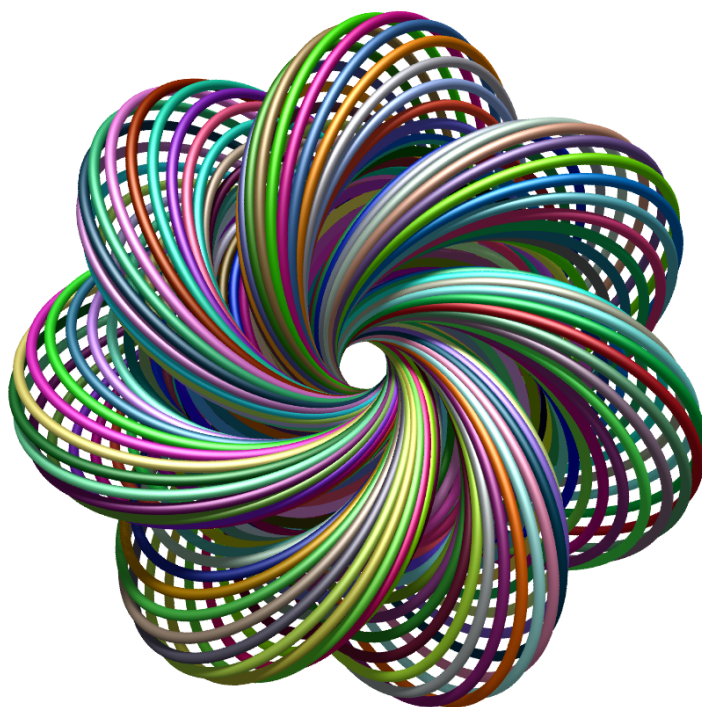


Essentials of Mathematical Methods:

Foundations, Principles, and Algorithms

Yuguang Yang

version 3.0



God used beautiful mathematics in creating the world. –Paul Dirac

*Dedicated to
those who appreciate the power of mathematical methods
and enjoy learning it.*

License statement

You are free to redistribute the material in any medium or format under the following terms:

- **Attribution:** You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- **NonCommercial:** You may not use the material for commercial purposes.

- * The licensor cannot revoke these freedoms as long as you follow the license terms.
- * This license is created via creative commons (<https://creativecommons.org>)
- * If you have any questions regarding the license, please contact the author.

Preface

Objective

Today, mathematical methods, models, and computational algorithms are playing increasingly significant roles in addressing major challenges arising from scientific research and technological developments. Although many novel methods and algorithms, such as deep learning and artificial intelligence, are emerging and reshaping various areas at an unprecedented pace, their core ideas and working mechanisms are inherently related to and deeply rooted in some essential mathematical foundations and principles. By performing an in-depth survey on the underlying foundations, principles, and algorithms, this book aims to navigate the vast landscape of mathematical methods widely used in diverse scientific and engineering domains.

This book starts with a survey of mathematical foundations, including essential concepts and theorems in real analysis, linear algebra, and related fundamentals. Then it examines a broad spectrum of applied mathematical methods, ranging from traditional ones such as optimizations and dynamical system modeling, to state-of-the-art such as machine learning, deep learning, and reinforcement learning. The emphasis is placed on methods for stochastic and dynamical system modeling, optimal decision-making, and statistical learning. For each topic, this book organizes fundamental definitions, theorems, methods, and algorithms in a logical and illuminating way.

Features and Highlights

- Comprehensive, essential, and self-contained.
- Concepts, theorems, and discussions are developed to suit real-world applications.
- Key references and resources are provided on each topic.
- Comparisons and discussions on similar definitions and theorems.
- An evolving book with regular updates on [Github](#).

Acknowledgment

This book evolved from my study notes during my PhD studies at the Johns Hopkins University (JHU). I want to thank the following professors at JHU for their courses and valuable discussion: Daniel Robinson, Teresa Lebar, Andrea Prosperetti, Gregory Chirikjian, Michael Kahadan, James C. Spall, Marin Kobilarov, Suchi Saria, Michael Dimitz, Sean Sun, Ari Turner, Gregory Eyink, Amitabh Basu, John Miller and David Audley. I also want to thank Rachael Zhang for her editorial assistance.

Yuguang Yang, Fall 2019
yangyutu123@gmail.com

Notations

- \mathbb{R} : real numbers.
- \mathbb{R}_+ : nonnegative real numbers.
- \mathbb{R}_{++} : positive real numbers.
- $\bar{\mathbb{R}}$: extended real numbers.
- \mathbb{C} : complex numbers.
- \mathbb{F} : real or complex numbers.
- \mathbb{Q} : rational numbers.
- \mathbb{Z} : integer numbers.
- \mathbb{P} : positive numbers.
- \mathcal{P}_n : polynomial of degree of n .
- \mathbb{N} : natural numbers.
- $\mathcal{R}(A)$: the range of matrix A .
- $\mathcal{N}(A)$: the null space of matrix A .
- V : vector space.
- $\det(A)$: the determinant of matrix A .
- $\text{rank}(A)$: the rank of matrix A .
- $\|\cdot\|_2$: Euclidean 2 norm of a vector of a matrix.
- $\|\cdot\|_F$: Frobenius norm of a matrix.
- $\rho(A)$: the spectral radius of matrix A .
- $\text{Tr}(A)$: the trace of matrix A .
- $L^2[a, b]$: Lebesgue integrable function on $[a, b]$.
- $L^1[a, b]$: Lebesgue integrable function on $[a, b]$.
- $N(0, 1)$: standard Gaussian distribution.
- $N(\mu, \sigma^2)$: Gaussian distribution with mean μ and variance σ^2 .
- $MN(\mu, \Sigma)$: multivariate Gaussian distribution with mean vector μ and covariance matrix Σ .
- $\mathbf{1}(x), I(x)$ indicator function.
- $E[X], \mathbb{E}[X], \mathbb{E}[X]$ expectation of random variable X .
- $\text{Var}[X]$ variance of random variable X .

CONTENTS

i mathematical foundations

- 1 SETS, SEQUENCES, AND SERIES 2
- 2 METRIC SPACE AND TOPOLOGICAL SPACE 36
- 3 ADVANCED CALCULUS 58
- 4 BASIC ABSTRACT ALGEBRA 138
- 5 LINEAR ALGEBRA AND MATRIX ANALYSIS 154
- 6 BASIC FUNCTIONAL ANALYSIS 279

ii mathematical optimization methods

- 7 UNCONSTRAINED NONLINEAR OPTIMIZATION 326
- 8 CONSTRAINED NONLINEAR OPTIMIZATION 369
- 9 LINEAR OPTIMIZATION 415
- 10 CONVEX ANALYSIS AND CONVEX OPTIMIZATION 438
- 11 BASIC GAME THEORY 486

iii classical statistical methods

- 12 PROBABILITY THEORY 508
- 13 STATISTICAL DISTRIBUTIONS 612
- 14 STATISTICAL ESTIMATION THEORY 669
- 15 MULTIVARIATE STATISTICAL METHODS 729
- 16 LINEAR REGRESSION ANALYSIS 823
- 17 MONTE CARLO METHODS 919

iv dynamics modeling methods

- 18 MODELS AND ESTIMATION IN LINEAR DYNAMICAL SYSTEMS 957

19	STOCHASTIC PROCESS	1034
20	STOCHASTIC CALCULUS	1084
21	FOKKER-PLANCK EQUATION	1134
22	MARKOV CHAIN AND RANDOM WALK	1161
23	TIME SERIES ANALYSIS	1211

v statistical learning methods

24	SUPERVISED LEARNING PRINCIPLES	1293
25	LINEAR MODELS FOR REGRESSION	1340
26	LINEAR MODELS FOR CLASSIFICATION	1360
27	GENERATIVE MODELS	1413
28	K NEAREST NEIGHBORS	1433
29	TREE METHODS	1440
30	ENSEMBLE AND BOOSTING METHODS	1465
31	UNSUPERVISED STATISTICAL LEARNING	1494
32	NEURAL NETWORK AND DEEP LEARNING	1560

vi optimal control and reinforcement learning methods

33	CLASSICAL OPTIMAL CONTROL THEORY	1675
34	REINFORCEMENT LEARNING	1694

vii appendix

A	SUPPLEMENTAL MATHEMATICAL FACTS	1780
Alphabetical Index		1811

LIST OF ALGORITHMS

1	A generic line search algorithm	335
2	Backtracking-Armijo line search algorithm	346
3	Steepest decent Backtracking-Armijo line search algorithm	349
4	Modified Newton Backtracking-Armijo line search algorithm	349
5	Quasi Newton with Wolfe line search algorithm	349
6	A generic trust-region algorithm	351
7	A linear conjugate algorithm	359
8	Iteratively reweighted least squares for p norm least square	362
9	Gauss-Newton method for nonlinear least-square algorithm	364
10	Levenberg-Marquardt method for nonlinear least-square algorithm	365
11	Newton method for root finding	366
12	Primal active-set method for strictly convex quadratic programming	388
13	First order gradient projection algorithm	391
14	The Simplex algorithm (non-degenerate system)	429
15	Primal-dual long-step path-following algorithm	434
16	A generic subgradient algorithm	472
17	Gradient projection algorithm with constant step size	479
18	Gradient projection algorithm with adaptive size	480
19	Proximal gradient algorithm	481
20	Iterative Shrinkage-Thresholding Algorithm with constant step size for L ₁ optimization	483
21	Iterative reweighed estimation for multivariate normal distribution	738
22	Alternating sparse canonical correlation analysis algorithm	757
23	EM algorithm for least square with nonconstant variance	887
24	Accept-Reject algorithm for random number generation.	923

25	Importance sampling for Monte Carlo integration.	933
26	MCMC Metropolis-Hasting algorithm	937
27	MCMC Gibbs sampling algorithm	939
28	Recursive linear least square estimation of dynamical systems.	1023
29	Recursive nonlinear least square of dynamical systems	1025
30	Kalman filtering	1029
31	Extended Kalman filter	1030
32	Coordinate descent for Lasso regression.	1351
33	Iteratively reweighted least squares for logistic regression	1366
34	Perceptron learning algorithm	1397
35	Soft margin SVM algorithm	1406
36	Multinomial Naive Bayes classification	1419
37	KNN classification and regression algorithm	1434
38	A generic decision tree generation algorithm	1448
39	ID3 classification decision tree algorithm	1453
40	A regression tree growth algorithm	1461
41	A basic bagging algorithm	1469
42	Generic Adaboost classifier algorithm	1475
43	Adaboost regressor algorithm	1478
44	A generic additive model algorithm	1480
45	Generic gradient boosting algorithm	1483
46	Gradient tree boosting algorithm	1484
47	XGBoost algorithm	1490
48	Iterative reweighted least square PCA with outliers algorithm	1507
49	Random sample consensus PCA with outliers algorithm	1507
50	Orthogonal Matching Pursuit	1509
51	K-SVD for dictionary learning.	1511
52	Online dictionary learning	1512
53	Stochastic gradient descent for matrix factorization based recommender systems.	1522

54	Isomap algorithm	1531
55	Kernel PCA algorithm	1532
56	Laplacian Eigenmap algorithm	1537
57	Diffusion map algorithm	1540
58	K-means algorithm.	1545
59	DBSCAN algorithm	1548
60	General spectral clustering algorithm	1550
61	Gaussian mixture model EM algorithm	1554
62	Full batch gradient descent algorithm	1578
63	Minibatch stochastic gradient descent algorithm	1579
64	Adam stochastic gradient descent algorithm.	1584
65	Solve forward backward stochastic differential equation via deep learning. .	1606
66	Minibatch stochastic gradient descent training of GAN.	1659
67	Minibatch stochastic gradient descent training of conditional GAN.	1665
68	Minibatch stochastic gradient descent training of Wasserstein GAN.	1667
69	The policy iteration algorithm for MDP	1703
70	Value iteration algorithm for a finite state MDP	1705
71	First-visit MC value function estimation	1712
72	MC-based reinforcement learning control	1714
73	TD(o) estimation of a value function.	1715
74	SARSA learning	1716
75	Q-learning algorithm	1717
76	TD(n) estimation of a value function.	1720
77	n -step SARSA learning	1721
78	A generic batch policy-gradient algorithm (REINFORCE)	1732
79	A basic Monte-Carlo policy-gradient algorithm	1733
80	Actor-Critic policy-gradient method	1735
81	Policy-gradient method with a value function baseline	1740
82	Neural Fitted Q Iteration (NFQ)	1745
83	Deep Q-learning with experience replay	1747

84	Asynchronous Deep Q-Learning for each thread	1752
85	Deep Q-learning with universal value function approximator	1753
86	Deep deterministic policy gradient algorithm (DDPG)	1756
87	Twin-delayed deep deterministic policy gradient (TD3)	1758
88	Trust Region Policy Optimization	1761
89	Proximal Policy Optimization	1764
90	Soft Actor-Critic (SAC) policy optimization	1767
91	Isotropic multivariate Gaussian evolution strategies for reinforcement learning	1769
92	Isotropic multivariate Gaussian parallelized evolution strategies for reinforcement learning	1769
93	Q learning with prioritized experience replay	1772
94	Deep Q-learning with hindsight experience replay	1773

