

Advanced Machine Learning

Singular Value Decomposition (SVD)

The diagram illustrates the Singular Value Decomposition (SVD) of a matrix X . On the left, a large matrix is shown in block form, enclosed in large parentheses. It consists of a top-left block labeled 0 , a top-right block labeled X (highlighted in pink), a bottom-left block labeled X^T (highlighted in blue), and a bottom-right block labeled 0 . To the right of this matrix is a vertical bar, followed by a column vector labeled u (highlighted in blue) and another column vector labeled v (highlighted in pink). This is followed by an equals sign and a scalar s . To the right of s is another vertical bar, followed by a column vector labeled u (highlighted in blue) and another column vector labeled v (highlighted in pink). The entire expression represents the equation $\begin{pmatrix} 0 & X \\ X^T & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = s \begin{pmatrix} u \\ v \end{pmatrix}$.

Singular Valued Decomposition, SVD, general linear maps

Outline

1. **Singular Value Decomposition**
2. General Linear Mappings
3. Linear Regression Revisited

$$\begin{pmatrix} 0 & X \\ X^T & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = s \begin{pmatrix} u \\ v \end{pmatrix}$$

Singular Valued Decomposition

- Consider an arbitrary $n \times m$ matrix \mathbf{X} , and construct the $(n + m) \times (n + m)$ symmetric matrix, \mathbf{B} ,

$$\begin{pmatrix} 0 & \mathbf{X} \\ \mathbf{X}^\top & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} = s \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}$$

$\begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}$ is an eigenvector of \mathbf{B} with eigenvalue s

- We observe that

$$\begin{aligned} \mathbf{X}\mathbf{v} &= s\mathbf{u} & \mathbf{X}^\top\mathbf{u} &= s\mathbf{v} \\ \mathbf{X}^\top\mathbf{X}\mathbf{v} &= s\mathbf{X}^\top\mathbf{u} = s^2\mathbf{v} & \mathbf{X}\mathbf{X}^\top\mathbf{u} &= s\mathbf{X}\mathbf{v} = s^2\mathbf{u} \end{aligned}$$

Eigenvectors

- Note that as $\mathbf{X}\mathbf{v} = s\mathbf{u}$ and $\mathbf{X}^\top\mathbf{u} = s\mathbf{v}$ then

$$\mathbf{X}(-\mathbf{v}) = (-s)\mathbf{u} \qquad \mathbf{X}^\top\mathbf{u} = (-s)(-\mathbf{v})$$

if $\begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}$ is an eigenvector of \mathbf{B} with eigenvalue s then so is $\begin{pmatrix} \mathbf{u} \\ -\mathbf{v} \end{pmatrix}$ with eigenvalue $-s$ ■

- If $n < m$ then $\mathbf{X}^\top\mathbf{X}$ is not full rank so some eigenvalues are zero■
- As a consequence $m - n$ vectors exist such that $\mathbf{X}\mathbf{v} = 0$ ■
- The eigenvalues and eigenvectors are

$$n \times \left(s_i, \begin{pmatrix} \mathbf{u}_i \\ \mathbf{v}_i \end{pmatrix} \right) \quad n \times \left(-s_i, \begin{pmatrix} \mathbf{u}_i \\ -\mathbf{v}_i \end{pmatrix} \right) \quad m - n \times \left(0, \begin{pmatrix} 0 \\ \mathbf{v}_k \end{pmatrix} \right) \quad \blacksquare$$

Matrix Decomposition

- Stacking the eigenvectors into a matrix

$$\begin{pmatrix} 0 & \mathbf{x} \\ \mathbf{x}^T & 0 \end{pmatrix} = \begin{pmatrix} \mathbf{u} & \mathbf{u} & 0 \\ \mathbf{v} & -\mathbf{v} & \mathbf{v}_0 \end{pmatrix} \begin{pmatrix} \mathbf{u} & \mathbf{u} & 0 \\ \mathbf{v} & -\mathbf{v} & \mathbf{v}_0 \end{pmatrix} \begin{pmatrix} \mathbf{S} & 0 & 0 \\ 0 & -\mathbf{S} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

