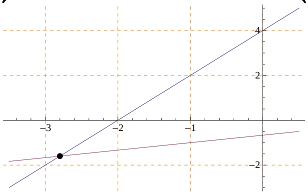
$$\overbrace{\begin{matrix} ax + by & = & r \\ cx + dy & = & s \end{matrix}}^{\text{rows of matrix}} \leftrightarrow \overbrace{x \begin{pmatrix} a \\ c \end{pmatrix} + y \begin{pmatrix} b \\ d \end{pmatrix} = \begin{pmatrix} r \\ s \end{pmatrix}}^{\text{columns of matrix}}$$

The Geometrical Picture: An example

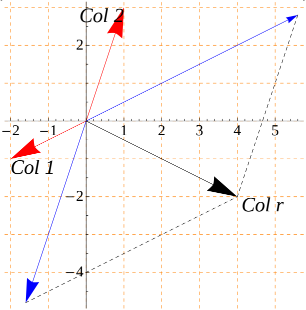
- Solve set of equations:

$$\begin{pmatrix} -2 & 1 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 4 \\ -2 \end{pmatrix}$$

rows of matrix



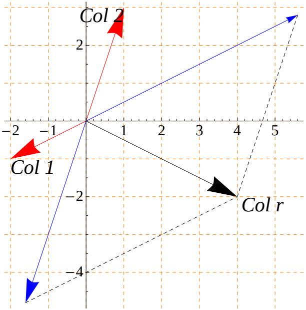
red vectors are columns of matrix



- Solution of $y - 2x = 4, 3y - x = -2$, is $(x, y) = (-2.8, -1.6)$.

Fundamental operations on vectors – multiply by scalars and perform addition

Multiply red vectors by numbers (elements of a field) and add vectors together



- For **linear regression**: find linear combination of columns of design matrix to get vector closest to output

Examples of vector spaces

- Add vectors $(1.0, -2.0) + (3.0, 4.0) = (4.0, 2.0)$, where the entries are in this case real.
- Multiply vectors by numbers (scalars) $3.2(1.0, -3.0) = (3.2, -9.6)$
- For any field $\mathbb{F}$ (such as the reals $\mathbb{R}$ or complex numbers $\mathbb{C}$, $\mathbb{F}$-valued n-tuples

$$\mathbb{F}^n = \{(a_1, \ldots, a_n) | a_i \in \mathbb{F}, i = 1, \ldots, n\}$$

form a vector space; $\mathbb{R}$-valued 3-tuples such as $\mathbf{v}_1 = (-1.2,\ 2.0,\ 5.5)$ locate points in 3D. Written as rows or columns $\mathbf{v}_1^T = \begin{pmatrix} -1.2 \\ 2.0 \\ 5.5 \end{pmatrix}$.

Even matrices form a vector space

- Matrices form a vector space: you can multiply $n \times m$ matrices A over $\mathbb{F}$ with entries $a_{ij} \in \mathbb{F}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$ by scalars and add any two such matrices together:

$$3 \begin{pmatrix} -2 & 1 \\ -1 & 4 \end{pmatrix} - 2 \begin{pmatrix} 2 & 2 \\ -1 & 6 \end{pmatrix} = \begin{pmatrix} -10 & -1 \\ -1 & 0 \end{pmatrix}.$$

Functions constitute vector spaces

- $\mathbb{F}[x]$, the space of polynomials $\sum_m a_m x^m$, where $a_m \in \mathbb{F}$ forms a vector space:

$$\begin{aligned}(a_0 + a_1 x + a_2 x^2) + (b_0 + b_1 x) &= (a_0 + b_0) + (a_1 + b_1)x + a_2 x \\ &=: (c_0 + c_1 x + c_2 x^2).\end{aligned}$$

  Monomials as basis elements:

$$(a_0, a_1, a_2) + (b_0, b_1, 0) = (a_0 + b_0, a_1 + b_1, a_2)$$

- Similarly, the set $\mathbb{F}[x_1, \ldots, x_k]$ of polynomials in k variables forms a vector space.
- Set of functions of the form $\sum_{|n|<N} a_n e^{in\theta}$ (Fourier series).
- Extension – replace sums (where the summation index is from a discrete set) by integrals (where the index being summed over is now continuous)

Formal definition of vector space

- A vector space V over a field $\mathbb{F}$ is a collection of objects (vectors) upon which two operations can be performed – addition amongst the vectors, and multiplication by elements of the field $\mathbb{F}$ (scalars). Upon addition of any two vectors $\mathbf{v}_1 \in V, \mathbf{v}_2 \in V$, the resulting vector $\mathbf{v}_1 + \mathbf{v}_2$ must also belong to V (closure). The binary product for $a \in \mathbb{F}$ (scalar) and $\mathbf{v} \in V$ (vector)

$$\begin{aligned} m : \mathbb{F} \times V &\rightarrow V \\ m(a, \mathbf{v}) &\mapsto a\mathbf{v} \end{aligned}$$

  is defined, and satisfies
  - $1\mathbf{v} = \mathbf{v}$, $1 \in \mathbb{F}$, $\mathbf{v} \in V$
  - $(ab)\mathbf{v} = a(b\mathbf{v})$, $a, b \in \mathbb{F}$, $\mathbf{v} \in V$
  - $(a+b)\mathbf{v} = a\mathbf{v} + b\mathbf{v}$ and $a(\mathbf{v}+\mathbf{w}) = a\mathbf{v} + a\mathbf{w}$, $a, b \in \mathbb{F}$, $\mathbf{v}, \mathbf{w} \in V$

On notation:

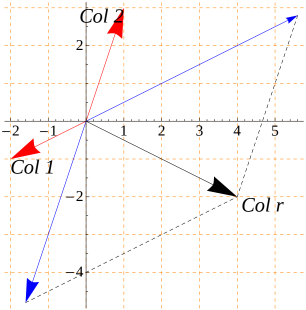
$A \times B$ is the Cartesian product of the sets A and B. This means that, for example, if $A = \{a_1, a_2, a_3\}$ and $B = \{b_1, b_2\}$, then

$$A \times B = \{(a_1, b_1), (a_1, b_2), (a_2, b_1), (a_2, b_2), (a_3, b_1), (a_3, b_2).\}$$

In the definition of m, the symbol $\rightarrow$ refers to the map between the sets, while $\mapsto$ takes a particular pair from $\mathbb{F} \times V$ to produce an output vector from V.

Reminder: Linear combination and dependence

Linear combination of vectors: $\mathbf{v} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i,\ \alpha_i \in \mathbb{F}, \mathbf{v}_i \in V$.



