1. **What is the primary goal of numerical optimization in machine learning?**
   - A) To find exact solutions for all problems
   - B) To approximate solutions that generalize well to unseen data
   - C) To eliminate the need for training data
   - D) To avoid using algorithms like gradient descent

   Answer: B
2. **Why is an analytic solution often preferred in machine learning?**
   - A) It is computationally expensive
   - B) It provides exact and fast solutions
   - C) It works for all data matrices, including non-square ones
   - D) It guarantees generalization to new data

   Answer: B
3. **What is a limitation of the analytic solution $\theta = X^{-1} \cdot y$?**
   - A) It only works for square matrices X
   - B) It cannot handle linear models
   - C) It is robust to outliers
   - D) It requires no initialization

   Answer: A
4. **For an overdetermined system (m > n), what algebraic method is used to find an approximate solution?**
   - A) Least norm method: $\theta = X^T(XX^T)^{-1}y$
   - B) Least squares method: $\theta = (X^TX)^{-1}X^Ty$
   - C) Matrix inversion: $\theta = X^{-1}y$
   - D) Gradient descent

   Answer: B
5. **What is a disadvantage of the algebraic solution for large n?**
   - A) It cannot handle nonlinear relationships
   - B) Computing $(X^TX)^{-1}$ becomes expensive
   - C) It requires labeled data
   - D) It is insensitive to outliers

   Answer: B

6. **Which of the following is a step in the numerical solution process for linear regression?**

- A) Randomly shuffle data once
- B) Initialize parameters to zero
- C) Use exact derivatives instead of gradients
- D) Skip the loss function evaluation

Answer: B

7. **The L1 norm (Manhattan norm) is directly related to which loss function?**

- A) Mean Squared Error (MSE)
- B) Mean Absolute Error (MAE)
- C) Cross-Entropy Loss
- D) Hinge Loss

Answer: B

8. **What is the formula for the L2 norm (Euclidean norm)?**

- A) $\sum_{i=1}^{N} |x_i|$
- B) $\sum_{i=1}^{N} |x_i|^2$
- C) $\max_{i=1}^{N} |x_i|$
- D) $\frac{1}{N}\sum_{i=1}^{N} x_i$

Answer: B

9. **Why is MAE robust to outliers?**

- A) It squares the errors, amplifying outliers
- B) It uses absolute values, reducing outlier impact
- C) It ignores errors beyond a threshold
- D) It normalizes the data first

Answer: B

10. **What is a drawback of MSE compared to MAE?**

- A) It is computationally expensive
- B) It is sensitive to outliers
- C) It cannot be used for regression
- D) It has a constant gradient

Answer: B

11. **What does the gradient represent in optimization?**
   - A) The error value
   - B) The slope of the loss function
   - C) The learning rate
   - D) The number of iterations

   Answer: B
12. **How is the derivative of $f(x)=x^2$ calculated?**
   - A) $2x$
   - B) x
   - C) 2
   - D) x^2

   Answer: A
13. **In gradient descent, how are parameters updated?**
   - A) $\theta_{t+1}=\theta_t+\alpha\cdot\text{gradient}$
   - B) $\theta_{t+1}=\theta_t-\alpha\cdot\text{gradient}$
   - C) $\theta_{t+1}=\theta_t\cdot\text{gradient}$
   - D) $\theta_{t+1}=\theta_t/\text{gradient}$

   Answer: B
14. **What happens if the learning rate α is too large?**
   - A) The model converges slowly
   - B) The model may overshoot the minimum
   - C) The gradient becomes zero
   - D) The loss function becomes convex

   Answer: B
15. **What is the purpose of the learning rate in gradient descent?**
   - A) To control the size of parameter updates
   - B) To compute the loss function
   - C) To initialize the parameters
   - D) To normalize the data

   Answer: A
16. **What is a local minimum in optimization?**

- A) The lowest point in the entire function domain
- B) A point where the function value is smaller than nearby points but not necessarily the smallest overall
- C) A saddle point
- D) A point with zero gradient

