Linear Algebra

Basics

Linearity

- the operation preserving addition and scalar multiplication
- allows models to be analyzed using linear algebra
    - makes optimization more tractable
    - linear models → convex objectives
- nonlinearity can be added via activation functions

Basis

- a set of linearly independent vectors spanning a vector space
  - defines a coordinate system for representing vectors
  - special bases: eigenvectors, singular vectors
    - eigen decomposition ≃ change of basis

Rank

- dimension of the space spanned by the colums (or rows) of a matrix
  - intrinsic dimensionality of a transformation
- number of linearly independent columns (or rows)

Hessian

- matrix of second-order partial derivatives
- can capture curvature information of the loss surface
- **Symmetric matrix**
  - equal to its transpose
  - all eigenvalues → real number
  - all eigenvectors → orthogonal basis

Pseudo-Inverse

- applicable to any non-square matrix
  - standard inverse matrix → full rank & square matrix
- defined via SVD (A^+)
  - $A = U\Sigma V^T$

    $A^+ = V\Sigma^+ U^T$
  - Σ: invert only signular value
- gives a solution to a minimum-square problem

- $\min_x \|Ax - b\|^2 \Rightarrow x = A^+ b$

Determinant

- a scalar value associated with a square matrix
- measures volume scaling of a linear transformation
- determinant = 0 ↔ rank-deficient ↔ non-invertible
  - from eigen decomposition → determinant = product of eigenvalues

Taylor Expansion

- local approximation of a function using its derivatives
  - around a point using polynomial terms
- First-order (linear approximation)
  - $f(x) \approx f(x_0) + \nabla f(x_0)^\top (x - x_0)$
  - giving the direction of steepest local change
- Second-order (quadratic approximation)
  - capture curvature by using Hessian
  - **Newton method**
    - $w_{t+1} = w_t - H^{-1} \nabla L(w_t)$
    - Hessian matrix H of loss function → eigenvalue ≃ steepness
      - able to control the step size
    - steep ↑ → size ↓
  - **v.s. Gradient descent**
    - gradient descent: single learning rate for all directions
    - advantage: zig-zag ↓ & quadratic convergence ↑
    - drawback: computiation of second-order matrix ↑

Norm

- measure the magnitude or length of a vector
- **L2 norm**
  - $\|w\|_2^2 = \sum_i w_i^2$
  - Euclidean distance
  - smooth → differentiable
  - discourage large weights → stable optimization
    - make magnitudes of parameters smaller
  - may be sensitive to outliers
  - Gaussian prior (MAP perspective)
- **L1 norm**
  - $\|w\|_1 = \sum_i |w_i|$
  - promote sparsity
    - useful where there is fewer features
  - implicit feature selection → robust to outliers but unstable optimization
  - non-differentiable at 0
  - Laplace prior (MAP persepctive)

Cosine Similarity

$\cos(x, y) = \frac{x^\top y}{\|x\|\|y\|}$

- measure how aligned or related two vectors are
  - focus on directions (rather than magnitude)
- scale-invariant: invariant to absolute scale
- valid range from -1 to 1
- **v.s. dot product**
  - doc product is affected by magnitude
- cons
  - unstable if vector norm is very small

Matrix Decomposition

- decomposing a matrix into simpler components
- reveals principal directions & effective dimensionality
- useful for understanding optimization geometry & parameter efficiency

Diagonalization

$A = PDP^{-1}$

- transforming a matrix into a diagonal form via a similarity transform
- diagonal matrix D
  - contains scaling factors along independent directions
- change of coordinate system
  - where the linear transformation becomes independent per dimension
- diagonalizable: the matrix has a full set of linearly independent eigenvectors
- allow per-direction analysis of curvature and step size

Eigen Decomposition

diagonalizable

- $A = Q\Lambda Q^{-1}$
  - applicable only to square matrices

$\det(A-\lambda I)=0 \rightarrow \lambda = ?$

$(A-\lambda I)v = 0 \rightarrow v = ?$

- similar to the change of bais
  - transforms the problem into independent directions
- **Eigenvectors**: principal directions of transformation
- **Eigenvalues**: scaling along each direction
  - Hessian analysis: eigenvalues ≃ curvature
    - second-order curvature from Taylor expansion
    - along principal directions of the loss surface

        - large eigenvalue → steep direction
    - **Spectral Theorem**
        - all of the real symmetric matrix can be decomposed by eigen decomposition
            - where leftmost and rightmost matrices have orthonormal eigenvector
        - Hessian eigen decomposition

Singular Value Decomposition (SVD)

- $W = U\Sigma V^T$
- sum of rank-1 component matrices, each weighted by a singular value
    - applicable to any matrix
    - singular vectors U, V: having orthonormal columns
- **Singular values**: importance (energy) of each rank-1 component
- **Singular vectors**: dominant input & output directions
- **Truncated SVD**
    - keep top-r singular values and corresponding singular vectors
        - discard small singular values → often corresponding to noise
        - suppress directions associated with small singular values
    - optimal low-rank approximation that minimizes mean squared reconstruction error
- widely used for compression and dimensionality reduction
    - e.g., PCA (Principal Components Analysis)

Low-Rank Adaptation (LoRA)

- assumption: weight updates lie in a low-rank subspace
- instead of updating full weight matrix
    - freezing pretrained weight matrix W
    - updating $\Delta W = BA$ with a small rank

- provides convergence guarantees for first-order methods

Principal Component Analysis (PCA)

- a linear dimensionality reduction technique
- addresses the observation that
  - high-dimensional data often lies near a low-dimensional subspace
- **Goals**
  - represent data with fewer dimensions
  - while preserving as much information as possible
- PCA finds a low-dimensional linear subspace that maximizes variance
- **Data preprocessing**
  - $X \leftarrow X - \mu$
  - centering the data
    - assuming zero-mean data

+Standardization

- **Covariance Matrix**
  - $\Sigma = \frac{1}{N} X^\top X$
  - measures correlations between features
  - $\mathrm{Var}(Xw) = w^\top \Sigma w$
  - variance along a direction w
- **Optimization**
  - first principal component solves:
    - $\max_{\|w\|=1} w^\top \Sigma w$
    - maximize projected variance
    - unit-norm constraint avoids trivial solutions

covariance matrix에 관한 eigen decomposition

  - solution:
    - eigenvector of Σ with the largest eigenvalue

가장 넓은 분산을 갖는 eigenvalue의 축으로 projection