- The vectors in the figure are linear combinations of $\mathbf{e}_1 = \binom{1}{0}$ and $\mathbf{e}_2 = \binom{0}{1}$. They are in the **span** of $\{\mathbf{e}_1, \mathbf{e}_2\}$.
- $\mathbf{v} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 = \binom{a_1}{a_2}$ can be zero iff $a_1 = 0 = a_2$.

Reminder: Linear independence & Basis

- A set of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are called **linearly independent** if none of them can be represented as a linear combination of the others:
$$\mathbf{v}_k \neq \sum_{i \neq k} c_i \mathbf{v}_i, \quad \text{for any } c_i \in \mathbb{F}$$
- Equivalently, condition for a set of vectors $\{\mathbf{v}_i\}_i$ to be linearly independent:
$$\text{If } \sum_{i=1}^{n} \alpha_i \mathbf{v}_i = 0, \text{ then } \alpha_i = 0 \text{ for all } i$$
.
- A **basis** for V is a set $B \subset V$ which is both spanning and independent. A finite dimensional vector space has a finite basis, and its dimension $\dim V$ is the number of elements in B.

Dot Products, Orthogonality and Norms

- We can associate, with two vectors $\mathbf{v}$ and $\mathbf{w}$ an element of $\mathbb{F}$ called their **scalar** (or **dot**) **product**:

$$\begin{aligned} \text{dot} : V \times V &\to \mathbb{F} \\ \text{dot}(\mathbf{v}, \mathbf{w}) = \mathbf{v} \cdot \mathbf{w} = \mathbf{v}^T \mathbf{w} &\mapsto a \end{aligned}$$

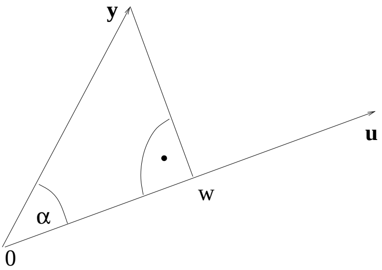
- Two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are called *orthogonal* if their dot product is zero, i.e. $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$. If k vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ are mutually orthogonal, ie. $\mathbf{v}_i \cdot \mathbf{v}_j = 0$ for $i \neq j$, they are called an **orthogonal set**.
- Euclidean norm: for $\mathbf{v} \in V$, $\dim V = N$,

$$||\mathbf{v}||_2 := \sqrt{\mathbf{v}^T \mathbf{v}} = \sqrt{v_1^2 + v_2^2 + \cdots + v_N^2} = |\mathbf{v}|$$

- If all vectors are of unit length $|\mathbf{v}_i| = 1$, the set is called **orthonormal**.

Using dot products to introduce projections

- **Project** a vector **y** on a direction given by a vector **u**



- The **projection** is given by the value w (length $\overline{0w}$). Note, w could be negative if α is bigger than 90^0. From the figure, we see that

$$w = |\mathbf{y}| \cos \alpha = \frac{\mathbf{y} \cdot \mathbf{u}}{|\mathbf{u}|} \ ,$$

because the **dot** product is $\mathbf{y} \cdot \mathbf{u} = |\mathbf{u}||\mathbf{y}| \cos \alpha$.

Expanding a vector in a set of orthogonal vectors – an example

Let $\mathbf{e}_1 = \binom{1}{0}$,$\mathbf{e}_2 = \binom{0}{1}$. We would like to expand $\mathbf{v} = \binom{-5}{3}$ as a *linear combination* of the set $\{\mathbf{e}_i\}$, ie. we would like to find numbers α_1, α_2 such that

$$\mathbf{v} = \sum_{i=1}^{n} \alpha_i \mathbf{e}_i \tag{1}$$

Solution: Multiply $\mathbf{v}$ by $\mathbf{e}_j$ and use the orthogonality ($\mathbf{e}_1 \cdot \mathbf{e}_2 = 0$):
$\mathbf{e}_1 \cdot \binom{-5}{3} = -5$, $\mathbf{e}_2 \cdot \binom{-5}{3} = 3$.

$$\begin{pmatrix} -5 \\ 3 \end{pmatrix} = (-5)\begin{pmatrix} 1 \\ 0 \end{pmatrix} + (3)\begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Expanding a vector in a set of orthogonal vectors

Suppose, $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are an orthogonal set of $n \times 1$ column vectors and $\mathbf{v}$ an arbitrary $n \times 1$ column vector. We would like to expand $\mathbf{v}$ as a *linear combination* of the set $\{\mathbf{v}_i\}$, ie. we would like to find numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that

$$\mathbf{v} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i. \tag{2}$$

Solution: Multiply by $\mathbf{v}_j$ and use the orthogonality to get

$$\begin{aligned} \mathbf{v}_j \cdot \mathbf{v} &= \alpha_1 \mathbf{v}_j \cdot \mathbf{v}_1 + \cdots + \alpha_j \mathbf{v}_j \cdot \mathbf{v}_j + \cdots \alpha_n \mathbf{v}_j \cdot \mathbf{v}_n \\ &= \alpha_j \mathbf{v}_j \cdot \mathbf{v}_j \end{aligned}$$

Hence

$$\alpha_j = \frac{\mathbf{v}_j \cdot \mathbf{v}}{\mathbf{v}_j \cdot \mathbf{v}_j} = \frac{\mathbf{v}_j \cdot \mathbf{v}}{|\mathbf{v}_j|^2} \tag{3}$$

Orthonormal bases (where all basis vectors have length 1) are very useful.

Linear mappings between vector spaces

- For V, W vector spaces over $\mathbb{F}$, a map $T : V \to W$ is **linear** if for all vectors $\mathbf{v}_i \in V$ and scalars $a_i \in \mathbb{F}$,

$$T(a_1\mathbf{v}_1 + a_2\mathbf{v}_2) = a_1 T\mathbf{v}_1 + a_2 T\mathbf{v}_2.$$

- Example (derivative): $T \equiv (\frac{d}{dx})$

$$(\frac{d}{dx})(af(x) + bg(x)) = a(\frac{d}{dx})f(x) + b(\frac{d}{dx})g(x)$$

- Example (verify):

$$T : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 \mapsto \begin{pmatrix} 5x_1 - x_2 \\ x_1 + x_2 \\ x_1 - x_2 \end{pmatrix} \in \mathbb{R}^3$$

- *Hint:* Let $\mathbf{v}_1 = (x_1, x_2)^T$ and $\mathbf{v}_2 = (y_1, y_2)^T$. Thus, $a_1\mathbf{v}_2 + a_2\mathbf{v}_2 = (a_1x_1 + a_2y_1, a_1x_2 + a_2y_2)^T$.

