### 1. Which of the following types of gradient descent uses the entire dataset to compute the gradient and update the parameters?

a) Stochastic Gradient Descent (SGD)  
b) Mini-Batch Gradient Descent  
c) Batch Gradient Descent  
d) None of the above

**Answer**: c) **Batch Gradient Descent**

**Explanation**: Batch Gradient Descent uses the entire dataset to calculate the gradient of the cost function and updates the parameters after processing all samples.

### 2. In which type of gradient descent is the model updated after processing each individual training sample?

a) Mini-Batch Gradient Descent  
b) Stochastic Gradient Descent (SGD)  
c) Batch Gradient Descent  
d) Gradient Descent with Momentum

**Answer**: b) **Stochastic Gradient Descent (SGD)**

**Explanation**: In Stochastic Gradient Descent, the model parameters are updated after each individual training example.

### 3. Which gradient descent method is typically the most efficient in terms of memory usage when dealing with large datasets?

a) Batch Gradient Descent  
b) Stochastic Gradient Descent (SGD)  
c) Mini-Batch Gradient Descent  
d) None of the above

**Answer**: b) **Stochastic Gradient Descent (SGD)**

**Explanation**: SGD is memory-efficient as it processes one sample at a time and does not require storing the entire dataset at once.

### 4. Which of the following is a key advantage of ****Mini-Batch Gradient Descent**** over ****Batch Gradient Descent****?

a) Converges to a global minimum faster  
b) More stable updates  
c) More computationally expensive  
d) Can be parallelized and optimized using GPUs

**Answer**: d) **Can be parallelized and optimized using GPUs**

**Explanation**: Mini-Batch Gradient Descent allows for faster computation and can be efficiently parallelized, especially on GPUs, improving performance with large datasets.

### 5. What is one of the main disadvantages of ****Stochastic Gradient Descent****?

a) It is computationally expensive  
b) It can cause noisy updates and oscillations  
c) It always converges to the optimal solution  
d) It requires large amounts of memory

**Answer**: b) **It can cause noisy updates and oscillations**

**Explanation**: SGD updates the parameters after each sample, which can lead to noisy gradients and oscillations, making it less stable than Batch Gradient Descent.

### 6. What is typically the effect of a ****high learning rate**** on Gradient Descent algorithms?

a) Faster convergence  
b) Slow convergence  
c) Risk of overshooting the minimum  
d) No effect

**Answer**: c) **Risk of overshooting the minimum**

**Explanation**: A high learning rate can cause the updates to overshoot the optimal point, resulting in poor convergence or divergence of the algorithm.

### 7. In which case would ****Batch Gradient Descent**** be preferred over other methods?

a) When the dataset is small  
b) When computational resources are limited  
c) When parallelization is required  
d) When the dataset is large

**Answer**: a) **When the dataset is small**

**Explanation**: Batch Gradient Descent works well when the dataset is small, as it processes the entire dataset at once and ensures stable convergence.

### 8. Which of the following is a characteristic of ****Mini-Batch Gradient Descent**** compared to ****Stochastic Gradient Descent****?

a) It uses only one data point at a time  
b) It updates parameters after processing the entire dataset  
c) It uses a small subset of the dataset for each update  
d) It is slower than Batch Gradient Descent

**Answer**: c) **It uses a small subset of the dataset for each update**

**Explanation**: Mini-Batch Gradient Descent divides the dataset into small batches, and updates parameters after processing each mini-batch, combining the benefits of both batch and stochastic methods.

### 9. ****Gradient Descent with Momentum**** is primarily used to:

a) Avoid local minima  
b) Increase the learning rate  
c) Reduce oscillations and speed up convergence  
d) Optimize memory usage

**Answer**: c) **Reduce oscillations and speed up convergence**

**Explanation**: Momentum helps smooth out the updates by accumulating previous gradients, which reduces oscillations and accelerates convergence.

### 10. What is the typical effect of ****Mini-Batch Size**** on the performance of ****Mini-Batch Gradient Descent****?

a) Smaller mini-batches tend to reduce computational efficiency  
b) Larger mini-batches lead to faster convergence but more memory usage  
c) Mini-batch size does not affect convergence speed  
d) Smaller mini-batches reduce the learning rate

**Answer**: b) **Larger mini-batches lead to faster convergence but more memory usage**

**Explanation**: Larger mini-batches provide more stable updates, which can lead to faster convergence, but they require more memory to process.

#### ****1. What does Singular Value Decomposition (SVD) decompose a given matrix into?****

A) Three diagonal matrices  
B) A product of three matrices: an orthogonal, a diagonal, and another orthogonal matrix  
C) A product of two orthogonal matrices  
D) A product of three triangular matrices

✅ **Answer:** B) A product of three matrices: an orthogonal, a diagonal, and another orthogonal matrix

📖 **Explanation:** SVD decomposes a matrix AA into A=UΣVTA = U \Sigma V^T, where UU and VV are orthogonal (or unitary for complex matrices) and Σ\Sigma is a diagonal matrix.

#### ****2. If**** AA ****is an**** m×nm \times n ****matrix, what is the size of the diagonal matrix**** Σ\Sigma ****in SVD?****

A) m×nm \times n  
B) n×mn \times m  
C) m×mm \times m  
D) m×nm \times n or n×mn \times m, depending on rank

✅ **Answer:** A) m×nm \times n

📖 **Explanation:** The diagonal matrix Σ\Sigma in SVD has the same dimensions as AA and contains singular values along its diagonal.

#### ****3. Which of the following is NOT true about the singular values of a matrix?****

A) They are always non-negative  
B) They are the square roots of the eigenvalues of ATAA^T A  
C) They can be complex numbers  
D) They are arranged in decreasing order in Σ\Sigma

✅ **Answer:** C) They can be complex numbers

📖 **Explanation:** Singular values are always non-negative real numbers.

#### ****4. Which of the following properties holds for matrices**** UU ****and**** VV ****in SVD**** A=UΣVTA = U \Sigma V^T****?****

A) They are always diagonal  
B) They are always symmetric  
C) They are always orthogonal (or unitary if complex)  
D) They always have the same dimensions as AA

✅ **Answer:** C) They are always orthogonal (or unitary if complex)

📖 **Explanation:** UU and VV are orthogonal matrices, meaning UTU=IU^T U = I and VTV=IV^T V = I.

#### ****5. What is the primary use of Singular Value Decomposition (SVD) in data science?****

A) Matrix inversion only  
B) Dimensionality reduction, noise filtering, and principal component analysis (PCA)  
C) Solving linear equations only  
D) None of the above

✅ **Answer:** B) Dimensionality reduction, noise filtering, and principal component analysis (PCA)

📖 **Explanation:** SVD is widely used in PCA for reducing the dimensionality of datasets and in applications like image compression.

#### ****6. If**** AA ****is a square and symmetric matrix, how does SVD relate to the Eigen decomposition?****

A) SVD and Eigen decomposition are completely unrelated  
B) The left singular vectors and right singular vectors are different  
C) The singular values are equal to the absolute values of the eigenvalues  
D) SVD cannot be applied to symmetric matrices

✅ **Answer:** C) The singular values are equal to the absolute values of the eigenvalues

📖 **Explanation:** For a symmetric matrix AA, the singular values of AA are the absolute values of its eigenvalues.

#### ****7. What happens when a matrix is rank-deficient (i.e., has less than full rank)?****

A) Some of its singular values will be zero  
B) All of its singular values will be zero  
C) The matrix cannot have an SVD decomposition  
D) The matrix becomes invertible

✅ **Answer:** A) Some of its singular values will be zero

📖 **Explanation:** A rank-deficient matrix will have fewer nonzero singular values than its full rank counterpart.

#### ****8. In SVD, what does the condition number of a matrix measure?****

A) The number of nonzero singular values  
B) The sum of the singular values  
C) The ratio of the largest singular value to the smallest nonzero singular value  
D) The determinant of the singular value matrix

✅ **Answer:** C) The ratio of the largest singular value to the smallest nonzero singular value

📖 **Explanation:** The condition number is used to measure the numerical stability of a matrix.

#### ****9. What is the computational complexity of computing the full SVD of an**** m×nm \times n ****matrix?****

A) O(m+n)O(m+n)  
B) O(m2n)O(m^2 n)  
C) O(mnlog⁡n)O(mn \log n)  
D) O(mn2)O(mn^2)

✅ **Answer:** D) O(mn2)O(mn^2)

📖 **Explanation:** Computing the full SVD typically requires O(mn2)O(mn^2) operations, making it computationally expensive for large matrices.

