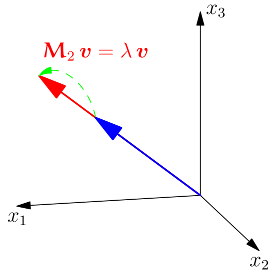
Advanced Machine Learning

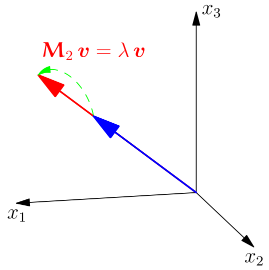
Eigensystems



Eigenvectors, Orthogonal Matrices, Eigenvector Decomposition, Rank

Outline

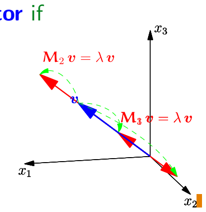
1. **Eigenvectors**
2. Orthogonal Matrices
3. Eigen Decomposition
4. Low Rank Approximation



Eigenvector equation

- Eigen-systems help us to understand mappings
- A vector $\boldsymbol{v}$ is said to be an **eigenvector** if

$$\mathbf{M}\boldsymbol{v} = \lambda\boldsymbol{v}$$



- $\mathbf{M}$ is square (i.e. $n \times n$)
- Where the number λ is the **eigenvalue**
- Eigenvalues play a fundamental role in understanding operators

Symmetric Matrices

- If $\mathbf{M}$ is an $n \times n$ **symmetric** matrix then it has n real orthogonal eigenvectors with real eigenvalues
- We denote the i^{th} eigenvector by $\boldsymbol{v}_i$ and the corresponding eigenvalue by λ_i so that

$$\mathbf{M}\boldsymbol{v}_i = \lambda_i\boldsymbol{v}_i$$

- Orthogonal means that if $i \neq j$ then

$$\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{v}_j = 0$$

- (We can always normalise eigenvectors if we want)

Proof of Orthogonality

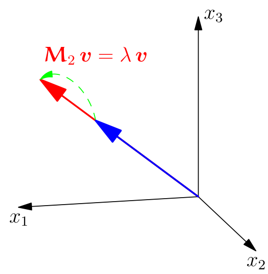
- $\left(\mathbf{M}\boldsymbol{v}_i = \lambda_i\boldsymbol{v}_i\right)^{\mathsf{T}}$ implies $\boldsymbol{v}_i^{\mathsf{T}}\mathbf{M}^{\mathsf{T}} = \lambda_i\boldsymbol{v}_i^{\mathsf{T}}$
- When $\mathbf{M}$ is symmetric then $\mathbf{M}\boldsymbol{v}_i = \lambda_i\boldsymbol{v}_i \Rightarrow \boldsymbol{v}_i^{\mathsf{T}}\mathbf{M} = \lambda_i\boldsymbol{v}_i^{\mathsf{T}}$
- Consider two eigenvectors $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ of $\mathbf{M}$

$$\begin{aligned}\boldsymbol{v}_i^{\mathsf{T}}\mathbf{M}\boldsymbol{v}_j &= (\boldsymbol{v}_i^{\mathsf{T}}\mathbf{M})\boldsymbol{v}_j = \lambda_i\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{v}_j \\ &= \boldsymbol{v}_i^{\mathsf{T}}(\mathbf{M}\boldsymbol{v}_j) = \lambda_j\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{v}_j\end{aligned}$$

- So either $\lambda_i = \lambda_j$ or $\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{v}_j = 0$
- If $\lambda_i = \lambda_j$ then any linear combination of $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ is an eigenvector $(\mathbf{M}(a\boldsymbol{v}_i + b\boldsymbol{v}_j) = \lambda_i(a\boldsymbol{v}_i + b\boldsymbol{v}_j))$ So I can choose two eigenvectors that are orthogonal to each other.