Linear mappings between vector spaces

- For V, W vector spaces over $\mathbb{F}$, a map $T : V \to W$ is **linear** if for all vectors $\mathbf{v}_i \in V$ and scalars $a_i \in \mathbb{F}$,

$$T(a_1\mathbf{v}_1 + a_2\mathbf{v}_2) = a_1 T\mathbf{v}_1 + a_2 T\mathbf{v}_2.$$

- By induction, this extends over any (finite) sum of vectors.
- Thus, a linear map from V to W is completely defined by values it assigns to basis elements of V, and these values can be arbitrary vectors in W.

$$T : x_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \cdots ?$$

Linear mappings as matrices

- For V, W vector spaces over $\mathbb{F}$ with bases $\{\mathbf{v}_1, \ldots, \mathbf{v}_q\}$ and $\{\mathbf{w}_1, \ldots, \mathbf{w}_p\}$, a map $T : V \to W$ is completely specified by scalars $a_{ij} \in \mathbb{F}$, $(i = 1, \ldots, p,\ j = 1, \ldots, q)$, such that

$$T\mathbf{v}_j = \sum_{i=1}^{p} a_{ij}\mathbf{w}_i.$$

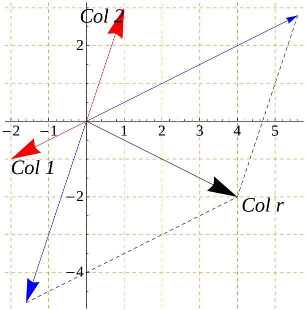
  Let $\mathbf{x} \in \mathbb{F}^q$ with components $x_1, \ldots, x_q$, $\mathbf{y} \in \mathbb{F}^p$ with components $y_1 \ldots, y_p$. Then, $T\mathbf{x}$ becomes

$$T\left(\sum_{j=1}^{q} x_j\mathbf{v}_j\right) = \sum_{i=1}^{p}\left(\sum_{j=1}^{q} x_j a_{ij}\right)\mathbf{w}_i = \sum_{i=1}^{p} y_i\mathbf{w}_i \Leftrightarrow T_A\mathbf{x} = \mathbf{y}$$

- The entries $a_{ij} \in \mathbb{F}^{p\times q}$ are determined by the action of T on the basis vectors.

Column space and Range of a matrix

Thus $A\mathbf{x} = \mathbf{y}$ can be solved if and only if $\mathbf{y}$ is a linear combination of columns of A.



- The column space of a matrix A (denoted col A) is the subspace spanned by all linear combinations of the columns of A.
- This is also the range of the linear map: range(A)=$AV = \{\mathbf{w} \in W : \mathbf{w} = A\mathbf{v} \text{ for some } \mathbf{v} \in V\}$

Examples illustrating linear dependence and nullspace

- Let $B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$. For vector $\mathbf{v}$ in direction $\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$, $B\mathbf{v} = 0.\mathbf{v}$ in *nullspace* or *kernel* of B.
- For $A = \begin{pmatrix} 2 & 1 & 3 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$, col(1) + col(2) = col(3),so

$$\begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} - 1\begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Rightarrow \ker(A) = c\begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$$

Show $\ker(A^T) = c\begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix}$.

Kernel or Null space of a matrix

- In the previous example $A = \begin{pmatrix} 2 & 1 & 3 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$, there are 3 variables $\mathbf{v}$ in $A\mathbf{v} = \mathbf{y}$ but only two independent equations.
- If $A\mathbf{v} = \mathbf{y}$ and $\mathbf{x} \in \ker(A)$ then $A(\mathbf{v} + \mathbf{x}) = \mathbf{y}$. Either there are no solutions or there are (infinitely) many solutions.
- The kernel of a map (or matrix) $\ker(A) = \text{nullspace } A = \{\mathbf{v} \in V : A\mathbf{v} = 0\}$.
- Let A be a $3 \times q$ matrix.

$$A = \begin{pmatrix} — & \mathbf{u} & — \\ — & \mathbf{v} & — \\ — & \mathbf{w} & — \end{pmatrix},$$

where $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ are q-dim row vectors. Then, $\mathbf{x} \in \ker(A) \Leftrightarrow A\mathbf{x} = 0$. This means $\mathbf{x} \perp \{\mathbf{u}, \mathbf{v}, \mathbf{w}\}$.

Rank of a matrix = number of independent equations

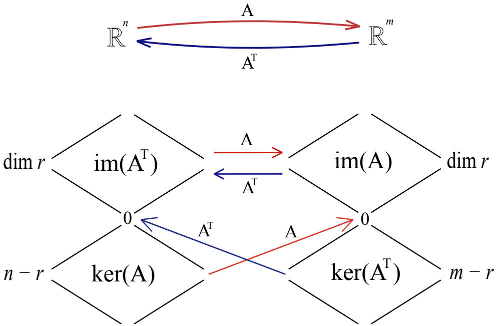
- The **rank** (column rank) of A is the dimension of the column space of A.
- A vector space is partitioned into its range and null spaces:

$$\dim V = \underbrace{\text{dim ker}(A)}_{\textbf{nullity}} + \underbrace{\text{dim range}(A)}_{\textbf{rank}}.$$

- We can do the same for the transpose: $\text{col}(A^T)$ and $\ker(A^T)$.
- 4 fundamental subspaces: $\text{col}(A)$, $\ker(A^T)$, $\text{col}(A^T)$ and $\ker(A^T)$

Four fundamental subspaces of a matrix

http://en.wikipedia.org/wiki/Fundamental_theorem_of_linear_alge



Multiplying vectors by matrices iteratively introduces linear dependence

- Let $\mathbf{A}$ be a $n \times n$ (square) matrix. The multiplication of $\mathbf{A}$ with vectors

$$\mathbf{w} = \mathbf{A}\mathbf{v}$$

  defines a mapping (or transformation) of vectors $\mathbf{v} \in V$ into vectors $\mathbf{w} \in W$. For square matrix $\mathbf{A}$, $V = W$.
- For $\mathbf{v} \neq 0$, the $(n+1)$ vectors $\mathbf{v}, \mathbf{A}\mathbf{v}, \mathbf{A}^2\mathbf{v}, \ldots \mathbf{A}^n\mathbf{v}$ cannot all be linearly independent if the rank of $\mathbf{A}$ is n.
- Therefore there must be scalars $a_0, a_1, \ldots, a_n$ such that

$$(a_0\mathbf{I} + a_1\mathbf{A}^1 + \cdots + a_n\mathbf{A}^n)\mathbf{v} = 0, \text{ for } \mathbf{v} \neq 0.$$