Answer: B

17. **Why is convexity important in optimization?**

- A) It guarantees multiple local minima
- B) It ensures the existence of a single global minimum
- C) It makes the loss function non-differentiable
- D) It eliminates the need for gradient descent

Answer: B

18. **Which loss function is commonly used in logistic regression?**

- A) Mean Squared Error (MSE)
- B) Cross-Entropy Loss
- C) L1 Norm
- D) Hinge Loss

Answer: B

19. **What is the main challenge of non-convex optimization?**

- A) It always converges to the global minimum
- B) It may get stuck in local minima or saddle points
- C) It requires no gradient computation
- D) It is faster than convex optimization

Answer: B

20. **What is the intuition behind gradient descent?**

- A) Climbing a hill to find the maximum
- B) Taking steps in the direction of the steepest descent to minimize the loss
- C) Randomly searching the parameter space
- D) Using algebraic solutions exclusively

Answer: B

---

21. **What is the primary purpose of calculating the gradient of a multivariable function in optimization?**
   - A) To find the maximum value of the function
   - B) To determine the direction of steepest ascent
   - C) To eliminate the need for partial derivatives
   - D) To avoid using contour plots

   Answer: B

22. **The gradient vector ∇f for a function $f(x,y)$ is defined as:**
   - A) $[\partial x \partial f, \partial y \partial f]$
   - B) $[\partial y \partial f, \partial x \partial f]$
   - C) $\partial x \partial f + \partial y \partial f$
   - D) $\partial x \partial f - \partial y \partial f$

   Answer: A

23. **In gradient descent, the direction of steepest descent is:**
   - A) The same as the gradient vector
   - B) The opposite of the gradient vector
   - C) Perpendicular to the gradient vector
   - D) Independent of the gradient vector

   Answer: B

24. **What does a contour plot visually represent?**
   - A) The gradient vector at each point
   - B) Lines connecting points of equal function value
   - C) The partial derivatives of the function
   - D) The learning rate of gradient descent

   Answer: B

25. **For the function $z=f(x,y)=x2+y2$, the gradient at any point (x,y) is:**
   - A) $[2x,2y]$
   - B) [x^2, y^2]
   - C) $[2,2]$
   - D) $[0,0]$

   Answer: A

26. **What is the hypothesis function for single-variable linear regression?**

- A) $h(x)=\theta_0+\theta_1 x$
- B) $h_\theta(x)=\theta_0+\theta_1 x$
- C) $h_\theta(x)=\theta_1 x$
- D) $h_\theta(x)=\theta_0 x+\theta_1$

Answer: B

27. **The cost function $J(\theta_0,\theta_1)$ for linear regression is given by:**

- A) $\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})$
- B) $\frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2$
- C) $\sum_{1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2$
- D) $\frac{1}{m}\sum_{i=1}^{m}|h_\theta(x^{(i)})-y^{(i)}|$

Answer: B

28. **In gradient descent for linear regression, how are the parameters θ_0 and θ_1 updated?**

- A) Sequentially, one after the other
- B) Simultaneously
- C) Only θ_0 is updated
- D) Only θ_1 is updated

Answer: B

29. **What is the role of the learning rate (α) in gradient descent?**

- A) To compute the cost function
- B) To control the size of parameter updates
- C) To initialize the parameters
- D) To normalize the data

Answer: B

30. **Why is feature scaling important in gradient descent?**

- A) To ensure all features contribute equally to the gradient
- B) To eliminate the need for a learning rate
- C) To reduce the number of iterations to convergence
- D) Both A and C

Answer: D

31. **Which of the following is a disadvantage of using a very large learning rate in gradient descent?**
   - A) The model converges too slowly
   - B) The model may overshoot the minimum
   - C) The gradient becomes zero
   - D) The cost function becomes non-convex

   Answer: B
32. **For multivariate linear regression, the hypothesis function is:**
   - A) $h_\theta(x)=\theta_0+\theta_1 x_1$
   - B) $h_\theta(x)=\theta_0 x_0+\theta_1 x_1+\cdots+\theta_n x_n$
   - C) $h_\theta(x)=\theta_0+\theta_1 x_1 2$
   - D) $h_\theta(x)=\theta_0\cdot\theta_1 x_1$