#### ****10. Which of the following applications heavily relies on SVD?****

A) Image compression  
B) Cryptography  
C) Sorting algorithms  
D) None of the above

✅ **Answer:** A) Image compression

📖 **Explanation:** SVD is used in image compression techniques like JPEG to retain significant image features while discarding unnecessary data

#### ****1. What is the primary objective of optimization in machine learning?****

A) To minimize or maximize an objective function  
B) To eliminate all errors in the model  
C) To increase the number of parameters in the model  
D) To reduce the number of training samples

✅ **Answer:** A) To minimize or maximize an objective function

📖 **Explanation:** Optimization in machine learning aims to minimize loss functions (e.g., Mean Squared Error) or maximize performance metrics (e.g., accuracy).

#### ****2. Which of the following is a first-order optimization algorithm commonly used in deep learning?****

A) Newton’s Method  
B) Gradient Descent  
C) Simulated Annealing  
D) Genetic Algorithms

✅ **Answer:** B) Gradient Descent

📖 **Explanation:** Gradient Descent is a first-order optimization algorithm that updates parameters in the direction of the negative gradient to minimize loss.

#### ****3. Which variant of gradient descent uses the entire dataset to compute the gradient before updating weights?****

A) Stochastic Gradient Descent (SGD)  
B) Mini-Batch Gradient Descent  
C) Batch Gradient Descent  
D) Momentum-based Gradient Descent

✅ **Answer:** C) Batch Gradient Descent

📖 **Explanation:** Batch Gradient Descent computes the gradient using the entire dataset before updating the model parameters.

#### ****4. Which optimization algorithm is designed to overcome the problem of slow convergence in standard Gradient Descent?****

A) Momentum-based Gradient Descent  
B) Random Search  
C) Simulated Annealing  
D) Newton’s Method

✅ **Answer:** A) Momentum-based Gradient Descent

📖 **Explanation:** Momentum helps accelerate gradient descent by accumulating past gradients, leading to faster convergence and reduced oscillations.

#### ****5. Adam optimizer combines which two optimization techniques?****

A) SGD and RMSProp  
B) Momentum and RMSProp  
C) Newton’s Method and Gradient Descent  
D) Simulated Annealing and Gradient Descent

✅ **Answer:** B) Momentum and RMSProp

📖 **Explanation:** Adam (Adaptive Moment Estimation) combines the benefits of Momentum (for smooth updates) and RMSProp (for adaptive learning rates).

#### ****6. Which of the following learning rate schedules dynamically adjusts the learning rate during training?****

A) Fixed Learning Rate  
B) Step Decay  
C) Batch Normalization  
D) Softmax

✅ **Answer:** B) Step Decay

📖 **Explanation:** Step Decay reduces the learning rate at fixed intervals to ensure better convergence during training.

#### ****7. What problem does L1 regularization (Lasso) help address in machine learning models?****

A) Overfitting by reducing model complexity  
B) Underfitting by adding more features  
C) Increasing computation time  
D) Converting categorical data into numerical data

✅ **Answer:** A) Overfitting by reducing model complexity

📖 **Explanation:** L1 regularization encourages sparsity by shrinking some coefficients to zero, effectively selecting important features.

#### ****8. Which of the following is NOT an issue that optimization algorithms attempt to address?****

A) Local minima  
B) Vanishing and exploding gradients  
C) Gradient updates in high-dimensional spaces  
D) Increasing the number of features in a dataset

✅ **Answer:** D) Increasing the number of features in a dataset

📖 **Explanation:** Optimization focuses on improving training efficiency, handling local minima, and stabilizing gradients, not adding features.

#### ****9. Which optimization algorithm uses a "temperature" parameter to escape local minima?****

A) Simulated Annealing  
B) Gradient Descent  
C) Adam Optimizer  
D) Genetic Algorithms

✅ **Answer:** A) Simulated Annealing

📖 **Explanation:** Simulated Annealing uses a probabilistic approach to escape local minima by allowing worse solutions with a decreasing probability.

#### ****10. What is the main disadvantage of Stochastic Gradient Descent (SGD) compared to Batch Gradient Descent?****

A) It is computationally expensive  
B) It requires the entire dataset at once  
C) It introduces noise in the gradient updates  
D) It cannot handle deep learning problems

✅ **Answer:** C) It introduces noise in the gradient updates

📖 **Explanation:** SGD updates model parameters using individual samples, leading to noisy updates but faster convergence compared to batch methods.

#### ****11. What is the role of the learning rate in optimization?****

A) It determines how quickly or slowly model parameters are updated  
B) It removes irrelevant features  
C) It directly affects the accuracy of the model  
D) It converts categorical data into numerical form

✅ **Answer:** A) It determines how quickly or slowly model parameters are updated

📖 **Explanation:** A high learning rate can cause the model to overshoot optimal values, while a low learning rate can lead to slow convergence.

#### ****12. In deep learning, which problem is often encountered in deep networks that use backpropagation with gradient descent?****

A) Vanishing and Exploding Gradients  
B) Lack of Features  
C) Feature Scaling Issues  
D) Missing Data

✅ **Answer:** A) Vanishing and Exploding Gradients

📖 **Explanation:** In deep networks, gradients can become very small (vanishing) or very large (exploding), leading to unstable training.

#### ****13. Which of the following methods is NOT used to improve optimization in deep learning?****

A) Batch Normalization  
B) Dropout  
C) Learning Rate Scheduling  
D) One-Hot Encoding

✅ **Answer:** D) One-Hot Encoding

📖 **Explanation:** One-Hot Encoding is a preprocessing technique for categorical data, not an optimization method.

#### ****14. Which optimizer dynamically adjusts learning rates for different parameters?****

A) SGD  
B) RMSProp  
C) Newton’s Method  
D) Lasso Regression

✅ **Answer:** B) RMSProp

📖 **Explanation:** RMSProp adjusts learning rates per parameter to handle non-stationary objectives in deep learning.

#### ****15. Which optimization method mimics the process of natural selection?****

A) Gradient Descent  
B) Genetic Algorithms  
C) Simulated Annealing  
D) Newton’s Method

✅ **Answer:** B) Genetic Algorithms

📖 **Explanation:** Genetic Algorithms use selection, crossover, and mutation to evolve solutions over generations, mimicking natural selection.

#### ****1. What is an eigenvector of a square matrix**** AA****?****

A) A vector that is transformed into a scalar multiple of itself when multiplied by AA  
B) A vector that is always perpendicular to AA  
C) A vector that is orthogonal to all the columns of AA  
D) A vector that is multiplied element-wise by AA

✅ **Answer:** A) A vector that is transformed into a scalar multiple of itself when multiplied by AA

📖 **Explanation:** An eigenvector vv of matrix AA satisfies the equation Av=λvA v = \lambda v, where λ\lambda is the eigenvalue.

#### ****2. What is an eigenvalue of a square matrix**** AA****?****

A) A scalar that satisfies the equation Av=λvA v = \lambda v for some nonzero vector vv  
B) A vector that is always equal to zero  
C) A matrix that can be decomposed into diagonal form  
D) The determinant of matrix AA

✅ **Answer:** A) A scalar that satisfies the equation Av=λvA v = \lambda v for some nonzero vector vv

📖 **Explanation:** The eigenvalue λ\lambda is a scalar such that multiplying AA with an eigenvector vv results in vv being scaled by λ\lambda.

#### ****3. How are eigenvalues of a matrix related to its determinant?****

A) The determinant is the sum of the eigenvalues  
B) The determinant is the product of the eigenvalues  
C) The determinant is always equal to one of the eigenvalues  
D) The determinant is unrelated to the eigenvalues

✅ **Answer:** B) The determinant is the product of the eigenvalues

📖 **Explanation:** If a matrix AA has eigenvalues λ1,λ2,...,λn\lambda\_1, \lambda\_2, ..., \lambda\_n, then det⁡(A)=λ1λ2...λn\det(A) = \lambda\_1 \lambda\_2 ... \lambda\_n.

#### ****4. In Principal Component Analysis (PCA), why are eigenvectors important?****

A) They represent the directions of maximum variance in the data  
B) They determine the number of clusters in the data  
C) They minimize the loss function  
D) They are used to normalize data

✅ **Answer:** A) They represent the directions of maximum variance in the data

📖 **Explanation:** PCA uses eigenvectors of the covariance matrix to identify the directions (principal components) along which variance is maximized.

#### ****5. Which property must an eigenvector satisfy?****

A) It must be a unit vector  
B) It must be nonzero  
C) It must be equal to its corresponding eigenvalue  
D) It must have only integer entries

✅ **Answer:** B) It must be nonzero

📖 **Explanation:** Eigenvectors are defined as nonzero vectors satisfying Av=λvA v = \lambda v.

