Module 9

Bias-Variance Decomposition / Optimization / Linear Algebra

Kentaro Nakamura

GOV 2003

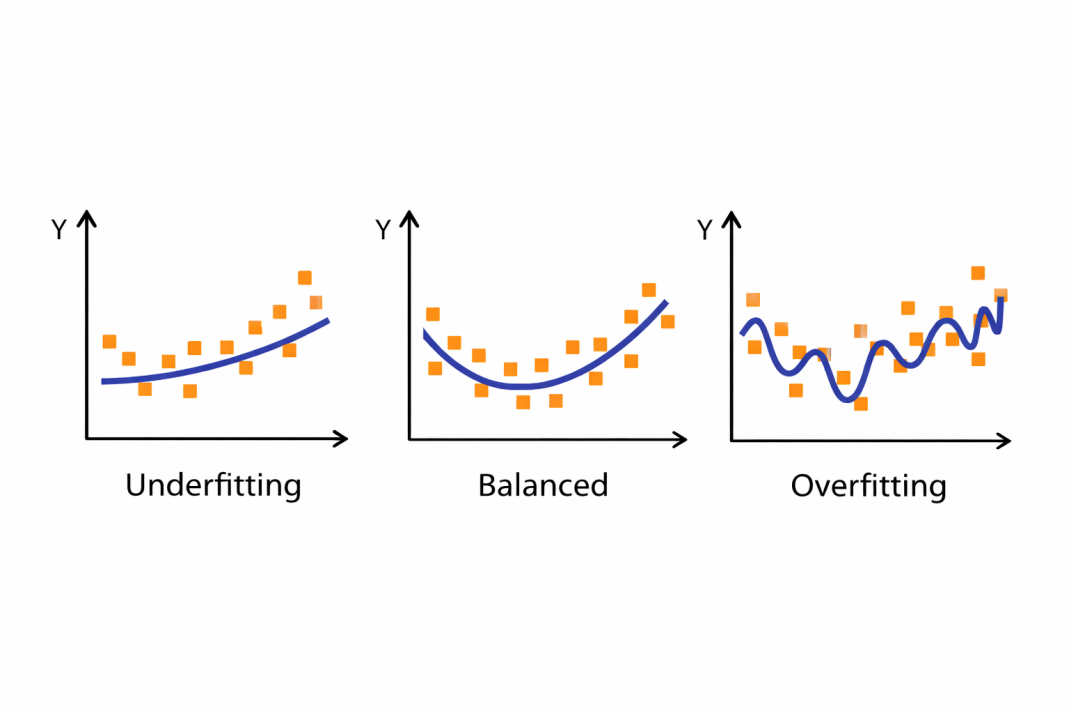
April 10th, 2026

Agenda

- Bias-Variance Decomposition
- Optimization under Constraints
  - Lagrangian multiplier
  - KKT condition
- Linear algebra
  - Eigenvalue and eigenvector
  - Eigenvalue decomposition (EVD)
  - Singular value decomposition (SVD)

Motivation

- We want to make the function flexible enough to minimize the bias
- However, we do not want the overfitting



- As figure illustrates, overfitting is caused by high variance
  - There is a trade-off: underfitting happens because of restrictive model, whereas overfitting happens because of complex model

Bias Variance Decomposition (1)

- Suppose that the true model is

$$Y_i = f(\boldsymbol{X}_i) + \epsilon_i \quad \mathbb{E}[\epsilon_i \mid \boldsymbol{X}_i] = 0$$

- Then, mean squared error is

$$\begin{aligned}
&\mathbb{E}[(Y_i - \hat{f}(\boldsymbol{X}_i))^2] \\
&= \mathbb{E}[(Y_i - f(\boldsymbol{X}_i))^2] + \mathbb{E}[(f(\boldsymbol{X}_i) - \hat{f}(\boldsymbol{X}_i))^2] \\
&\qquad + 2 \cdot \mathbb{E}[(Y_i - f(\boldsymbol{X}_i))(f(\boldsymbol{X}_i) - \hat{f}(\boldsymbol{X}_i))]
\end{aligned}$$

- Now, the last cross term is zero because

$$\begin{aligned}
&\mathbb{E}[(Y_i - f(\boldsymbol{X}_i))(f(\boldsymbol{X}_i) - \hat{f}(\boldsymbol{X}_i))] \\
&= \mathbb{E}\big[\underbrace{\mathbb{E}\{Y_i - f(\boldsymbol{X}_i) \mid \boldsymbol{X}_i\}}_{=0}(f(\boldsymbol{X}_i) - \hat{f}(\boldsymbol{X}_i))\big] = 0
\end{aligned}$$