   Answer: B
33. **In batch gradient descent, when are the parameters updated?**
   - A) After each training example
   - B) After processing all training examples
   - C) Only once at the beginning
   - D) Randomly during training

   Answer: B
34. **Which feature scaling method scales features to the range [0, 1]?**
   - A) Mean normalization
   - B) Min-Max normalization
   - C) Robust scaling
   - D) Standardization

   Answer: B
35. **Which algorithm is most affected by the scale of features?**
   - A) Linear regression with gradient descent
   - B) Decision trees
   - C) Random forests
   - D) All of the above

   Answer: A

36. **What is the main advantage of batch gradient descent?**
   - A) It updates parameters frequently
   - B) It guarantees convergence to the global minimum for convex functions
   - C) It is computationally efficient for large datasets
   - D) It does not require a learning rate

   Answer: B
37. **What does the gradient vector point to for a multivariable function?**
   - A) The direction of steepest descent
   - B) The direction of steepest ascent
   - C) The global minimum
   - D) A saddle point

   Answer: B
38. **Which of the following is true about contour plots?**
   - A) They show the gradient at each point
   - B) They connect points with the same function value
   - C) They are only used for convex functions
   - D) They replace the need for gradient descent

   Answer: B
39. **What is the primary goal of gradient descent in optimization?**
   - A) To maximize the cost function
   - B) To minimize the cost function
   - C) To compute partial derivatives
   - D) To visualize the loss function

   Answer: B

---

41. **What is the primary disadvantage of Batch Gradient Descent (Vanilla GD)?**
   - A) Frequent parameter updates
   - B) Slow convergence for large datasets
   - C) High sensitivity to learning rate

- D) Inability to handle non-convex functions

  Answer: B

42. **Stochastic Gradient Descent (SGD) updates parameters:**

- A) After processing the entire dataset
- B) For each individual training example
- C) Only once per epoch
- D) Using a fixed learning rate

  Answer: B

43. **Mini-batch GD balances trade-offs between:**

- A) Speed (SGD) and stability (Batch GD)
- B) Memory usage and computational complexity
- C) Convex and non-convex optimization
- D) Local and global minima

  Answer: A

44. **A common batch size for Mini-batch GD is:**

- A) 1
- B) 32 (power of 2)
- C) Equal to the dataset size
- D) Randomly chosen

  Answer: B

45. **The main challenge of SGD is:**

- A) High computational cost per epoch
- B) Noisy updates causing oscillations
- C) Inability to use vectorization
- D) Both B and C

  Answer: D

46. **Choosing a learning rate too large in GD may cause:**

- A) Slow convergence
- B) Overshooting the minimum
- C) Vanishing gradients

- D) Exact solutions

  Answer: B

47. **To select a good learning rate, you should:**

- A) Always use 0.01
- B) Test values (e.g., 0.001, 0.01, 0.1) and plot cost vs. epochs
- C) Use the same rate for all models
- D) Ignore the cost function

  Answer: B

48. **Local minima are problematic in GD because:**

- A) They are global optima
- B) The algorithm may converge to suboptimal solutions
- C) They only occur in convex functions
- D) They accelerate convergence

  Answer: B

49. **Vanishing gradients occur when:**

- A) Gradients become extremely large
- B) Gradients approach zero, slowing learning
- C) The learning rate is too high
- D) Using Mini-batch GD

  Answer: B

50. **Exploding gradients are common in:**

- A) Shallow networks
- B) Deep networks with multiplicative gradients
- C) Linear regression
- D) Convex functions

  Answer: B

51. **Momentum-based GD addresses:**

- A) Slow convergence in flat regions
- B) Exact gradient calculations
- C) Linear separability

- D) Feature scaling

  Answer: A

52. **The momentum term (γ) typically ranges:**

- A) 0 to 1
- B) -1 to 1
- C) 1 to 10
- D) Arbitrary values

  Answer: A

53. **Momentum update rule includes:**

- A) Only the current gradient
- B) A weighted sum of past gradients
- C) Fixed step size
- D) Random noise

