Section 3: Systems of Linear Equations (Expanded)

Applications of Linear Systems in Machine Learning

1. **Linear Regression**:
   - The equation $y = X\beta + \epsilon$ (where X is the feature matrix, β are coefficients, and ϵ is noise) can be solved to find β using $\beta = (X^T X)^{-1} X^T y$.
2. **Network Flows**:
   - Optimization problems in transportation or communication networks often reduce to solving linear systems.

Section 4: Eigenvalues and Eigenvectors

Eigenvalues and Eigenvectors are foundational for understanding transformations in vector spaces, with applications in **dimensionality reduction**, **PCA**, and more.

Core Concepts

1. **Eigenvalues and Eigenvectors**:
   - For a square matrix A, if $A\vec{v} = \lambda\vec{v}$, then:
     - $\vec{v}$: Eigenvector (unchanged direction under transformation).
     - λ: Eigenvalue (scales the eigenvector).
2. **Characteristic Equation**:
   - To find eigenvalues, solve $\det(A - \lambda I) = 0$.
3. **Diagonalization**:

- A matrix A is diagonalizable if $A = PDP^{-1}$, where D is a diagonal matrix of eigenvalues, and P contains eigenvectors as columns.

Applications in Machine Learning

1. **Principal Component Analysis (PCA)**:
   - Reduces dimensionality by finding eigenvectors of the covariance matrix, selecting those with the largest eigenvalues.
2. **Spectral Clustering**:
   - Uses eigenvectors of the graph Laplacian to group data points.

Example

1. Matrix:

$$A = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}.$$

2. Find eigenvalues (λ):
   - Solve $\det(A - \lambda I) = 0$:

$$\det \begin{bmatrix} 4 - \lambda & 1 \\ 2 & 3 - \lambda \end{bmatrix} = 0.$$

   Results: $\lambda_1 = 5$, $\lambda_2 = 2$.

3. Eigenvectors:
   - For $\lambda = 5$, solve $(A - 5I)\vec{v} = 0$ to get $\vec{v}_1 = [1, 2]^T$.

- For $\lambda = 2$, solve $(A - 2I)\vec{v} = 0$ to get $\vec{v}_2 = [-1, 1]^T$.

Section 5: Singular Value Decomposition (SVD)

SVD is a matrix factorization method that generalizes eigenvalue decomposition to non-square matrices.

Core Concepts

1. **Definition**:
   - For a matrix A of size $m \times n$:

   $$A = U\Sigma V^T,$$

   where:

     - U: Orthogonal matrix of size $m \times m$ (left singular vectors).
     - Σ: Diagonal matrix of singular values (σ_i).
     - V^T: Transpose of an orthogonal matrix of size $n \times n$ (right singular vectors).
2. **Properties**:
   - Singular values σ_i are the square roots of eigenvalues of A^TA or AA^T.
   - U and V provide the basis for column and row spaces, respectively.

Applications in Machine Learning

1. **Dimensionality Reduction**:
   - In Latent Semantic Analysis (LSA) for text data, SVD reduces matrix dimensionality.
2. **Image Compression**:
   - SVD approximates an image using the largest singular values.

Example

1. Matrix:

$$A = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}.$$

2. Compute $A^T A$:

$$A^T A = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}^T \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix} = \begin{bmatrix} 13 & 12 \\ 12 & 13 \end{bmatrix}.$$

3. Find eigenvalues of $A^T A$: $\lambda_1 = 25, \lambda_2 = 1$.
4. Singular values: $\sigma_1 = 5, \sigma_2 = 1$.

Section 6: Optimization Techniques

Optimization is the backbone of machine learning, helping minimize or maximize objective functions.

Core Concepts

1. **Gradient Descent**:
   - Iteratively updates parameters in the direction of the negative gradient:
     $$\theta \leftarrow \theta - \eta \nabla J(\theta),$$
     where:
     - θ: Parameters,
     - η: Learning rate,
     - $J(\theta)$: Cost function.
2. **Convex Optimization**:
   - If $J(\theta)$ is convex, gradient descent converges to the global minimum.
3. **Lagrange Multipliers**:
   - Used for optimization with constraints.

Applications in Machine Learning

1. **Training Neural Networks**:
   - Backpropagation employs gradient descent to minimize loss functions.
2. **Support Vector Machines (SVMs)**:
   - Quadratic optimization is used to find the maximum-margin hyperplane.

Example

Minimize:

$$J(\theta) = (x-2)^2 + (y-3)^2.$$

1. Gradient:

$$\nabla J = [2(x-2), 2(y-3)].$$

2. Iterative updates:

$$x \leftarrow x - \eta \cdot 2(x-2), \quad y \leftarrow y - \eta \cdot 2(y-3).$$

Section 7: Probability and Statistics in Machine Learning

Core Concepts

1. **Bayes' Theorem**:
   - Used for probabilistic reasoning:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

2. **Probability Distributions**:
   - **Gaussian Distribution**:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

3. **Expectations and Variance**:
   - Expectation: $\mathrm{E}[X] = \sum_x xP(x)$.
   - Variance: $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$.

Applications in Machine Learning

1. **Naive Bayes Classifier**:
   - Based on Bayes' theorem.
2. **Clustering**:
   - GMMs (Gaussian Mixture Models) use probabilistic distributions.

Example

Calculate the likelihood of $A|B$:

- Given $P(B|A) = 0.8, P(A) = 0.5, P(B) = 0.6$:

$$P(A|B) = \frac{0.8 \cdot 0.5}{0.6} = 0.6667.$$