#### ****6. What is the sum of the eigenvalues of a square matrix**** AA****?****

A) The determinant of AA  
B) The rank of AA  
C) The trace of AA (sum of diagonal elements)  
D) The number of nonzero elements in AA

✅ **Answer:** C) The trace of AA (sum of diagonal elements)

📖 **Explanation:** The sum of eigenvalues of a matrix is equal to the sum of its diagonal elements (trace of the matrix).

#### ****7. If an eigenvalue of a matrix is zero, what does this imply?****

A) The matrix is invertible  
B) The matrix is singular (non-invertible)  
C) The eigenvectors are all zero  
D) The matrix is diagonalizable

✅ **Answer:** B) The matrix is singular (non-invertible)

📖 **Explanation:** A zero eigenvalue means the determinant is zero, implying the matrix is singular and not invertible.

#### ****8. What happens to the eigenvalues of a matrix if it is multiplied by a scalar**** cc****?****

A) The eigenvalues remain the same  
B) Each eigenvalue is multiplied by cc  
C) Each eigenvalue is squared  
D) The eigenvalues are inverted

✅ **Answer:** B) Each eigenvalue is multiplied by cc

📖 **Explanation:** If AA has eigenvalues λi\lambda\_i, then the matrix cAcA has eigenvalues cλic\lambda\_i.

#### ****9. Which type of matrices always have real eigenvalues?****

A) Symmetric matrices  
B) Skew-symmetric matrices  
C) Singular matrices  
D) Non-square matrices

✅ **Answer:** A) Symmetric matrices

📖 **Explanation:** Symmetric matrices always have real eigenvalues, making them useful in applications like PCA.

#### ****10. If an eigenvector of a matrix is scaled by a nonzero constant, what happens to its corresponding eigenvalue?****

A) It remains the same  
B) It is also scaled by the same constant  
C) It is squared  
D) It becomes zero

✅ **Answer:** A) It remains the same

📖 **Explanation:** Scaling an eigenvector does not change its eigenvalue, as the equation Av=λvA v = \lambda v remains valid.

#### ****11. What is the effect of taking the inverse of a matrix on its eigenvalues?****

A) The eigenvalues are unchanged  
B) The eigenvalues are squared  
C) The eigenvalues are inverted (reciprocals)  
D) The eigenvalues become complex

✅ **Answer:** C) The eigenvalues are inverted (reciprocals)

📖 **Explanation:** If λ\lambda is an eigenvalue of AA, then 1/λ1/\lambda is an eigenvalue of A−1A^{-1}, assuming AA is invertible.

#### ****12. What does it mean if a matrix has complex eigenvalues?****

A) The matrix is necessarily complex  
B) The matrix cannot be diagonalized  
C) The matrix may not be symmetric  
D) The matrix has no real eigenvectors

✅ **Answer:** C) The matrix may not be symmetric

📖 **Explanation:** Non-symmetric matrices can have complex eigenvalues, while symmetric matrices always have real eigenvalues.

#### ****13. If a matrix has distinct eigenvalues, what can be said about its eigenvectors?****

A) The eigenvectors are linearly dependent  
B) The eigenvectors are linearly independent  
C) The eigenvectors are all identical  
D) The eigenvectors must be complex

✅ **Answer:** B) The eigenvectors are linearly independent

📖 **Explanation:** A matrix with distinct eigenvalues has linearly independent eigenvectors, meaning it can be diagonalized.

#### ****14. In machine learning, eigenvalues and eigenvectors are primarily used in which of the following techniques?****

A) Gradient Descent  
B) Principal Component Analysis (PCA)  
C) Decision Trees  
D) k-Nearest Neighbors

✅ **Answer:** B) Principal Component Analysis (PCA)

📖 **Explanation:** PCA relies on eigenvalues and eigenvectors to determine the principal components that explain variance in the data.

#### ****15. How do eigenvalues relate to the stability of a system in dynamical systems analysis?****

A) If all eigenvalues have negative real parts, the system is stable  
B) If all eigenvalues are zero, the system is stable  
C) If all eigenvalues are positive, the system is stable  
D) Eigenvalues do not affect system stability

✅ **Answer:** A) If all eigenvalues have negative real parts, the system is stable

📖 **Explanation:** In dynamical systems, negative real eigenvalues indicate stability, while positive eigenvalues suggest instability.

#### ****1. What is a linear transformation in the context of matrices?****

A) A transformation that preserves vector addition and scalar multiplication  
B) A transformation that changes the rank of a matrix  
C) A transformation that increases the determinant of a matrix  
D) A transformation that only applies to square matrices

✅ **Answer:** A) A transformation that preserves vector addition and scalar multiplication

📖 **Explanation:** A linear transformation TT satisfies T(a+b)=T(a)+T(b)T(a + b) = T(a) + T(b) and T(ca)=cT(a)T(ca) = cT(a), preserving linear structure.

#### ****2. How is a linear transformation represented using matrices?****

A) As a system of nonlinear equations  
B) As a square matrix multiplication with a vector  
C) As a determinant equation  
D) As a polynomial function

✅ **Answer:** B) As a square matrix multiplication with a vector

📖 **Explanation:** A linear transformation is represented as T(x)=AxT(x) = Ax, where AA is a matrix and xx is a vector.

#### ****3. What happens to a vector when a transformation matrix has a determinant of zero?****

A) The vector remains unchanged  
B) The vector is mapped to a lower-dimensional space  
C) The vector is inverted  
D) The vector’s magnitude increases

✅ **Answer:** B) The vector is mapped to a lower-dimensional space

📖 **Explanation:** A determinant of zero means the transformation compresses the space, leading to rank deficiency and loss of dimensions.

#### ****4. Which of the following is an example of a linear transformation?****

A) Rotation  
B) Translation  
C) Exponential scaling  
D) Logarithmic mapping

✅ **Answer:** A) Rotation

📖 **Explanation:** Rotations are linear transformations since they preserve vector addition and scalar multiplication, whereas translations are not linear.

#### ****5. What type of matrix represents a reflection transformation?****

A) Diagonal matrix  
B) Orthogonal matrix with determinant -1  
C) Symmetric matrix  
D) Upper triangular matrix

✅ **Answer:** B) Orthogonal matrix with determinant -1

📖 **Explanation:** A reflection matrix is an orthogonal matrix whose determinant is -1, flipping the orientation of vectors.

#### ****6. What is the effect of applying an identity matrix as a transformation?****

A) It scales the vector  
B) It reflects the vector  
C) It rotates the vector  
D) It leaves the vector unchanged

✅ **Answer:** D) It leaves the vector unchanged

📖 **Explanation:** The identity matrix II satisfies Ix=xI x = x, meaning it does not change the vector.

#### ****7. Which of the following operations is NOT a linear transformation?****

A) Scaling  
B) Rotation  
C) Reflection  
D) Translation

✅ **Answer:** D) Translation

📖 **Explanation:** Translation does not preserve the origin, making it a non-linear transformation.

#### ****8. What is the geometric interpretation of a matrix with eigenvalues greater than 1?****

A) It shrinks vectors  
B) It rotates vectors  
C) It scales vectors outward  
D) It reflects vectors

✅ **Answer:** C) It scales vectors outward

📖 **Explanation:** Eigenvalues determine scaling. If greater than 1, they stretch vectors in the direction of the eigenvectors.

#### ****9. If a transformation matrix is singular, what does this imply?****

A) The transformation is invertible  
B) The transformation collapses dimensions  
C) The transformation rotates vectors  
D) The transformation preserves distances

✅ **Answer:** B) The transformation collapses dimensions

📖 **Explanation:** A singular matrix has a determinant of zero, meaning it maps vectors to a lower-dimensional space.

#### ****10. Which property must a matrix satisfy to be considered an orthogonal transformation?****

A) Its determinant must be positive  
B) Its columns must be linearly dependent  
C) Its transpose must be equal to its inverse  
D) It must have all eigenvalues equal to zero

✅ **Answer:** C) Its transpose must be equal to its inverse

📖 **Explanation:** A matrix AA is orthogonal if ATA=IA^T A = I, meaning it preserves vector norms and angles.

#### ****11. What happens when a transformation matrix is applied multiple times to a vector?****

A) The effect is additive  
B) The effect is compounded through matrix multiplication  
C) The vector remains unchanged  
D) The vector's norm always increases

✅ **Answer:** B) The effect is compounded through matrix multiplication

📖 **Explanation:** Applying a transformation multiple times is equivalent to raising the matrix to a power, compounding the effect.

#### ****12. Which matrix property ensures that a linear transformation preserves angles and distances?****

A) Being a diagonal matrix  
B) Having a determinant of zero  
C) Being an orthogonal matrix  
D) Having only positive eigenvalues

✅ **Answer:** C) Being an orthogonal matrix

📖 **Explanation:** Orthogonal matrices satisfy ATA=IA^T A = I, preserving angles and vector norms.