Outline

1. Eigenvectors
2. **Orthogonal Matrices**
3. Eigen Decomposition
4. Low Rank Approximation



Orthogonal Matrices

- We can construct an **orthogonal** matrix $\mathbf{V}$ from the eigenvectors

$$\mathbf{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n)$$

- Matrix $\mathbf{V}$ is an $n \times n$ matrix
- Because of the orthogonality of the vectors $\boldsymbol{v}_i$

$$\mathbf{V}^{\mathsf{T}}\mathbf{V} = \begin{pmatrix} \boldsymbol{v}_1^{\mathsf{T}}\boldsymbol{v}_1 & \boldsymbol{v}_1^{\mathsf{T}}\boldsymbol{v}_2 & \cdots & \boldsymbol{v}_1^{\mathsf{T}}\boldsymbol{v}_n \\ \boldsymbol{v}_2^{\mathsf{T}}\boldsymbol{v}_1 & \boldsymbol{v}_2^{\mathsf{T}}\boldsymbol{v}_2 & \cdots & \boldsymbol{v}_2^{\mathsf{T}}\boldsymbol{v}_n \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{v}_n^{\mathsf{T}}\boldsymbol{v}_1 & \boldsymbol{v}_n^{\mathsf{T}}\boldsymbol{v}_2 & \cdots & \boldsymbol{v}_n^{\mathsf{T}}\boldsymbol{v}_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = \mathbf{I}$$

The Other Way Around

- We have shown that $\mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I}$
- Thus multiply both sides on the left by $\mathbf{V}$

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

- $\mathbf{V}$ will have an inverse, $\mathbf{V}^{-1}$, such that $\mathbf{V}\mathbf{V}^{-1} = \mathbf{I}$
- Multiplying the equation on the right by $\mathbf{V}^{-1}$

$$\begin{aligned}(\mathbf{V}\mathbf{V}^{\mathsf{T}})\mathbf{V}\mathbf{V}^{-1} &= \mathbf{V}\mathbf{V}^{-1} \\ \mathbf{V}\mathbf{V}^{\mathsf{T}} &= \mathbf{I}\end{aligned}$$

- Note that, $\mathbf{V}^{-1} = \mathbf{V}^{\mathsf{T}}$ (definition of orthogonal matrix)

Invertible Matrices

- A matrix, $\mathbf{M}$, will be singular (uninvertible) if there exists a vector $\boldsymbol{x}$ ($\neq \mathbf{0}$) such that

$$\mathbf{M}\boldsymbol{x} = \mathbf{0}$$

- Now if there exists such a vector such that $\mathbf{V}\boldsymbol{x} = \mathbf{0}$ then multiply by $\mathbf{V}^{\mathsf{T}}$ we get

$$\mathbf{V}^{\mathsf{T}}\mathbf{V}\boldsymbol{x} = \mathbf{V}^{\mathsf{T}}\mathbf{0}$$
$$\boldsymbol{x} = \mathbf{0}$$

since $\mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I}$

- Thus $\boldsymbol{V}$ is invertible

Rotations

- Orthogonal matrices satisfy $\mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{V}\mathbf{V}^{\mathsf{T}} = \mathbf{I}$
- As a consequent they define rotations (and possibly a reflection)
- Consider a vector $\boldsymbol{x}$ and $\boldsymbol{x}' = \mathbf{V}\boldsymbol{x}$, now

$$\|\boldsymbol{x}'\|_2^2 = \boldsymbol{x}'^{\mathsf{T}}\boldsymbol{x}' = (\mathbf{V}\boldsymbol{x})^{\mathsf{T}}(\mathbf{V}\boldsymbol{x}) = \boldsymbol{x}^{\mathsf{T}}\mathbf{V}^{\mathsf{T}}\mathbf{V}\boldsymbol{x} = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} = \|\boldsymbol{x}\|_2^2$$

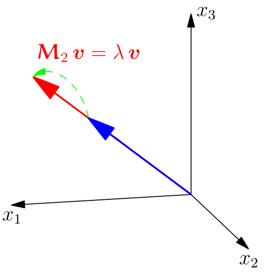
- Similarly if additionally $\boldsymbol{y}' = \mathbf{V}\boldsymbol{y}$ then

$$\langle \boldsymbol{x}', \boldsymbol{y}' \rangle = (\mathbf{V}\boldsymbol{x})^{\mathsf{T}}(\mathbf{V}\boldsymbol{y}) = \boldsymbol{x}^{\mathsf{T}}\mathbf{V}^{\mathsf{T}}\mathbf{V}\boldsymbol{y} = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{y} = \langle \boldsymbol{x}, \boldsymbol{y} \rangle = \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2 \cos(\theta)$$