- Finally, we have

$$\underbrace{\mathbb{E}[(Y_i - \hat{f}(\boldsymbol{X}_i))^2]}_{\text{Mean Squaed Err}} = \underbrace{\mathbb{E}[(Y_i - f(\boldsymbol{X}_i))^2]}_{\text{Irreducible Error}} + \underbrace{\mathbb{E}[(f(\boldsymbol{X}_i) - \hat{f}(\boldsymbol{X}_i))^2]}_{\text{Modeling Error}}$$

Bias Variance Decomposition (2)

- Now, suppose that we use $\mathcal{D} = \{\boldsymbol{X}_i, Y_i\}_{i=1}^n$ to train $\hat{f}(\cdot; \mathcal{D})$ and we validate it at new observation $\{x_{\text{new}}, y_{\text{new}}\}$.
  - Notice that $\hat{f}$ depends on training data $\mathcal{D}$

- Now, using the result from the previous page,

$$\begin{aligned}\text{MSE}(x_{\text{new}}, y_{\text{new}}) &= \mathbb{E}[(y_{\text{new}} - \hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D}))^2] \\ &= \mathbb{E}[(y_{\text{new}} - f(\boldsymbol{x}_{\text{new}}))^2] + \mathbb{E}[(f(\boldsymbol{x}_{\text{new}}) - \hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D}))^2]\end{aligned}$$

Bias Variance Decomposition (3)

- And the model error is written as

$$\mathbb{E}[(f(\boldsymbol{x}_{\text{new}}) - \hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D}))^2]$$

$$= \mathbb{E}\left[\left(f(\boldsymbol{x}_{\text{new}}) - \mathbb{E}[\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})] + \mathbb{E}[\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})] - \hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})\right)^2\right]$$

$$= \mathbb{E}\left[\left(f(\boldsymbol{x}_{\text{new}}) - \mathbb{E}[\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})]\right)^2\right]$$

$$+ \mathbb{E}\left[\left(\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D}) - \mathbb{E}[\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})]\right)^2\right]$$

$$+ 2\mathbb{E}\left[\left(f(\boldsymbol{x}_{\text{new}}) - \mathbb{E}[\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})]\right)\left(\mathbb{E}[\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})] - \hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})\right)\right]$$

Bias Variance Decomposition (4)

- Now, the first term is

$$\mathbb{E}\left[\left(f(\boldsymbol{x}_{\text{new}}) - \mathbb{E}[\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})]\right)^2\right]$$

$$= \mathbb{E}\left[f(\boldsymbol{x}_{\text{new}})^2\right] - 2\mathbb{E}\left[f(\boldsymbol{x}_{\text{new}})\mathbb{E}[\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})]\right] + \mathbb{E}\left[\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})^2\right]$$

$$= \left(f(x) - \mathbb{E}[\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})]\right)^2 = \text{Bias}(\boldsymbol{x}_{\text{new}})^2$$

- On the other hand, the last term is

$$\mathbb{E}\left[\left(f(\boldsymbol{x}_{\text{new}}) - \mathbb{E}[\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})]\right)\left(\mathbb{E}[\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})] - \hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})\right)\right]$$

$$= \mathbb{E}\Big[f(\boldsymbol{x}_{\text{new}})\mathbb{E}[\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})] - f(\boldsymbol{x}_{\text{new}})\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})$$

$$- \mathbb{E}[\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})]^2 + \mathbb{E}[\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})]\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})\Big] = 0$$

  - Notice that there is outer expectation. As a result, the first two terms and the last two terms become 0.

Bias Variance Decomposition (5)