- Since the vectors $\begin{pmatrix} \mathbf{u}_i \\ \mathbf{v}_i \end{pmatrix}$ are eigenvectors of a symmetric matrix they form an orthogonal matrix if they are normalised.■
- Multiply on the right by the transpose of the orthogonal matrix

$$\begin{pmatrix} 0 & \mathbf{x} \\ \mathbf{x}^T & 0 \end{pmatrix} = \begin{pmatrix} \mathbf{u} & \mathbf{u} & 0 \\ \mathbf{v} & -\mathbf{v} & \mathbf{v}_0 \end{pmatrix} \begin{pmatrix} \mathbf{S} & 0 & 0 \\ 0 & -\mathbf{S} & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}^T & \mathbf{v}^T \\ \mathbf{u}^T & -\mathbf{v}^T \\ 0 & \mathbf{v}_0^T \end{pmatrix}$$

Normalisation Subtlety

$$\begin{pmatrix} 0 & \mathbf{X} \\ \mathbf{X}^T & 0 \end{pmatrix} = \begin{pmatrix} \mathbf{U} & \mathbf{U} & 0 \\ \mathbf{V} & -\mathbf{V} & \mathbf{V}_0 \end{pmatrix} \begin{pmatrix} \mathbf{S} & 0 & 0 \\ 0 & -\mathbf{S} & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{U}^T & \mathbf{V}^T \\ \mathbf{U}^T & -\mathbf{V}^T \\ 0 & \mathbf{V}_0^T \end{pmatrix}$$

- Multiplying out we have

$$\mathbf{X} = 2\mathbf{U}\mathbf{S}\mathbf{V}^T$$

$$\mathbf{X}^T = 2\mathbf{V}\mathbf{S}\mathbf{U}^T$$

- Now the vectors \mathbf{u}_i and \mathbf{v}_i form an orthogonal set as it satisfy

$$\mathbf{X}^T \mathbf{X} \mathbf{v} = s^2 \mathbf{v}$$

$$\mathbf{X} \mathbf{X}^T \mathbf{u} = s^2 \mathbf{u}$$

- But they are not normalised (since $\begin{pmatrix} \mathbf{u}_i \\ \mathbf{v}_i \end{pmatrix}$ is normalised). If we define $\tilde{\mathbf{U}} = \sqrt{2}\mathbf{U}$ and $\tilde{\mathbf{V}} = \sqrt{2}\mathbf{V}$ we find

$$\mathbf{X} = \tilde{\mathbf{U}}\mathbf{S}\tilde{\mathbf{V}}^T$$

$$\mathbf{X}^T = \tilde{\mathbf{V}}\mathbf{S}\tilde{\mathbf{U}}^T$$

SVD

- Any matrix, \mathbf{X} , can be written as $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$
 - ★ \mathbf{U} , \mathbf{V} are orthogonal matrices
 - ★ $\mathbf{S} = \text{diag}(s_1, s_2, \dots, s_n)$ ■
- s_i can always be chosen to be positive and are known as **singular values**■
- Singular value decomposition applies to both square and non-square matrices—they describe general linear mappings■

Finding SVD

- Most libraries will compute the SVD for you■
- They can do this by choosing the smaller of two matrices $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$ and then compute the eigenvalues■
- The singular values are the square root of the eigenvalues (notice that $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$ are both positive semi-definite so the eigenvalues will be non-negative)■
- It can compute the \mathbf{U} matrix or \mathbf{V} matrix by multiplying through by \mathbf{X} or \mathbf{X}^T ($\mathbf{U} = \mathbf{X}\mathbf{V}\mathbf{S}^{-1}$ and $\mathbf{V} = \mathbf{X}^T\mathbf{U}\mathbf{S}^{-1}$)■
- In practice to perform PCA most people subtract the mean from their data and then perform SVD■

Economical Forms of SVD

- Often the rows or columns of the orthogonal matrices \mathbf{U} and \mathbf{V} that are not associated with a singular value are ignored

$$\mathbf{X} = \mathbf{U} \mathbf{S} \mathbf{V}^T$$

$$\mathbf{X} = \mathbf{U} \mathbf{S} \mathbf{V}^T$$

$$\mathbf{X} = \mathbf{U} \mathbf{S} \mathbf{V}^T$$

$$\mathbf{X} = \mathbf{U} \mathbf{S} \mathbf{V}^T$$

- In Matlab these are obtained using

```
>> [U, S, V] = svd(X)
>> [U, S, V] = svd(X, 'econ')
```

Outline

1. Singular Value Decomposition
2. **General Linear Mappings**
3. Linear Regression Revisited

$$\begin{pmatrix} 0 & X \\ X^T & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = s \begin{pmatrix} u \\ v \end{pmatrix}$$