- Rotations and reflections preserve lengths and angles

Outline

1. Eigenvectors
2. Orthogonal Matrices
3. **Eigen Decomposition**
4. Low Rank Approximation



Matrix Decomposition

- Taking the matrix of eigenvectors, $\boldsymbol{V}$, then

$$\mathbf{M}\mathbf{V} = \mathbf{M}(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n) = (\lambda_1 \boldsymbol{v}_1, \lambda_2 \boldsymbol{v}_2, \ldots, \lambda_n \boldsymbol{v}_n) = \mathbf{V}\boldsymbol{\Lambda}$$

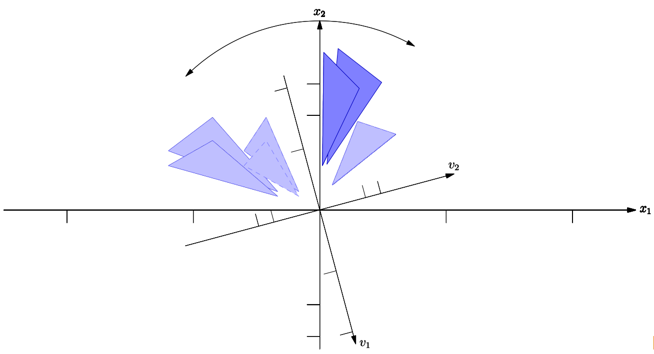
- where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$
- Now

$$\mathbf{M} = \mathbf{M}\mathbf{V}\mathbf{V}^{\mathsf{T}} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathsf{T}}$$

- Very important *similarity transform*

Mappings by Symmetric Matrices

$$\mathbf{M} = \begin{pmatrix} 0.83 & -0.31 \\ -0.31 & 1.9 \end{pmatrix} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathsf{T}} = \begin{pmatrix} \cos(-75) & \sin(-75) \\ -\sin(-75) & \cos(-75) \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 0.75 \end{pmatrix} \begin{pmatrix} \cos(75) & \sin(75) \\ -\sin(75) & \cos(75) \end{pmatrix}$$



Inverses

- For any square matrix

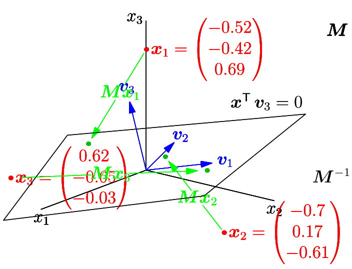
$$\mathbf{M} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathsf{T}} \qquad \mathbf{M}^{-1} = \mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^{\mathsf{T}}$$

- Where $\boldsymbol{\Lambda}^{-1} = \mathrm{diag}(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \ldots, \frac{1}{\lambda_n}) = \begin{pmatrix} \frac{1}{\lambda_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\lambda_n} \end{pmatrix}$
- Since

$$\mathbf{M}\mathbf{M}^{-1} = (\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathsf{T}})(\mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^{\mathsf{T}}) = \mathbf{V}\boldsymbol{\Lambda}(\mathbf{V}^{\mathsf{T}}\mathbf{V})\boldsymbol{\Lambda}^{-1}\mathbf{V}^{\mathsf{T}}$$
$$= \mathbf{V}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1}\mathbf{V}^{\mathsf{T}} = \mathbf{V}\mathbf{V}^{\mathsf{T}} = \mathbf{I}$$

- I.e, Small eigenvalues become large eigenvalues and visa verse

Ill-Conditioning Again



$$\boldsymbol{M} = \begin{pmatrix} 0.80 & 0.01 & 0.25 \\ 0.01 & 0.85 & 0.48 \\ 0.25 & 0.48 & 0.5 \end{pmatrix}$$
$$= \boldsymbol{V} \begin{pmatrix} 0.8 & 0 & 0 \\ 0 & 1.25 & 0 \\ 0 & 0 & 0.1 \end{pmatrix} \boldsymbol{V}^{\mathsf{T}}$$
$$\boldsymbol{M}^{-1} = \begin{pmatrix} 1.9 & 1.17 & -2.1 \\ 1.17 & 3.34 & -3.8 \\ -2.1 & -3.8 & 6.8 \end{pmatrix}$$
$$= \boldsymbol{V} \begin{pmatrix} 1.25 & 0 & 0 \\ 0 & 0.8 & 0 \\ 0 & 0 & 10 \end{pmatrix} \boldsymbol{V}^{\mathsf{T}}$$