Linear dependence determines eigenvalues from matrix polynomials

- For a polynomial $f(x) = \sum_{i=0}^{n} a_i x^n, x \in \mathbb{F}$ define a **matrix polynomial** $f(\mathbf{A}) = \sum_{i=0}^{n} a_i \mathbf{A}^n$ where $\mathbf{A}$ is a square matrix.
- Therefore if $f(x) = (x - \lambda_1)(x - \lambda_2) \cdots (x - \lambda_n)$, we can express $f(\mathbf{A}) = (\mathbf{A} - \lambda_1 \mathbf{I})(\mathbf{A} - \lambda_2 \mathbf{I}) \cdots (\mathbf{A} - \lambda_n \mathbf{I})$ where $\mathbf{I}$ is an $n \times n$ identity matrix.
- Therefore, since

  $$0 = (a_0 \mathbf{I} + a_1 \mathbf{A}^1 + \cdots + a_n \mathbf{A}^n)\mathbf{v} = (\mathbf{A} - \lambda_1 \mathbf{I})(\mathbf{A} - \lambda_2 \mathbf{I}) \cdots (\mathbf{A} - \lambda_n \mathbf{I})\mathbf{v},$$

  at least one of $(\mathbf{A} - \lambda_k \mathbf{I})$ maps a non-zero vector in that space to 0.
- This means that $\ker(\mathbf{A} - \lambda_k \mathbf{I}) \neq \{0\}$ and $(\mathbf{A} - \lambda_k \mathbf{I})$ is not invertible.
- $\lambda_1, \ldots, \lambda_n$ are the **eigenvalues** of $\mathbf{A}$.

Matrix polynomials and eigenvalues: example

- For polynomials $f_1(x) = x^2 - 5x - 2$ and $f_2(x) = x^2 - 2x - 5$ compute $f_1(\mathbf{A})$ and $f_2(\mathbf{A})$ for $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}$:
- Check that $\mathbf{A}^2 = \begin{pmatrix} 7 & 4 \\ 6 & 7 \end{pmatrix}$
- $f_1(\mathbf{A}) = \begin{pmatrix} 0 & -6 \\ -9 & 0 \end{pmatrix}$ and $f_2(\mathbf{A}) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$
- $f_2(\mathbf{A})$ is the **characteristic polynomial** of $\mathbf{A}$. $f_2(x) = (x - \lambda_1)(x - \lambda_2)$. $\lambda_{1,2}$ are the **eigenvalues** of the matrix $\mathbf{A}$. (In this example, $\lambda_{1,2} = (1 \pm \sqrt{6})$.)
- In general, the characteristic polynomial of a matrix $\mathbf{A}$ is denoted $\chi_{\mathbf{A}}(x)$. So $\chi_{\mathbf{A}}(x) = f_2(x)$ for this example.

Determinants and characteristic polynomials of matrices

- The **determinant** of a matrix $\mathbf{T}$, denoted $\det(\mathbf{T})$ or $|\mathbf{T}|$ is the product of its eigenvalues.
- The characteristic polynomial $\chi_{\mathbf{T}}(x)$ of a matrix $\mathbf{T}$ equals $\det(x\mathbf{I} - \mathbf{T})$.
- You should all know the Laplace expansion of $\det(\mathbf{T})$.
- Check that $|x\mathbf{I} - \mathbf{A}|$ for $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}$ is indeed $f_2(\mathbf{A})$.

Eigenvectors: vectors $\mathbf{x}$ whose lengths are scaled by eigenvalue λ upon action of $\mathbf{A}$

- The eigenvalue problem $\mathbf{Ax} = \lambda\mathbf{x}$: find, for a matrix $\mathbf{A}$, the eigenvectors $\mathbf{x}$ and eigenvalues λ.
- Example: Find the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & -2 \\ -1 & 2 & 1 \\ 0 & 1 & -1 \end{pmatrix}$$

- **STEP I:** Compute the characteristic polynomial of $\mathbf{A}$ and find its roots:

$$-\chi_{\mathbf{A}}(\lambda) = \begin{vmatrix} 1-\lambda & 1 & -2 \\ -1 & 2-\lambda & 1 \\ 0 & 1 & -1-\lambda \end{vmatrix} = (1-\lambda)(2-\lambda)(-1-\lambda)$$

$\chi_{\mathbf{A}}(\lambda) = 0 \Rightarrow \lambda_1 = 2$, $\lambda_2 = 1$ and $\lambda_3 = -1$.

...continuing...the calculation of the eigensystem

STEP II:

For each eigenvalue, we need to compute the corresponding eigenvectors. We demonstrate this only for $\lambda_1 = 2$. Setting $\lambda = 2$, denoting the corresponding eigenvector by $\mathbf{v}_1 = (x\ y\ z)^T \in \ker(\mathbf{A} - \lambda_1 \mathbf{I})$, compute

$$(\mathbf{A} - 2I)\,\mathbf{v}_1 = \begin{pmatrix} -1 & 1 & -2 \\ -1 & 0 & 1 \\ 0 & 1 & -3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = 0$$

This is equivalent to the set of equations

$$\begin{array}{rrrl} -x & +y & -2z & = 0 \\ -x & & +z & = 0 \\ & y & -3z & = 0 \end{array}$$

Solution $\mathbf{v}_1 = c(1\ 3\ 1)^T$

...continuing...(and how to find the result in python)

Often, one computes the *normalized* eigenvectors $\hat{\mathbf{v}}$, which have unit length $|\hat{\mathbf{v}}| = 1$. In our case,

$$|\hat{\mathbf{v}}_1| = \frac{\mathbf{v}_1}{|\mathbf{v}_1|} = \frac{\mathbf{v}_1}{\sqrt{11}} = \frac{1}{\sqrt{11}}(1\ 3\ 1)^T$$

Proceeding in a similar way with the two other eigenvalues, we get the set of eigenvectors $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}$ $\mathbf{v}_2 = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$ and $\mathbf{v}_3 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$.

In python (seek the online Help):

```
np.linalg.eig?
```

Example: eigenvectors of repeated eigenvalues

You may not always get distinct eigenvalues (like we did in the previous case). Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 2 \\ 0 & 2 & 1 \\ -1 & 2 & 2 \end{pmatrix} \tag{4}$$

Solving $\chi_{\mathbf{A}}(\lambda) = 0$ gives the eigenvalues $\lambda_1 = \lambda_2 = 2$ and $\lambda_3 = 1$. For the eigenvalue λ_3 we find $\mathbf{v}_3 = (1\;1\;-1)^T$ as an eigenvector. If we set $\lambda_{1,2} = 2$ and $\mathbf{v}_{1,2} = (x\;y\;z)^T$ the equation $(\mathbf{A} - 2\mathbf{I})\mathbf{v}_{1,2} = 0$ gives

$$\begin{array}{rrrl} -x & +2y & +2z & = 0 \\ & & z & = 0 \\ -x & +2y & & = 0 \end{array}$$

yielding $z = 0$ and $x = 2y$ and $\mathbf{v}_{1,2} = (2\;1\;0)^T$. Hence, we get only two linear independent eigenvectors.