  Answer: B

54. **A drawback of Momentum GD is:**

- A) Oscillations near minima
- B) Inability to handle non-convex functions
- C) High memory usage
- D) Requires exact gradients

  Answer: A

55. **Nesterov Accelerated GD (NAG) improves Momentum GD by:**

- A) Computing gradients at a "look-ahead" point
- B) Using larger learning rates
- C) Ignoring past gradients
- D) Eliminating the need for learning rates

  Answer: A

56. **In NAG, the correction step is applied:**

- A) Before the momentum jump
- B) After the momentum jump
- C) Only at initialization

- D) Randomly

  Answer: B

57. **NAG reduces oscillations by:**

- A) Slowing convergence
- B) Correcting gradients after momentum steps
- C) Increasing batch size
- D) Using second-order derivatives

  Answer: B

58. **For non-convex functions, GD variants may get stuck in:**

- A) Global minima only
- B) Saddle points or local minima
- C) Exact solutions
- D) Linear regions

  Answer: B

59. **Feature scaling helps GD by:**

- A) Ensuring features contribute equally to gradients
- B) Reducing dataset size
- C) Eliminating the need for learning rates
- D) Making all features binary

  Answer: A

60. **The key advantage of Mini-batch GD over SGD is:**

- A) Fewer hardware requirements
- B) Smoother convergence due to reduced noise
- C) No need for gradient calculations
- D) Fixed parameter updates

  Answer: B

---

61. **Adagrad adapts the learning rate based on:**

- A) The magnitude of the current gradient
- B) The accumulated sum of squared past gradients
- C) The number of training epochs

- D) Random noise

  Answer: B

62. **The primary advantage of Adagrad is:**

- A) Faster convergence for dense features
- B) Larger updates for sparse features
- C) Fixed learning rates
- D) Elimination of gradient noise

  Answer: B

63. **A key disadvantage of Adagrad is:**

- A) Learning rate becomes too small over time
- B) Inability to handle sparse data
- C) High computational cost per update
- D) Requires manual tuning of momentum

  Answer: A

64. **The term ϵ in Adagrad's update rule is used to:**

- A) Accelerate convergence
- B) Prevent division by zero
- C) Increase the learning rate
- D) Introduce randomness

  Answer: B

65. **Adagrad's update rule is:**

- A) $w_{t+1}=w_t-\eta\nabla w_t$
- B) $w_{t+1}=w_t-{\scriptstyle G_t+\epsilon\eta}\nabla w_t$
- C) $w_{t+1}=w_t-\eta\nabla w_t$
- D) $w_{t+1}=w_t-\eta\sum\nabla w_t$

  Answer: B

66. **RMSProp improves Adagrad by:**

- A) Using a fixed learning rate
- B) Exponentially decaying the accumulated gradients
- C) Ignoring past gradients

- D) Increasing the learning rate for dense features

  Answer: B

67. **The parameter β in RMSProp controls:**

- A) The weight of past gradients in the moving average
- B) The initial learning rate
- C) The sparsity of features
- D) The number of epochs

  Answer: A

68. **RMSProp's update rule includes:**

- A) A momentum term
- B) An exponentially weighted average of squared gradients
- C) A fixed denominator
- D) No learning rate

  Answer: B

69. **A typical value for β in RMSProp is:**

- A) 0.1
- B) 0.5
- C) 0.9
- D) 1.0

  Answer: C

70. **RMSProp prevents the rapid growth of $v(t)$ by:**

- A) Resetting v(t)$v(t)$ to zero periodically
- B) Using only the current gradient
- C) Weighting past gradients less heavily
- D) Ignoring small gradients

  Answer: C

71. **Adam combines the ideas of:**

- A) Momentum and Adagrad
- B) Momentum and RMSProp
- C) SGD and Mini-batch GD

- D) NAG and Adagrad

  Answer: B

72. **Adam's update rule includes:**

- A) Only adaptive learning rates
- B) Adaptive learning rates and adaptive momentum
- C) Fixed momentum
- D) No bias correction