General Matrix

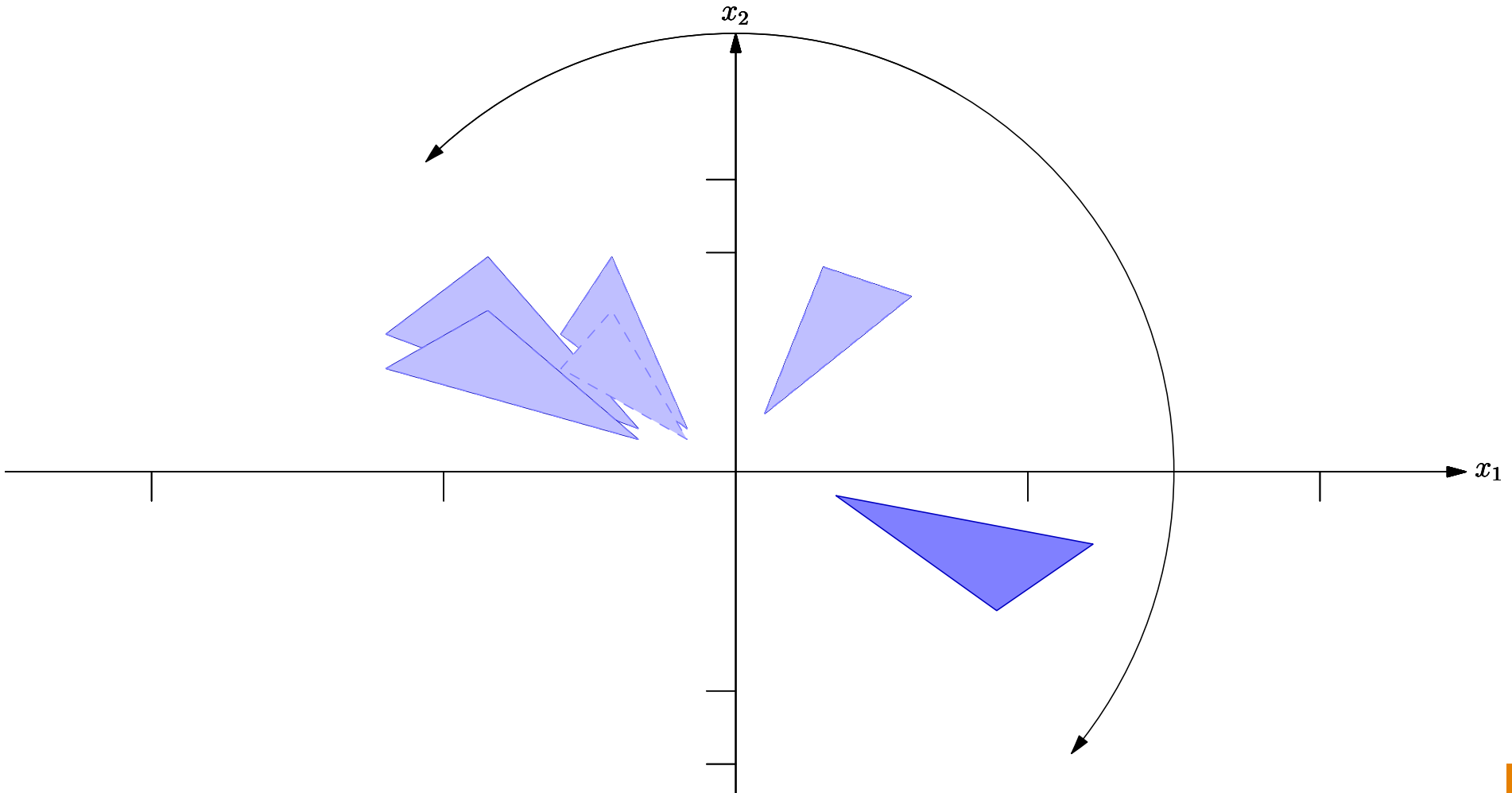
- Recall that we can compute the SVD for any matrix, \mathbf{X}
- As matrices describe the most general linear mapping

$$\mathbf{v} \rightarrow \mathcal{T}[\mathbf{v}] = \mathbf{X}\mathbf{v}$$

- We can use SVD to understand any linear mapping
- Thus any linear mapping can be seen as a rotation followed by a squashing or expansion independently in each coordinate followed by another rotation