Linear independence of eigenvectors*

- If matrix $\mathbf{A}$ has m **distinct eigenvalues** $\lambda_1, \ldots, \lambda_m$, then the corresponding eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$ are linearly independent.
- **Proof** (*by contradiction*): *Assume* linear dependence: $\exists a_1, \ldots, a_m$, all $a_i \neq 0$ such that $\mathbf{w} := a_1\mathbf{v}_1 + \cdots a_m\mathbf{v}_m = 0$. For $k = 1, \ldots, m$ we apply to the zero vector $\mathbf{w}$ the operators $\check{\mathbf{A}}_k$ defined thus (missing k-th factor):

$$\check{\mathbf{A}}_k = (\mathbf{A} - \lambda_1\mathbf{I})(\mathbf{A} - \lambda_2\mathbf{I}) \cdots (\mathbf{A} - \lambda_{k-1}\mathbf{I})(\mathbf{A} - \lambda_{k+1}\mathbf{I}) \cdots (\mathbf{A} - \lambda_m).$$

For example, for $k = 1$ we have

$$0 = \check{\mathbf{A}}_1\mathbf{w} = a_1(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3) \cdots (\lambda_1 - \lambda_m)\mathbf{v}_1 \Rightarrow a_1 = 0.$$

Similarly, $0 = \check{\mathbf{A}}_k\mathbf{w} \Rightarrow a_k = 0$. For $\mathbf{w} = 0$ all $a_k = 0$ necessarily. *Contradicts assumption.* $\square$

Complex eigenvalues*

Eigenvalues may not be real numbers, even if the matrix elements are real:

$$\mathbf{A} = \begin{pmatrix} 1 & -2 \\ 2 & 0 \end{pmatrix}$$

Setting $|\mathbf{A} - \lambda I| = \lambda^2 - \lambda + 4 = 0$, we obtain $\lambda_{1,2} = (1 \pm \sqrt{-15})/2 = (1 \pm i\sqrt{15})/2$ with $i = \sqrt{-1}$.

Real symmetric and hermitian matrices

- A matrix $\mathbf{A}$ is called *real symmetric* if all matrix elements are real numbers and $\mathbf{A}^T = \mathbf{A}$, where $\mathbf{A}^T$ is the transpose of $\mathbf{A}$.
- For a matrix $\mathbf{A}$ with elements $(\mathbf{A})_{ij} = a_{ij}$, $\cdot^T$ is defined as $(\mathbf{A}^T)_{ij} = a_{ji}$.
- $\cdot^\dagger$ is an operation called hermitian conjugation, defined as $(\mathbf{A}^\dagger)_{ij} = a^*_{ji}$, and $\mathbf{A}^\dagger$ is referred to as "A-dagger."
- **NB:** For real matrices, $\mathbf{A}^\dagger = \mathbf{A}^T$.
- A matrix $\mathbf{A}$ is called *hermitian* if all matrix elements are complex numbers and $\mathbf{A}^\dagger := (\mathbf{A}^*)^T = \mathbf{A}$, where $\mathbf{A}^*$ is the matrix whose elements are complex conjugates of those in $\mathbf{A}$.
- *The eigenvalues of a hermitian matrix are real.*
- Certain types of symmetric matrices (covariance matrices and kernel/gram matrices) are often generated from data sets, and are used extensively in ML algorithms. You'll see a few examples as we go through the course.

Eigenvectors of real symmetric matrices

We show the following:

- *Let* $\mathbf{A}$ *be a real symmetric matrix. Then eigenvectors associated with distinct eigenvalues are orthogonal.*
- **Proof:** Let $\mathbf{u}$ and $\mathbf{v}$ be two eigenvectors with distinct eigenvalues λ and μ respectively, i.e $\mathbf{Au} = \lambda\mathbf{u}$ and $\mathbf{Av} = \mu\mathbf{v}$. We shall prove $\mathbf{u}^T\mathbf{v} = 0$.
  Since $\mathbf{A}$ is symmetric, $(\mathbf{Au})^T = \mathbf{u}^T\mathbf{A}$.Therefore $\mathbf{Au} = \lambda\mathbf{u} \Leftrightarrow \mathbf{u}^T\mathbf{A}^T = \lambda\mathbf{u}^T$. Multiply on the right with $\mathbf{v}$:

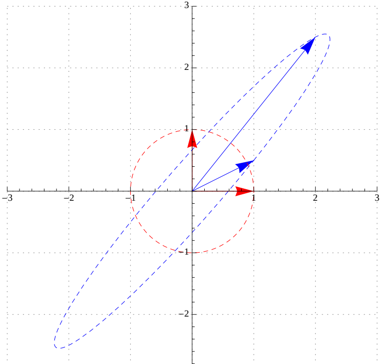
  $$\lambda\mathbf{u}^T\mathbf{v} = \mathbf{u}^T\mathbf{Av} = \mu\mathbf{u}^T\mathbf{v}.$$

  Hence $(\lambda - \mu)\mathbf{u}^T\mathbf{v} = 0$ and, since $\mu \neq \lambda$, $\mathbf{u}^T\mathbf{v} = 0$.
- With some more effort one can show (even with repeated eigenvalues!) that for any real symmetric $n \times n$ matrix we can find n orthogonal eigenvectors.

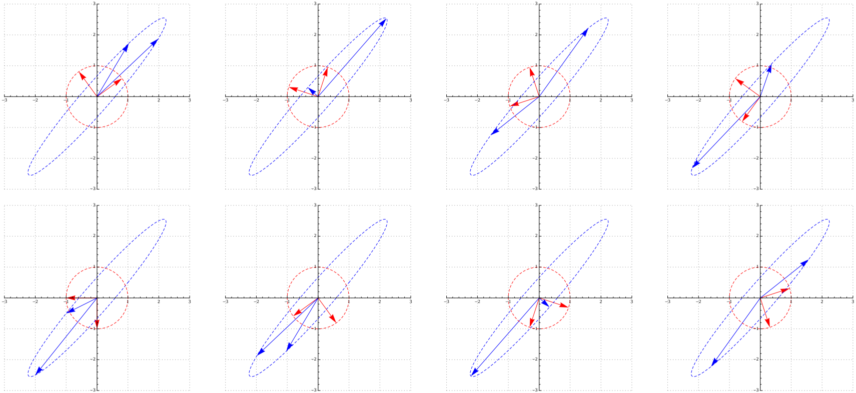
Singular Value Decomposition (SVD) of a Matrix

- The action of an arbitrary matrix on a vector space can be pieced together from its action on an orthonormal basis in that vector space.
- SVD measures how a circle is mapped into an ellipse; how an *n*-dimensional hyper-sphere is mapped into an *n*-dimensional hyper-ellipse.

