Assignment Project Exam Help

COMP 3223:
Foundations of Machine Learning
Linear Algebra
Singular Value Decomposition

https://powcoder.com

Srinandan Dasmahapatra

Add WeChat powcoder

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
- Eigenvalues and eigenvectors

Linear Regression: approximately solving equations

- In linear regression the model is $\widehat{\mathbf{y}} = \mathbf{A}\mathbf{w}$:

$$\begin{pmatrix} w_0\phi_0(x_1) + w_1\phi_1(x_1) + \cdots + w_p\phi_p(x_1) \\ w_0\phi_0(x_2) + w_1\phi_1(x_2) + \cdots + w_p\phi_p(x_2) \\ \vdots \\ w_0\phi_0(x_N) + w_1\phi_1(x_N) + \cdots + w_p\phi_p(x_N) \end{pmatrix} + \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

- Find weights $\{w_i\}$ that make residuals $\|\mathbf{r}\|$ small.

Linear dependence & Linear Regression

- In linear regression the model is $\widehat{\mathbf{y}} = \mathbf{A}\mathbf{w}$:

$$w_0 \underbrace{\begin{pmatrix} \phi_0(x_1) \\ \phi_0(x_2) \\ \vdots \\ \phi_0(x_N) \end{pmatrix}}_{\mathrm{col}_0(\mathbf{A})} + w_1 \underbrace{\begin{pmatrix} \phi_1(x_1) \\ \phi_1(x_2) \\ \vdots \\ \phi_1(x_N) \end{pmatrix}}_{\mathrm{col}_1(\mathbf{A})} + \cdots + w_p \underbrace{\begin{pmatrix} \phi_p(x_1) \\ \phi_p(x_2) \\ \vdots \\ \phi_p(x_N) \end{pmatrix}}_{\mathrm{col}_p(\mathbf{A})} = \begin{pmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_N \end{pmatrix}$$

- Find **linear combination of columns** of design matrix that spans $\mathbf{y}$.
- Residual $\mathbf{r}$ not in space spanned by columns of $\mathbf{A}$ – residual orthogonal to each column:
- $\sum_n r_n \phi_i(x_n) = 0 \ldots$ is where gradient of squared loss vanishes.

Design matrix has information on patterns in data

- In linear regression the model is $\widehat{\mathbf{y}} = \mathbf{A}\mathbf{w}$:

$$\begin{pmatrix} 1 & \phi_1(x_1) & \cdots & \phi_p(x_1) \\ 1 & \phi_1(x_2) & \cdots & \phi_p(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \phi_1(x_N) & \cdots & \phi_p(x_N) \end{pmatrix}$$

- Idea of this lecture: decompose matrix using transformations and data appropriate descriptive bases

Reminder: Solving Linear Equations – Geometrical Picture

- Solve set of equations:

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

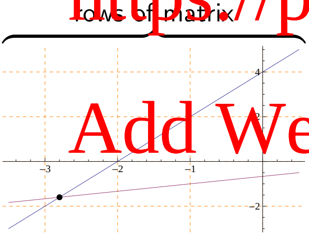
- Geometrically viewed as intersection of linear linear combination of vectors:

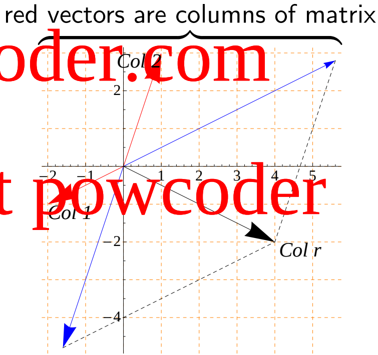
$$\overbrace{\begin{array}{rcl} ax + by & = & r \\ cx + dy & = & s \end{array}}^{\text{rows of matrix}} \leftrightarrow \overbrace{x \begin{pmatrix} a \\ c \end{pmatrix} + y \begin{pmatrix} b \\ d \end{pmatrix} = \begin{pmatrix} r \\ s \end{pmatrix}}^{\text{columns of matrix}}$$

The Geometrical Picture: An example

- Solve set of equations:

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

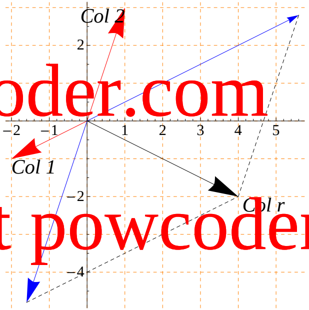




- Solution of $y - 2x = 4, 3y - x = -2$, is $(x, y) = (-2.8, -1.6)$.

