Advanced Machine Learning

Singular Value Decomposition (SVD)

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

Singular Valued Decomposition, SVD, general linear maps

Outline

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

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

Singular Valued Decomposition

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

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

  $\binom{\boldsymbol{u}}{\boldsymbol{v}}$ is an eigenvector of $\mathbf{B}$ with eigenvalue s

- We observe that

$$\mathbf{X}\boldsymbol{v} = s\boldsymbol{u} \qquad\qquad \mathbf{X}^\mathsf{T}\boldsymbol{u} = s\boldsymbol{v}$$

$$\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{v} = s\mathbf{X}^\mathsf{T}\boldsymbol{u} = s^2\boldsymbol{v} \qquad\qquad \mathbf{X}\mathbf{X}^\mathsf{T}\boldsymbol{u} = s\mathbf{X}\boldsymbol{v} = s^2\boldsymbol{u}$$

Eigenvectors

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

$$\mathbf{X}(-\boldsymbol{v}) = (-s)\boldsymbol{u} \qquad\qquad \mathbf{X}^\mathsf{T}\boldsymbol{u} = (-s)(-\boldsymbol{v})$$

  if $\binom{\boldsymbol{u}}{\boldsymbol{v}}$ is an eigenvector of $\mathbf{B}$ with eigenvalue s then so is $\binom{\boldsymbol{u}}{-\boldsymbol{v}}$ with eigenvalue $-s$

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

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

Matrix Decomposition

- Stacking the eigenvectors into a matrix

$$\begin{pmatrix} 0 & X \\ X^{\mathsf{T}} & 0 \end{pmatrix}\begin{pmatrix} U & U & 0 \\ V & -V & V_0 \end{pmatrix} = \begin{pmatrix} U & U & 0 \\ V & -V & V_0 \end{pmatrix}\begin{pmatrix} S & 0 & 0 \\ 0 & -S & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

- Since the vectors $\binom{\boldsymbol{u}_i}{\boldsymbol{v}_i}$ are eigenvectors of a symmetric matrix they from an orthogonal matrix if they are normalised.

- Multiply on the right by the transpose of the orthogonal matrix

$$\begin{pmatrix} 0 & X \\ X^{\mathsf{T}} & 0 \end{pmatrix} = \begin{pmatrix} U & U & 0 \\ V & -V & V_0 \end{pmatrix}\begin{pmatrix} S & 0 & 0 \\ 0 & -S & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} U^{\mathsf{T}} & V^{\mathsf{T}} \\ U^{\mathsf{T}} & -V^{\mathsf{T}} \\ 0 & V_0^{\mathsf{T}} \end{pmatrix}$$

Normalisation Subtlety

$$\begin{pmatrix} 0 & X \\ X^{\mathsf{T}} & 0 \end{pmatrix} = \begin{pmatrix} U & U & 0 \\ V & -V & V_0 \end{pmatrix}\begin{pmatrix} S & 0 & 0 \\ 0 & -S & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} U^{\mathsf{T}} & V^{\mathsf{T}} \\ U^{\mathsf{T}} & -V^{\mathsf{T}} \\ 0 & V_0^{\mathsf{T}} \end{pmatrix}$$

- Multiplying out we have

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

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

$$\mathbf{X}^{\mathsf{T}}\mathbf{X}\boldsymbol{v} = s^2\boldsymbol{v} \qquad \mathbf{X}\mathbf{X}^{\mathsf{T}}\boldsymbol{u} = s^2\boldsymbol{u}$$

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

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

SVD

- Any matrix, $\mathbf{X}$, can be written as $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathsf{T}}$
  - ⋆ $\mathbf{U}$, $\mathbf{V}$ are orthogonal matrices
  - ⋆ $\mathbf{S} = \mathrm{diag}(s_1, s_2, \ldots, s_n)$

- 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 $\mathbf{X}\mathbf{X}^{\mathsf{T}}$ and $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ and then compute the eigenvalues