Matrices

$$\mathbf{M} = \begin{pmatrix} -0.45 & 1.9 \\ -0.77 & -0.025 \end{pmatrix} = \mathbf{U}\mathbf{S}\mathbf{V}^T = \begin{pmatrix} \cos(-175) & \sin(-175) \\ -\sin(-175) & \cos(-175) \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 0.75 \end{pmatrix} \begin{pmatrix} \cos(75) & \sin(75) \\ -\sin(75) & \cos(75) \end{pmatrix}$$



Determinants

- The determinant, $|\mathbf{M}|$ of a matrix \mathbf{M} is defined for square matrices
- It describes the change in volume under the mapping
- Now for any two matrices $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$
- Thus

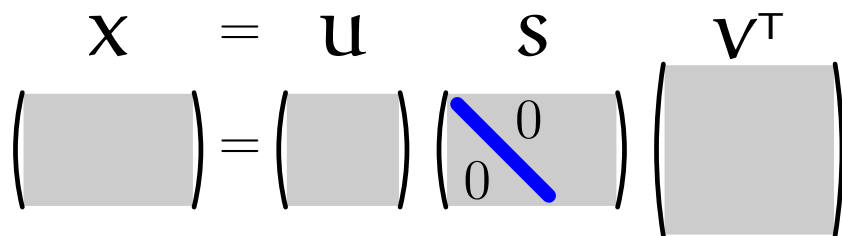
$$|\mathbf{M}| = |\mathbf{U}||\mathbf{S}||\mathbf{V}^T|$$

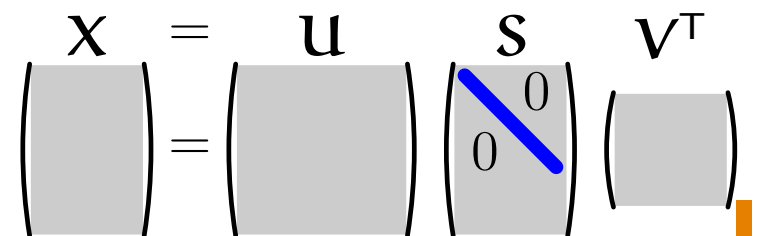
- For an orthogonal matrix $|\mathbf{U}| = \pm 1$
- Thus

$$|\mathbf{M}| = \pm |\mathbf{S}| = \pm \prod_i s_i$$

Non-Square Matrices

- When the matrices are non-square then the matrix of singular value matrix will either
 - ★ Squash some directions to zero
 - ★ Introduce new dimensions orthogonal to the vector

$$\mathbf{X} = \mathbf{U} \mathbf{S} \mathbf{V}^T$$


$$\mathbf{X} = \mathbf{U} \mathbf{S} \mathbf{V}^T$$


- The rank of an arbitrary matrix is the number of non-zero singular values (also number of linearly independent rows or columns)

Duality Revisited

- If $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ then

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

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

- If \mathbf{X} is an $p \times m$ matrix then $\mathbf{S}\mathbf{S}^T$ is a $p \times p$ diagonal matrix with elements $S_{ii}^2 = s_i^2$
- $\mathbf{S}^T\mathbf{S}$ is an $m \times m$ matrix with elements $S_{ii}^2 = s_i^2$
- \mathbf{U} and \mathbf{V} are matrices of eigenvectors for \mathbf{C} and \mathbf{D}
- The eigenvalues are $\lambda_i = S_{ii}^2 = s_i^2$

SS^T and $S^T S$

$$S = \begin{pmatrix} s_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_m & 0 & 0 \cdots & 0 \end{pmatrix} \quad \blacksquare$$

$$S^T S = \begin{pmatrix} s_1^2 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & s_2^2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_m^2 & 0 & 0 \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 \cdots & 0 \end{pmatrix} \quad \blacksquare$$

$$S S^T = \begin{pmatrix} s_1^2 & 0 & \cdots & 0 \\ 0 & s_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_m^2 \end{pmatrix} \quad \blacksquare$$

Having A Go

- It's really easy to verify this in MATLAB or OCTAVE

```
>> X = rand(3,2)
>> [U, S, V] = svd(X)
>> U*S*V'
>> U(:,1)'*U(:,2)
>> U'*U
>> U*U'
>> [Ua,L] = eig(X*X')
>> S*S'
```

- Test yourself!