Action of $\begin{pmatrix} 1.0 & 2.0 \\ 0.5 & 2.5 \end{pmatrix}$ on unit vectors $\binom{1}{0}$ and $\binom{0}{1}$. The lengths of the semi-major axes of the hyper-ellipse are properties of the map.



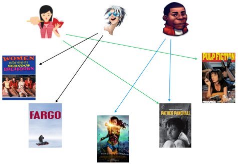
In pictures: mapping a unit circle into an ellipse



- Even when the vectors in the domain and range of the map change, their locus displays the geometrical character of the transformation enacted by the matrix.
- While the displayed pairs of vectors in the domain (red) are orthogonal by construction, the pairs they map to (blue) are

Example of SVD for recommender matrix

- $\mathbf{U} = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \end{pmatrix}$
- $\Sigma = \begin{pmatrix} \sqrt{3} & 0 & 0 & 0 & 0 \\ 0 & \sqrt{2} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$
- $\mathbf{V} =$
$$\begin{pmatrix} 0 & \frac{1}{\sqrt{2}} & 0 & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & 0 & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} & 0 \\ \frac{1}{\sqrt{6}} & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} \\ \sqrt{\frac{2}{3}} & 0 & 0 & \frac{1}{\sqrt{3}} & 0 \end{pmatrix}$$



$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

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

Example of SVD

- The action of an arbitrary matrix on a vector space can be pieced together from its action on an orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2\}$ in that vector space, here 2-dimensional. So, $\mathbf{x} = (\mathbf{v}_1^T\mathbf{x})\mathbf{v}_1 + (\mathbf{v}_2^T\mathbf{x})\mathbf{v}_2$.
- The set of vector equations $\mathbf{A}\mathbf{v}_j = \sigma_j \mathbf{u}_j$ for $j = 1, 2$ becomes:

$$\begin{aligned} \mathbf{A}\mathbf{x} &= (\mathbf{v}_1^T\mathbf{x})\mathbf{A}\mathbf{v}_1 + (\mathbf{v}_2^T\mathbf{x})\mathbf{A}\mathbf{v}_2 \\ &= (\mathbf{v}_1^T\mathbf{x})\sigma_1\mathbf{u}_1 + (\mathbf{v}_2^T\mathbf{x})\sigma_2\mathbf{u}_2 \\ \Rightarrow \mathbf{A} &= \mathbf{v}_1^T\sigma_1\mathbf{u}_1 + \mathbf{v}_2^T\sigma_2\mathbf{u}_2 \end{aligned}$$

- Express that as $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$, with $\mathbf{U}$ containing the columns of $\mathbf{u}_i$, $\mathbf{V}$ the columns of $\mathbf{v}_i$, and Σ a diagonal matrix with σ_i along the diagonal.

The reduced SVD – the range may not have a basis

- The action of an arbitrary matrix on a vector space can be pieced together from its action on an orthonormal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ in that vector space. The set of vector equations $\mathbf{A}\mathbf{v}_j = \sigma_j \mathbf{u}_j$ for $j = 1, \ldots, n$ may be expressed as a matrix equation $\mathbf{AV} = \hat{\mathbf{U}}\hat{\Sigma}$:

$$\begin{pmatrix} & & \\ & \mathbf{A} & \\ & & \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{v}_1 & \cdots & \mathbf{v}_n \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_n \\ | & & | \end{pmatrix} \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix}$$

- Since $\mathbf{V}$ is an orthogonal (unitary) matrix, $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}$ ($\mathbf{V}^\dagger\mathbf{V} = \mathbf{V}\mathbf{V}^\dagger = \mathbf{I}$),

$$\mathbf{A} = \hat{U}\hat{\Sigma}\mathbf{V}^\dagger$$

- The columns of $\hat{\mathbf{U}}$ are n orthonormal vectors in $\mathbb{C}^m$ ($m \geq n$).

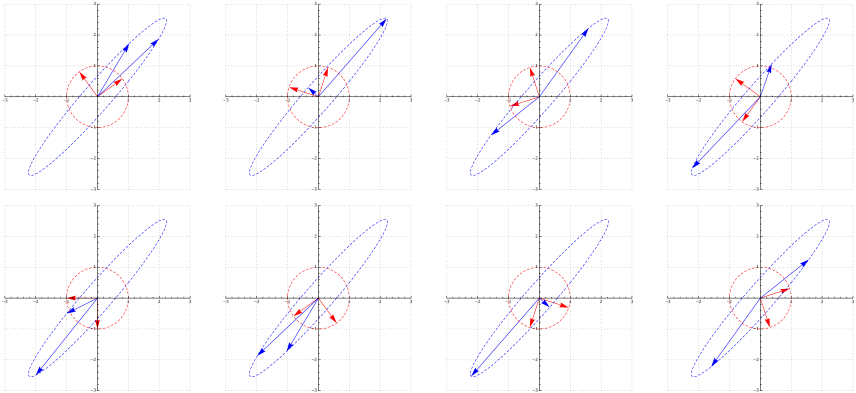
The full SVD describes both the domain and range of a matrix by orthonormal bases

- Extend the size of the vector space in the range from n to m by adding columns to $\hat{\mathbf{U}}$ to yield a $m \times m$ *unitary* (for complex) or *orthogonal* (for real) matrix $\mathbf{U}$.
- To maintain the same value for the product of matrices (after all, we need to recover $\mathbf{A}$ from its factors), extend matrix $\hat{\Sigma}$ by adding zeros along the diagonal to obtain matrix Σ.
- For an arbitrary matrix $\mathbf{A} \in \mathbb{C}^{m \times n}$ we have an $n \times n$ matrix $\mathbf{V}$ and a $m \times m$ matrix $\mathbf{U}$ that are both orthonormal, and a $m \times n$ matrix Σ whose non-zero entries $\sigma_i = \Sigma_{ii}$ are along the diagonal:

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

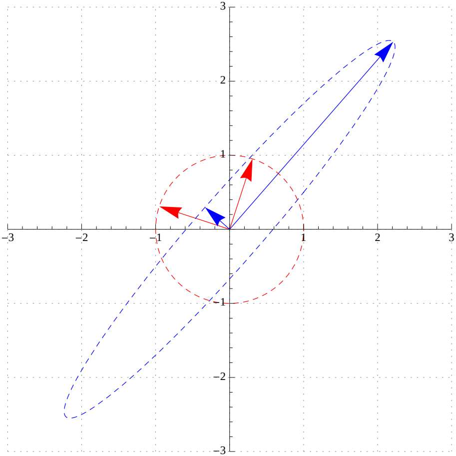
- The columns of $\mathbf{V}$ and $\mathbf{U}$ are the right and left singular vectors, and the diagonal entries of Σ are the singular values of $\mathbf{A}$.