#### ****13. In machine learning, which technique relies heavily on linear transformations?****

A) Decision Trees  
B) Principal Component Analysis (PCA)  
C) k-Nearest Neighbors  
D) Random Forest

✅ **Answer:** B) Principal Component Analysis (PCA)

📖 **Explanation:** PCA uses eigenvectors and eigenvalues to apply linear transformations for dimensionality reduction.

#### ****14. If a transformation scales a vector by a factor of**** kk****, what happens to the determinant of the transformation matrix?****

A) It remains unchanged  
B) It is scaled by knk^n, where nn is the matrix dimension  
C) It becomes zero  
D) It is always equal to one

✅ **Answer:** B) It is scaled by knk^n, where nn is the matrix dimension

📖 **Explanation:** Scaling transformations multiply the determinant by knk^n, where nn is the size of the square matrix.

#### ****15. If a matrix has a determinant of -1, what does this tell us about the transformation?****

A) It is a rotation  
B) It is a reflection  
C) It is a scaling transformation  
D) It is singular

✅ **Answer:** B) It is a reflection

📖 **Explanation:** A determinant of -1 indicates a transformation that flips orientation, which is characteristic of reflections.

#### ****1. What is the purpose of vector norms in machine learning?****

A) To compute the number of features in a dataset  
B) To measure the length or magnitude of a vector  
C) To convert a vector into a matrix  
D) To find the eigenvalues of a vector

✅ **Answer:** B) To measure the length or magnitude of a vector

📖 **Explanation:** Norms quantify the size of a vector, which is useful for optimization, regularization, and distance computation.

#### ****2. The**** L1L\_1 ****norm of a vector is also known as:****

A) Euclidean norm  
B) Manhattan norm  
C) Frobenius norm  
D) Spectral norm

✅ **Answer:** B) Manhattan norm

📖 **Explanation:** The L1L\_1 norm, defined as ∥x∥1=∑∣xi∣\|x\|\_1 = \sum |x\_i|, represents the sum of absolute values and is also called the Manhattan norm.

#### ****3. The**** L2L\_2 ****norm is also known as:****

A) Taxicab norm  
B) Chebyshev norm  
C) Euclidean norm  
D) Infinity norm

✅ **Answer:** C) Euclidean norm

📖 **Explanation:** The L2L\_2 norm, ∥x∥2=∑xi2\|x\|\_2 = \sqrt{\sum x\_i^2}, measures the standard Euclidean distance from the origin.

#### ****4. Which norm is commonly used in Ridge Regression (L2 regularization)?****

A) L1L\_1 norm  
B) L2L\_2 norm  
C) L∞L\_{\infty} norm  
D) Frobenius norm

✅ **Answer:** B) L2L\_2 norm

📖 **Explanation:** Ridge regression uses the L2L\_2 norm to penalize large weights, improving generalization.

#### ****5. The**** L∞L\_{\infty} ****norm (maximum norm) is defined as:****

A) The square root of the sum of squared elements  
B) The sum of absolute values of elements  
C) The maximum absolute value among the elements  
D) The sum of squared elements

✅ **Answer:** C) The maximum absolute value among the elements

📖 **Explanation:** The L∞L\_{\infty} norm is given by ∥x∥∞=max⁡∣xi∣\|x\|\_{\infty} = \max |x\_i|, making it useful in worst-case scenarios.

#### ****6. Which norm is most sensitive to outliers?****

A) L1L\_1 norm  
B) L2L\_2 norm  
C) L∞L\_{\infty} norm  
D) Frobenius norm

✅ **Answer:** B) L2L\_2 norm

📖 **Explanation:** The L2L\_2 norm squares each component, making large values (outliers) disproportionately affect the result.

#### ****7. What is the main difference between**** L1L\_1 ****and**** L2L\_2 ****norms in optimization?****

A) L1L\_1 norm encourages sparsity, while L2L\_2 norm does not  
B) L2L\_2 norm encourages sparsity, while L1L\_1 norm does not  
C) Both norms encourage sparsity  
D) Both norms avoid overfitting

✅ **Answer:** A) L1L\_1 norm encourages sparsity, while L2L\_2 norm does not

📖 **Explanation:** The L1L\_1 norm is used in Lasso regression to shrink some coefficients to zero, promoting sparsity.

#### ****8. In high-dimensional spaces, which norm is least affected by the "curse of dimensionality"?****

A) L1L\_1 norm  
B) L2L\_2 norm  
C) L∞L\_{\infty} norm  
D) All norms are equally affected

✅ **Answer:** C) L∞L\_{\infty} norm

📖 **Explanation:** The L∞L\_{\infty} norm depends only on the largest component, reducing sensitivity to increasing dimensions.

#### ****9. Which norm is used in Support Vector Machines (SVM) for margin maximization?****

A) L1L\_1 norm  
B) L2L\_2 norm  
C) L∞L\_{\infty} norm  
D) Frobenius norm

✅ **Answer:** B) L2L\_2 norm

📖 **Explanation:** SVM maximizes the margin using the L2L\_2 norm to measure distances between support vectors and the hyperplane.

#### ****10. The Frobenius norm is used to measure the magnitude of:****

A) A vector  
B) A matrix  
C) A scalar  
D) A probability distribution

✅ **Answer:** B) A matrix

📖 **Explanation:** The Frobenius norm measures matrix magnitude and is defined as ∥A∥F=∑∣aij∣2\|A\|\_F = \sqrt{\sum |a\_{ij}|^2}.

#### ****11. Which norm is best suited for feature selection in machine learning models?****

A) L1L\_1 norm  
B) L2L\_2 norm  
C) L∞L\_{\infty} norm  
D) Spectral norm

✅ **Answer:** A) L1L\_1 norm

📖 **Explanation:** The L1L\_1 norm (used in Lasso regression) forces some coefficients to zero, effectively selecting important features.

#### ****12. When computing norms, what is the effect of normalizing a vector?****

A) It increases the norm  
B) It reduces all components to zero  
C) It makes the vector unit length (norm = 1)  
D) It removes negative values

✅ **Answer:** C) It makes the vector unit length (norm = 1)

📖 **Explanation:** Normalizing a vector scales it to have unit norm, preserving direction but not magnitude.

#### ****13. In neural networks, weight regularization often uses:****

A) L1L\_1 and L2L\_2 norms  
B) Only the L∞L\_{\infty} norm  
C) The determinant of the weight matrix  
D) The Hessian matrix

✅ **Answer:** A) L1L\_1 and L2L\_2 norms

📖 **Explanation:** L1L\_1 regularization promotes sparsity, while L2L\_2 prevents large weights, improving model generalization.

#### ****14. Which norm is used in the K-Nearest Neighbors (KNN) algorithm to measure distance?****

A) L1L\_1 norm  
B) L2L\_2 norm  
C) L∞L\_{\infty} norm  
D) Frobenius norm

✅ **Answer:** B) L2L\_2 norm

📖 **Explanation:** KNN typically uses the Euclidean distance (L2L\_2 norm) to find the nearest neighbors in a dataset.

#### ****15. If a vector norm is used as a loss function in machine learning, what does it measure?****

A) The total variation in the data  
B) The difference between predicted and actual values  
C) The dimensionality of the data  
D) The correlation between features

✅ **Answer:** B) The difference between predicted and actual values

📖 **Explanation:** Norms measure error magnitudes, often serving as loss functions in regression and classification

#### ****1. What does it mean for a set of vectors to be linearly independent?****

A) They cannot be represented as a linear combination of each other  
B) They have the same direction  
C) They form an orthogonal basis  
D) They all have the same magnitude

✅ **Answer:** A) They cannot be represented as a linear combination of each other

📖 **Explanation:** A set of vectors is **linearly independent** if no vector in the set can be written as a linear combination of the others.

#### ****2. How do you determine if a set of vectors is linearly dependent?****

A) If their dot product is zero  
B) If the determinant of their coefficient matrix is nonzero  
C) If at least one vector can be written as a combination of others  
D) If they are orthonormal

✅ **Answer:** C) If at least one vector can be written as a combination of others

📖 **Explanation:** A set of vectors is **linearly dependent** if at least one vector is a linear combination of others, meaning they do not add new information.

#### ****3. If a matrix has linearly dependent columns, what does this imply about its determinant?****

A) The determinant is positive  
B) The determinant is negative  
C) The determinant is zero  
D) The determinant is nonzero

✅ **Answer:** C) The determinant is zero

📖 **Explanation:** If a matrix has linearly dependent columns, then its determinant is **zero**, indicating the matrix is singular and not invertible.