Outline

1. Singular Value Decomposition
2. General Linear Mappings
3. **Linear Regression Revisited**

$$\begin{pmatrix} 0 & X \\ X^T & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = s \begin{pmatrix} u \\ v \end{pmatrix}$$

Linear Regression

- Given a set of data $\mathcal{D} = \{(\mathbf{x}_i, y_i) | k = 1, 2, \dots, m\}$ ■
- In linear regression we try to fit a linear model

$$f(\mathbf{x}|\mathbf{w}) = \mathbf{x}^\top \mathbf{w} \blacksquare$$

- Which we fit by minimising the squared error loss

$$L(\mathbf{w}) = \sum_{k=1}^m (f(\mathbf{x}_i|\mathbf{w}) - y_i)^2 \blacksquare$$

Matrix Form

- In matrix form we write $L(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_m^\top \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

- Then $\nabla L(\mathbf{w}^*) = 0$ implies

$$\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \mathbf{X}^+ \mathbf{y}$$

- This is known as the pseudo-inverse

Using SVD

- Using $X = \mathbf{U}\mathbf{S}\mathbf{V}^T$ then

$$\begin{aligned}
 \mathbf{X}^+ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \\
 &= (\mathbf{V} \mathbf{S}^T \mathbf{S} \mathbf{V}^T)^{-1} \mathbf{V} \mathbf{S}^T \mathbf{U}^T \\
 &= \mathbf{V} (\mathbf{S}^T \mathbf{S})^{-1} \mathbf{V}^T \mathbf{V} \mathbf{S}^T \mathbf{U}^T \\
 &= \mathbf{V} (\mathbf{S}^T \mathbf{S})^{-1} \mathbf{S}^T \mathbf{U}^T = \mathbf{V} \mathbf{S}^+ \mathbf{U}^T
 \end{aligned}$$

- If $m > p$

$$\mathbf{X}^T = \begin{pmatrix} | & | & | & | & | & | & | & | \\ & & & & & & & \end{pmatrix}, \mathbf{S}^T = \begin{pmatrix} s_1 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & s_2 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & s_3 & \dots & 0 & & 0 \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & s_p & 0 & 0 \dots & 0 \end{pmatrix}$$

Pseudo-Inverse of \mathbf{S}

$$\mathbf{S}^T \mathbf{S} = \begin{pmatrix} s_1^2 & 0 & \dots & 0 \\ 0 & s_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_p^2 \end{pmatrix} \quad (\mathbf{S}^T \mathbf{S})^{-1} = \begin{pmatrix} s_1^{-2} & 0 & \dots & 0 \\ 0 & s_2^{-2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_p^{-2} \end{pmatrix}$$

$$\mathbf{S}^+ = (\mathbf{S}^T \mathbf{S})^{-1} \mathbf{S}^T = \begin{pmatrix} s_1^{-1} & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & s_2^{-1} & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & s_3^{-1} & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & s_p^{-1} & 0 & 0 & \dots & 0 \end{pmatrix}$$

III-Conditioned Data Matrix

- Recall that

$$\mathbf{w}^* = \mathbf{X}^+ \mathbf{y} = \mathbf{V} \mathbf{S}^+ \mathbf{U}^\top \mathbf{y}$$

- If any of the singular values of \mathbf{X} are small then \mathbf{S}^+ will magnify components in that direction
- Any errors in the target \mathbf{y} will be magnified
- This leads to poor weights

Regularisation

- Consider linear regression with a regulariser

$$\begin{aligned}\mathcal{L}(\mathbf{w}) &= \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \eta\|\mathbf{w}\|^2 \\ &= \mathbf{w}^\top (\mathbf{X}^\top \mathbf{X} + \eta \mathbf{I}) \mathbf{w} - 2\mathbf{w}^\top \mathbf{X}^\top \mathbf{y} + \mathbf{y}^\top \mathbf{y}\end{aligned}$$

- Thus

$$\nabla \mathcal{L}(\mathbf{w}) = 2(\mathbf{X}^\top \mathbf{X} + \eta \mathbf{I}) \mathbf{w} - 2\mathbf{X}^\top \mathbf{y}$$

- and $\nabla \mathcal{L}(\mathbf{w}^*) = 0$ gives

$$\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X} + \eta \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$$