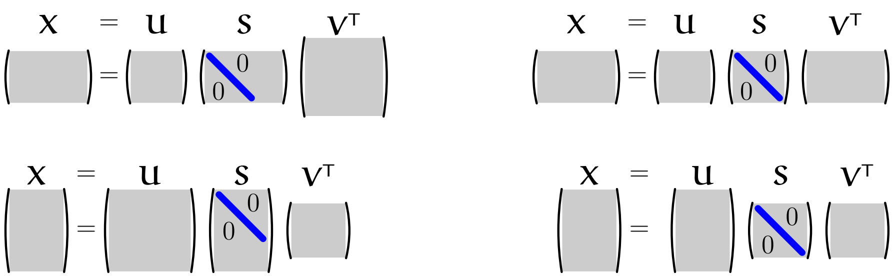
- The singular values are the square root of the eigenvalues (notice that $\mathbf{X}\mathbf{X}^{\mathsf{T}}$ and $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ are both positive semi-definite so the eigenvalues will be non-negative)

- It can compute the $\mathbf{U}$ matrix or $\boldsymbol{V}$ matrix by multiplying through by $\mathbf{X}$ or $\mathbf{X}^{\mathsf{T}}$ ($\mathbf{U} = \mathbf{X}\mathbf{V}\mathbf{S}^{-1}$ and $\mathbf{V} = \mathbf{X}^{\mathsf{T}}\mathbf{U}\mathbf{S}^{-1}$)

- 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

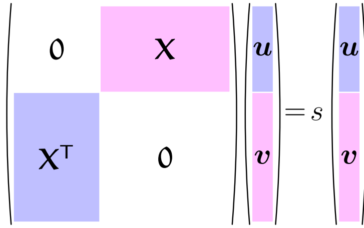


- In Matlab these are obtained using

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


Outline

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



General Matrix

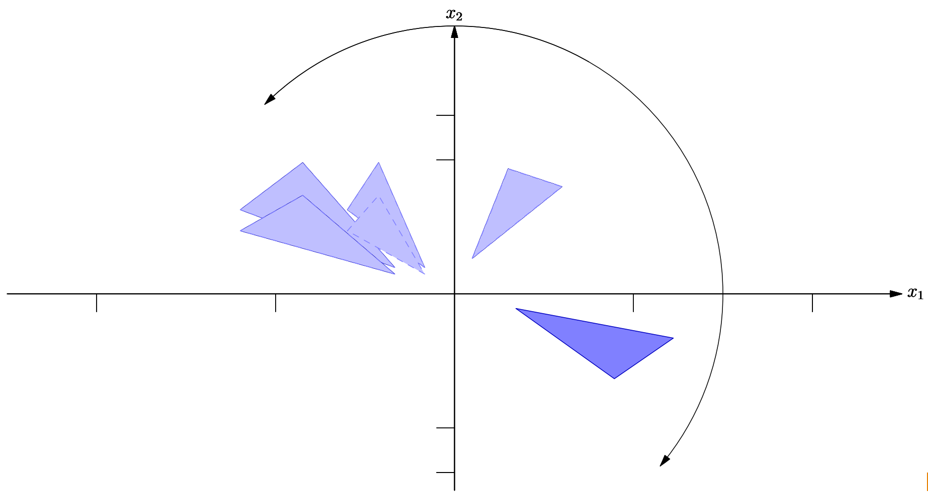
- Recall that we can compute the SVD for any matrix, $\mathbf{X}$
- As matrices describe the most general linear mapping

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

- 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

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



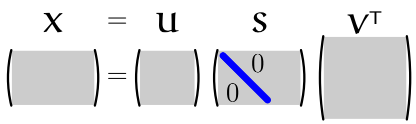
Determinants

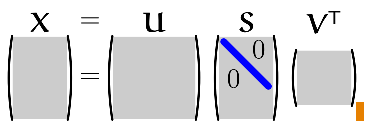
- The determinant, $|\mathbf{M}|$ of a matrix $\mathbf{M}$ is defined for square matrices
- It describes the change in volume under the mapping
- Now for any two matrices $|\mathbf{A}\mathbf{B}| = |\mathbf{A}||\mathbf{B}|$
- Thus $|\mathbf{M}| = |\mathbf{U}||\mathbf{S}||\mathbf{V}^\mathsf{T}|$
- For and orthogonal matrix $|\mathbf{U}| = \pm 1$ since $\mathbf{U}\mathbf{U}^\mathsf{T} = \mathbf{I} \Rightarrow |\mathbf{U}\mathbf{U}^\mathsf{T}| = |\mathbf{I}| \Rightarrow |\mathbf{U}||\mathbf{U}^\mathsf{T}| = 1$ or $|\mathbf{U}|^2 = 1$
- 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