#### ****4. If a set of**** nn ****vectors in**** Rn\mathbb{R}^n ****is linearly independent, what is the rank of the matrix formed by these vectors?****

A) 0  
B) Less than nn  
C) Exactly nn  
D) Greater than nn

✅ **Answer:** C) Exactly nn

📖 **Explanation:** The rank of a matrix is the maximum number of **linearly independent** columns. If nn vectors in Rn\mathbb{R}^n are independent, the rank is **n**.

#### ****5. How can you check for linear dependence using the determinant?****

A) If the determinant is nonzero, the vectors are linearly dependent  
B) If the determinant is zero, the vectors are linearly dependent  
C) If the determinant is positive, the vectors are independent  
D) The determinant does not determine linear dependence

✅ **Answer:** B) If the determinant is zero, the vectors are linearly dependent

📖 **Explanation:** If the determinant of a matrix formed by a set of vectors is **zero**, the vectors are **linearly dependent**.

#### ****6. Which of the following is always true for linearly independent vectors?****

A) Their Gram matrix is singular  
B) Their determinant is zero  
C) Their dot product is always zero  
D) Their rank equals the number of vectors

✅ **Answer:** D) Their rank equals the number of vectors

📖 **Explanation:** The rank of a matrix formed by independent vectors is **equal** to the number of vectors.

#### ****7. In machine learning, why is it important for feature vectors to be linearly independent?****

A) To increase computation time  
B) To ensure unique feature representations  
C) To make the model overfit  
D) To maximize bias in the model

✅ **Answer:** B) To ensure unique feature representations

📖 **Explanation:** Linearly independent features provide **unique** and **useful** information, reducing redundancy in the model.

#### ****8. What is the rank of a matrix if its columns are linearly dependent?****

A) Full rank  
B) Zero  
C) Less than the number of columns  
D) Equal to the number of columns

✅ **Answer:** C) Less than the number of columns

📖 **Explanation:** If the columns of a matrix are **linearly dependent**, the rank is **less than the number of columns**.

#### ****9. If a set of vectors is linearly dependent, what happens when one vector is removed?****

A) The set remains dependent  
B) The set becomes independent  
C) The rank of the matrix increases  
D) The determinant becomes nonzero

✅ **Answer:** A) The set remains dependent

📖 **Explanation:** Removing **one vector** from a linearly dependent set may still result in dependence unless the remaining vectors become independent.

#### ****10. What happens if a matrix has more columns than rows?****

A) The columns must be linearly independent  
B) The matrix is always invertible  
C) The columns must be linearly dependent  
D) The determinant is always nonzero

✅ **Answer:** C) The columns must be linearly dependent

📖 **Explanation:** In an m×nm \times n matrix where n>mn > m, there are **more vectors than dimensions**, so the columns must be **linearly dependent**.

#### ****11. What does it mean if the null space of a matrix contains only the zero vector?****

A) The matrix has linearly dependent columns  
B) The matrix has linearly independent columns  
C) The matrix is singular  
D) The determinant of the matrix is zero

✅ **Answer:** B) The matrix has linearly independent columns

📖 **Explanation:** If the **null space** contains only the **zero vector**, it means the matrix has **linearly independent** columns.

#### ****12. How does linear dependence affect feature selection in machine learning?****

A) It reduces computation efficiency  
B) It introduces redundant features  
C) It makes models more interpretable  
D) It improves generalization

✅ **Answer:** B) It introduces redundant features

📖 **Explanation:** Linearly **dependent** features provide **redundant** information, which can negatively impact the model's performance.

#### ****13. In Principal Component Analysis (PCA), what happens to linearly dependent features?****

A) They are removed by selecting principal components  
B) They increase the number of components  
C) They make PCA more effective  
D) They increase the variance

✅ **Answer:** A) They are removed by selecting principal components

📖 **Explanation:** PCA removes redundancy by selecting only **linearly independent** components.

#### ****14. What is the geometric interpretation of linearly dependent vectors?****

A) They span the entire space  
B) They lie in the same subspace  
C) They form an orthonormal basis  
D) They always have the same magnitude

✅ **Answer:** B) They lie in the same subspace

📖 **Explanation:** Linearly dependent vectors do not **span** the entire space; instead, they lie within the same lower-dimensional subspace.

#### ****15. If a dataset contains highly correlated features, what does this indicate about the feature vectors?****

A) They are linearly independent  
B) They are linearly dependent  
C) They are orthogonal  
D) They are always normalized

✅ **Answer:** B) They are linearly dependent

📖 **Explanation:** Highly correlated features indicate **linear dependence**, meaning one feature can be predicted from another.

### ****1. Why is a basis important in machine learning models?****

A) It helps in reducing overfitting  
B) It ensures that feature representations are minimal and independent  
C) It increases the number of dimensions for better learning  
D) It makes models more complex

✅ **Answer:** **B) It ensures that feature representations are minimal and independent**

📖 **Explanation:** In machine learning, a **basis** ensures that features are **linearly independent**, avoiding redundancy and improving model efficiency.

### ****2. What does it mean for a set of feature vectors to span a space?****

A) Any feature in the space can be written as a linear combination of these vectors  
B) The vectors are always linearly dependent  
C) The vectors must be orthogonal  
D) The vectors always form a square matrix

✅ **Answer:** **A) Any feature in the space can be written as a linear combination of these vectors**

📖 **Explanation:** In **machine learning**, if a set of feature vectors **spans a space**, it means all possible feature vectors in that space can be **constructed** from these vectors.

### ****3. Why do machine learning algorithms benefit from working in a lower-dimensional subspace?****

A) It improves computational efficiency and reduces overfitting  
B) It makes the model more complex  
C) It always leads to better accuracy  
D) It increases the number of features

✅ **Answer:** **A) It improves computational efficiency and reduces overfitting**

📖 **Explanation:** Using a **lower-dimensional subspace** (e.g., through **PCA**) removes redundant features, speeds up training, and reduces overfitting.

### ****4. Principal Component Analysis (PCA) finds a new basis for data. What is special about this basis?****

A) It consists of correlated vectors  
B) It is formed by eigenvectors of the covariance matrix  
C) It always has dependent vectors  
D) It does not span the space

✅ **Answer:** **B) It is formed by eigenvectors of the covariance matrix**

📖 **Explanation:** PCA finds **principal components**, which are the **eigenvectors** of the covariance matrix, forming a new **basis** for the data.

### ****5. What is the main advantage of having an orthonormal basis in feature space?****

A) It reduces model accuracy  
B) It makes computations easier and prevents redundancy  
C) It increases feature correlations  
D) It increases feature redundancy

✅ **Answer:** **B) It makes computations easier and prevents redundancy**

📖 **Explanation:** An **orthonormal basis** (e.g., PCA transformation) simplifies computations (dot products, norms) and removes **feature redundancy**.

### ****6. If the columns of a feature matrix form a basis, what does this imply?****

A) The features are dependent  
B) The features are redundant  
C) The features are linearly independent and span the space  
D) The determinant of the matrix is zero

✅ **Answer:** **C) The features are linearly independent and span the space**

📖 **Explanation:** A **basis** means the features **span** the space and are **independent**, which is ideal for feature engineering.

### ****7. Why is feature selection related to the concept of a basis in machine learning?****

A) Feature selection removes redundant dimensions  
B) Feature selection adds extra dependent features  
C) Feature selection makes the dataset more complex  
D) Feature selection always reduces accuracy

✅ **Answer:** **A) Feature selection removes redundant dimensions**

📖 **Explanation:** Selecting a **basis** for the data (e.g., via **PCA, Lasso Regression**) helps remove redundant, **linearly dependent** features.

### ****8. In high-dimensional data, why is it beneficial to find a lower-dimensional spanning set?****

A) It helps visualize the data and improves generalization  
B) It always increases computation time  
C) It reduces variance in the model  
D) It removes all important features

✅ **Answer:** **A) It helps visualize the data and improves generalization**

📖 **Explanation:** Reducing dimensionality (e.g., PCA, t-SNE) **helps in visualization** and improves **model generalization** by removing noise.

### ****9. What does a basis vector represent in an image recognition model?****

A) A single pixel value  
B) A fundamental pattern or feature in the dataset  
C) A redundant part of the image  
D) A noise component

✅ **Answer:** **B) A fundamental pattern or feature in the dataset**

📖 **Explanation:** In image recognition (e.g., PCA in face recognition), basis vectors represent **patterns** (edges, textures) that form images.

### ****10. If a set of feature vectors does not span the entire feature space, what happens in a machine learning model?****

A) The model may not generalize well to unseen data  
B) The model has too many features  
C) The model will always overfit  
D) The model becomes perfectly accurate

✅ **Answer:** **A) The model may not generalize well to unseen data**

📖 **Explanation:** If feature vectors **do not span** the space, the model may **fail** to represent new data properly, leading to **poor generalization**.

