Contents

Preface xi

1 Vectors 1
- 1.1 Vectors . . . 1
- 1.2 Subspaces . . . 4
- 1.3 Span and Linear Independence . . . 7
- 1.4 Basis and Dimension . . . 11

2 Inner Product, Orthogonality, Norm 19
- 2.1 Inner Products . . . 20
- 2.2 Inequalities . . . 26
  - 2.2.1 The Cauchy–Schwarz Inequality . . . 26
  - 2.2.2 The Triangle Inequality . . . 28
- 2.3 Orthogonal Vectors and Orthogonal Bases . . . 29
- 2.4 Orthogonal Projection and the Closest Point . . . 33
- 2.5 The Gram–Schmidt Process . . . 36
- 2.6 Orthogonal Subspaces and Complements . . . 42
- 2.7 Norms . . . 45
  - 2.7.1 Basic Examples . . . 45
  - 2.7.2 Spheres and Balls . . . 46
  - 2.7.3 Equivalence of Norms . . . 47
  - 2.7.4 Metrics and Distance . . . 49

3 Matrices 53
- 3.1 Matrices and Matrix Arithmetic . . . 53
- 3.2 Transposes and Symmetric Matrices . . . 60
- 3.3 Linear Systems and Vectors . . . 62
- 3.4 Image, Kernel, Rank, Nullity . . . 65
- 3.5 Superposition Principles for Linear Systems . . . 69
- 3.6 Matrix Inverses . . . 73
- 3.7 Linear and Affine Functions . . . 76
  - 3.7.1 Linear Functions . . . 76
  - 3.7.2 Affine Functions . . . 79

4 How Matrices Interact with Inner Products and Norms 81
- 4.1 Symmetric Positive Definite Matrices . . . 81
- 4.2 Gram Matrices . . . 86
- 4.3 Adjoints . . . 90
  - 4.3.1 Self-Adjoint and Positive Definite Matrices . . . 91
- 4.4 The Fundamental Matrix Subspaces . . . 94
  - 4.4.1 Applications to Self-Adjoint Matrices . . . 95
  - 4.4.2 Applications to Linear Systems . . . 96
- 4.5 Orthogonal and Norm–Preserving Matrices . . . 100
  - 4.5.1 Rigid Motions . . . 103
- 4.6 Projection Matrices . . . 104

4.7 The General QR Factorization and the Solution of Linear Systems . . . 107
4.7.1 The QR Factorization of a Matrix . . . 108
4.7.2 Solutions to Linear Systems and Least Squares . . . 111
4.8 Matrix Norms . . . 115
4.8.1 Natural Matrix Norms . . . 115
4.8.2 The Frobenius Inner Product and Norm . . . 118

5 Eigenvalues and Singular Values 121
5.1 Eigenvalues and Eigenvectors . . . 122
5.2 Eigenvector Bases . . . 127
5.2.1 Powers of Matrices and the Spectral Radius . . . 131
5.2.2 Connections with Matrix Norms . . . 133
5.3 Eigenvalues of Self-Adjoint Matrices . . . 135
5.3.1 The Spectral Theorem . . . 138
5.3.2 Powers of Self-Adjoint Matrices . . . 139
5.3.3 The Schur Product Theorem . . . 141
5.3.4 Generalized Eigenvalues and Eigenvectors . . . 142
5.4 Optimization Principles . . . 144
5.4.1 Intermediate Eigenvalues and the Min-Max Theorem . . . 147
5.4.2 Eigenvalue Inequalities . . . 151
5.5 Linear Iterative Systems . . . 153
5.5.1 Affine Iterative Systems . . . 155
5.5.2 Markov Processes . . . 156
5.6 Numerical Computation of Eigenvalues . . . 161
5.6.1 The Power Method . . . 161
5.6.2 Orthogonal Iteration . . . 165
5.7 Singular Values . . . 172
5.7.1 The Singular Value Decomposition . . . 173
5.7.2 The Euclidean Matrix Norm . . . 176
5.7.3 Condition Number and Rank . . . 177

6 Basics of Optimization 181
6.1 The Objective Function . . . 182
6.2 Minimization of Quadratic Functions . . . 185
6.2.1 Unconstrained Minimization . . . 186
6.2.2 Least Squares . . . 188
6.2.3 Constrained Minimization . . . 190
6.3 The Gradient and Critical Points . . . 193
6.3.1 The Gradient . . . 193
6.3.2 Critical Points . . . 197
6.4 Gradient Descent . . . 199
6.4.1 Proximal Gradient Descent . . . 203
6.5 The Conjugate Gradient Method . . . 205
6.6 The Second Derivative Test . . . 210
6.7 Convex Functions . . . 215
6.7.1 Some Inequalities . . . 220
6.7.2 Strong Convexity . . . 222
6.8 Lipschitz Continuity . . . 227
6.9 Basic Convergence Results . . . 233
6.10 Newton's Method . . . 239