Condition Number

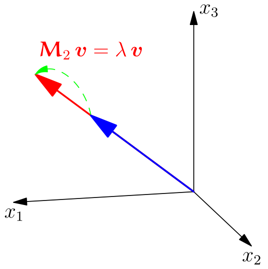
- Taking matrix inverses can be inherently unstable
- Any small error can be amplified by taking the inverse
- The stability of the inverse depends on the ratio of smallest eigenvalue to the largest eigenvalue (i.e. the biggest possible amplification compared to the smallest)
- Note that the Hilbert-norm of a matrix is the absolute value of the largest eigenvalue
- The condition number is given by

$$\|\mathbf{M}\|_H \times \|\mathbf{M}^{-1}\|_H = \frac{|\lambda_{\mathsf{max}}|}{|\lambda_{\mathsf{min}}|}$$

- Large condition number implies very ill-conditioned

Outline

1. Eigenvectors
2. Orthogonal Matrices
3. Eigen Decomposition
4. **Low Rank Approximation**



Rank of a Matrix

- The rank of a matrix, $\mathbf{M}$, is the number of non-zero eigenvalues
- The space spanned by the eigenvectors $\boldsymbol{v}_a$, $\boldsymbol{v}_b$, etc. with zero eigenvalue forms a **null space**
- Any vector in the null space will get projected to the zero vector

$$\mathbf{M}(a\boldsymbol{v}_a + b\boldsymbol{v}_b + \cdots) = \mathbf{0}$$

- A square matrix is said to be **rank deficient** if it has any eigenvectors with eigenvalue equal to 0
- This happens when the columns of the matrix are not linearly independent

"Inverting" Rank Deficient Matrices

- Rank deficient matrices are non-invertible (i.e. we don't know the vector $\boldsymbol{x}$ such that $\mathbf{M}\boldsymbol{x} = \boldsymbol{b}$) as we don't know the component of the $\boldsymbol{x}$ in the null space
- Although we don't know $\boldsymbol{x}$ we can find a vector, $\boldsymbol{x}$, that satisfies $\mathbf{M}\boldsymbol{x} = \boldsymbol{b}$
- Given a symmetric $n \times n$ matrix with k non-zero eigenvalues λ_1, λ_2, …, λ_k we can construct a "pseudo inverse" $\mathbf{M}^+$ as $\mathbf{V}\boldsymbol{\Lambda}^+\mathbf{V}^\mathsf{T}$ where $\boldsymbol{\Lambda}^+ = \mathrm{diag}(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \ldots, \frac{1}{\lambda_k}, 0, \ldots, 0)$
- This finds the vector $\boldsymbol{x}$ with no component in the null space (it is the solution with the smallest norm)
- This is a different to the pseudo inverse for non-square matrices

Low Rank Approximation

- Recall that matrices with large and small eigenvectors are ill-conditions so the inverse has the potential to greatly amplify any measurement error
- One work around is to set all small eigenvalues to zero and use the pseudo inverse
- Setting small eigenvalues to zero reduces the rank of the matrix and is an example of a low rank approximation
- Low rank approximations are much used to obtain approximate models for arrays of data (we will revisit this when we look at SVD)

Summary

- Linear mappings are commonly used in machine learning algorithms such as regression
- We can understand symmetric operators by looking at their eigenvectors
- Any symmetric matrix can be decomposed as $\mathbf{M} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\mathsf{T}$
  - ⋆ where $\mathbf{V}$ are orthogonal matrices whose rows are the eigenvector
  - ⋆ and $\boldsymbol{\Lambda}$ is a diagonal matrix of the eigenvalues
- This decomposition allows us to understand inverse mappings