### ****11. If a dataset has redundant features, which technique can be used to find an optimal basis?****

A) Adding more features  
B) PCA (Principal Component Analysis)  
C) Increasing the model complexity  
D) Using a larger dataset

✅ **Answer:** **B) PCA (Principal Component Analysis)**

📖 **Explanation:** **PCA** helps **find an optimal basis** by selecting principal components, reducing **feature redundancy**.

### ****12. What happens if a dataset contains too many dependent features?****

A) The determinant of the feature matrix is nonzero  
B) The model may suffer from multicollinearity and overfitting  
C) The model generalizes better  
D) The features become independent

✅ **Answer:** **B) The model may suffer from multicollinearity and overfitting**

📖 **Explanation:** Dependent features lead to **multicollinearity**, making it hard for the model to assign correct weights, reducing performance.

### ****13. Why is it beneficial for feature vectors to form an orthogonal basis in machine learning?****

A) It minimizes feature redundancy and simplifies computations  
B) It increases the complexity of matrix operations  
C) It makes data representation harder  
D) It requires more training data

✅ **Answer:** **A) It minimizes feature redundancy and simplifies computations**

📖 **Explanation:** **Orthogonal basis** simplifies transformations (e.g., PCA, SVD), making models more efficient and interpretable.

### ****14. Which of the following methods finds a new basis for representing data in machine learning?****

A) PCA  
B) k-Means clustering  
C) Naïve Bayes classifier  
D) Decision trees

✅ **Answer:** **A) PCA**

📖 **Explanation:** **PCA** finds a new **basis** (principal components) to represent data efficiently.

### ****15. In deep learning, why are basis vectors important in layer transformations?****

A) They define how features are transformed in each layer  
B) They do not affect the learning process  
C) They always remain unchanged throughout training  
D) They prevent weight updates

✅ **Answer:** **A) They define how features are transformed in each layer**

📖 **Explanation:** Neural networks learn **basis transformations** at each layer to extract important patterns in data.

### ****1. What does it mean for a vector to be a normal vector in machine learning?****

A) It is always a zero vector  
B) It is perpendicular to a given surface or plane  
C) It has only positive components  
D) It is always an eigenvector

✅ **Answer:** **B) It is perpendicular to a given surface or plane**

📖 **Explanation:** A **normal vector** is **perpendicular** to a surface or hyperplane and is commonly used in **SVM decision boundaries** and **plane equations**.

### ****2. What is the key property of an orthonormal set of vectors?****

A) The vectors are linearly dependent  
B) Each vector has unit length and is perpendicular to the others  
C) The vectors have a determinant of zero  
D) The vectors are parallel

✅ **Answer:** **B) Each vector has unit length and is perpendicular to the others**

📖 **Explanation:** An **orthonormal set** consists of vectors that are **orthogonal** (perpendicular) and **normalized** (unit length).

### ****3. How is the dot product of two orthonormal vectors defined?****

A) It is always 1  
B) It is always 0  
C) It is always greater than 1  
D) It is always negative

✅ **Answer:** **B) It is always 0**

📖 **Explanation:** Orthonormal vectors are **orthogonal**, meaning their **dot product is zero** if they are different. If they are the same vector, the dot product is 1.

### ****4. Which of the following machine learning techniques relies heavily on orthonormal vectors?****

A) k-Nearest Neighbors (k-NN)  
B) Principal Component Analysis (PCA)  
C) Decision Trees  
D) Naïve Bayes

✅ **Answer:** **B) Principal Component Analysis (PCA)**

📖 **Explanation:** **PCA** finds an **orthonormal basis** for the data by computing the eigenvectors of the covariance matrix.

### ****5. If two vectors are orthonormal, what is the angle between them?****

A) 0°  
B) 45°  
C) 90°  
D) 180°

✅ **Answer:** **C) 90°**

📖 **Explanation:** **Orthonormal** vectors are **orthogonal**, meaning they are at **90 degrees** to each other.

### ****6. In a Support Vector Machine (SVM), what is the significance of the normal vector?****

A) It determines the direction of the decision boundary  
B) It classifies new data points  
C) It always points toward the origin  
D) It is ignored during training

✅ **Answer:** **A) It determines the direction of the decision boundary**

📖 **Explanation:** The **normal vector** of the decision boundary in SVMs helps **separate** different classes by defining the **hyperplane’s orientation**.

### ****7. Which matrix has orthonormal columns?****

A) Diagonal matrix  
B) Singular matrix  
C) Orthogonal matrix  
D) Symmetric matrix

✅ **Answer:** **C) Orthogonal matrix**

📖 **Explanation:** A matrix is **orthogonal** if its **columns are orthonormal**. It satisfies ATA=IA^T A = I.

### ****8. Why is it beneficial to use an orthonormal basis in feature space?****

A) It increases computational complexity  
B) It ensures features are independent and simplifies calculations  
C) It makes models more complex  
D) It removes important information

✅ **Answer:** **B) It ensures features are independent and simplifies calculations**

📖 **Explanation:** **Orthonormal** vectors help **remove feature redundancy**, improving **model efficiency** and reducing **correlations**.

### ****9. What is the length (norm) of an orthonormal vector?****

A) 0  
B) 1  
C) It depends on the vector  
D) Infinity

✅ **Answer:** **B) 1**

📖 **Explanation:** **Orthonormal vectors** are **unit vectors**, meaning their **length (norm) is always 1**.

### ****10. How are orthonormal vectors related to Singular Value Decomposition (SVD)?****

A) The left and right singular vectors are orthonormal  
B) The singular values are always zero  
C) The vectors are always linearly dependent  
D) The determinant of the matrix is always zero

✅ **Answer:** **A) The left and right singular vectors are orthonormal**

📖 **Explanation:** **SVD** decomposes a matrix A=UΣVTA = U \Sigma V^T, where **U and V** have **orthonormal columns**.

### ****11. How does Gram-Schmidt Orthogonalization work?****

A) It converts a set of linearly dependent vectors into an orthonormal set  
B) It computes the determinant of a matrix  
C) It finds eigenvalues of a matrix  
D) It normalizes all data points

✅ **Answer:** **A) It converts a set of linearly dependent vectors into an orthonormal set**

📖 **Explanation:** **Gram-Schmidt** is an algorithm that **constructs an orthonormal basis** from a set of vectors.

### ****12. Why are orthonormal vectors preferred in machine learning optimization problems?****

A) They make gradient descent slower  
B) They allow for more stable numerical computations  
C) They decrease accuracy  
D) They increase data redundancy

✅ **Answer:** **B) They allow for more stable numerical computations**

📖 **Explanation:** **Orthonormal vectors** reduce numerical **instability**, improving performance in **optimization and regression**.

### ****13. If a dataset is projected onto an orthonormal basis, what happens to variance?****

A) It is distributed optimally across the components  
B) It is minimized  
C) It becomes zero  
D) It does not change

✅ **Answer:** **A) It is distributed optimally across the components**

📖 **Explanation:** **PCA** uses an **orthonormal basis** to distribute **variance across principal components**, ensuring maximum data retention.

### ****14. If a feature matrix has orthonormal columns, what happens when it is multiplied by its transpose?****

A) It results in an identity matrix  
B) It results in a zero matrix  
C) It results in a singular matrix  
D) It results in a diagonal matrix with random values

✅ **Answer:** **A) It results in an identity matrix**

📖 **Explanation:** **Orthonormal matrices** satisfy ATA=IA^T A = I, where **I is the identity matrix**.

### ****15. What happens if non-orthonormal feature vectors are used in machine learning models?****

A) Feature redundancy increases, leading to multicollinearity  
B) The model performs better  
C) The model always converges faster  
D) The number of features decreases

✅ **Answer:** **A) Feature redundancy increases, leading to multicollinearity**

📖 **Explanation:** Non-orthonormal features can introduce **multicollinearity**, causing **unstable parameter estimates** in regression models.

### ****1. What is the main purpose of PCA in machine learning?****

A) To increase the number of features in a dataset  
B) To reduce the dimensionality while preserving as much variance as possible  
C) To remove missing values from the dataset  
D) To make a dataset categorical

✅ **Answer:** **B) To reduce the dimensionality while preserving as much variance as possible**

📖 **Explanation:** PCA helps **reduce the number of features** while keeping most of the **important information** by finding the directions (principal components) that maximize variance.

### ****2. PCA transforms the original features into a new set of features called \_\_\_\_\_\_.****

A) Covariance components  
B) Principal components  
C) Regression coefficients  
D) Singular values

✅ **Answer:** **B) Principal components**

📖 **Explanation:** PCA creates **new orthogonal features** called **principal components**, which are **linear combinations** of the original features.