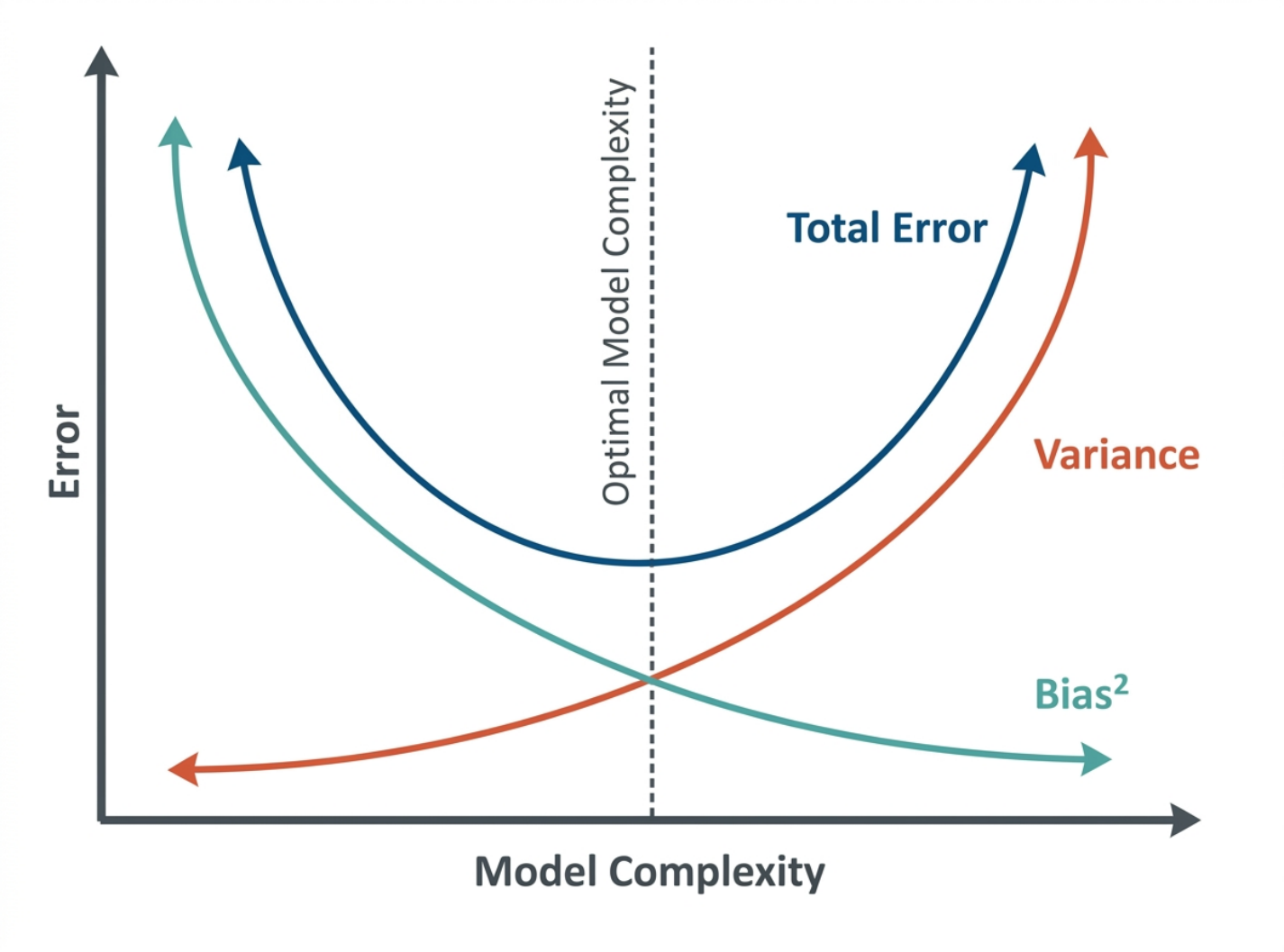
- As a result, we finally get

$$\begin{aligned}
\mathrm{MSE}(x_{\mathrm{new}}, y_{\mathrm{new}}) &:= \mathbb{E}[(y_{\mathrm{new}} - \hat{f}(\boldsymbol{x}_{\mathrm{new}}; \mathcal{D}))^2] \\
&= \underbrace{\mathbb{E}[(y_{\mathrm{new}} - f(\boldsymbol{x}_{\mathrm{new}}))^2]}_{\text{Irreduciable Error}} \\
&\quad + \underbrace{\left(f(\boldsymbol{x}_{\mathrm{new}}) - \mathbb{E}[\hat{f}(\boldsymbol{x}_{\mathrm{new}}; \mathcal{D})] \right)^2}_{\text{Bias}(x_{\mathrm{new}})^2} \\
&\quad + \underbrace{\mathbb{E}\left[\left(\hat{f}(\boldsymbol{x}_{\mathrm{new}}; \mathcal{D}) - \mathbb{E}[\hat{f}(\boldsymbol{x}_{\mathrm{new}}; \mathcal{D})] \right)^2 \right]}_{\text{Variance}(x_{\mathrm{new}})}
\end{aligned}$$

- We can only control model error as $f(\boldsymbol{x}_{\mathrm{new}})$ is the truth we cannot change