Regularisation Continued

- Using $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$

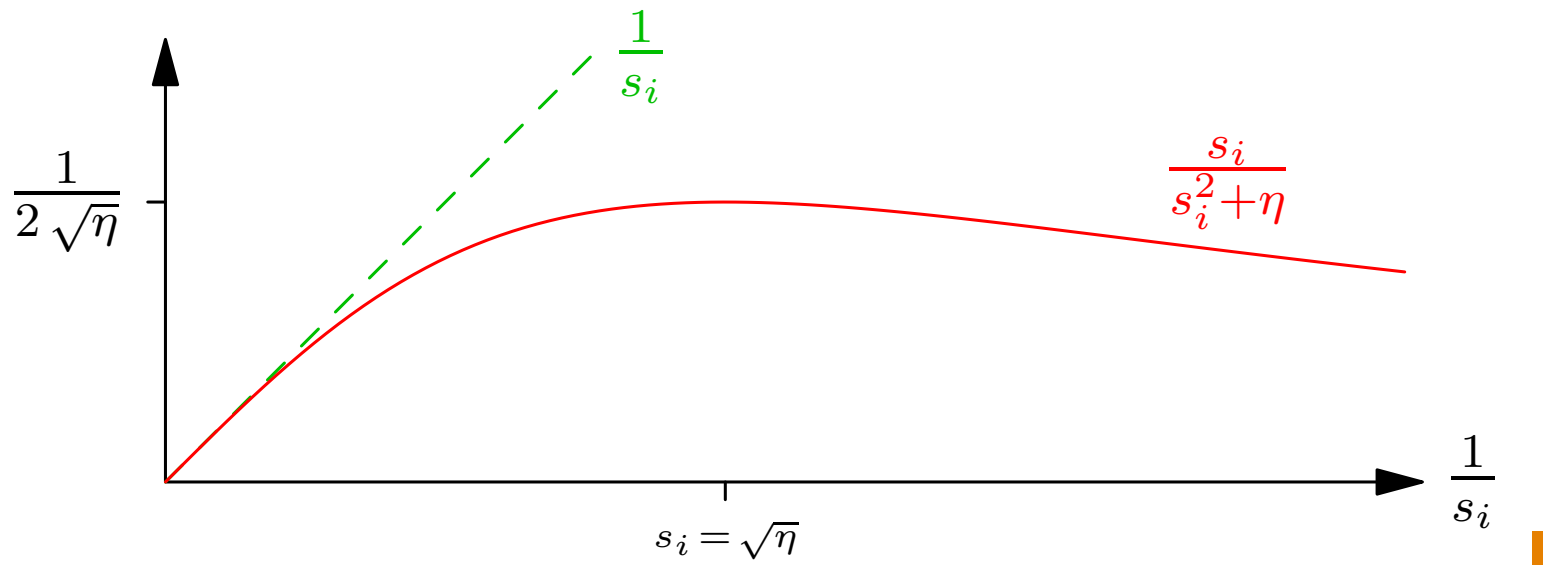
$$\begin{aligned} \mathbf{w}^* &= (\mathbf{X}^\top \mathbf{X} + \eta \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} \\ &= \mathbf{V} (\mathbf{S}^\top \mathbf{S} + \eta \mathbf{I})^{-1} \mathbf{S}^\top \mathbf{U}^\top \mathbf{y} \end{aligned}$$

- where

$$(\mathbf{S}^\top \mathbf{S} + \eta \mathbf{I})^{-1} \mathbf{S}^\top = \begin{pmatrix} \frac{s_1}{s_1^2 + \eta} & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \frac{s_2}{s_2^2 + \eta} & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & \frac{s_3}{s_3^2 + \eta} & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{s_p}{s_p^2 + \eta} & 0 & 0 \dots & 0 \end{pmatrix}$$

Effect of Regularisation

- Without regularisation if $s_i = 0$ the problem would be ill-posed (even \mathbf{S}^+ does not exist since s_i^{-1} would be ill defined) and if s_i is small then \mathbf{S}^+ is ill conditioned
- Using $\hat{\mathbf{S}}^+ = (\mathbf{S}^\top \mathbf{S} + \eta)^{-1} \mathbf{S}^\top$ instead of \mathbf{S}^+ then



- Regularisation makes the machine much more stable (reduces the variance)

Summary

- Any matrix can be decomposed as $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ where
 - ★ \mathbf{U} and \mathbf{V} are orthogonal (rotation matrices)■
 - ★ $\mathbf{S} = \text{diag}(s_1, \dots, s_n)$ is a diagonal matrix of positive singular values■
- This describes the most general linear transform■
- The transform exploits the duality between $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$ ■
- In linear regression the pseudo-inverse involves the reciprocal of the singular values, which can lead to poor generalisation■
- Regularisation improves the conditioning of the “inverse” matrix■