LIST OF FIGURES

Figure 3.1.1	An example 3D curve generated by $z = t, x = t \cos(t), y = t \sin(t)$. 69
Figure 3.1.2	An example smooth surface generated by $z = x \exp(-2x^2 - y^2)$ 72
Figure 5.9.1	Demonstration of SVD for matrices of different shapes. The dashed lines highlight the compact form SVD. 222
Figure 5.12.1	Illustration of different quadratic forms. 243
Figure 7.1.1	Demonstration of local minimizer (red, green, and blue), strict local minimizer (red and blue), and global minimizer (blue). 329
Figure 7.1.2	A complex objective function in unconstrained optimization. 330
Figure 7.1.3	Illustration of different cases in unconstrained quadratic optimization. 334
Figure 7.2.1	Drawbacks of steepest gradient descent. 337
Figure 7.2.2	Demonstration on the step choices on the iterative algorithm. (a) Large step size. (b) Small step size. 342
Figure 7.2.3	Armijo sufficient decrease condition. 346
Figure 7.3.1	Demonstration of the dogleg path as an approximation to the exact solution path in the trust-region subproblem. 354
Figure 7.4.1	Demonstration of coordinate descent procedures when A is diagonal and non-diagonal. 357
Figure 8.1.1	Demonstration of KKT condition at a local minimal for $f(x_1, x_2) = x_1^2 + x_2^2$ under constraint $x_1 + x_2 = 1$. 373
Figure 8.2.1	Demonstration of KKT condition at a local minimal for $f(x_1, x_2) = x_1^2 + x_2^2$ under constraint $x_1 + x_2 \geq 1$. 383
Figure 9.2.1	The geometry of linear programming. (a) The feasible region is an open space extending to infinity if A is not full column rank. (b-d) Example feasible regions if $\text{rank}(A) = n, m \geq n$. Red arrows are direction of $-c$. When moving along $-c$ in the feasible region, the objective function will decrease. 419
Figure 9.2.2	Demonstration on multiple minimizers forming a convex set. 421
Figure 9.3.1	Overview of geometry approach to linear programming. 423
Figure 10.1.1	Example 2D affine hull and 3D affine hull. 440
Figure 10.1.2	Affine subspace and linear subspace. 442
Figure 10.2.1	(left) A convex set. (right) A non-convex set. 446

- Figure 10.2.2 (left) The affine hull of two points in a plane is a line passing through them. (right) The convex hull of two points in a plane is a line segment containing them. 448
- Figure 10.2.3 An illustration of separating hyperplane theorem for two convex bodies. 450
- Figure 10.2.4 An illustration of Farkas' lemma. (left) When b lies outside the cone (that is, $Ax = b, x \geq 0$ has no solution), there exists a hyperplane, characterized by normal vector y , separating b and the cone. (right) When b lies inside the cone (that is, $Ax = b, x \geq 0$ has a solution), there does not exist a hyperplane, characterized by normal vector y , separating b and the cone. 452
- Figure 10.2.5 An illustration of Farkas' lemma variant where the cone is open set. (left) When b lies outside the cone (that is, $Ax = b, x > 0$ has no solution), there exists a hyperplane, characterized by normal vector y , separating b and the cone. (right) When b lies inside the cone (that is, $Ax = b, x > 0$ has a solution), there does not exist a hyperplane, characterized by normal vector y , separating b and the cone. 453
- Figure 10.3.1 Demonstration of convex functions. (a) A convex function satisfying $f(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha f(x_1) + (1 - \alpha)f(x_2)$. (b) A non-convex function where the green points does not satisfy the relation. 454
- Figure 10.3.2 The epigraph (green area) of a convex function. 456
- Figure 10.3.3 Illustration of linear underestimator. 459
- Figure 10.5.1 Demonstration of optimality condition $\nabla f(x^*)^T(x - x^*) \geq 0, \forall x \in X$ when x^* lies on the boundary of X . 469
- Figure 12.1.1 An illustration of measurable functions. 517
- Figure 13.1.1 Comparison of Laplace distribution and normal distribution. 618
- Figure 13.1.2 Density of $LN(0,1)$ and $LN(0,0.5)$. Note the positive skewness. 625
- Figure 13.1.3 Density of $NLN(0,1)$ and $NLN(0,0.5)$. Note the negative skewness. 626
- Figure 13.2.1 Distributions with left-skewness (black) and right-skewness (red). 659
- Figure 13.2.2 Distributions with zero excess Kurtosis (Normal distribution, black), positive excess kurtosis (Laplace distribution, red), and negative excess Kurtosis (Uniform distribution, blue). 660
- Figure 14.1.1 An example of biased estimator with smaller variance than unbiased estimator 675
- Figure 14.7.1 Demonstration for rejection regions for upper-tailed one-sided hypothesis (a), two-sided hypothesis(b), and lower-tailed one-sided hypothesis (c). 711

Figure 14.8.1	QQ plot with different sample distributions, including normal, Laplace ($b = 4$), uniform $U([-1, 1])$ (d) Lognormal $LN(0, 1)$. Red solid lines are the fitted linear lines. 720
Figure 15.2.1	Principal components for 2D samples. 741
Figure 15.2.2	PCA eigenface analysis. 749
Figure 15.2.3	PCA eigen-digit analysis for MNIST dataset. 750
Figure 15.2.4	Demonstration of interest rate curve dynamics. 752
Figure 15.2.5	Demonstration of first three dominating PCA factor in the swap rate curve daily change. 752
Figure 15.4.1	Gaussian copula with different correlations. 774
Figure 15.4.2	Student T copula with different correlations. 778
Figure 15.4.3	Generated correlated default time via Gaussian copula with different correlations. The hazard rate for both parties is $h(t) = 0.01$. 795
Figure 15.5.1	Correlation structure for the Gaussian copula implied by the factor model. 806
Figure 15.5.2	Factor model using (a) external factors or (b) internal factors. 807
Figure 15.5.3	Scatter plot of AAPL return vs. market excess return, SMB excess return and HML excess return. 811
Figure 15.6.1	A fully connected graphical model representing the joint distribution $P(X_1, \dots, X_N)$. 814
Figure 15.6.2	Graphical model examples. 817
Figure 15.6.3	d-separation examples. 818
Figure 16.1.1	Demonstration of simple linear regression model $y = \beta_1 x + \beta_0 + \epsilon$ and multiple linear regression model $y = \beta_1 x_1 + \beta_2 x_2 + \beta_0 + \epsilon$. Scatter points are observed data. The solid line in the left and the plane in the right are the mean responses. 827
Figure 16.3.1	Demo of heteroskedasticity in linear regression. The noises are larger at larger x values. 885
Figure 16.3.2	Demonstrations on linear regression with auto-correlated error. Observations are generated by $y_i = x_i + \epsilon_i, \epsilon_i = \rho \epsilon_{i-1} + z, z \sim N(0, 1)$. 890
Figure 16.3.3	Illustration of an outlier, a high-leverage point, and a influential point. Left subfigure shows a red-colored outlier, which does not have high leverage and large influence on the regression result. Middle subfigure shows a red-colored high-leverage point, which is not an outlier or influential point due to its weak influence on the regression result. Right subfigure shows an influential point that is both an outlier and a high-leverage point. 895
Figure 16.3.4	Visual scatter and box plots to identify outliers. 896
Figure 16.3.5	Different function choice for M-estimation linear regression 901

Figure 16.3.6	Common linear regression diagnosis plots	904
Figure 16.4.1	Diagnosis plots for a toy linear regression example	907
Figure 16.4.2	Diagnosis plots for the Boston Housing example	909
Figure 17.4.1	Brownian motion interpolation Demo.	945
Figure 19.1.1	An illustration of a random walk mapping a sample point, ω , to a trajectory parameterized by time, where red trajectory sample point HHT, and blue trajectory has sample point THT.	1037
Figure 19.1.2	An illustration of left and right continuous functions.	1041
Figure 19.6.1	A typical realized trajectory from the Poisson process with jumps at T_1, T_2 , and T_3 .	1064
Figure 20.3.1	The variance function $Var[X(t)]$ for Brownian motion (red) and OU process(black) with $a = 0.5, \sigma = 1$.	1113
Figure 20.3.2	Representative trajectories from three OU processes with different k . k has the unit of inverse year. Mean level $\mu = 50$ and volatility $\sigma = 20$.	1115
Figure 20.4.1	Variance of X_t in a Brownian bridge	1128
Figure 22.1.1	Example Markov chains. Arrows and numbers are transition directions and probabilities.	1164
Figure 22.1.2	Markov chain diagram for a random walk on the state space \mathbb{Z} .	1165
Figure 22.2.1	Demonstration accessibility in a Markov chain. In chain (a), states A and B are accessible to each other or they can communicate. In chain (b), state A can access B but B cannot access A.	1167
Figure 22.2.2	Demonstration of partitioning state space by communicating classes. Green and orange states belong to different communicating classes. Note that a communicating class can consist of only one state.	1168
Figure 22.2.3	Classification of communicating classes into recurrent class and state space by communicating classes. Green states form a communicating class belonging to the transient class. Orange states form a communicating class belonging to the recurrent/closed/adsorbing class.	1175
Figure 22.2.4	Example periodic and aperiodic Markov chains.	1177
Figure 23.1.1	Example time series including a white noise process (upper left), a seasonable time series with periodicity 20, the US new privately owned housing [source], and the US GDP time series[source].	1215
Figure 23.1.2	Demonstration on using STL to decompose an example CO2 concentration time series.	1220
Figure 23.2.1	Example trajectories of AR(1) models ($X_t = aX_{t-1} + Z_t$.) with different choices of a .	1225

Figure 23.2.2	Example trajectories of MA(1) models (upper) and MA(2) models (lower). 1232
Figure 23.2.3	Representative simulated trajectories for AR(1) process with coefficient 1, which forms a unit-root process, and with coefficient -1. 1241
Figure 23.2.4	The ACF and PACF corlogram for a white noise process. 1248
Figure 23.2.5	The ACF and PACF corlogram for an AR(1) process with coefficient 0.8. 1248
Figure 23.2.6	The ACF and PACF corlogram for an MA(1) process. 1249
Figure 23.2.7	The ACF and PACF corlogram for an ARMA(1,1) process. 1249
Figure 23.2.8	Diagnosis plot of residuals for AR(2) model estimation. 1258
Figure 23.2.9	Diagnosis plot of residuals for AR(1) model fitted to time series generated by AR(2) ground truth model. 1259
Figure 23.4.1	Stock index SP500 daily return between 2014 and 2019. 1272
Figure 23.4.2	Simulated representative trajectories from ARCH(1) model with coefficients $a = 0.9$ and $a = 0.5$. 1274
Figure 24.1.1	Scheme of a supervised learning task. Training samples are fed into a learning system to obtain an optimized model, which will be further used in a prediction system for regression and classification tasks. 1296
Figure 24.1.2	Input feature data type examples. 1296
Figure 24.2.1	A simple regression problem illustrating underfitting and overfitting. Solid lines are models, and points are samples. 1300
Figure 24.2.2	The commonly observed phenomenon of overfitting and underfitting in machine learning. 1301
Figure 24.2.3	The performance of different types of models as training proceeds. 1304
Figure 24.3.1	Regression loss functions: MSE Loss, MAE Loss, Huber loss ($\delta = 0.1, 1, 3$), and Log-Cosh Loss. 1308
Figure 24.3.2	Common classification loss functions. 1310
Figure 24.3.3	Scheme for ROC curves diagram. A and B demote ROC curves of different model. 1315
Figure 24.4.1	Scheme for cross-validation error calculation procedure. 1318
Figure 24.4.2	Hyperparameter search via turning regularization parameter λ . At large λ , heavy regularization causes underfitting; at small λ , insufficient regularization causes overfitting. 1320