Bias Variance Decomposition (6)

- **Implication**: To minimize the out-of-sample prediction error, we want to minimize both bias and variance
  - Only minimizing bias might not be optimal for variane



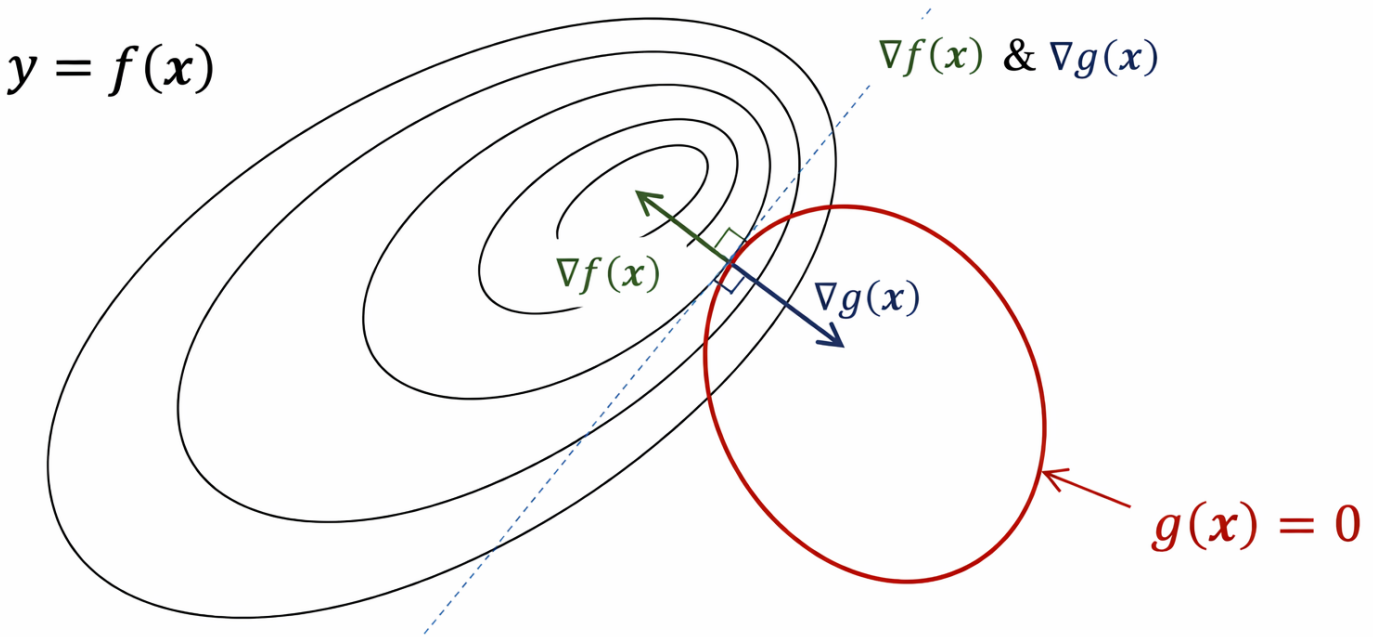
Optimization Under Equality Constraints

- **Setup**: Consider the following optimization problem

$$\arg\max_x \quad f(x)$$

$$\text{such that} \quad g(x) = 0$$

- **Intuition**: $\nabla f(x)$ should be parallel to $\nabla g(x)$ at the maximum



$$\nabla f(x) = -\lambda \nabla g(x)$$

Method of Lagrange Multiplier

Theorem (Lagrange Multiplier)

For the problem in the previous page, consider the Lagrangian

$$\mathcal{L}(x, \lambda) = f(x) + \lambda^\top g(x).$$

Suppose x^* is a local maximizer subject to the constraint $g(x^*) = 0$, and the gradient of the constraint is nonzero at x^*. Then, there exists a vector of multipliers λ^* such that

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = \nabla f(x^*) + \nabla g(x^*)^\top \lambda^* = 0 \quad \textit{and} \quad g(x^*) = 0.$$

- Therefore, instead of solving a constrained problem directly, we solve

$$\frac{\partial \mathcal{L}(x, \lambda)}{\partial x} = 0 \qquad \text{and} \qquad \frac{\partial \mathcal{L}(x, \lambda)}{\partial \lambda} = g(x) = 0.$$

- Lagrangian multiplier transform the original problem (**primal problem**) into easy-to-solve problem (**dual problem**)