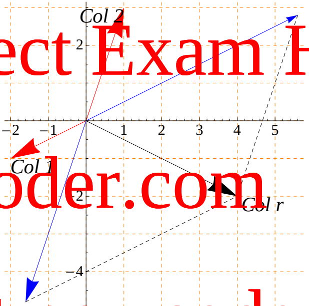
Fundamental operations on vectors – multiply by scalars and perform addition

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



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 v in direction $\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$, $Bv = 0$. 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 $\ker(A^\top) = c \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix}$.

Kernel or Null space of a matrix

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

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

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

Rank of a matrix = number of independent equations

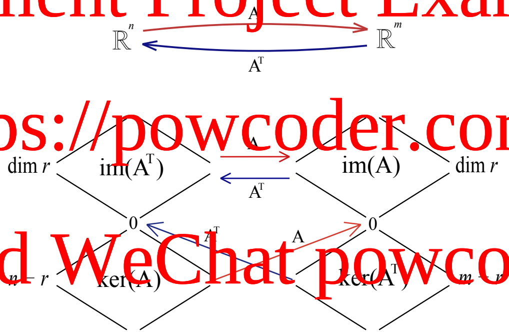
- The **rank** (column rank) of A is the dimension of the column space of A.
- A vector space is partitioned into its range and null spaces:

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

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

Four fundamental subspaces of a matrix

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



Linear regression with fixed functions of data

- Linear combinations of fixed functions $\phi_j(\mathbf{x}_n)$:

$$w_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + w_1 \begin{pmatrix} \phi_1(x_1) \\ \phi_1(x_2) \\ \vdots \\ \phi_1(x_N) \end{pmatrix} + \cdots + w_p \begin{pmatrix} \phi_p(x_1) \\ \phi_p(x_2) \\ \vdots \\ \phi_p(x_N) \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

- Sets of functions constitute vector spaces
- Approximate outputs/targets $\mathbf{y}$ by element of column space of the design matrix.

Functions constitute vector spaces

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

$$(a_0 + a_1 x + a_2 x^2) + (b_0 + b_1 x) = \underbrace{(a_0 + b_0)}_{c_0} + \underbrace{(a_1 + b_1)}_{c_1} x + \underbrace{a_2}_{c_2} x^2$$

Monomials as basis elements $\mathbf{a} + \mathbf{b} = \mathbf{c}$:

$$(a_0, a_1, a_2) + (b_0, b_1, 0) = (a_0 + b_0, a_1 + b_1, a_2) = (c_0, c_1, c_2)$$

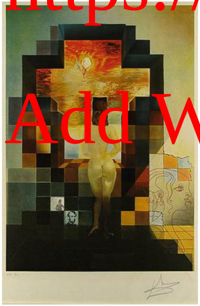
- Similarly the set $\mathbb{R}[x_1, \ldots, x_k]$ of polynomials in k variables forms a vector space.
- Set of functions of the form $\sum_{|n|<N} a_n e^{in\theta}$ (Fourier series).
- Extension – replace sums (where the summation index is from a discrete set) by integrals (where the index being summed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

Even matrices form a vector space

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

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

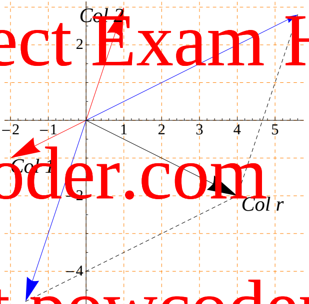
- image processing/computer vision



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

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 $\mathbb{R}$ called their **scalar** (or **dot**) **product**:

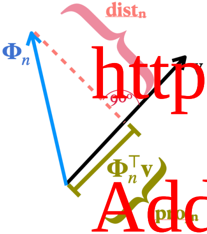
$$\begin{aligned} \text{dot} : V \times V &\to \mathbb{R} \\ \text{dot}(\mathbf{v}, \mathbf{w}) = \mathbf{v} \cdot \mathbf{w} = \mathbf{v}^\top \mathbf{w} &\mapsto a \end{aligned}$$

- 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 $\mathbf{v} \in V$, $\dim V = N$,

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

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

Using dot products to introduce projections



- **Project** vector $\boldsymbol{\Phi}_n$ on to direction given by vector $\mathbf{v}$
- Value of **projection** proj_n could be negative if α is bigger than 90°.

$$\text{Note, } \text{proj}_n = \|\boldsymbol{\Phi}_n\| \cos\alpha = \frac{\boldsymbol{\Phi}_n \cdot \mathbf{v}}{\|\mathbf{v}\|}$$

since **dot** product is $\boldsymbol{\Phi}_n \cdot \mathbf{v} = \|\mathbf{v}\|\|\boldsymbol{\Phi}_n\| \cos\alpha$.

Example: expand vector in orthogonal basis

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

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

- Solution: Multiply $\mathbf{v}$ by $\mathbf{e}_j$, use 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

- Let $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ be a set of orthogonal $n \times 1$ column vectors and $\mathbf{v}$ an arbitrary $n \times 1$ column vector.
- Task: Expand $\mathbf{v}$ as *linear combination* of set $\{\mathbf{v}_i\}_i$, ie. find numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$ s.t. $\mathbf{v} = \sum_{i=1}^n \alpha_i \mathbf{v}_i$.
- Solution: Multiply $\mathbf{v}_j \cdot (\mathbf{v} = \sum_{i=1}^n \alpha_i \mathbf{v}_i)$ use $\mathbf{v}_i \cdot \mathbf{v}_j = \delta_{ij} \|\mathbf{v}_i\|^2$ (orthogonality) to get $\mathbf{v}_j \cdot \mathbf{v} = \sum_{i=1}^n \alpha_i \delta_{ij} = \alpha_j$. Explicitly,

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

Hence

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

- Seek to characterise design matrix in terms of some orthonormal bases

Design matrix is not square

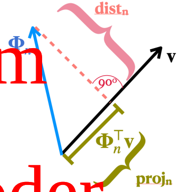
- The domain and range of matrix have different dimensions
- Need descriptive basis for each action of matrix on vector space pieced together from its action on orthonormal basis
- For instance, each feature corresponds to a direction represented by a unit vector:

$$\mathbf{Aw} = w_0\mathbf{A}\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix} + w_1\mathbf{A}\begin{pmatrix}0\\1\\\vdots\\0\end{pmatrix} + \cdots + w_p\mathbf{A}\begin{pmatrix}0\\0\\\vdots\\1\end{pmatrix}$$

- $\mathbf{Aw} = w_0\mathrm{col}_0(\mathbf{A}) + \cdots + w_p\mathrm{col}_p(\mathbf{A})$
- Introduce **singular value decomposition** (SVD) to find approximate subspaces
- Generalise notion of eigenvalue/eigenvector pair

Approximating subspaces

- Point $\boldsymbol{\Phi}_n = (\phi_{n1}, \phi_{n2}, \ldots, \phi_{np})$
- Project $\boldsymbol{\Phi}_n$ along $\mathbf{v}$
- $\|\boldsymbol{\Phi}_n\|^2 = (\text{proj}_{n,\mathbf{v}})^2 + (\text{dist}_{n,\mathbf{v}})^2$
- … sum of squares of distance to $\mathbf{v}$ and projection along $\mathbf{v}$
- $(\text{dist}_{n,\mathbf{v}})^2 = -(\text{proj}_{n,\mathbf{v}})^2 + \|\boldsymbol{\Phi}_n\|^2$
- $\boldsymbol{\Phi}_n$ is $\mathbf{v}$-independent
- Minimising distance to $\mathbf{v}$ equivalent to maximising projection along $\mathbf{v}$