  Answer: B

73. **The bias correction terms in Adam (m^ and v^) are used to:**

- A) Reduce noise in gradients
- B) Correct initial underestimation of moments
- C) Increase the learning rate
- D) Eliminate sparse features

  Answer: B

74. **Default values for Adam's hyperparameters are:**

- A) $\beta_1=0.5$, $\beta_2=0.9$
- B) $\beta_1=0.9$, $\beta_2=0.999$
- C) $\beta_1=0.99$, $\beta_2=0.9999$
- D) $\beta_1=0.1$, $\beta_2=0.01$

  Answer: B

75. **Adam is preferred for training deep neural networks because it:**

- A) Requires no hyperparameter tuning
- B) Combines adaptive learning rates and momentum
- C) Uses only first-order gradients
- D) Ignores past gradients

  Answer: B

76. **The Exponentially Weighted Moving Average (EWMA) is used in:**

- A) Adagrad only
- B) RMSProp and Adam
- C) SGD only

- D) Momentum GD only

Answer: B

77. **For $\beta=0.9$, EWMA averages over approximately:**

- A) 2 time steps
- B) 10 time steps
- C) 50 time steps
- D) 100 time steps

Answer: B

78. **In Adam, m_t represents:**

- A) The sum of squared gradients
- B) The exponentially weighted average of gradients
- C) The learning rate
- D) The bias correction term

Answer: B

79. **The primary challenge addressed by adaptive optimization methods is:**

- A) High computational cost
- B) Feature scaling
- C) Non-uniform gradients across parameters
- D) Large batch sizes

Answer: C

80. **Adam's update rule for parameter w is:**

- A) $w_{t+1}=w-\eta\nabla w_t$
- B) $w_{t+1}=w-\frac{\eta\hat{m}_t}{\hat{v}_t+\epsilon}$
- C) $w_{t+1}=w-\eta w_t$
- D) $w_{t+1}=w-\eta\nabla w_t$

Answer: B

---

81. **Newton's method is a ______ optimization technique.**

- A) First-order
- B) Second-order
- C) Zero-order

- D) Stochastic

  Answer: B

82. **The key advantage of Newton's method over gradient descent is:**

- A) No need to compute gradients
- B) Faster convergence (quadratic under ideal conditions)
- C) Works better for non-convex functions
- D) Lower computational cost per iteration

  Answer: B

83. **In Newton's method, the update rule for a single-variable function is:**

- A) $x_{k+1} = x_k - \alpha f(x_k)$
- B) $x_{k+1} = x_k - f'(x_k) / f(x_k)$
- C) $x_{k+1} = x_k - f'(x_k)$
- D) $x_{k+1} = x_k + f'(x_k) / f(x_k)$

  Answer: B

84. **For a multivariable function, Newton's method uses the ______ to compute the update step.**

- A) Gradient and Hessian
- B) Gradient only
- C) Learning rate and momentum
- D) Exponentially weighted average

  Answer: A

85. **The Hessian matrix is:**

- A) A vector of first derivatives
- B) A matrix of second partial derivatives
- C) A scalar value
- D) A diagonal matrix of gradients

  Answer: B

86. **A major drawback of Newton's method is:**

- A) It requires manual tuning of the learning rate
- B) High computational cost of inverting the Hessian
- C) It cannot handle convex functions

- D) It ignores gradient information

  Answer: B

87. **Newton's method may converge to a saddle point if:**

- A) The Hessian is positive definite
- B) The Hessian is indefinite
- C) The gradient is zero
- D) The learning rate is too small

  Answer: B

88. **A saddle point in optimization is characterized by:**

- A) Zero gradient and indefinite Hessian
- B) Zero gradient and positive definite Hessian
- C) Non-zero gradient and zero Hessian
- D) Non-zero gradient and negative definite Hessian

  Answer: A

89. **Newton's method is guaranteed to converge to a local minimum if:**

- A) The Hessian is positive definite at all points
- B) The learning rate is small
- C) The function is non-convex
- D) The gradient is stochastic

  Answer: A