- Figure 24.5.1 Left: A sample of nine real-world time series reveals a diverse range of temporal patterns [4, 5]. Right: Examples of different classes of methods for quantifying the different types of structure, such as those seen in time series on the left: (i) distribution (the distribution of values in the time series, regardless of their sequential ordering); (ii) autocorrelation properties (how values of a time series are correlated to themselves through time); (iii) stationarity (how statistical properties change across a recording); (iv) entropy (measures of complexity or predictability of the time series quantified using information theory); and (v) nonlinear time-series analysis (methods that quantify nonlinear properties of the dynamics).[9] 1330
- Figure 25.1.1 Correlation among features and the label MEDV. 1343
- Figure 25.1.2 Pair plot among features and the label. 1344
- Figure 25.2.1 Comparison different penalization: Lasso L_1 , elastic net, and Ridge L_2 . Orange regions are admissible set for parameter β . Contours are objective function value as a function of parameter β . Black solid circles are the minimizers $\hat{\beta}$ when there are no penalties, and red solid circles are the minimizers when penalties are applied. 1353
- Figure 25.2.2 Penalized regression path in a toy regression problem. 1354
- Figure 25.3.1 Linear regression with non-linear high order terms. The samples are generated via $y = 2x_1 + x_2 - 0.8x_1x_2 + 0.5x_1^2 + \epsilon$, where ϵ is noise. 1357
- Figure 25.3.2 Linear model enhancement flowchart. 1357
- Figure 26.1.1 Logistic regression for classification on the Iris data set. 1363
- Figure 26.1.2 Pair plot analysis results on South Africa heart disease problem. 1371
- Figure 26.1.3 L1 regularization path for South Africa heart disease problem. 1372
- Figure 26.1.4 Logistic regression result with L1 penalty. 1374
- Figure 26.1.5 Balanced accuracy score vs. inverse regularization strength in credit card fraud detection problem. 1375
- Figure 26.1.6 Logistic regression coefficients corresponding to each class. 1376
- Figure 26.2.1 Geometry of decision boundary in linear Gaussian discriminant model. 1380
- Figure 26.2.2 Comparison of Gaussian LDA and Gaussian QDA on binary classification. Decision boundary of LDA is simply a line in 2D input space and unable to discriminate difficult cases. Gaussian discrimination can have richer decision boundary geometry. LDA using polynomial features is a special case of GDA. 1384
- Figure 26.2.3 The decision boundary geometry of LDA and GDA can be understood via their decision functions 1385

Figure 26.3.1	The linear discriminant w that maximizes the separability for 2D sample points belonging to two classes. 1387
Figure 26.3.2	Fisher linear discriminate will fail to achieve class separability for complex data structures. 1391
Figure 26.4.1	Scheme of a hyperplane. 1396
Figure 26.4.2	Binary classification using the Perceptron learning algorithm. The hyperplane learned separates the two clusters. 1398
Figure 26.5.1	Left: existence of multiple separating hyperplanes in 2D binary classification problem. Right: hyperplanes with maximum margin. 1400
Figure 26.5.2	SVM classification with different regularization strength. Small C tends to emphasize the margin and ignore the outliers in the training data, while large C may tend to overfit the training data. 1403
Figure 26.5.3	SVM classification using Gaussian kernel. The original problem cannot be separated by linear kernel. 1408
Figure 26.5.4	Comparison of classification loss functions. 1410
Figure 27.2.1	Histogram of the features group by class label (fraud vs. genuine). 1423
Figure 27.2.2	Feature density similarity and predictive performance of model 1424
Figure 28.2.1	Binary classification via KNN algorithm with different choices $K = 1, 3, 5, 7$. Scattered points are training examples classified into two different classes (red and blue). Colored regions are corresponding decision boundaries. 1438
Figure 29.2.1	Demonstration of decision trees. 1447
Figure 29.2.2	Different types of impurity measure. Entropy function, Gini function and classification error function. 1450
Figure 29.2.3	Different splitting strategy when variables taking more than two discrete values. 1451
Figure 29.2.4	The visualization of the decision tree classifier for Iris data set. The tree grows until all examples are classified correctly. The splitting criterion is Gini impurity. 1456
Figure 29.2.5	The visualization of the decision tree classifier for Iris data set. The tree grows until examples in each node is smaller than 10. The splitting criterion is Gini impurity. 1457
Figure 29.2.6	The visualization of the decision tree classifier for Iris data set. The tree grows until examples in each node is smaller than 10. The splitting criterion is entropy and information gain. 1457
Figure 29.3.1	Demonstration of a tree and input space partitioning. 1460
Figure 29.3.2	2D input space partitions cannot be represented by a regression tree. 1460

Figure 29.3.3	Regression tree demonstration in a toy example. 1462
Figure 29.3.4	Variable importance from regression tree in the Boston Housing Pricing problem. 1463
Figure 30.1.1	The correctness probability of a majority vote is greater than the correctness probability of individual votes when individual accuracy probability is greater than 0.5. 1467
Figure 30.3.1	Illustration of adaptive boosting where sample weights are adjusted iteratively based on the classification error. 1474
Figure 31.1.1	Demonstration of SVD for two different types of matrices. The dashed lines highlight the compact form SVD. 1498
Figure 31.1.2	Principal components for 2D samples. 1502
Figure 31.2.1	Singular value spectrum of LSA on 20-news-group text data. 1517
Figure 31.2.2	A simplified illustration of the latent factor approach, which characterizes both users and movies using two axes—male versus female and serious versus escapist. [9] 1522
Figure 31.3.1	Triangle and Tetrahydron reconstructed from distance matrix by MDS method. 1529
Figure 31.3.2	Application of MDS, based on Euclidean distance, to Swiss Roll data set cannot fully reveal of the global structure. 1530
Figure 31.3.3	Isomap analysis of MNIST dataset. 1541
Figure 31.3.4	Isomap analysis of digit '5' in MNIST dataset 1542
Figure 31.4.1	Demonstration of k-means clustering on a data set with two blobs. 1544
Figure 31.4.2	K-means performance can be affected by a number of factors, including incorrect number of clusters/blobs, non-spherical clusters/blobs, clusters/blobs with unequal variance, and bad initial cluster centers. 1546
Figure 31.4.3	Clustering comparison between Kmeans and Kmeans++. 1547
Figure 31.4.4	DBSCAN demo with different ϵ . 1549
Figure 31.4.5	Spectral clustering demo. 1551
Figure 31.4.6	Clustering comparison between K-means and GMM. 1555
Figure 31.4.7	K-means application to image segmentation. 1556
Figure 32.1.1	Scheme of an artificial neuron. 1566
Figure 32.1.2	Common activation functions in artificial neural networks. 1567
Figure 32.1.3	Scheme for an artificial neural network. 1568
Figure 32.1.4	A four-layer feed-forward neural network. 1568
Figure 32.1.5	A three-layer feed-forward neural network with three output units. 1569
Figure 32.1.6	A four-layer feed-forward neural network. 1571
Figure 32.2.1	An example saddle point at $(0, 0, 0)$, which locally minimizes the x direction but maximizes the y direction.. 1578

Figure 32.2.2	SGD without momentum and with momentum. SGD with momentum can accumulate gradient/velocity in horizontal direction and move faster towards the minimum located at the center. 1582
Figure 32.3.1	Dropout technique for a simple feedforward neural networks. The original network (left) and the neural network after some neurons being dropped out (right). 1590
Figure 32.4.1	Neural network architecture for polynomial regression. 1592
Figure 32.4.2	Polynomial regression with degree $d = 1, 3, 6, 10$. 1593
Figure 32.4.3	Visualization of first layer weight for a one-layer linear multi-class classification neural network 1594
Figure 32.4.4	Example images from the Fashion MNIST dataset. 1595
Figure 32.4.5	The confusion matrix from fashion MNIST classification results. 1596
Figure 32.4.6	Classification results for a set of randomly selected samples. 1597
Figure 32.4.7	(a) Embedding layer maps large, sparse one-hot vectors to short, dense vectors. (b) Example of low dimensional embeddings that capture semantic meanings. 1598
Figure 32.4.8	(a) The Skip-gram architecture that predicts surrounding words given the central word. (b) The CBOW architecture that predicts the central word given its surrounding context words. The one-hot vector has size V ; the dense vector has length $D \ll V$. Also note that no nonlinearity activation is applied between input and hidden layer. 1599
Figure 32.4.9	A feed-forward neural network architecture for sentiment analysis. 1603
Figure 32.5.1	Comparison of receptive fields in fully-connected layer and local-connected layer in CNN. Credit 1607
Figure 32.5.2	Demo for one kernel 'convoluting' with an input image. 1608
Figure 32.5.3	Pooling layer demo. 1609
Figure 32.5.4	A typical CNN architecture for image classification tasks. 1610
Figure 32.5.5	Scheme for LeNet [24] 1610
Figure 32.5.6	Architecture of AlexNet. 1611
Figure 32.5.7	A typical VGG architecture: VGG-19 scheme. 1611
Figure 32.5.8	The depth of the configurations increases from the left (A) to the right (E), as more layers are added (the added layers are shown in bold). The convolutional layer parameters are denoted as "conv-receptive field size-number of channels" [26] 1612
Figure 32.5.9	Training error (left) and testing error (right) on CIFAR-10 with 20-layer and 56-layer vanilla CNN networks. The deeper network has both higher training error and testing error.[27] 1613
Figure 32.5.10	Scheme for a residual block. 1613

Figure 32.5.11	Scheme of a 18 layer ResNet 1614
Figure 32.6.1	Example images from CIFAR10 image dataset. 1615
Figure 32.6.2	Visualization of convolution layer. 1617
Figure 32.6.3	Grad-CAM method applied to understand image classification. Middle and right are class-discriminative localization map superimposed onto the original image. 1618
Figure 32.6.4	A CNN based autoencoder. An autoencoder consists of an encoder that transforms high-dimensional image into a low-dimensional code and a decoder that unfolds the code to reconstruct the image. 1620
Figure 32.6.5	Comparison of reconstruction performance on random samples from MNIST data set. Top row is the original data. Middle row is autoencoder result based on a 49 dimensional code. Bottom row is PCA result based on a 50 dimensional code. 1620
Figure 32.6.6	Denoising autoencoders applied to remove the noise in the MNIST dataset. 1621
Figure 32.6.7	Demonstration of neural style transfer with Van Gogh painting style. 1623
Figure 32.6.8	Demonstration of neural style transfer with Picasso painting style. 1624
Figure 32.6.9	Application of CNN in deep reinforcement learning in Q learning approach and Actor-Critic approach. Two streams of sensory inputs are fed to the neural network, including a pixel image of the robot's neighborhood fed into a convolutional layer and the target's position fed into a fully connected layer. 1626
Figure 32.6.10	Demonstration of a CNN filter applying to a sentence to produce a one-dimensional feature vector. 1627
Figure 32.6.11	CNN for sentence classification proposed in [43]. 1628
Figure 32.7.1	Scheme of recurrent units in a neural network (left). Recurrent neural network can be unrolled (right) 1630
Figure 32.7.2	Scheme for backpropogation through time in a simple RNN. Red arrows are backpropogation directions. 1631
Figure 32.7.3	Scheme of an LSTM cell. Modified from [45, p. 149]. 1633
Figure 32.7.4	Scheme of a GRU cell. Modified from [45, p. 152]. 1636
Figure 32.7.5	Different types of units in RNNs: (a) Vanilla RNN cell. (b) LSTM cell. (c) GRU cell. Credit 1637
Figure 32.7.6	Typical RNN connection to the output layer. 1638
Figure 32.7.7	Scheme for stacked RNN. 1639
Figure 32.7.8	Scheme for bidirectional RNN. 1639

- Figure 32.8.1 RNN architecture for time series prediction . (a) In the training phase, RNN are updated by minimizing the next step prediction error. (b) In the prediction phase, trained RNN is used to sequentially predict next step state value based on preceding predicted state value. [1641](#)
- Figure 32.8.2 RNN one-step forward and multiple forward prediction performance for Sine time series. [1642](#)
- Figure 32.8.3 RNN architecture for time series prediction with covariates. (a) In the training phase, RNN are updated by minimizing the next step prediction error. (b) In the prediction phase, trained RNN is used to sequentially predict next step state value based on preceding predicted state value. Note that covariate time series are assumed available for all time steps. [1643](#)
- Figure 32.8.4 DeepAR model architecture for time series prediction. Outputs are the parameters characterizing the condition distribution of x_{t+1} conditioned on histories of x_t and z_t . (a) In the training phase, RNN are updated by minimizing the negative log likelihood function. (b) In the prediction phase, a predicted \hat{x}_{t+1} are sampled from predicted conditional distribution and then used to predict next step conditional distribution. [1644](#)
- Figure 32.8.5 A RNN architecture for MNIST image recognition. [1646](#)
- Figure 32.8.6 RNN architecture for sentiment analysis. [1647](#)
- Figure 32.8.7 A RNN for character-level word classification. [1648](#)
- Figure 32.8.8 A RNN for character-level language model. During the training session, one-hot coded characters are directly fed into RNN and predict the next character in the word. [1649](#)
- Figure 32.8.9 A RNN for character-level language model. During the word generation session, the network starts with a user input character and continues the generation process with predicted character from the previous step. [1650](#)
- Figure 32.9.1 Seq2seq modeling for language transformation. The input and output sequences might have different lengths, and not in synchrony. [1652](#)
- Figure 32.9.2 The Encoder-decoder architecture for seq2seq language modeling. The input sequence is fed into the encoder RNN and terminated by an explicit <EOS> symbol. Then the decoder RNN starts with the context vector and the final prediction of the encoder to generate an output sequence until an explicit <EOS> symbol is produced. [1653](#)