Geometry of SVD: choice of basis vectors lying on circle and map



♡ Choose the pre-image of the orthogonal pair in the range of the map.

Singular vectors describe spheres and ellipsoids by semi-major axes



- There is one choice of vector pairs (basis) in the domain that gets mapped into an orthogonal pair along the major axes of the ellipse.
- These pairs are the **singular vectors** of the matrix. The lengths of the semi-major axes of the ellipse are the **singular values**.
- There will be left and right singular vectors

How is the SVD made useful in machine learning?

- Distance minimisation: matrix generalisation of the following
- To find a vector $\mathbf{y}$ from a set $\mathcal{Y}$ closest to $\mathbf{x}$ we perform

$$\mathbf{y} = \operatorname*{argmin}_{\mathbf{v} \in \mathcal{Y}} \|\mathbf{x} - \mathbf{v}\|.$$

- For $\mathbf{z} = (z_1, \ldots, z_n)$, $\|\mathbf{z}\|$ is a **norm** – several choices:
  - L_2 norm: $\|\mathbf{z}\|_2 = \sqrt{\mathbf{z} \cdot \mathbf{z}} = \sqrt{\sum_i z_i^2}$
  - L_1 norm: $\|\mathbf{z}\|_1 = \sum_i |z_i|$
  - L_p norm: $\|\mathbf{z}\|_p = (\sum_i |z_i|^p)^{1/p}$
  - L_0 norm: $\|\mathbf{z}\|_0 = \#(i|z_i \neq 0)$
  - L_∞ norm: $\|\mathbf{z}\|_\infty = max(|z_i|)$, $1 \leq i \leq n$.
- SVD helps find a matrix $\widetilde{\mathbf{X}}$ from a set $\mathcal{M}$ closest to given matrix $\mathbf{X}$:

$$\widetilde{\mathbf{X}} = \operatorname*{argmin}_{\mathbf{Y} \in \mathcal{M}} \|\mathbf{X} - \mathbf{Y}\|_2.$$

SVD gives low-rank approximation of matrices

- We seek $\widetilde{\mathbf{X}} = \mathrm{argmin}_{\mathbf{Y}\in\mathcal{M}} \|\mathbf{X} - \mathbf{Y}\|_2$.
- By partitioning the numerically ordered diagonal entries of Σ into the first k and the rest, we have (from the SVD)

$$\begin{aligned}
\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T &= \mathbf{U}_k\Sigma_k\mathbf{V}_k^T + \mathbf{U}_\perp\Sigma_\perp\mathbf{V}_\perp^T \\
&= (\mathbf{U}_k \;\; \mathbf{U}_\perp)\begin{pmatrix} \Sigma_k & \\ & \Sigma_\perp \end{pmatrix}\begin{pmatrix} \mathbf{V}_k^T \\ \mathbf{V}_\perp^T \end{pmatrix} \\
&\approx \mathbf{U}_k\Sigma_k\mathbf{V}_k^T \equiv \tilde{\mathbf{A}}_k
\end{aligned}$$

- $\mathbf{A}$ is replaced by the rank k matrix $\tilde{\mathbf{A}}_k$. Of all possible rank-k matrices $\mathbf{B} \in \mathcal{M}_k$, $\tilde{\mathbf{A}}_k$ constructed via the SVD gives the best approximation to $\mathbf{A}$ in the sense that it minimises the L_2-norm:

$$\tilde{\mathbf{A}}_k = \underset{\mathbf{B}\in\mathcal{M}_k}{\mathrm{argmin}} \|\mathbf{A} - \mathbf{B}\|_2.$$

Linear regression using SVD: find $\mathbf{w}$ for smallest $\|\mathbf{A}\mathbf{w} - \mathbf{y}\|_2$

- A vector $\mathbf{w}$ that is closest to target vector $\mathbf{y}$ along direction $\mathbf{u}$ is $\mathbf{w} = x^*\mathbf{v}$. Proof:

$$x^* = \operatorname*{argmin}_{x \in \mathbb{R}} \|\mathbf{y} - x\mathbf{u}\|_2 = \frac{\mathbf{y} \cdot \mathbf{u}}{\mathbf{u} \cdot \mathbf{u}} = \mathbf{y} \cdot \mathbf{u} \text{ projection.}$$

- Use SVD to find singular vectors $\mathbf{u}_i$ and find projections of $\mathbf{y}$ along each.
- Reminder: SVD expressed as

$$\begin{pmatrix} & & \\ & \mathbf{A} & \\ & & \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{v}_1 & \cdots & \mathbf{v}_n \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_n \\ | & & | \end{pmatrix} \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix}$$

Linear regression by SVD: express weights and targets in terms of singular vectors

- $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T = \sum_{k=1}^{r} \mathbf{u}_k \sigma_k \mathbf{v}_k^T$ implies $\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i$.
- Express $\mathbf{y} = \sum_k \beta_k \mathbf{u}_k$ and $\mathbf{w} = \sum_i \alpha_i \mathbf{v}_i$.
- Find projection of $\mathbf{y}$ along $\mathbf{u}_k$ for closest vectors $(\mathbf{u}_k \cdot \mathbf{y})\mathbf{u}_k$ to $\mathbf{y}$ along each direction $\mathbf{u}_k$. Choose $\beta_k = (\mathbf{u}_k^T \mathbf{y})$.
- The left hand side combines weighted features

$$\mathbf{A}\mathbf{w} = \mathbf{A}(\sum_i \alpha_i \mathbf{v}_i) = \sum_i \alpha_i (\mathbf{A}\mathbf{v}_i) = \sum_i \alpha_i \sigma_i \mathbf{u}_i.$$