7 Introduction to Machine Learning and Data **247**
7.1 Basics of Machine Learning and Data . . . 247
7.1.1 Mean, Variance, and Covariance . . . 250
7.1.2 Labels and Learning from Data . . . 255
7.1.3 Fully Supervised Learning . . . 256
7.1.4 Overfitting and Generalization . . . 257
7.1.5 The Train–Test Split and Hyperparameters . . . 259
7.1.6 Semi-supervised and Unsupervised Learning . . . 260
7.2 Linear Regression . . . 261
7.2.1 Ridge Regression . . . 263
7.2.2 Lasso Regression . . . 266
7.2.3 Optimization Aspects . . . 269
7.2.4 Kernel Regression . . . 271
7.3 Support Vector Machines (SVM) . . . 273
7.3.1 Optimization Aspects . . . 277
7.3.2 Multiclass Support Vector Machines . . . 277
7.3.3 Kernel Support Vector Machines . . . 280
7.4 k–Nearest Neighbor Classification . . . 282
7.4.1 Computational Aspects . . . 286
7.5 k–Means Clustering . . . 287
7.6 Kernel Methods . . . 298
7.6.1 Kernel Regression . . . 304
7.6.2 Kernel Support Vector Machines . . . 305

8 Principal Component Analysis **311**
8.1 The Principal Components . . . 311
8.1.1 Kernel Principal Component Analysis . . . 319
8.2 The Best Approximating Subspace . . . 322
8.2.1 Robust Subspace Recovery . . . 330
8.3 PCA-based Compression . . . 333
8.4 Linear Discriminant Analysis . . . 339
8.5 Multidimensional Scaling (MDS) . . . 346

9 Graph Theory and Graph-based Learning **357**
9.1 Graphs and Digraphs . . . 358
9.1.1 Graphs in Applications . . . 364
9.1.2 Similarity Graphs . . . 367
9.2 The Incidence Matrix . . . 372
9.3 The Graph Laplacian . . . 378
9.4 Binary Spectral Clustering . . . 386
9.4.1 Community Detection . . . 391
9.5 Distances on Graphs . . . 397
9.5.1 Computing the Shortest Path Distance . . . 399
9.5.2 Computing Shortest Paths via Dynamic Programming . . . 404
9.5.3 ISOMAP and Metric Multidimensional Scaling . . . 405
9.6 Diffusion on Graphs and Digraphs . . . 410
9.6.1 Jump Diffusion and PageRank . . . 418
9.7 Diffusion Maps and Spectral Embeddings . . . 426
9.7.1 Diffusion Distance . . . 426
9.7.2 Spectral Clustering . . . 433

9.8 t-SNE Embedding . . . 437
9.9 Graph-based Semi-supervised Learning . . . 446
9.9.1 Laplacian Regularization . . . 448
9.9.2 Label Propagation and Hard Constraints . . . 452
9.10 The Discrete Fourier Transform . . . 455
9.10.1 Complexification . . . 459
9.10.2 Roots of Unity . . . 462
9.10.3 The Complex Discrete Fourier Transform . . . 463
9.10.4 Sampling, Trigonometric Interpolation, and Aliasing . . . 465
9.10.5 Convolution and the DFT . . . 469
9.10.6 The Fast Fourier Transform . . . 476

10 Neural Networks and Deep Learning 483
10.1 Fully Connected Networks . . . 484
10.1.1 Training and Optimization . . . 488
10.1.2 Approximation of Functions . . . 490
10.1.3 Classification . . . 491
10.1.4 Connection to Kernel Methods . . . 494
10.2 Backpropagation and Automatic Differentiation . . . 496
10.3 Convolutional Neural Networks . . . 501
10.3.1 Transfer Learning . . . 508
10.4 Graph Convolutional Neural Networks . . . 510
10.4.1 Convolution on Graphs . . . 511
10.4.2 Numerical Experiments . . . 514
10.5 Transformers and Large Language Models . . . 517
10.5.1 The Attention Mechanism . . . 519
10.5.2 Positional Encoding . . . 521
10.5.3 Experiments with Character-based Models . . . 525
10.6 Universal Approximation . . . 526
10.6.1 Polynomials . . . 528
10.6.2 Trigonometric Polynomials . . . 529
10.6.3 Piecewise Affine Functions . . . 530
10.6.4 Two Layer Neural Networks . . . 532
10.6.5 Two Layer ReLU Networks . . . 534
10.6.6 Deep ReLU Networks . . . 537
10.6.7 Approximating Multivariate Functions . . . 541

11 Advanced Optimization 547
11.1 Linear Convergence of Gradient Descent . . . 547
11.2 The Heavy Ball Method . . . 552
11.3 Krylov Subspace Methods and Conjugate Gradients . . . 556
11.4 Nesterov's Accelerated Gradient Descent . . . 562
11.5 Stochastic Gradient Descent . . . 567
11.6 Continuum Analysis of Optimization . . . 577
11.7 Optimizing Neural Networks . . . 584

Bibliography 593

Index 609