- Figure 32.9.3 The Encoder-decoder architecture with attention mechanism for seq2seq modeling. During the encoding phase, all hidden states, rather than the final one, are saved to construct different context vectors via linear combination for the decoding stage. During the decoding phase, relevant context vectors are constructed and fed into the each hidden states in the decoder. [1655](#)
- Figure 32.9.4 A bidirectional RNN encoder system with attention mechanism. [\[51\]](#) [1656](#)
- Figure 32.10.1 Scheme for a canonical GAN. The discriminator is trained to distinguish between real and fake image, while the generator is trained to generate realistic image to 'fool' the discriminator. The generator usually uses a decoder-like neural network structure that generate a high-dimensional data from a sample point in the low-dimensional latent space. [1658](#)
- Figure 32.10.2 Generated image samples from a GAN consisting of feed-forward networks. [1660](#)
- Figure 32.10.3 DCGAN generator network architecture. A 100 dimensional uniform noise is passing through a series of fractionally-strided convolutions then is converted into a final image. Notably, no fully connected or pooling layers are used [\[56\]](#) [1663](#)
- Figure 32.10.4 Understanding DCGAN. Each row represents the image generated as we interpolate a random point z in the latent space. Images show smooth transitions from one bedroom and another bedroom.[\[56\]](#) [1663](#)
- Figure 32.10.5 Scheme for a conditional GAN. The discriminator is trained to distinguish between real and fake image given external label information, while the generator is trained to generate realistic image to 'fool' the discriminator given external label information. The generator usually uses a decoder-like neural network structure that generate a high-dimensional data from a sample point in the low-dimensional latent space. [1664](#)
- Figure 34.1.1 Policy iteration involves iteratively carrying out policy evaluation and policy improvement procedures. [1699](#)
- Figure 34.2.1 One core component in reinforcement learning is agent environment interaction. The agent takes actions based on observations on the environment and a decision-making module that maps observations to action. The environment model updates system state and provides rewards according to the action [1707](#)
- Figure 34.2.2 Scheme of the Atari game *breakout*. [1708](#)
- Figure 34.2.3 Policy evaluation and policy improvement framework in the context reinforcement learning. [1709](#)
-

- Figure 34.3.1 Neural network parameterization for a Gaussian policy (a) and (b) a generic stochastic policy. 1737
- Figure 34.4.1 A typical feed-forward neural network used in NFQ to approximate the Q function. 1744
- Figure 34.4.2 The network architecture for canonical deep Q learning. The network takes a state or an observation, denoted by s , as the input, and outputs multiple values corresponding $Q(s, a)$. The number of outputs 1746
- Figure 34.4.3 A typical single stream Q -network (top) in deep Q learning and a dueling Q -network (bottom). The dueling network has two streams to separately estimate (scalar) state-value and the advantage function for each action; the green output module synthesize the Q value from two streams. Both networks output Q -values for each action. [10]. 1749
- Figure 34.4.4 In a DQRN, recurrent layers are usually placed on the last layer before output. Unlike canonical DQN, we need to feed sequence of observation into the network and finally output Q values. Above scheme only unfolds to two steps.[11] 1750
- Figure 34.4.5 An example architecture that implements the asynchronous reinforcement learning paradigm. Multiple agents interact with multiple instances of environments in parallel. Agents collect experiences to train a globally shared network that learns control policies. 1751
- Figure 34.4.6 A comparison between a typical deep Q learning network (left) and a typical universal value function approximator network (right). 1753
- Figure 34.4.7 A typical deep neural network architecture for DDPG reinforcement learning. The actor network outputs control policy; the critic network outputs estimate of Q function. Through back-prorogation, the critic network improves estimation accuracy and the actor network improves policy. 1755
- Figure 34.5.1 Illustration of planning curriculum on a swiss roll manifold. Red targets can generating along the path connecting from an intended start point to the target goal position. 1774
- Figure 34.5.2 One representative trajectory on the curved surface via a control policy learned via curriculum learning on a low-dimensional manifold. 1775

LIST OF TABLES

Table 14.8.1	Test on mean with known variance σ^2	719
Table 14.8.2	Test on mean with unknown variance σ^2	721
Table 14.8.3	Test on variance	721
Table 14.8.4	Test on variance comparison between two samples	722
Table 15.2.1	Eigenvectors and eigenvalues for swap rate daily change	753
Table 15.5.1	statistics on Fama-French 3 factors from July 1963 to Dec. 1991.	810
Table 15.5.2	AAPL stock return modeled by the Fama-French 3 factor model.	812
Table 22.2.1	Summary of Markov chain state property[3, p. 140]	1179
Table 23.2.1	Summary of PACF and ACF for AR, MA, and ARMA processes.	1248
Table 25.1.1	Linear regression results.	1345
Table 26.1.1	Logistic regression results on South Africa heart disease problem.	1372
Table 31.2.1	Most frequent words in the top 8 topics	1518
Table 31.2.2	Rating matrix or utility matrix, where each row is a user's ratings for different movies. SW ₁ , SW ₂ are Star wars episodes; HP ₁ and HP ₂ are Harry Potter episodes; TW is Twilight; BM is Batman.	1520
Table 34.2.1	Estimating cumulative rewards $G_t^{(n)}$ of different steps n as the target for value function V . $G^{(1)}$ corresponds to temporal-difference TD(0) and $G^{(\infty)}$ corresponds to Monte-Carlo estimation. If the process terminates at K and $K < n$, then we use $G_t^{(n)} = G_t^{(K)}$. Trajectories are generated under policy π .	1719
Table A.9.1	Closed Newton-Cotes Formula	1801

CONTENTS

i mathematical foundations

1	SETS, SEQUENCES, AND SERIES	2
1.1	Sets	4
1.1.1	Definitions and basic properties	4
1.1.2	DeMorgan's Law	5
1.1.3	Set equivalence and partition	5
1.1.4	Countability	6
1.2	Functions	8
1.2.1	Basic concepts	8
1.2.2	Inverse image vs. inverse function	8
1.2.3	Set operations in function mapping	9
1.2.4	Parameter change of function	9
1.3	Real numbers	10
1.3.1	Rational numbers	10
1.3.2	Dense subset	10
1.3.3	Axiom of completeness	11
1.4	Sequence in \mathbb{R}	14
1.4.1	Basics	14
1.4.2	Cauchy criterion	16
1.4.3	Sequence characterization of dense subset	17
1.5	Monotone sequence	19
1.5.1	Fundamentals	19
1.5.2	Applications	19
1.6	Subsequence and limits	23
1.6.1	Subsequence	23
1.6.2	Bolzano-Weierstrass theorem	23
1.6.3	Subsequence limits	24
1.7	Infinite series	27
1.7.1	Fundamental results	27
1.7.2	Tests for convergence	28
1.7.3	Inequalities and l_2 series	30
1.7.3.1	Holder's and Minkowski's inequality	30
1.7.3.2	Cauchy-Schwarz inequality	31
1.7.4	Alternating series	32

1.8	Notes on bibliography	34
2	METRIC SPACE AND TOPOLOGICAL SPACE	36
2.1	Metric space	38
2.1.1	Definitions	38
2.1.2	metric space vs. normed (vector) space vs. Banach space	40
2.2	Sequences in metric space	41
2.3	Closed sets & open sets in metric space	43
2.3.1	Closed set	43
2.3.2	Open sets	43
2.3.3	Further characterization and properties	44
2.3.4	Open and closed sets in \mathbb{R}^n	46
2.4	Compact sets	48
2.4.1	Basic concepts	48
2.4.1.1	closed set vs. compact set	49
2.4.2	Compact sets in \mathbb{R}^N	49
2.4.3	The Heine-Borel Theorem and boundedness of continuous function	50
2.5	Completeness of metric space	51
2.5.1	Sequence and completeness	51
2.5.2	Completeness of \mathbb{R}^n	51
2.6	Topology space	53
2.6.1	Definitions	53
2.6.2	Continuous function, Homeomorphism in topological space	54
2.6.3	Subspaces of topological space	54
2.7	Notes on bibliography	56
3	ADVANCED CALCULUS	58
3.1	Continuous functions	61
3.1.1	Continuous function on \mathbb{R}	61
3.1.1.1	Basics	61
3.1.1.2	Continuity and inverse	64
3.1.2	Continuous function in metric space	65
3.1.3	Boundedness and extreme value theorems	66
3.1.4	More on extreme values	68
3.1.5	Curves and surfaces	68
3.1.5.1	Curvature	71
3.1.5.2	Surfaces	71
3.2	Uniform continuity	74
3.2.1	Uniform continuity on real line	74
3.2.1.1	Concepts	74
3.2.1.2	Lipschitz continuity	76
3.2.2	Uniform continuity on metric space	79

3.2.3	Locally and globally Lipschitz continuous	80
3.3	Differentiation	82
3.3.1	Differential function concept	82
3.3.2	Differential rules	83
3.3.3	Mean value theorem	85
3.4	Function sequence and series	88
3.4.1	Pointwise convergence, uniform convergence	88
3.4.2	Properties of uniform convergence	89
3.4.2.1	Uniform convergence preserve continuity	89
3.4.2.2	Exchange limits and integration	90
3.4.2.3	Exchange limits and differential	90
3.4.2.4	Linearity of uniform convergence	90
3.5	Power series	92
3.5.1	Fundamentals	92
3.5.2	Term-by-term operation	94
3.5.3	Power series and analytic function	94
3.5.4	Approximation by polynomials	95
3.6	Taylor polynomial and Taylor series	97
3.6.1	Taylor polynomial and approximation	97
3.6.2	Taylor series and Taylor's theorem	99
3.6.3	Common Taylor series	101
3.6.4	Useful approximations	103
3.7	Riemann Integral Theory	105
3.7.1	Construction of Riemann integral	105
3.7.2	Riemann integrability	106
3.7.2.1	Basics	106
3.7.2.2	Lebesgue characterization of integrability	107
3.7.2.3	limits and integrability	107
3.7.2.4	Algebraic properties	108
3.7.3	First Fundamental Theorem of Calculus	109
3.7.4	Second Fundamental Theorem of Calculus	109
3.7.4.1	Fundamentals	109
3.7.4.2	Differentiating definite integrals	112
3.7.4.3	Application to differential equation	113
3.7.5	Essential theorems	114
3.7.6	Integration rules	115
3.7.7	Improper Riemann integrals	115
3.8	Basic measure theory	118
3.8.1	Measurable space	118
3.8.1.1	σ algebra	118
3.8.1.2	Measurable space and positive measure	119

	3.8.1.3	Borel algebra and Lebesgue measure	119
	3.8.2	Measurable functions and properties	121
	3.8.2.1	Measurable function and measurability	121
	3.8.2.2	Properties	122
	3.8.3	Convergence of measurable functions	124
	3.8.4	Almost everywhere convergence	124
3.9		Lebesgue integral	126
	3.9.1	Simple function and its Lebesgue integral	126
	3.9.2	Lebesgue integral of measurable functions	128
	3.9.2.1	Integral of non-negative functions	128
	3.9.2.2	Integral of general functions	130
	3.9.3	Riemann vs. Lebesgue integrals	131
	3.9.4	Convergence theorems	131
	3.9.4.1	Applications	133
3.10		Notes on bibliography	136
4		BASIC ABSTRACT ALGEBRA	138
	4.1	Groups	139
	4.1.1	Definitions and examples	139
	4.1.2	Matrix groups	140
	4.1.3	Elementary properties of groups	140
	4.1.4	Homomorphism and isomorphism	141
	4.1.5	Subgroup	141
	4.1.6	Cyclic group	141
	4.1.7	Permutation groups	142
	4.2	Ring and field	144
	4.2.1	Ring	144
	4.2.2	Field	145
	4.3	Polynomials	148
	4.3.1	Polynomials: Basics	148
	4.3.2	Factorization of polynomial over \mathbb{C}	150
	4.3.3	Factorization of polynomial over \mathbb{R}	151
	4.4	Notes on bibliography	152
5		LINEAR ALGEBRA AND MATRIX ANALYSIS	154
	5.1	Theory for system of linear equations	159
	5.1.1	Overview	159
	5.1.2	Homogeneous systems	159
	5.1.3	Non-homogeneous systems	160
	5.1.4	Overdetermined vs. underdetermined systems	161
	5.1.5	Solution methods	162
	5.1.6	Error bounds in numerical solutions	166
	5.1.6.1	Condition number	166