Example: Method of Lagrange multiplier

Example

$$\max\ x^3 + y^3 \qquad \text{such that} \quad x^2 + y^2 = 1$$

- Lagrangian is $\mathcal{L}(x, y, \lambda) = x^3 + y^3 + \lambda(x^2 + y^2 - 1)$
- First-order conditions:

$$\frac{\partial \mathcal{L}}{\partial x} = 3x^2 + 2\lambda x = 0, \quad \frac{\partial \mathcal{L}}{\partial y} = 3y^2 + 2\lambda y = 0, \quad \frac{\partial \mathcal{L}}{\partial \lambda} = x^2 + y^2 - 1 = 0$$

- From the first two equations,

$$x(3x + 2\lambda) = 0, \qquad y(3y + 2\lambda) = 0$$

so either $x = 0$ or $3x + 2\lambda = 0$, and either $y = 0$ or $3y + 2\lambda = 0$, which gives candidate solutions

$$(1, 0),\ (-1, 0),\ (0, 1),\ (0, -1),\ \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right),\ \left(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$$

Optimization under Inequality Constraints (1)

- Consider the optimization problem

$$\min \quad f(x) \qquad \text{such that} \quad g(x) \leq 0$$

- **Case 1**: The optimal of $f(x)$ satisfies $g(x) < 0$
  - In this case, we only need $\nabla f(x) = 0$
  - Thus, $\lambda^* = 0$
- **Case 2**: The optimal of $f(x)$ satisfies $g(x) = 0$
  - This is the case of equality constraint
  - Thus, we need $-\nabla f(x^*) = \lambda^* \nabla g(x^*)$ with $\lambda^* > 0$

Optimization under Inequality Constraints (2)

- These give the necessary conditions for the optimal solution (x^*, λ^*)
  - These conditions are called **KKT condition**

Definition (KKT condition)

The following conditions are satisfied for the optimal solution (x^*, λ^*): (i) $-\nabla f(x^*) = \lambda^* \nabla g(x^*)$ (ii) $\lambda^* g(x^*) = 0$ (known as **complementary slackness**) (iii) $\lambda^* > 0$ (iv) $g(x^*) \leq 0$

- Check with previous page's two cases and see where these conditions come from.

- **Note**: This does not automatically give the necessary and sufficient conditions
  - You need to use the condition like Slater condition to argue that the optimal obtained from Lagrangian is sufficient.

Extra: Slater Condition (Constraint Qualification)

- When do we have necessary and sufficient answers?
  - KKT conditions are necessary conditions of answers, but not sufficient
  - One useful condition is **Slater condiiton**

Definition (Slater Condition)

Consider a **convex** optimization problem (i.e., $f(x)$ convex, $g(x)$ convex). Slater condition holds if there exists a strictly feasible point x such that $g(x) < 0$

- **Intuition**: Remove the edge cases

Theorem

If the problem is convex and Slater condition holds, then KKT conditions are necessary and sufficient for the optimal answer.

Eigenvalue problem

Definition (Eigenvalue & Eigenvector)

Let $A \in \mathbb{R}^{n \times n}$ be a square matrix. A scalar $\lambda \in \mathbb{R}$ and a nonzero vector $v \in \mathbb{R}^n$ satisfying

$$Av = \lambda v$$

are called an **eigenvalue** and its corresponding **eigenvector** of A, respectively.

- Rearranging: $(A - \lambda I)v = 0$ has a nonzero solution iff

$$\det(A - \lambda I) = 0.$$

  This is the **characteristic equation** of A.
  - Recall that matrix is invertible if determinant is non-zero. And if $A - \lambda I$ is invertible, then $v = (A - \lambda I)^{-1}0 = 0$ and thus it only admits $v = 0$ (trivial solution)
- The set of all eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ is called the **spectrum** of A.

Example: Eigenvalue and Eigenvector

Example

Find the eigenvalues and eigenvectors of $A = \begin{pmatrix} 3 & 1 \\ 0 & 2 \end{pmatrix}$.

- Solve **characteristic equation**

$$\det(A - \lambda I) = (3 - \lambda)(2 - \lambda) = 0 \implies \lambda_1 = 3, \quad \lambda_2 = 2.$$

- For $\lambda_1 = 3$: $(A - 3I)v = 0 \Rightarrow \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} v = 0 \Rightarrow v_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.
- For $\lambda_2 = 2$: $(A - 2I)v = 0 \Rightarrow \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} v = 0 \Rightarrow v_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$.