- The best fit vector to $\mathbf{y}$ along each $\mathbf{u}_i$ is $\beta_i \mathbf{u}_i$. The vector in the column space of $\mathbf{A}$ along direction $\mathbf{u}_i$ is $\alpha_i \sigma_i \mathbf{u}_i$.
- The coefficients α_i of the optimal weight vector $\mathbf{w}$ along each of the singular vectors $\mathbf{v}_i$ are obtained from

$$\alpha_i \sigma_i = \beta_i = \mathbf{u}_i^T \mathbf{y} \implies \alpha_i = \frac{\mathbf{u}_i^T \mathbf{y}}{\sigma_i}.$$

Linear regression by SVD: small singular values are unwelcome

- The best fit weight vector is
$$\mathbf{w} = \sum_i \left(\frac{\mathbf{u}_i^T \mathbf{y}}{\sigma_i} \right) \mathbf{v}_i.$$
- What is the relationship between this expression and
$$\mathbf{w} = \left(\mathbf{A}^T \mathbf{A}\right)^{-1} \mathbf{A}^T \mathbf{y}?$$
- Verify
$$\left(\mathbf{A}^T \mathbf{A}\right)^{-1} \mathbf{A}^T \mathbf{y} = \mathbf{V} \Sigma^{-1} \mathbf{U}^T \mathbf{y}.$$
- Very small (zero) singular values cause problems. The large (infinite) components of the weight vectors track noise in the targets, not useful signals. This leads to the subject of **regularisation**.

Relationship between singular vectors/values and eigen- vectors/values

Since the eigenvectors of a matrix can be used as a basis for a vector space, it will be important to show how these constructs are related.

Represent matrix by its eigenvectors and eigenvalues

- Suppose $\mathbf{A}$ has n linear independent eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. Its stacked column vectors

$$\mathbf{Q} = (\mathbf{v}_1 \mathbf{v}_2 \cdots \mathbf{v}_n)$$

  give representation of the matrix $\mathbf{A}$

$$\mathbf{A} = \mathbf{Q}\,\mathbf{\Lambda}\,\mathbf{Q}^{-1} \text{ if nonsingular } \mathbf{Q},$$

  where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ of $\mathbf{A}$.
- **Proof:** The eigenvalue equations for the $\mathbf{v}_i$ can be written as $\mathbf{AQ} = \mathbf{Q\Lambda}$. Multiplying by $\mathbf{Q}^{-1}$ from the right gives the result.

Relationship between SVD and eigen-analysis

- Representation of $\mathbf{A}$ (real symmetric) $\mathbf{A} = \mathbf{Q}\,\boldsymbol{\Lambda}\,\mathbf{Q}^T$.
- For SVD of $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$,

$$
\begin{aligned}
\mathbf{X}\mathbf{X}^T = (\mathbf{U}\Sigma\mathbf{V}^T)(\mathbf{U}\Sigma\mathbf{V}^T)^T &= (\mathbf{U}\Sigma\mathbf{V}^T)((\mathbf{V}^T)^T\Sigma^T\mathbf{U}^T) \\
&= (\mathbf{U}\Sigma(\mathbf{V}^T\mathbf{V})\Sigma^T\mathbf{U}^T) \\
&= \mathbf{U}(\Sigma\Sigma^T)\mathbf{U}^T
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{X}^T\mathbf{X} = (\mathbf{U}\Sigma\mathbf{V}^T)^T(\mathbf{U}\Sigma\mathbf{V}^T) &= ((\mathbf{V}^T)^T\Sigma^T\mathbf{U}^T)(\mathbf{U}\Sigma\mathbf{V}^T) \\
&= (\mathbf{V}\Sigma^T(\mathbf{U}^T\mathbf{U})\Sigma\mathbf{V}^T) \\
&= \mathbf{V}(\Sigma^T\Sigma)\mathbf{V}^T
\end{aligned}
$$

- Right singular vectors of $\mathbf{X}$ are eigenvectors of $\mathbf{X}^T\mathbf{X}$;left singular vectors of $\mathbf{X}$ are eigenvectors of $\mathbf{X}\mathbf{X}^T$.Eigenvalues are σ_i^2 where $\sigma_i = \Sigma_{ii}$.

Representation for real symmetric matrices

- If $\mathbf{A}$ is a real symmetric matrix $\mathbf{A}$ we can construct $\mathbf{Q}$ from the n *orthonormal* eigenvectors $\hat{\mathbf{v}}_i$ (i.e. the eigenvectors must also be normalized to unit length) as $\mathbf{Q} = (\hat{\mathbf{v}}_1 \hat{\mathbf{v}}_2 \cdots \hat{\mathbf{v}}_n)$. We can show that $\mathbf{Q}$ is an *orthogonal matrix* ie

$$\mathbf{Q}^{-1} = \mathbf{Q}^T \ .$$

- This is easily proved from the fact that $\hat{\mathbf{v}}_i \cdot \hat{\mathbf{v}}_j = 0$ for $i \neq j$ and $|\hat{\mathbf{v}}_i| = 1$, which can be written as $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}$. Hence, we get

$$\mathbf{A} = \mathbf{Q} \, \mathbf{\Lambda} \, \mathbf{Q}^T \ .$$

Summary SVD/eigenvalues/vectors

- SVD for matrix $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$; columns of $\mathbf{U}$, $\mathbf{V}$ orthonormal, Σ has only diagonal entries non-zero $\sigma_i = \Sigma_{ii}$ (singular values) .
- Definition: For a square matrix $\mathbf{A}$, find nontrivial vectors $\mathbf{v}$ (eigenvectors) such that matrix multiplication behaves like scalar multiplication: $\mathbf{Av} = \lambda\mathbf{v}$ for scalars (eigenvalues) λ.
- For **real symmetric matrices** $n \times n$ matrices, eigenvalues λ are real numbers and we can always find n orthogonal eigenvectors $\mathbf{v}_i$, for $i = 1, \ldots, n$. This means that $\mathbf{v}_i \cdot \mathbf{v}_j = 0$ for $j \neq i$.
- Representation of $\mathbf{A}$ (real symmetric)

$$\mathbf{A} = \mathbf{Q}\,\mathbf{\Lambda}\,\mathbf{Q}^T \ .$$

  where $\mathbf{Q} = (\hat{\mathbf{v}}_1 \hat{\mathbf{v}}_2 \cdots \hat{\mathbf{v}}_n)$ and $\mathbf{\Lambda}$ a diagonal matrix containing the eigenvalues.
- Right singular vectors of $\mathbf{X}$ are eigenvectors of $\mathbf{X}^T\mathbf{X}$,left singular vectors of $\mathbf{X}$ are eigenvectors of $\mathbf{X}\mathbf{X}^T$,with eigenvalues σ_i^2.