	5.1.6.2	Error bounds	167
5.2		Vector space theory	168
	5.2.1	Vector space	168
	5.2.2	Subspace	169
	5.2.3	Sum and direct sum	170
	5.2.4	Basis and dimensions	172
	5.2.5	Complex vector space vs. real vector space	174
5.3		Linear maps & linear operators	176
	5.3.1	Basic concepts of linear maps	176
	5.3.2	Fundamental theorem of linear maps	177
	5.3.3	Isomorphism	179
	5.3.4	Coordinate map properties	181
	5.3.5	Change of basis and similarity	182
	5.3.5.1	Change of basis for coordinate vector	182
	5.3.5.2	Change of basis for linear maps	182
	5.3.6	Linear maps and matrices	182
	5.3.6.1	Similarity	184
5.4		Fundamental theorems of ranks and linear algebra	185
	5.4.1	Basics of ranks	185
	5.4.2	Fundamental theorem of ranks	186
	5.4.3	Fundamental theorem of linear algebra	187
5.5		Complementary subspaces and projections	189
	5.5.1	General complementary subspaces	189
	5.5.2	Orthogonal complementary spaces and projections	191
	5.5.3	Decomposition of orthogonal projectors	195
5.6		Orthonormal basis and projections	198
	5.6.1	Gram-Schmidt Procedure	198
	5.6.2	Orthogonal-triangular decomposition	198
	5.6.3	Orthonormal basis for linear operators	199
	5.6.4	Riesz representation theorem	200
5.7		Eigenvectors and eigenvalues of Matrices: general theory	201
	5.7.1	Existence and properties of eigenvalues	201
	5.7.2	Properties of eigenvectors	203
	5.7.3	Right and left eigenvectors	204
	5.7.4	Diagonalizable matrices	205
5.8		Eigenvalue and eigenvectors of matrices: case studies	208
	5.8.1	Real diagonalizable matrix	208
	5.8.2	Real symmetric matrix	209
	5.8.2.1	Spectral properties	209
	5.8.2.2	Rayleigh quotients	211
	5.8.2.3	Pointcare inequality	214

5.8.3	Hermitian matrix	215
5.8.4	Matrix congruence	217
5.8.5	Complex symmetric matrix	218
5.8.6	Unitary, orthonormal & rotation matrix	218
5.9	Singular Value Decomposition theory	220
5.9.1	SVD fundamentals	220
5.9.2	SVD and matrix norm	222
5.9.3	SVD vs. eigendecomposition	223
5.9.4	SVD low rank approximation	224
5.9.4.1	Frobenius norm low rank approximation	224
5.9.4.2	Two-norm low rank approximation	226
5.10	Generalized eigenvectors and Jordan normal forms	228
5.10.1	Generalized eigenvectors	228
5.10.2	Upper triangle matrix and nilpotent matrix	231
5.10.3	Jordan normal forms	233
5.11	Matrix factorization	237
5.11.1	Orthogonal-triangular decomposition	237
5.11.2	LU decomposition	238
5.11.3	Cholesky decomposition	238
5.12	Positive definite matrices and quadratic forms	240
5.12.1	Quadratic forms	240
5.12.2	Real symmetric non-negative definite matrix	241
5.12.2.1	Characterization	241
5.12.2.2	Decomposition and transformation	244
5.12.2.3	Matrix square root	245
5.12.2.4	Maximization of quadratic forms	246
5.12.2.5	Gramian matrix	248
5.12.3	Completing the square	249
5.13	Matrix norm and spectral estimation	250
5.13.1	Basics	250
5.13.2	Singularity from matrix norm and spectral radius	251
5.13.3	Gerschgorin theorem	252
5.13.4	Irreducible matrix and stronger results	253
5.14	Pseudoinverse of matrix	254
5.14.1	Pseudoinverse for full rank system	254
5.14.2	Pseudoinverse for general matrix	256
5.14.3	Application in linear systems	258
5.15	Multilinear forms	261
5.15.1	Bilinear forms	261
5.15.2	Multilinear forms	262
5.16	Determinant	265

5.16.1	Basic properties	265
5.16.2	Vandermonde matrix and determinant	271
5.17	Numerical iteration analysis	273
5.17.1	Numerical linear equation solution	273
5.17.1.1	Goals and general principles	273
5.17.1.2	Jacobi algorithm	273
5.17.1.3	Gauss Seidel algorithm	274
5.17.2	Power method for eigen-decomposition	274
5.18	Notes on bibliography	277
6	BASIC FUNCTIONAL ANALYSIS	279
6.1	Normed vector space	281
6.1.1	Basic properties	281
6.1.2	Equivalence of norms	283
6.2	Contraction mapping and fixed point theorems	285
6.2.1	Complete normed space (Banach space)	285
6.2.2	Contraction mapping	286
6.2.3	Banach fixed point theorem	287
6.2.4	Applications in root finding	288
6.2.5	Application to numerical linear equations	288
6.2.6	Applications to integral and differential equations	289
6.3	Inner product space and Hilbert space	292
6.3.1	Inner product space (pre-Hilbert space) and Hilbert space	292
6.3.1.1	Foundations	292
6.3.2	Hilbert spaces	294
6.3.2.1	Basics	294
6.3.3	Orthogonal decomposition of Hilbert spaces	295
6.3.3.1	Orthogonality	295
6.3.4	Projection and orthogonal decomposition	296
6.4	Approximations in Hilbert space	299
6.4.1	Approximation via projection	299
6.4.2	Application examples	300
6.4.2.1	Orthogonal projection and normal equations in \mathbb{R}^n	300
6.4.2.2	Approximation by continuous polynomials	302
6.4.2.3	Legendre polynomial via Gram-Schmidt process	303
6.5	Orthonormal systems	304
6.5.1	Basic definitions	304
6.5.2	Gram-Schmidt process	304
6.5.3	Properties of orthonormal systems	304
6.5.4	Orthonormal expansion in Hilbert space	306
6.5.5	Complete orthonormal system	307
6.5.5.1	Weierstrass approximation theorem for polynomials	309

6.5.5.2	Examples of complete orthonormal function set	309
6.6	Theory for trigonometric Fourier Series	311
6.6.1	Basic definitions	311
6.6.2	Completeness of Fourier series	313
6.6.3	Complex representation	314
6.7	Fourier transform	316
6.7.1	Definitions and basic concepts	316
6.7.2	Convolution theorem	318
6.7.3	Fourier transform and Fourier series	319
6.7.4	Discrete Fourier transform	320
6.7.4.1	Properties	320
6.8	Notes on bibliography	323

ii mathematical optimization methods

7	UNCONSTRAINED NONLINEAR OPTIMIZATION	326
7.1	Optimality conditions	328
7.1.1	Optimality concepts	328
7.1.2	Necessary and sufficient conditions	330
7.1.3	Special case: unconstrained quadratic programming	333
7.2	Line search method	335
7.2.1	A generic algorithm	335
7.2.2	Theory and computation of descent directions	336
7.2.2.1	Gradient descent direction and properties	336
7.2.2.2	Curvature-modified descent direction	337
7.2.2.3	Quasi-Newton method	339
7.2.2.4	Subspace optimization in quadratic forms	341
7.2.3	Theory and computation of step length	342
7.2.3.1	Overview	342
7.2.3.2	Lipschitz bounded convex functions	342
7.2.3.3	Backtracking-Armijo step size search	345
7.2.3.4	Wolfe condition	347
7.2.4	Complete algorithms	348
7.3	Trust region method	350
7.3.1	Motivation and the framework	350
7.3.2	Cauchy point method	351
7.3.3	Exact solution method	353
7.3.4	Approximate method	353
7.4	Conjugate gradient method	356
7.4.1	Motivating problems	356
7.4.2	Theory conjugate direction	357
7.4.3	Linear conjugate gradient algorithm	358

7.5	Least square problems	360
7.5.1	Linear least square theory and algorithm	360
7.5.1.1	Linear least square problems	360
7.5.1.2	SVD methods	361
7.5.1.3	Extension to L^p norm optimization	361
7.5.2	nonlinear least square problem	362
7.5.3	Line search Gauss-Newton method	363
7.5.4	Trust region method	365
7.5.5	Application: roots for nonlinear equation	365
7.6	Notes on bibliography	367
8	CONSTRAINED NONLINEAR OPTIMIZATION	369
8.1	Quadratic optimization I: equality constraints	371
8.1.1	Problem formulation	371
8.1.2	Optimality condition	371
8.1.2.1	General case	371
8.1.2.2	Positive semi-definitive quadratic programming	374
8.1.3	Solving KKT systems	375
8.1.3.1	Factorization approach	375
8.1.3.2	Range space approach	376
8.1.4	Linear least square with linear constraints	376
8.1.4.1	Least norm problem	376
8.1.5	Application: Markovitz Portfolio Optimization Model	378
8.2	Quadratic optimization II: inequality constraints	381
8.2.1	Problem formulation	381
8.2.2	Optimality conditions	381
8.2.2.1	Pure inequality case	381
8.2.2.2	General constrained optimization	383
8.2.2.3	Positive semi-definitive quadratic programming	384
8.2.3	Primal active-set method	385
8.2.4	Gradient projection method	390
8.2.5	Dual convex quadratic programming	391
8.3	General equality constrained optimization	393
8.3.1	Feasible path and optimality	393
8.3.2	Constraint qualification and Lagrange theory	394
8.3.3	Second order condition	397
8.4	General inequality constrained optimization	401
8.4.1	Feasible path and optimality	401
8.4.2	Constraint qualifications and KKT conditions	402
8.4.3	Second order conditions	406
8.5	Envelope theorem and sensitive analysis	410
8.6	Notes on bibliography	413

9	LINEAR OPTIMIZATION	415
9.1	Equality constrained linear programming	416
9.2	Inequality constrained linear programming	418
9.2.1	Linear optimization with inequality constraints	418
9.2.2	Geometry of linear programming	418
9.2.3	Optimality property and condition	420
9.2.4	Standard form of linear programming	421
9.2.5	Application examples	422
9.3	Linear programming geometry and simplex algorithm	423
9.3.1	Geometrical approach to linear programming	423
9.3.1.1	Overview	423
9.3.1.2	Vertex and optimality	423
9.3.1.3	Descent direction at a vertex	427
9.3.1.4	Stepping along a descent direction	428
9.3.2	The simplex algorithm	429
9.4	Interior point method	430
9.4.1	Optimality condition	430
9.4.2	Newton step and perturbed system	432
9.4.3	Algorithms	434
9.5	Notes on bibliography	436
10	CONVEX ANALYSIS AND CONVEX OPTIMIZATION	438
10.1	Affine sets	440
10.1.1	Basic concepts	440
10.1.2	Affine independence and dimensions	442
10.2	Convex sets and properties	446
10.2.1	Concepts of convex sets	446
10.2.2	Projection theorems	448
10.2.3	Separation theorems	449
10.2.3.1	Separating hyperplane theorem	449
10.2.3.2	Farka's lemma	451
10.3	Convex functions	454
10.3.1	Basic concepts	454
10.3.2	Connection to convex set	456
10.3.3	Strongly convex functions	456
10.3.4	Operations preserve convexity	457
10.3.5	Convexity and derivatives	458
10.3.6	Subgradient	460
10.4	Duality theory	462
10.5	Convex optimization and optimality conditions	466
10.5.1	Local optimality vs. global optimality	466
10.5.2	Unconstrained optimization optimality conditions	467

10.5.3	Constrained optimization optimality conditions	467
10.6	Subgradient methods	472
10.6.1	A generic algorithm for unconstrained problem	472
10.6.2	Convergence under Lipschitz smoothness	474
10.6.3	Projected gradient methods	477
10.6.3.1	Foundations	477
10.6.3.2	Algorithms	479
10.6.4	Proximal gradient methods	480
10.6.4.1	Foundations	480
10.6.4.2	Algorithms	481
10.6.4.3	Case study: sparsity regularization problem	482
10.7	Notes on bibliography	484
11	BASIC GAME THEORY	486
11.1	Static normal form game	487
11.1.1	Normal form game concepts	487
11.1.2	Pure strategy and equilibrium	487
11.1.2.1	Solution concepts	487
11.1.3	Mixed strategy and equilibrium	491
11.1.4	Pareto optimality	492
11.2	Zero-sum matrix game	494
11.2.1	Fundamentals	494
11.2.2	Optimal strategy and Nash equilibrium	495
11.2.3	Saddle points as solutions	497
11.2.4	Maxmin strategies and Nash equilibrium	498
11.2.5	Linear programming approach to optimal strategy	501
11.3	Notes on bibliography	505
 iii classical statistical methods		
12	PROBABILITY THEORY	508
12.1	Sigma algebra	512
12.1.1	sigma algebra concepts	512
12.1.2	Generation of sigma algebra	512
12.1.3	Partition of sample space	513
12.1.4	Filtration & information	513
12.1.5	Borel σ algebra	514
12.1.6	Measurable set and measurable space	515
12.2	Probability space	518
12.2.1	Event, sample point and sample space	518
12.2.2	Probability space	518
12.2.3	Properties of probability measure	520
12.2.4	Conditional probability	520