Eigenvalue Decomposition (EVD)

- From the eigenvalue problem, we can derive the **eigenvalue decomposition**
- We have n eigenvalue equations $Av_k = \lambda_k v_k$. Stacking them column-wise:

$$A \underbrace{[v_1 \mid \cdots \mid v_n]}_{Q} = \underbrace{[v_1 \mid \cdots \mid v_n]}_{Q} \underbrace{\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}}_{\Lambda} \quad \Longrightarrow \quad AQ = Q\Lambda$$

- If Q is invertible (i.e., eigenvectors are linearly independent),

$$A = Q\Lambda Q^{-1}$$

- EVD requires Q to be invertible, i.e., the n eigenvectors must be *linearly independent*.
  - But it can fail, so EVD might not exist for some square matrix

Spectoral Theorem

- In the case of square matrix, EVD always exists

Theorem (Spectral Theorem)

*Let $A \in \mathbb{R}^{n \times n}$ be a **symmetric** matrix. Then A can be decomposed as*

$$\boxed{A = Q \Lambda Q^\top}$$

where

- $Q = [v_1 \mid v_2 \mid \cdots \mid v_n] \in \mathbb{R}^{n \times n}$ *is **orthogonal** ($Q^\top Q = I$), with columns being the eigenvectors of A, and*
- $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ *contains the corresponding real **eigenvalues**.*

- Every symmetric matrix is diagonalizable with an orthonormal eigenbasis.
- $Q^\top = Q^{-1}$, so the decomposition can also be written $Q^\top A Q = \Lambda$ (*diagonalization*).

Properties of Eigenvalues

- Let $A \in \mathbb{R}^{n \times n}$ have eigenvalues $\lambda_1, \ldots, \lambda_n$
- **Trace**:

$$\text{tr}(A) = \text{tr}(Q\Lambda Q^\top) = \text{tr}(\Lambda Q^\top Q) = \text{tr}(\Lambda) = \sum_{i=1}^{n} \lambda_i$$

- **Determinant**

$$\det(A) = \det(Q\Lambda Q^\top) = \det(Q)\det(\Lambda)\det(Q^\top) = \det(\Lambda) = \prod_{i=1}^{n} \lambda_i$$

  where we used $\det(Q)\det(Q^\top) = \det(QQ^\top) = \det(I) = 1$.
  - Recall that non-zero determinant $\leftrightarrow$ invertible $\leftrightarrow$ full-rank

- **Positive definiteness**: For any nonzero $x \in \mathbb{R}^n$, let $z = Q^\top x \neq 0$. Then

$$x^\top Ax = x^\top Q\Lambda Q^\top x = z^\top \Lambda z = \sum_{i=1}^{n} \lambda_i z_i^2.$$

  Since $z_i^2 > 0$, we have $x^\top Ax > 0 \; \forall x \iff \lambda_i > 0$ **for all** i.

Singular Value Decomposition (SVD)

- EVD is applied only to square matrix, so we want to generalize it.

Theorem (SVD)

Let $A \in \mathbb{R}^{m \times n}$ be any matrix. Then A can be decomposed as

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

where

- $U = [u_1 \mid \cdots \mid u_m] \in \mathbb{R}^{m \times m}$ *is orthogonal, with columns called* ***left singular vectors***,
- $V = [v_1 \mid \cdots \mid v_n] \in \mathbb{R}^{n \times n}$ *is orthogonal, with columns called* ***right singular vectors***, *and*
- $\Sigma \in \mathbb{R}^{m \times n}$ *is diagonal with non-negative entries* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 = \cdots = 0$, *called* ***singular values***, *where* $r = rank(A)$.

SVD: Connection to EVD

- **SVD reduces to EVD for symmetric PSD matrices**
  - When A is symmetric with EVD $A = Q\Lambda Q^\top$, setting $U = V = Q$ and $\Sigma = \Lambda$ gives
$$A = Q\Lambda Q^\top = U\Sigma V^\top.$$
  - So EVD is a **special case** of SVD.

- **Singular values are eigenvalues of $A^\top A$ and $AA^\top$**
  - Substituting $A = U\Sigma V^\top$:
$$A^\top A = V(\Sigma^\top \Sigma)V^\top, \qquad AA^\top = U(\Sigma\Sigma^\top)U^\top.$$
  - So V, U are eigenvectors of $A^\top A$, $AA^\top$ respectively, and
$$\sigma_i = \sqrt{\lambda_i(A^\top A)} = \sqrt{\lambda_i(AA^\top)} \geq 0.$$
  - This guarantees singular values are always *real and non-negative*, even when A is not symmetric.