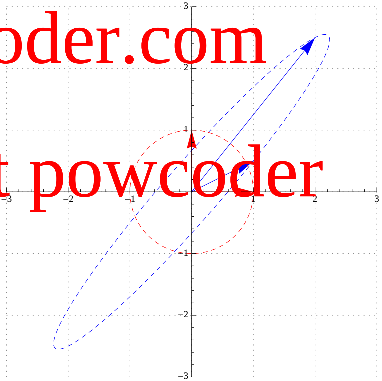
Two interpretations of best fit subspaces

1. minimises sum of squares of distances of data points to the subspace
2. maximises the sum of squares of projections to the subspace (subspace contains maximal relevant information of data, among all subspaces of same dimension)

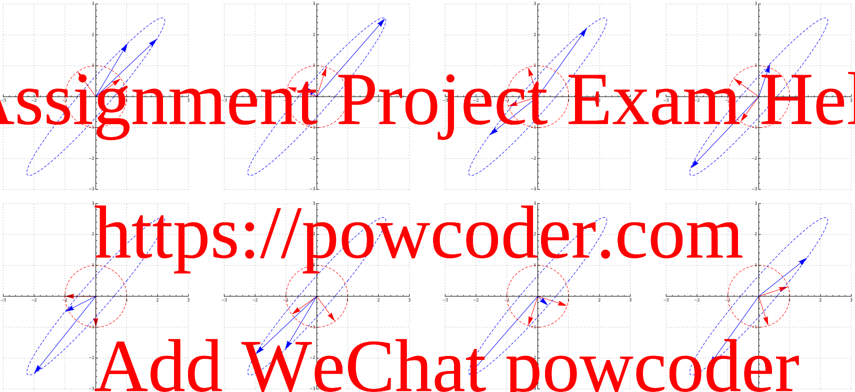
Singular Value Decomposition (SVD) of a Matrix

- 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;
- If matrix A is n-by-m, SVD of A characterises how an m-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 usually not.

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, $\mathbf{w} = (\mathbf{v}_1^\top \mathbf{w})\mathbf{v}_1 + (\mathbf{v}_2^\top \mathbf{w})\mathbf{v}_2$.
- The range of $\mathbf{A}$ is spanned by $\mathbf{u}_1, \mathbf{u}_2$ with $\mathbf{A}\mathbf{v}_j = \sigma_j \mathbf{u}_j$ and σ_j scalars for $j = 1,2$
- Action of $\mathbf{A}$ on vector $\mathbf{w}$

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

- Express that as $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$, 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.

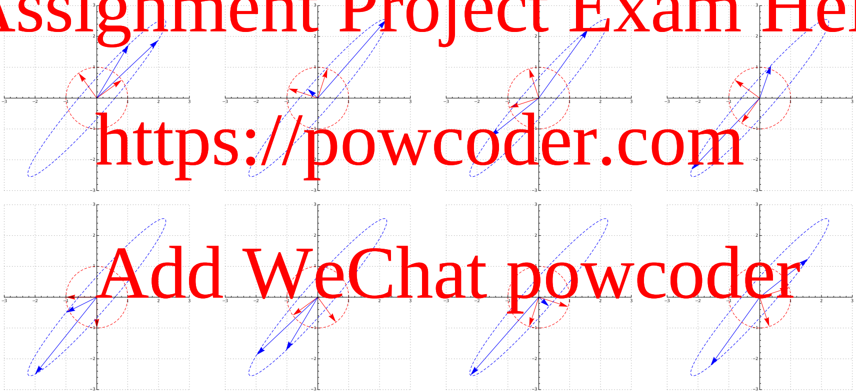
The full SVD describes both the domain and range of a matrix by orthonormal bases

- For an arbitrary matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ 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:

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

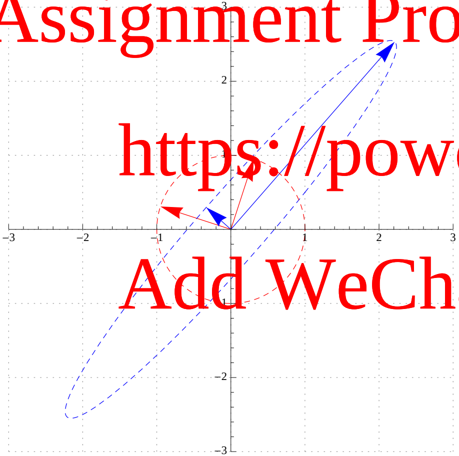
- The columns of $\mathbf{V}$ and $\mathbf{U}$ are the right and left singular vectors, and the diagonal entries of Σ are the singular values of $\mathbf{A}$.

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**.
- Left and right singular vectors obtained by finding vectors $\mathbf{v}, \mathbf{u}$ that yield maximum lengths lengths of $\mathbf{Av}$ and $\mathbf{A}^\top\mathbf{u}$

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

- A vector $\boldsymbol{\phi}$ that is closest to target vector $\mathbf{y}$ along direction $\mathbf{u}$ is $\boldsymbol{\phi} = \alpha^* \mathbf{u}$. Proof: all vectors in direction $\mathbf{u}$ of the form $\alpha \mathbf{u}$,

$$\alpha^* = \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} \|\mathbf{y} - \alpha \mathbf{u}\|^2 = \frac{\mathbf{y}^\top \mathbf{u}}{\mathbf{u}^\top \mathbf{u}}$$

- Recall linear regression: slope $w_1 = (\mathbf{y}^\top \mathbf{x})/(\mathbf{x}^\top \mathbf{x})$

- Projection along $\mathbf{u}$.
- Use SVD to find singular vectors $\mathbf{u}_i$ and find projections $\mathbf{y} \cdot \mathbf{u}_i$.

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

- $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top = \sum_{k=1}^{r} \mathbf{u}_k \sigma_k \mathbf{v}_k^\top$ or, equivalently, $\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i$.
- Weight space spanned by $\mathbf{v}_i \in \mathbb{R}^{p+1}$, output space spanned by $\mathbf{u}_k \in \mathbb{R}^N$:

$$\mathbf{w} = \sum_i \alpha_i \mathbf{v}_i \text{ and } \widehat{\mathbf{y}} = \sum_k \beta_k \mathbf{u}_k.$$

- Nearest vector to $\mathbf{y}$ along direction $\mathbf{u}_k$ is $(\mathbf{u}_k^\top \mathbf{y})\mathbf{u}_k$. Let $\beta_k = (\mathbf{u}_k^\top \mathbf{y})$.
- Model prediction $\mathbf{A}\mathbf{w}$ combines weighted features

$$\mathbf{A}\mathbf{w} = \mathbf{A}(\sum_k \alpha_k \mathbf{v}_k) = \sum_k \alpha_k (\mathbf{A}\mathbf{v}_k) = \sum_k \alpha_k \sigma_k \mathbf{u}_k.$$

- Best fit vector to $\mathbf{y}$ along each $\mathbf{u}_k$ is $\beta_k \mathbf{u}_k$. Vector in column space of $\mathbf{A}$ along direction $\mathbf{u}_k$ is $\alpha_k \sigma_k \mathbf{u}_k$.
- Equating coefficients along $\mathbf{u}_i$,

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

Linear regression by SVD: small singular values are unwelcome

- The best fit weight vector is

$$\mathbf{w} = \sum_i \left(\frac{\mathbf{u}_i^\top \mathbf{y}}{\sigma_i} \right) \mathbf{v}_i.$$

- Singular values $\sigma_i \approx 0$ cause problems: large components weight vectors could track "noise" in targets of training set, not "signal" which will generalise.

- Requires **regularisation**: add positive constant to denominator from minimising

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} \|\mathbf{y} - \mathbf{A}\mathbf{w}\|^2 + \lambda \|\mathbf{w}\|^2 \implies \mathbf{w}^* = \sum_i \frac{\sigma_i}{\sigma_i^2 + \lambda} (\mathbf{u}_i^\top \mathbf{y}) \mathbf{v}_i.$$