12.2.4.1	Basics	520
12.2.4.2	Independence of events and sigma algebra	522
12.3	Measurable map and random variable	524
12.3.1	Random variable	524
12.3.2	Image measure	525
12.3.3	σ algebra of random variables	526
12.3.4	Independence of random variables	526
12.4	Distributions of random variables	528
12.4.1	Basic concepts	528
12.4.1.1	Probability mass function	528
12.4.1.2	Distributions on \mathbb{R}^n	528
12.4.1.3	Probability density function	529
12.4.1.4	Conditional distributions	530
12.4.1.5	Bayes law	531
12.4.2	Independence	532
12.4.3	Conditional independence	534
12.4.4	Transformations	534
12.4.4.1	Transformation for univariate distribution	534
12.4.4.2	Location-scale transformation	535
12.4.4.3	Transformation for multivariate distribution	537
12.5	Expectation	540
12.5.1	Failure of elementary approach	540
12.5.2	Formal definitions	540
12.5.3	Properties of expectation	541
12.6	Variance and covariance	543
12.6.1	Basic properties	543
12.6.2	Conditional variance	544
12.7	Characteristic function and Moment generating functions	545
12.7.1	Moment generating function	545
12.7.2	Characteristic function	546
12.7.3	Joint moment generating functions for random vectors	548
12.7.4	Probability generating function	548
12.7.5	Cumulants	550
12.8	Conditional expectation	553
12.8.1	General intuitions & comments	553
12.8.2	Formal definitions	553
12.8.3	Different versions of conditional expectation	555
12.8.3.1	Conditioning on an event	555
12.8.3.2	Conditioning on a discrete random variable as a new random variable	555
12.8.3.3	Condition on random variable vs. event vs σ algebra	556

12.8.4	Properties	556
12.8.4.1	Linearity	556
12.8.4.2	Taking out what is known	557
12.8.4.3	Law of iterated expectations	557
12.8.4.4	Conditioning on independent random variable/ σ algebra	558
12.8.4.5	Least Square minimizing property	559
12.9	The Hilbert space of random variables	560
12.9.1	Definitions	560
12.9.2	Subspaces, projections, and approximations	560
12.9.3	Connection to conditional expectation	565
12.10	Probability inequalities	568
12.10.1	Some common inequalities	568
12.10.2	Chernoff bounds	573
12.11	Convergence of random variables	574
12.11.1	Different levels of equivalence among random variables	574
12.11.2	Convergence almost surely	574
12.11.3	Convergence in probability	575
12.11.3.1	Basics	575
12.11.3.2	Algebraic properties	576
12.11.4	Mean square convergence	577
12.11.5	Convergence in r th mean	578
12.11.6	Convergence in distribution	578
12.11.6.1	Convergence in probability vs in distribution	578
12.12	Finite sampling models	580
12.12.1	Counting principles	580
12.12.2	Matching problem	583
12.12.3	Birthday problem	585
12.12.4	Coupon collection problem	586
12.12.5	Balls into bins model	587
12.13	Law of Large Number and Central Limit theorem	590
12.13.1	Law of Large Numbers	590
12.13.2	Central limit theorem	591
12.13.3	Delta method & generalized CLT	593
12.14	Order statistics	596
12.15	Information theory	600
12.15.1	Concept of entropy	600
12.15.2	Entropy maximizing distributions	601
12.15.3	KL divergence	605
12.15.4	Conditional entropy and mutual information	606
12.15.5	Cross-entropy	607

12.16	Notes on bibliography	609
13	STATISTICAL DISTRIBUTIONS	612
13.1	Common distributions and properties	614
13.1.1	Bernoulli distribution	614
13.1.2	Normal distribution	614
13.1.3	Half-normal distribution	616
13.1.4	Laplace distribution	617
13.1.5	Multivariate Gaussian/normal distribution	618
13.1.5.1	Basic definitions	618
13.1.5.2	Affine transformation and its consequences	620
13.1.5.3	Marginal and conditional distribution	621
13.1.5.4	Box Muller transformation	623
13.1.6	Lognormal distribution	623
13.1.6.1	Univariate lognormal distribution	623
13.1.6.2	Extension to univariate lognormal distribution	625
13.1.6.3	Moment matching approximation	627
13.1.6.4	Multivariate lognormal distribution	629
13.1.7	Exponential distribution	629
13.1.8	Poisson distribution	631
13.1.9	Gamma distribution	632
13.1.10	Geometric distribution	634
13.1.11	Binomial distribution	635
13.1.12	Hypergeometric distribution	637
13.1.13	Beta distribution	638
13.1.14	Multinomial distribution	640
13.1.15	Dirichlet distribution	641
13.1.16	χ^2 -distribution	643
13.1.16.1	Basic properties	643
13.1.16.2	Quadratic forms and chi-square distribution	644
13.1.16.3	Noncentral chi-squared distribution	647
13.1.17	Wishart distribution	647
13.1.18	t -distribution	648
13.1.18.1	Standard t distribution	648
13.1.18.2	classical t distribution	649
13.1.18.3	Multivariate t distribution	650
13.1.18.4	Student's Theorem	650
13.1.19	F -distribution	652
13.1.20	Empirical distributions	653
13.1.21	Heavy-tailed distributions	653
13.1.21.1	Basic characterization	653
13.1.21.2	Pareto and power distribution	654

13.1.21.3	Student t distribution family	654
13.1.21.4	Gaussian mixture distributions	655
13.2	Characterizing distributions	658
13.2.1	Skewness and kurtosis	658
13.2.2	Quantiles and percentiles	660
13.2.2.1	Basics	660
13.2.2.2	Cornish-Fisher expansion	661
13.2.3	Exponential families	662
13.3	Cochran's theorem	664
13.4	Notes on bibliography	667
14	STATISTICAL ESTIMATION THEORY	669
14.1	Estimator theory	672
14.1.1	Overview	672
14.1.2	Statistic	672
14.1.3	Estimators properties	673
14.1.3.1	Basic concepts	673
14.1.3.2	Variance-bias decomposition	674
14.1.3.3	Consistence	676
14.1.3.4	Efficiency	678
14.1.4	Robust statistics	679
14.2	Method of moments	681
14.3	Maximum likelihood estimation	683
14.3.1	Basic concepts	683
14.3.2	Examples	683
14.4	Information and efficiency	686
14.4.1	Fish information	686
14.4.2	Information matrix for common distributions	688
14.4.2.1	Bernoulli distribution	688
14.4.2.2	Normal distribution	688
14.4.3	Cramer-Rao lower bound	689
14.4.3.1	Information inequality	689
14.4.3.2	Cramer-Rao lower bound: univariate case	690
14.4.3.3	Cramer-Rao lower bound: multivariate case	691
14.4.3.4	Efficient estimator	693
14.4.4	Fisher information characterization of MLE	695
14.4.4.1	Properties of score function	695
14.4.4.2	Fisher information and MLE	696
14.4.4.3	MLE efficiency	696
14.4.5	MLE for normal distribution	698
14.4.6	Asymptotic properties of MLE	700
14.5	Sufficiency and data reduction	703

14.5.1	Sufficient estimators	703
14.5.2	Factorization theorem	704
14.6	Bootstrap method	707
14.7	Hypothesis testing general theory	709
14.7.1	Basics	709
14.7.2	Characterizing errors and power	712
14.7.3	Power of a statistical test	713
14.7.4	Common statistical tests	715
14.7.4.1	Chi-square goodness-of-fit test	715
14.7.4.2	Chi-square test for statistical independence	717
14.7.4.3	Kolmogorov-Smirnov goodness-of-fit test	718
14.8	Hypothesis testing on normal distributions	719
14.8.1	Normality test	719
14.8.2	Sample mean with known variance	719
14.8.3	Sample mean with unknown variance	721
14.8.4	Variance test	721
14.8.5	Variance comparison test	722
14.8.6	Person correlation t test	722
14.8.7	Two sample tests	723
14.8.7.1	Two-sample z test	723
14.8.7.2	Two-sample t test	723
14.8.7.3	Paired Data	724
14.8.8	Interval estimation for normal distribution	724
14.9	Notes on bibliography	726
15	MULTIVARIATE STATISTICAL METHODS	729
15.1	Multivariate data and distribution	732
15.1.1	Sample statistics	732
15.1.2	Multivariate Gaussian distribution	733
15.1.3	Estimation methods	735
15.1.3.1	Maximum likelihood estimation	735
15.1.3.2	Weighted estimation	738
15.2	Principal component analysis (PCA)	739
15.2.1	Statistical fundamentals of PCA	739
15.2.1.1	PCA for random vectors	739
15.2.1.2	Sample principal components	740
15.2.2	Geometric fundamentals of PCA	743
15.2.2.1	Optimization approach	743
15.2.2.2	Properties	745
15.2.3	Probabilistic PCA	746
15.2.4	Applications	748
15.2.4.1	Eigenfaces and eigendigits	748

15.2.4.2	Interest rate curve dynamics modeling	750
15.3	Canonical correlation analysis	754
15.3.1	Basics	754
15.3.2	Sparse CCA	756
15.4	Copulas and dependence modeling	758
15.4.1	Definitions and properties	758
15.4.2	Copulas and distributions	762
15.4.2.1	Fundamentals	762
15.4.2.2	Survival copula	768
15.4.2.3	Partial differential and conditional distribution	769
15.4.3	Common copula functions	773
15.4.3.1	Gaussian copula	773
15.4.3.2	t copula	778
15.4.3.3	Common copula functions: other copula	778
15.4.4	Dependence and copula	779
15.4.4.1	Linear correlations	779
15.4.4.2	Rank correlations	781
15.4.4.3	Tail dependence	787
15.4.5	Estimating copula function	789
15.4.5.1	Empirical copula method	789
15.4.5.2	Maximum likelihood method	790
15.4.6	Applications of copula	791
15.4.6.1	Generating correlated uniform random number	791
15.4.6.2	Generating general correlated random number	793
15.4.6.3	Multivariate distribution approximation with Gaussian copula	797
15.5	Covariance structure and factor analysis	798
15.5.1	The orthogonal factor model	798
15.5.1.1	Motivation and factor models	798
15.5.1.2	Covariance structure implied by factor model	798
15.5.2	Parameter estimation	800
15.5.2.1	Data collection and preparation	800
15.5.2.2	PCA method	801
15.5.2.3	Maximum likelihood method	802
15.5.3	Factor score estimation	802
15.5.4	Application I: Joint default modeling	803
15.5.4.1	Single factor model	803
15.5.4.2	Multiple factor model	806
15.5.5	Application II: factor models for stock return	807
15.5.5.1	Overview	807
15.5.5.2	The Fama-French 3 factor model	809

15.6	Graphical models	813
15.6.1	Fundamentals	813
15.7	Notes on Bibliography	820
16	LINEAR REGRESSION ANALYSIS	823
16.1	Linear regression analysis: basics	826
16.1.1	Linear regression models	826
16.1.2	Ordinary least square (OLS) solutions	829
16.1.2.1	Review on orthogonal projections	829
16.1.2.2	OLS results	829
16.1.2.3	OLS results with demeaned data	835
16.1.2.4	Orthogonal input and successive regression	838
16.1.2.5	Frisch-Waugh-Lovell(FWL) theorem and partial regression	839
16.1.2.6	Gauss-Markov theorem	843
16.1.2.7	Residual and variance estimation	845
16.1.2.8	Forecasting analysis with normality assumption	846
16.1.3	Hypothesis testing and analysis of variance	848
16.1.3.1	Distribution of coefficients	849
16.1.3.2	t test and normality test of single coefficients	851
16.1.3.3	F lack-of-fit test	853
16.1.3.4	χ^2 test for variance	855
16.1.4	Maximum likelihood method with normality assumption	856
16.1.5	Asymptotic properties of least square solutions	858
16.1.5.1	Asymptotic properties of standard OLS	858
16.1.5.2	Asymptotic efficiency of standard OLS	860
16.1.6	Partial and multiple correlation	860
16.1.6.1	Multiple correlation coefficient, R^2	860
16.1.6.2	Partial correlation coefficient	863
16.1.7	Generalized linear regression (GLR)	864
16.1.7.1	Linear regression with structural error	864
16.1.7.2	Generalized least square solution	865
16.1.7.3	Gauss-Markov theorem for GLR	867
16.1.7.4	Feasible GLS	868
16.1.8	Linear structure in joint distributions	868
16.2	Model specification and selection	871
16.2.1	Model order mis-specification	871
16.2.1.1	Omission of relevant regressors	871
16.2.1.2	Inclusion of irrelevant regressors	872
16.2.2	Model selection methods	874
16.2.2.1	Adjusted R square method	874
16.2.2.2	F test method	874