### ****3. What mathematical concept is used to compute the principal components in PCA?****

A) Gradient Descent  
B) Eigenvalues and Eigenvectors  
C) Decision Trees  
D) k-Means Clustering

✅ **Answer:** **B) Eigenvalues and Eigenvectors**

📖 **Explanation:** **Eigenvalues and eigenvectors** of the **covariance matrix** determine the principal components. The eigenvectors define directions, while eigenvalues indicate the amount of variance captured.

### ****4. In PCA, the principal components are ordered based on what criterion?****

A) The magnitude of their eigenvalues  
B) The number of missing values  
C) The mean of each feature  
D) The correlation between variables

✅ **Answer:** **A) The magnitude of their eigenvalues**

📖 **Explanation:** The **eigenvalue** of each principal component shows how much **variance** it captures. Higher eigenvalues indicate more important components.

### ****5. Which of the following is a key assumption of PCA?****

A) The dataset has normally distributed features  
B) The dataset has linearly correlated features  
C) The dataset is categorical  
D) The dataset contains missing values

✅ **Answer:** **B) The dataset has linearly correlated features**

📖 **Explanation:** PCA works best when features are **correlated**, allowing it to find meaningful directions in the data.

### ****6. What is the role of the covariance matrix in PCA?****

A) It measures the relationship between features  
B) It finds clusters in the data  
C) It predicts output values  
D) It eliminates outliers

✅ **Answer:** **A) It measures the relationship between features**

📖 **Explanation:** The **covariance matrix** shows **how features are related**. PCA diagonalizes this matrix to find the principal components.

### ****7. Which step is necessary before applying PCA?****

A) Normalizing the features (zero mean, unit variance)  
B) Converting categorical variables to numerical values  
C) Removing outliers  
D) Both A and B

✅ **Answer:** **D) Both A and B**

📖 **Explanation:** PCA works best when data is **centered (mean = 0)** and **scaled**. Categorical variables must also be **encoded**.

### ****8. If we apply PCA to a dataset with 100 features, what is the maximum number of principal components we can get?****

A) 10  
B) 50  
C) 100  
D) Unlimited

✅ **Answer:** **C) 100**

📖 **Explanation:** The maximum number of **principal components** equals the number of **original features**.

### ****9. Which of the following methods is an alternative to PCA for dimensionality reduction?****

A) Singular Value Decomposition (SVD)  
B) Decision Trees  
C) Logistic Regression  
D) Naïve Bayes

✅ **Answer:** **A) Singular Value Decomposition (SVD)**

📖 **Explanation:** **SVD** is a mathematical technique closely related to PCA, often used for matrix factorization.

### ****10. What happens when we retain only the first k principal components?****

A) We lose all the variance in the dataset  
B) We retain most of the variance while reducing the number of features  
C) The dataset becomes linearly dependent  
D) The dataset becomes more complex

✅ **Answer:** **B) We retain most of the variance while reducing the number of features**

📖 **Explanation:** The first **k** principal components capture the **most important patterns**, reducing dimensionality while preserving information.

### ****11. How do we determine the optimal number of principal components to retain?****

A) By using a correlation matrix  
B) By plotting a scree plot and looking for the "elbow"  
C) By setting a fixed number of components  
D) By using k-Means clustering

✅ **Answer:** **B) By plotting a scree plot and looking for the "elbow"**

📖 **Explanation:** A **scree plot** shows the **variance explained** by each component. The "elbow point" indicates where adding more components has **diminishing returns**.

### ****12. Can PCA be applied to categorical data directly?****

A) Yes  
B) No, because PCA assumes numerical data  
C) Yes, but only if the categorical variables are one-hot encoded  
D) No, because PCA requires normally distributed data

✅ **Answer:** **C) Yes, but only if the categorical variables are one-hot encoded**

📖 **Explanation:** PCA requires **numerical** inputs. Categorical data must be **encoded** before applying PCA.

### ****13. In which machine learning applications is PCA commonly used?****

A) Image compression  
B) Feature extraction for classification  
C) Noise reduction  
D) All of the above

✅ **Answer:** **D) All of the above**

📖 **Explanation:** PCA is widely used for **image compression, feature extraction**, and **denoising**.

### ****14. What is a limitation of PCA?****

A) It cannot be used on large datasets  
B) It assumes that principal components are independent, which is not always true  
C) It increases the dimensionality of the dataset  
D) It requires labeled data

✅ **Answer:** **B) It assumes that principal components are independent, which is not always true**

📖 **Explanation:** PCA assumes that principal components are **uncorrelated**, but in some cases, dependencies can still exist.

### ****15. What happens if we apply PCA to a dataset where all features are already uncorrelated?****

A) PCA will still be effective  
B) PCA will not provide any dimensionality reduction  
C) PCA will create new correlated features  
D) PCA will increase accuracy

✅ **Answer:** **B) PCA will not provide any dimensionality reduction**

📖 **Explanation:** If the features are **already uncorrelated**, PCA will not find meaningful lower-dimensional representations.

### ****1. What does the directional derivative of a function measure?****

A) The rate of change of the function in a specific direction  
B) The second derivative of the function  
C) The gradient of the function at a point  
D) The minimum value of the function

✅ **Answer:** **A) The rate of change of the function in a specific direction**

📖 **Explanation:** The **directional derivative** tells us how a function changes **along a given direction** at a particular point.

### ****2. Which of the following is the mathematical formula for the directional derivative**** Duf(x)D\_u f(x)****?****

A) ∇f(x)⋅u\nabla f(x) \cdot u  
B) ∂f∂x\frac{\partial f}{\partial x}  
C) ∂f∂y\frac{\partial f}{\partial y}  
D) ∇f(x)×u\nabla f(x) \times u

✅ **Answer:** **A)** ∇f(x)⋅u\nabla f(x) \cdot u

📖 **Explanation:** The **directional derivative** in the direction of a unit vector uu is given by:

Duf(x)=∇f(x)⋅uD\_u f(x) = \nabla f(x) \cdot u

where ∇f(x)\nabla f(x) is the gradient vector and uu is the direction.

### ****3. In machine learning, directional derivatives are most commonly used in which optimization algorithm?****

A) k-Means Clustering  
B) Gradient Descent  
C) Naïve Bayes  
D) Decision Trees

✅ **Answer:** **B) Gradient Descent**

📖 **Explanation:** In **Gradient Descent**, we move in the direction **opposite** to the **gradient** (steepest ascent) to minimize the function.

### ****4. If the directional derivative**** Duf(x)D\_u f(x) ****is zero, what does it mean?****

A) The function is increasing in that direction  
B) The function is decreasing in that direction  
C) The function does not change in that direction  
D) The function has reached its maximum

✅ **Answer:** **C) The function does not change in that direction**

📖 **Explanation:** A **zero directional derivative** means there is **no change** in function value in that direction.

### ****5. The directional derivative is maximized in the direction of which vector?****

A) The unit vector  
B) The Hessian matrix  
C) The gradient vector  
D) The Laplacian operator

✅ **Answer:** **C) The gradient vector**

📖 **Explanation:** The **gradient vector** points in the **direction of steepest ascent**, meaning the directional derivative is maximized in this direction.

### ****6. What happens when we take the directional derivative in the opposite direction of the gradient?****

A) The function increases the fastest  
B) The function remains constant  
C) The function decreases the fastest  
D) The function diverges

✅ **Answer:** **C) The function decreases the fastest**

📖 **Explanation:** Moving **opposite** to the gradient vector leads to the **steepest descent**, which is the basis of **Gradient Descent**.

### ****7. In machine learning, directional derivatives help in which of the following tasks?****

A) Finding local minima of loss functions  
B) Computing eigenvalues of a matrix  
C) Measuring dataset bias  
D) Predicting categorical labels

✅ **Answer:** **A) Finding local minima of loss functions**

📖 **Explanation:** In **optimization**, we use **directional derivatives** to determine the fastest way to **reduce** a function’s value.

### ****8. What is the relationship between the directional derivative and the gradient of a function?****

A) The directional derivative is the dot product of the gradient and direction vector  
B) The directional derivative is the cross product of the gradient and direction vector  
C) The directional derivative is always greater than the gradient  
D) The directional derivative is independent of the gradient

✅ **Answer:** **A) The directional derivative is the dot product of the gradient and direction vector**

📖 **Explanation:** The **directional derivative** is calculated using:

Duf(x)=∇f(x)⋅uD\_u f(x) = \nabla f(x) \cdot u

which is the **dot product** of the gradient and the direction vector.