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

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

- If $\mathbf{X}$ is an $p \times m$ matrix then $\mathbf{S}\mathbf{S}^\mathsf{T}$ is a $p \times p$ diagonal matrix with elements $S_{ii}^2 = s_i^2$
- $\mathbf{S}^\mathsf{T}\mathbf{S}$ 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 $S^\mathsf{T}S$

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

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

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

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**

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

Linear Regression

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

$$f(\boldsymbol{x}|\boldsymbol{w}) = \boldsymbol{x}^{\mathsf{T}} \boldsymbol{w}$$

- Which we fit by minimising the squared error loss

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

Matrix Form

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

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1^{\mathsf{T}} \\ \boldsymbol{x}_2^{\mathsf{T}} \\ \vdots \\ \boldsymbol{x}_m^{\mathsf{T}} \end{pmatrix} \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2, \\ \vdots \\ y_m \end{pmatrix}$$

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

$$\boldsymbol{w}^* = \left(\mathrm{X}^{\mathsf{T}}\mathrm{X}\right)^{-1} \mathrm{X}^{\mathsf{T}} \boldsymbol{y} = \mathrm{X}^{+} \boldsymbol{y}$$

- This is known as the pseudo-inverse

Using SVD

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

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

- If $m > p$

$\mathbf{X}^\mathsf{T} =$ 

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

Pseudo-Inverse of $\mathbf{S}$

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

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

Ill-Conditioned Data Matrix

- Recall that

$$\boldsymbol{w}^* = \mathbf{X}^+\boldsymbol{y} = \mathbf{V}\mathbf{S}^+\mathbf{U}^\mathsf{T}\boldsymbol{y}$$

- 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 $\boldsymbol{y}$ will be magnified
- This leads to poor weights

Regularisation

- Consider linear regression with a regulariser

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

- Thus

$$\boldsymbol{\nabla}\mathcal{L}(\boldsymbol{w}) = 2\left(\mathbf{X}^\mathsf{T}\mathbf{X} + \eta\mathbf{I}\right)\boldsymbol{w} - 2\mathbf{X}^\mathsf{T}\boldsymbol{y}$$

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

$$\boldsymbol{w}^* = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \eta\mathbf{I}\right)^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{y}$$

Regularisation Continued

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

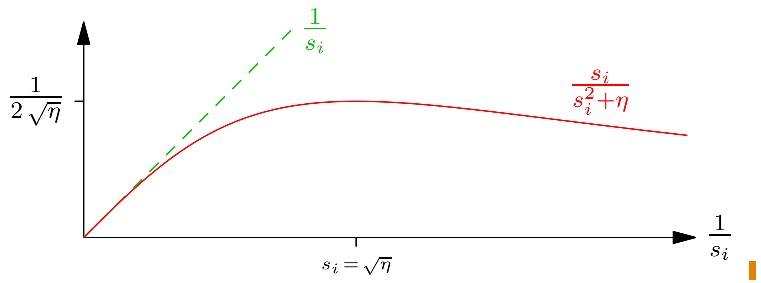
$$
\begin{aligned}
\boldsymbol{w}^* &= \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \eta \mathbf{I}\right)^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{y} \\
&= \mathbf{V}\left(\mathbf{S}^\mathsf{T}\mathbf{S} + \eta \mathbf{I}\right)^{-1}\mathbf{S}^\mathsf{T}\mathbf{U}^\mathsf{T}\boldsymbol{y}
\end{aligned}
$$

- where

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

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 $\hat{\mathbf{S}}^+ = (\mathbf{S}^\mathsf{T}\mathbf{S} + \eta)^{-1}\mathbf{S}^\mathsf{T}$ instead of $\mathbf{S}^+$ then



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

Summary

- Any matrix can be decomposed as $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\mathsf{T}$ where
  - $\mathbf{U}$ and $\mathbf{V}$ are orthogonal (rotation matrices)
  - $\mathbf{S} = \mathrm{diag}(s_1, \ldots, s_n)$ is a diagonal matrix of positive singular values
- This describes the most general linear transform
- The transform exploits the duality between $\mathbf{X}\mathbf{X}^\mathsf{T}$ and $\mathbf{X}^\mathsf{T}\mathbf{X}$
- 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