16.2.2.3	Information criterion methods	876
16.2.2.4	Bayesian information criterion (BIC)	877
16.2.3	Test for structure change	878
16.3	Linear regression analysis: diagnostics & solutions	880
16.3.1	Multi-collinearity	880
16.3.1.1	Detection and characterization	880
16.3.1.2	Regressor linear regression and variance inflation factor	880
16.3.1.3	Principal component linear regression (PCLR)	883
16.3.2	Rank deficiency and rigid regression	883
16.3.3	Heteroskedasticity	885
16.3.3.1	Test for heteroskedasticity	885
16.3.3.2	Heteroskedasticity robust estimator	886
16.3.3.3	Feasible weighted least square	886
16.3.4	Residual normality test	888
16.3.4.1	Jarque-Bera test	888
16.3.4.2	D'Agostino's K^2 test	888
16.3.5	Autocorrelation of errors	889
16.3.5.1	Motivation and general remarks	889
16.3.5.2	Test of autocorrelation of errors	890
16.3.5.3	Models with known autocorrelation	892
16.3.5.4	Transformation to generalized linear regression	893
16.3.6	Outliers analysis and robust linear regression	895
16.3.6.1	Outliers and influential points	895
16.3.6.2	Outlier impact analysis	896
16.3.6.3	Robust M-estimation linear regression	899
16.3.7	Visual diagnosis	902
16.4	Linear regression case studies	905
16.4.1	Standard linear regression	905
16.4.2	Boston Housing example	907
16.5	Multivariate multiple linear regression (MMLR)	910
16.5.1	Canonical MMLR	910
16.5.1.1	Motivation and model	910
16.5.1.2	Ordinary least square solution	911
16.5.2	Reduced rank regression	912
16.6	Notes on Bibliography	917
17	MONTE CARLO METHODS	919
17.1	Generating random variables	921
17.1.1	Inverse transform method	921
17.1.2	Box-Muller method for standard normal random variable	923
17.1.3	Acceptance-rejection method	923

17.1.4	Composition approach	925
17.1.5	Generate dependent continuous random variables	927
17.1.5.1	Multivariate normal and lognormal distribution	927
17.1.5.2	Multivariate student t distribution	927
17.1.5.3	General joint distribution	928
17.1.6	Generate discrete random variables	928
17.1.6.1	Generate single discrete random variables	928
17.1.6.2	Generate correlated discrete random variables	929
17.2	Monte Carlo integration	932
17.2.1	Naive approach	932
17.2.2	Importance sampling	933
17.3	Markov chain Monte Carlo	936
17.3.1	Basics	936
17.3.1.1	Markov chain Monte Carlo (MCMC)	936
17.3.2	Metropolis-Hasting algorithm	937
17.3.3	Gibbs sampling	939
17.4	Monte Carlo for random processes	940
17.4.1	Simulating stochastic differential equations	940
17.4.1.1	Simulating Brownian motion	940
17.4.1.2	Simulating linear arithmetic SDE	940
17.4.1.3	Simulating linear geometric SDE	941
17.4.1.4	Simulation mean-reversion(OU) process	941
17.4.2	Stochastic interpolation	942
17.4.2.1	Interpolating Gaussian processes	942
17.4.2.2	Interpolating one Dimensional Brownian motions	942
17.4.2.3	Interpolating multi-dimensional Brownian motions	945
17.5	Monte Carlo variance reduction	948
17.5.1	Antithetic sampling	948
17.5.1.1	Basic principles	948
17.5.1.2	Methods and analysis	949
17.5.2	Control variates	951
17.5.2.1	Basic principles	951
17.5.2.2	Multiple control variates	953
17.6	Notes on bibliography	954

iv dynamics modeling methods

18	MODELS AND ESTIMATION IN LINEAR DYNAMICAL SYSTEMS	957
18.1	Difference equation	960
18.1.1	Existence and uniqueness of solutions	960
18.1.2	Linear difference equations	960
18.1.3	Solution to non-homogeneous equation	962

18.1.4	Linear equations with constant coefficients	963
18.2	Differential equations	967
18.2.1	Existence & uniqueness of solution	967
18.2.2	Linear differential equations	968
18.2.2.1	Concepts	968
18.2.2.2	Wronskian and linear independence	969
18.2.2.3	General solution theory	973
18.2.3	Linear homogeneous differential equations with constant coefficients	976
18.2.3.1	The key identity	976
18.2.3.2	The case of real roots	977
18.2.3.3	The case of complex roots	978
18.2.3.4	The complete solution set	979
18.2.4	Solution to non-homogeneous ODEs	981
18.2.4.1	General principles	981
18.2.4.2	Key identity approach	982
18.2.5	First order linear differential equation	988
18.3	Linear system	990
18.3.1	Solution space for linear homogeneous system	990
18.3.2	Linear independence and the Wronskian	992
18.3.3	The fundamental system and solution method	994
18.3.4	The non-homogeneous linear equation	995
18.3.5	Conversion of linear differential/difference equation to linear systems	998
18.3.6	Solution method for discrete system	1000
18.4	Linear system with constant coefficients	1002
18.4.1	General solutions	1002
18.4.2	System eigenvector method: continuous-time system	1002
18.4.2.1	Diagonalizable system	1002
18.4.2.2	two-by-two non-diagonalizable system	1009
18.4.2.3	Non-diagonalizable system	1010
18.4.3	System eigenvector method: discrete-time system	1012
18.4.3.1	Discrete-time system	1012
18.4.4	Equilibrium point	1013
18.4.4.1	Discrete-time system	1013
18.4.4.2	Continuous-time system	1014
18.4.5	Stability	1015
18.4.6	Complex eigenvalues/eigenvectors	1017
18.4.7	Boundedness of linear systems	1017
18.4.8	One dimensional Nonlinear dynamical system analysis	1018
18.5	Least square estimation of constant vectors	1020

18.5.1	linear static estimation from single measurement with no prior information	1020
18.5.2	linear static estimation from single measurement with prior information	1021
18.5.3	Batch and recursive least square estimation with multiple measurements	1022
18.5.4	Nonlinear least square estimation	1024
18.6	Kalman filter	1026
18.6.1	Preliminary: error propagation in linear systems	1026
18.6.1.1	Discrete-time system	1026
18.6.1.2	Continuous-time system	1026
18.6.2	Batch estimation	1027
18.6.3	From batch estimation to Kalman filter	1028
18.6.4	Extended Kalman filter for nonlinear system	1029
18.7	Notes on bibliography	1031
19	STOCHASTIC PROCESS	1034
19.1	Stochastic process	1036
19.1.1	Basic definition and concepts	1036
19.1.2	Filtration and adapted process	1037
19.1.3	Natural filtration of a stochastic process	1038
19.1.4	Continuity of sample path	1040
19.1.5	Predictable process	1040
19.2	Stationary process	1042
19.2.1	Stationarity concepts	1042
19.2.2	Random phase and amplitude	1044
19.3	Gaussian process and Finite dimension distributions	1048
19.3.1	One-dimensional Gaussian process	1048
19.3.1.1	Definitions and properties	1048
19.3.1.2	Stationarity	1048
19.3.1.3	Examples	1049
19.3.2	finite dimensional distribution	1050
19.3.3	Gaussian process generated by Brownian motion	1050
19.4	Markov process	1053
19.5	Wiener process (Brownian motion)	1054
19.5.1	Basics	1054
19.5.2	Filtration for Brownian motion	1055
19.5.3	Quadratic variation	1056
19.5.4	Symmetries and scaling laws	1058
19.5.5	Non-differentiability and unbounded variation of path	1058
19.5.6	The reflection principle	1058
19.5.6.1	Driftless case	1058

19.5.6.2	Drifting case	1060
19.5.7	Asymptotic behaviors	1060
19.5.8	Levy characterization of Brownian motion	1061
19.5.9	Discrete-time approximations	1063
19.6	Poisson process	1064
19.6.1	Basics	1064
19.6.2	Arrival and Inter-arrival Times	1065
19.7	Martingale theory	1067
19.7.1	Basics	1067
19.7.2	Exponential martingale	1069
19.7.3	Martingale transformation	1071
19.8	Stopping time	1073
19.8.1	Stopping time examples	1073
19.8.1.1	First passage time	1073
19.8.1.2	Trivial stopping time	1073
19.8.1.3	Counter example: last exit time	1074
19.8.2	Wald's equation	1074
19.8.3	Optional stopping	1075
19.8.4	Stopping time analysis of Wiener processes	1075
19.8.4.1	Minimum and maximum of a Wiener process	1075
19.8.4.2	Martingale method	1077
19.8.4.3	General method via Feynman Kac formula	1078
19.9	Notes on bibliography	1081
20	STOCHASTIC CALCULUS	1084
20.1	Ito integral	1086
20.1.1	Construction of Ito integral	1086
20.1.2	Properties of Ito integral	1088
20.1.3	Wiener integral and Riemman integral with Wiener process	1089
20.1.4	Quadratic variations	1092
20.2	Stochastic differential equations	1093
20.2.1	Ito Stochastic differential equations	1093
20.2.2	Ito's lemma	1094
20.2.2.1	one-dimensional version	1094
20.2.2.2	Multi-dimensional version	1094
20.2.2.3	Product rule and quotient rule	1094
20.2.2.4	Logorithm and exponential	1095
20.2.2.5	Integrals of Ito process	1096
20.2.2.6	Ito Integral by parts	1097
20.2.2.7	Fundamental theorem of Ito stochastic calculus	1098
20.2.3	Solutions to Ito stochastic differential equations	1098
20.2.4	Solution method to linear SDE	1099

20.2.4.1	State-independent linear arithmetic SDE	1099
20.2.4.2	State-independent linear geometric SDE	1100
20.2.4.3	Integral of state-independent linear arithmetic SDE	1102
20.2.4.4	Multiple dimension extension	1105
20.2.5	Exact SDE	1107
20.2.6	Calculation mean and variance from SDE	1108
20.2.7	Multi-dimensional Ito stochastic differential equations	1110
20.3	Ornstein-Uhlenbeck(OU) process	1112
20.3.1	OU process	1112
20.3.1.1	Constant coefficient OU process	1112
20.3.1.2	Time-dependent coefficient OU process	1116
20.3.1.3	Integral of OU process	1118
20.3.2	Exponential OU process	1122
20.3.3	Parameter estimation for OU process	1124
20.3.4	Multiple factor extension	1124
20.4	Brownian motion variants	1127
20.4.1	Brownian bridge	1127
20.4.1.1	Constructions	1127
20.4.1.2	Simulation	1130
20.4.1.3	Applications	1130
20.4.2	Geometric Brownian motion	1131
20.5	Notes on bibliography	1132
21	FOKKER-PLANCK EQUATION	1134
21.1	Fokker-Planck equations	1135
21.1.1	Formulations:one dimension	1135
21.1.1.1	Formulations: multiple dimension	1136
21.1.2	Steady state & detailed balance	1137
21.1.3	Averages and adjoint operator	1138
21.1.4	Backward equation	1140
21.1.5	Mean first passage time problem	1143
21.2	Smoluchowski/advection-diffusion equation	1145
21.3	Feynman Kac theorem and backward equation	1147
21.3.1	Feynman Kac theorem	1147
21.3.2	Backward equation	1150
21.3.3	Application to first hitting probability	1153
21.4	Advanced analysis for Brownian motion	1154
21.4.1	Kramers problem: barrier escape	1157
21.5	Notes on bibliography	1159
22	MARKOV CHAIN AND RANDOM WALK	1161
22.1	Discrete-time Markov chain	1163
22.1.1	The model	1163

22.1.2	Evolution of discrete chain	1165
22.2	Classification of states	1167
22.2.1	accessibility and communicating classes	1167
22.2.2	Transient and recurrent states and classes	1169
22.2.2.1	Transient and recurrent states	1169
22.2.2.2	From states to classes	1173
22.2.2.3	Qualitative classification of recurrent and transient classes	1174
22.2.3	Periodicity	1175
22.2.4	Positive and null recurrent	1177
22.2.5	Summary	1179
22.3	Absorption analysis	1180
22.3.1	Matrix structure for adsorption analysis	1180
22.3.2	Absorbing Markov chains	1182
22.3.3	Hitting and return analysis	1184
22.3.4	Examples	1186
22.3.4.1	Consecutive coin toss game	1186
22.4	Limiting behavior & distributions	1188
22.4.1	Preliminary: eigenvalue propoties of stochastic matrices	1188
22.4.1.1	Preliminary: Frobenius-Perron matrix theory	1188
22.4.1.2	More general situations	1190
22.4.2	Limiting theorem	1192
22.4.2.1	Limiting distribution	1192
22.4.2.2	Extensions via Long run return analysis	1196
22.4.3	Application: PageRank algorithm	1198
22.5	Detailed balance and spectral properties	1200
22.6	Random walk	1202
22.6.1	Basic concepts and properties	1202
22.6.2	Persistent random walk	1202
22.6.3	Asymptotic properties	1204
22.6.4	Gambler's ruin problems	1204
22.7	Notes on bibliography	1209
23	TIME SERIES ANALYSIS	1211
23.1	Overview of time series analysis	1214
23.1.1	Introduction to time series	1214
23.1.2	Stationarity	1215
23.1.2.1	Stationarity concept	1215
23.1.2.2	Rolling analysis	1217
23.1.3	Remove trend and seasonality	1218
23.2	Linear stationary process theory	1221
23.2.1	Preliminaries: the lag operator and polynomial	1221