### ****9. If the gradient of a function is**** ∇f(x)=(3,4)\nabla f(x) = (3,4)****, what is the directional derivative in the direction of unit vector**** u=(0.6,0.8)u = (0.6, 0.8)****?****

A) 3.6  
B) 4.2  
C) 5  
D) 6

✅ **Answer:** **C) 5**

📖 **Explanation:**  
Using the formula:

Duf(x)=∇f(x)⋅uD\_u f(x) = \nabla f(x) \cdot u =(3,4)⋅(0.6,0.8)=(3×0.6)+(4×0.8)=1.8+3.2=5= (3,4) \cdot (0.6, 0.8) = (3 \times 0.6) + (4 \times 0.8) = 1.8 + 3.2 = 5

### ****10. What is the directional derivative when the direction vector is orthogonal to the gradient?****

A) Maximum  
B) Minimum  
C) Zero  
D) Undefined

✅ **Answer:** **C) Zero**

📖 **Explanation:** If the direction vector is **orthogonal** (perpendicular) to the gradient, the **dot product is zero**, meaning no change occurs in that direction.

### ****11. Which of the following is an extension of directional derivatives to multiple variables?****

A) Hessian matrix  
B) Jacobian matrix  
C) Taylor series  
D) Laplacian operator

✅ **Answer:** **B) Jacobian matrix**

📖 **Explanation:** The **Jacobian matrix** generalizes directional derivatives for **multivariable functions**.

### ****12. How is the Hessian matrix related to directional derivatives?****

A) It measures the second-order directional derivatives  
B) It calculates the first derivative  
C) It finds the loss function  
D) It computes the variance of a dataset

✅ **Answer:** **A) It measures the second-order directional derivatives**

📖 **Explanation:** The **Hessian matrix** consists of **second-order partial derivatives**, which measure **curvature** in different directions.

### ****13. If a function has a zero gradient at a point, what can we conclude?****

A) It is at a local maximum  
B) It is at a local minimum  
C) It is at a saddle point  
D) It could be any of the above

✅ **Answer:** **D) It could be any of the above**

📖 **Explanation:** A **zero gradient** indicates a **critical point**, which could be a **minimum, maximum, or saddle point**.

### ****14. Directional derivatives are useful in training which type of models?****

A) k-Means clustering  
B) Neural networks  
C) Decision trees  
D) Rule-based models

✅ **Answer:** **B) Neural networks**

📖 **Explanation:** In **deep learning**, directional derivatives are used in **backpropagation** to update weights.

### ****15. In which scenario is the concept of directional derivatives NOT useful?****

A) Training deep learning models  
B) Finding optimal hyperparameters  
C) Computing gradients for non-differentiable functions  
D) Optimizing convex functions

✅ **Answer:** **C) Computing gradients for non-differentiable functions**

📖 **Explanation:** Directional derivatives require **differentiable functions**. If a function is **non-differentiable**, subgradients or alternative methods must be used.

### ****MCQs on Backpropagation and Automatic Differentiation with Solutions****

Backpropagation and **automatic differentiation** are fundamental techniques in **deep learning** and optimization. These MCQs will test your understanding of how these methods work in **training neural networks**.

### ****1. What is the primary purpose of backpropagation in neural networks?****

A) To compute the output of a neural network  
B) To update the weights by computing gradients of the loss function  
C) To initialize the neural network  
D) To perform feature selection

✅ **Answer:** **B) To update the weights by computing gradients of the loss function**

📖 **Explanation:** Backpropagation computes gradients using the **chain rule**, which helps **adjust weights** to minimize loss.

### ****2. Backpropagation is based on which mathematical rule?****

A) Chain rule of differentiation  
B) Bayes' theorem  
C) Central limit theorem  
D) Markov property

✅ **Answer:** **A) Chain rule of differentiation**

📖 **Explanation:** Backpropagation relies on the **chain rule** to compute **gradients** for each layer in a neural network.

### ****3. What is the main drawback of backpropagation?****

A) It is only used in supervised learning  
B) It cannot handle non-differentiable functions  
C) It does not work for large datasets  
D) It requires a lot of labeled data

✅ **Answer:** **B) It cannot handle non-differentiable functions**

📖 **Explanation:** Backpropagation requires the **loss function** and **activation functions** to be **differentiable**.

### ****4. In backpropagation, what role does the learning rate play?****

A) It controls how much weights are updated  
B) It determines the number of hidden layers  
C) It selects the best features in the dataset  
D) It prevents overfitting

✅ **Answer:** **A) It controls how much weights are updated**

📖 **Explanation:** The **learning rate** controls the **step size** of weight updates during training.

### ****5. Which of the following problems can occur in backpropagation?****

A) Vanishing gradients  
B) Exploding gradients  
C) Overfitting  
D) All of the above

✅ **Answer:** **D) All of the above**

📖 **Explanation:**

**Vanishing gradients** → Gradients become too small, slowing learning.

**Exploding gradients** → Gradients grow too large, leading to instability.

**Overfitting** → The model learns noise instead of general patterns.

### ****6. How does ReLU activation help in backpropagation?****

A) It prevents gradient vanishing  
B) It makes neural networks faster  
C) It increases the depth of networks  
D) It helps in data normalization

✅ **Answer:** **A) It prevents gradient vanishing**

📖 **Explanation:** The **ReLU (Rectified Linear Unit)** activation function avoids the **vanishing gradient problem** by keeping gradients **non-zero** for positive inputs.

### ****7. Automatic differentiation is mainly used to compute:****

A) Partial derivatives  
B) Integrals  
C) Matrix multiplication  
D) Random noise

✅ **Answer:** **A) Partial derivatives**

📖 **Explanation:** **Automatic differentiation (AutoDiff)** computes **derivatives** efficiently, which is crucial for **optimizing deep learning models**.

### ****8. Which technique is commonly used for automatic differentiation in deep learning?****

A) Symbolic differentiation  
B) Numerical differentiation  
C) Computational graphs  
D) Finite difference method

✅ **Answer:** **C) Computational graphs**

📖 **Explanation:** **Computational graphs** track operations to efficiently compute derivatives using **automatic differentiation**.

### ****9. What is the main advantage of automatic differentiation over numerical differentiation?****

A) It avoids rounding errors  
B) It is more accurate and efficient  
C) It can handle non-differentiable functions  
D) It requires fewer training examples

✅ **Answer:** **B) It is more accurate and efficient**

📖 **Explanation:** **Automatic differentiation** is **faster and more precise** than numerical differentiation, which suffers from **rounding errors**.

### ****10. Which deep learning framework does NOT support automatic differentiation?****

A) TensorFlow  
B) PyTorch  
C) NumPy  
D) JAX

✅ **Answer:** **C) NumPy**

📖 **Explanation:** NumPy does not have built-in **automatic differentiation**, whereas TensorFlow, PyTorch, and JAX do.

### ****11. Which type of automatic differentiation is used in backpropagation?****

A) Forward mode differentiation  
B) Backward mode differentiation  
C) Finite difference method  
D) Central difference method

✅ **Answer:** **B) Backward mode differentiation**

📖 **Explanation:** **Backward mode differentiation** computes gradients **efficiently** for functions with **many inputs and one output**, like **neural networks**.

### ****12. What is the computational complexity of backpropagation in a fully connected neural network with L layers?****

A) O(L)O(L)  
B) O(L2)O(L^2)  
C) O(n⋅L)O(n \cdot L)  
D) O(Ln2)O(Ln^2)

✅ **Answer:** **C)** O(n⋅L)O(n \cdot L)

📖 **Explanation:** Backpropagation requires computing gradients for **each layer**, leading to a complexity of O(n⋅L)O(n \cdot L).

### ****13. What is a key benefit of using PyTorch for automatic differentiation?****

A) It uses static computation graphs  
B) It supports dynamic computation graphs  
C) It does not require GPUs  
D) It only supports small neural networks

✅ **Answer:** **B) It supports dynamic computation graphs**

📖 **Explanation:** PyTorch uses **dynamic computation graphs**, allowing more **flexibility** in model design.

### ****14. Why is automatic differentiation preferred over symbolic differentiation?****

A) It avoids computing large symbolic expressions  
B) It is less accurate  
C) It requires more memory  
D) It is only used for small functions

✅ **Answer:** **A) It avoids computing large symbolic expressions**

📖 **Explanation:** **Symbolic differentiation** can produce **very large expressions**, while **automatic differentiation** computes derivatives more **efficiently**.

### ****15. If we change the architecture of a neural network, what happens to the computation graph used in automatic differentiation?****

A) It remains the same  
B) It must be recompiled  
C) It does not affect training  
D) The loss function is unchanged

✅ **Answer:** **B) It must be recompiled**

📖 **Explanation:** The **computation graph** is **dynamically updated** whenever the model structure changes.