23.2.2	Linear process	1222
23.2.3	Autoregressive (AR) process	1224
23.2.3.1	Basics	1224
23.2.3.2	Stationarity and invertibility condition	1227
23.2.3.3	Forecasting	1228
23.2.4	Moving average (MA) process	1231
23.2.4.1	Basics	1231
23.2.4.2	Stationarity and invertibility	1234
23.2.4.3	Forecasting	1235
23.2.5	ARMA process	1239
23.2.5.1	Basic properties	1239
23.2.6	Unit root AR process	1240
23.2.6.1	Unit root process	1240
23.2.6.2	Trend stationarity vs. unit root process	1242
23.2.6.3	Unit root test	1242
23.2.6.4	Forecasting	1243
23.2.7	Correlation analysis	1243
23.2.7.1	Autocorrelation statistical analysis	1243
23.2.7.2	Partial autocorrelation function theory	1245
23.2.7.3	Correlogram analysis example	1247
23.2.8	Model analysis and calibration	1249
23.2.8.1	Order selection	1249
23.2.8.2	Yule-Walker equations and related methods	1250
23.2.8.3	Linear regression approach	1253
23.2.8.4	Maximum likelihood estimation	1255
23.2.8.5	Example: a toy example	1256
23.2.9	Wold Representation theorem	1259
23.3	Extensions to multivariate time series	1262
23.3.1	Introduction	1262
23.3.2	Vector autoregressive models	1263
23.3.2.1	VAR(1) model	1263
23.3.2.2	VAR(2) model	1265
23.3.2.3	VAR(p) model	1266
23.3.3	Vector moving-average model	1269
23.4	Autoregressive conditional heteroscedastic model	1272
23.4.1	ARCH models	1272
23.4.1.1	The motivation and the model	1272
23.4.1.2	Statistical properties	1274
23.4.1.3	Variance forecasting	1280
23.4.1.4	Detect ARCH effect	1283
23.4.1.5	Parameter estimation	1284

23.4.2	GARCH models	1284
23.4.2.1	The model	1284
23.4.2.2	Connecting GARCH to ARCH	1287
23.4.2.3	Variance forecasting	1287
23.5	Notes on Bibliography	1290

v statistical learning methods

24	SUPERVISED LEARNING PRINCIPLES	1293
24.1	The supervised learning problem	1295
24.1.1	Concepts	1295
24.1.2	Framework	1297
24.2	Variance bias trade-off	1299
24.2.1	Underfitting and Overfitting	1299
24.2.2	Variance and bias trade-off	1301
24.2.3	Examples	1304
24.2.3.1	Linear regression	1304
24.2.4	No free lunch theorem	1305
24.3	Model loss and evaluation	1307
24.3.1	Common loss functions	1307
24.3.2	Model evaluation metrics	1311
24.3.2.1	Regression metrics	1311
24.3.2.2	Classification metrics	1312
24.3.2.3	ROC and PRC metrics	1314
24.3.2.4	Metrics for imbalanced data	1316
24.4	Model selection methods	1317
24.4.1	The training-validation-testing idea	1317
24.4.2	Cross-validation	1317
24.5	Data and feature engineering	1322
24.5.1	Data preprocessing	1322
24.5.1.1	Data standardization	1322
24.5.1.2	Data normalization	1323
24.5.1.3	Handle categorical data	1323
24.5.1.4	Handle missing values	1324
24.5.1.5	Dimensional reduction	1324
24.5.1.6	Centering kernel matrix	1324
24.5.2	Feature engineering I: basic routines	1325
24.5.2.1	Nonlinear transformation	1325
24.5.2.2	Polynomial features	1325
24.5.2.3	Binning	1325
24.5.3	Feature engineering II: feature selection	1326
24.5.3.1	Filtering methods	1326

24.5.3.2	Recursive elimination methods	1327
24.5.3.3	Regularization methods	1327
24.5.4	Feature engineering III: feature extraction	1327
24.5.4.1	Text analytics	1328
24.5.4.2	Image	1328
24.5.4.3	Time series	1329
24.5.5	Imbalanced data	1330
24.5.5.1	Motivations	1330
24.5.5.2	Data resampling: undersampling	1331
24.5.5.3	Data resampling: upsampling	1331
24.5.5.4	Choice of loss functions, algorithms, and metrics	1332
24.6	Kernel methods	1333
24.6.1	Basic concepts of kernels and feature maps	1333
24.6.2	Mercer's theorem	1333
24.6.3	Common kernels	1335
24.6.4	Kernel trick and elementary algorithms using kernels	1337
24.6.5	Elementary algorithms	1337
24.7	Note on bibliography	1339
25	LINEAR MODELS FOR REGRESSION	1340
25.1	Standard linear regression	1341
25.1.1	Ordinary linear regression	1341
25.1.2	Application examples	1342
25.1.2.1	Boston housing prices	1342
25.2	Penalized linear regression	1346
25.2.1	Ridge regression	1346
25.2.1.1	Basics	1346
25.2.1.2	Dual form of ridge regression	1349
25.2.2	Lasso regression	1349
25.2.3	Elastic net	1352
25.2.4	Shrinkage Comparison	1352
25.2.5	Effective degree of freedom	1355
25.3	Basis function extension	1356
25.4	Note on bibliography	1358
26	LINEAR MODELS FOR CLASSIFICATION	1360
26.1	Logistic regression	1362
26.1.1	Logistic regression model	1362
26.1.2	Parameter estimation via maximum likelihood estimation	1363
26.1.3	Logistic regression with regularization	1367
26.1.4	Feature augmentation strategies	1368
26.1.5	Multinomial logistic regression	1369
26.1.6	Application examples	1370

26.1.6.1	South Africa heart disease	1370
26.1.6.2	Credit card fraud detection	1373
26.1.6.3	MNIST	1375
26.2	Gaussian discriminate analysis	1377
26.2.1	Linear Gaussian discriminant model	1377
26.2.1.1	The model	1377
26.2.1.2	Model parameter estimation	1378
26.2.1.3	Geometry of decision boundary	1378
26.2.2	Quadratic Gaussian discriminant model	1380
26.2.2.1	The model	1380
26.2.2.2	Model parameter estimation	1382
26.2.3	Application examples	1382
26.2.3.1	A toy example	1382
26.3	Fisher Linear discriminate analysis (Fisher LDA)	1386
26.3.1	One dimensional linear discriminant	1386
26.3.1.1	Basics	1386
26.3.1.2	Application in classification	1388
26.3.1.3	Possible issues	1390
26.3.2	Multi-dimensional linear discriminate	1391
26.3.2.1	Basics	1391
26.3.2.2	Application in classification	1392
26.3.3	Supervised dimensional reduction via Fisher LDA	1394
26.4	Separating hyperplane and Perceptron learning algorithm	1396
26.4.1	Basic geometry of hyperplanes	1396
26.4.2	The Perceptron learning algorithm	1397
26.5	Support vector machine classifier	1399
26.5.1	Motivation and formulation	1399
26.5.2	Optimality condition and dual form	1400
26.5.3	Soft margin SVM	1402
26.5.3.1	Basics	1402
26.5.3.2	Optimality condition for soft margin SVM	1403
26.5.3.3	Algorithm	1405
26.5.4	SVM with kernels	1407
26.5.5	A unified perspective from loss functions	1408
26.6	Note on bibliography	1411
27	GENERATIVE MODELS	1413
27.1	Naive Bayes classifier (NBC)	1414
27.1.1	Overview	1414
27.1.2	Binomial NBC	1414
27.1.3	Multinomial NBC	1416
27.1.4	Gaussian NBC	1419

27.1.5	Discussion	1421
27.2	Application	1422
27.2.1	Classifying documents using bag of words	1422
27.2.2	Credit card fraud prediction	1422
27.3	Supporting mathematical results	1425
27.3.1	Beta-binomial model	1425
27.3.1.1	The model	1425
27.3.1.2	Parameter inference	1426
27.3.2	Dirichlet-multinomial model	1428
27.3.2.1	The model	1428
27.3.2.2	Parameter inference	1431
28	K NEAREST NEIGHBORS	1433
28.1	Principles	1434
28.1.1	The algorithm	1434
28.1.2	Metrics and features	1435
28.2	Application examples	1437
29	TREE METHODS	1440
29.1	Preliminaries: entropy concepts	1441
29.1.1	Concept of entropy	1441
29.1.2	Conditional entropy and mutual information	1442
29.2	Classification tree	1446
29.2.1	Basic concepts of decision tree learning	1446
29.2.2	A generic tree-growth algorithm	1447
29.2.3	Splitting criterion	1448
29.2.4	Tree pruning	1452
29.2.5	Practical algorithms	1452
29.2.6	Examples	1455
29.2.6.1	Tree structures in Iris data classification	1455
29.3	Regression tree	1458
29.3.1	Basics	1458
29.3.2	Practical algorithms	1461
29.3.3	Examples	1461
29.3.3.1	A toy example	1461
29.3.3.2	Boston Housing prices	1462
30	ENSEMBLE AND BOOSTING METHODS	1465
30.1	Motivation and overview	1466
30.2	Bagging Methods	1468
30.2.1	A basic bagging method	1468
30.2.2	Tree bagging	1469
30.2.3	Random Forest	1471

30.3	Adaboost	1474
30.3.1	Adaboost classifier	1474
30.3.2	Adaboost regressor	1478
30.3.3	Additive model framework	1479
30.3.3.1	Generic additive model algorithm	1479
30.3.3.2	Adaboost as a special additive model	1480
30.4	Gradient boosting machines	1482
30.4.1	Fundamental	1482
30.4.2	Gradient boosting tree	1483
30.5	XGBoost	1488
30.6	Notes on Bibliography	1492
31	UNSUPERVISED STATISTICAL LEARNING	1494
31.1	Singular value decomposition (SVD) and matrix factorization	1496
31.1.1	SVD theory	1496
31.1.1.1	SVD fundamentals	1496
31.1.1.2	SVD and matrix norm	1498
31.1.1.3	SVD low rank approximation	1499
31.1.2	Principal component analysis (PCA)	1501
31.1.2.1	Statistical perspective of PCA	1501
31.1.2.2	Geometric fundamentals of PCA	1504
31.1.2.3	Robust PCA with outliers	1506
31.1.3	Sparse coding and dictionary learning	1508
31.1.3.1	Sparse coding	1508
31.1.3.2	Dictionary learning	1509
31.1.3.3	Online dictionary learning	1511
31.1.4	Non-negative matrix factorization	1513
31.2	Advanced applications of matrix factorization methods	1515
31.2.1	Latent semantic analysis	1515
31.2.2	Collaborative filtering in recommender systems	1518
31.2.3	Co-occurrence based word embedding	1523
31.3	Manifold learning	1525
31.3.1	Overview	1525
31.3.2	Preliminary: multidimensional scaling (MDS)	1525
31.3.2.1	Motivation	1525
31.3.2.2	Solution to classical MDS	1526
31.3.3	Isomap	1529
31.3.4	Kernel PCA	1531
31.3.5	Laplacian eigenmap	1533
31.3.5.1	Preliminary: graph Laplacian	1533
31.3.5.2	Laplacian eigenmap	1535
31.3.6	Diffusion map	1538

31.3.7	Application examples	1541
31.3.7.1	MNIST	1541
31.4	Clustering	1543
31.4.1	Overview	1543
31.4.2	K-means	1543
31.4.2.1	Canonical K-means	1543
31.4.2.2	K means++	1546
31.4.2.3	Kernel K means	1547
31.4.3	Density-based spatial clustering of applications with noise (DB-SCAN)	1548
31.4.4	Spectral clustering	1550
31.4.5	Gaussian mixture models (GMM)	1551
31.4.5.1	Preliminaries: Expectation Maximization (EM) algorithm	1551
31.4.5.2	The GMM model and algorithm	1553
31.4.6	Application examples	1555
31.4.6.1	Image segmentation	1555
31.5	Notes on Bibliography	1557
32	NEURAL NETWORK AND DEEP LEARNING	1560
32.1	Neural network foundations	1564
32.1.1	From machine learning to deep learning	1564
32.1.2	Neurons and neural networks	1565
32.1.2.1	Artificial neurons	1565
32.1.2.2	Artificial neural networks	1567
32.1.3	Universal approximation	1570
32.1.4	Training via backpropagation	1571
32.2	Optimization algorithms	1577
32.2.1	Motivation	1577
32.2.2	Full Batch gradient descent	1578
32.2.3	Minibatch stochastic gradient descent	1579
32.2.4	Adaptive gradient method	1580
32.2.4.1	Adaptive gradient (AdaGrad)	1580
32.2.4.2	RMSProp & AdaDelta	1580
32.2.5	Momentum method	1582
32.2.6	Combined together: adaptive momentum (Adam)	1583
32.3	Training and regularization techniques	1585
32.3.1	Choices of activation functions	1585
32.3.2	Weight initialization	1585
32.3.2.1	Motivation	1585
32.3.2.2	Xvaier initialization	1586
32.3.2.3	He initialization	1587

32.3.3	Data normalization	1587
32.3.3.1	Initial data standardization	1587
32.3.3.2	Batch normalization	1588
32.3.4	Regularization	1589
32.3.4.1	L_p regularization	1589
32.3.4.2	Weight decay	1589
32.3.4.3	Early stopping	1590
32.3.4.4	Dropout	1590
32.3.4.5	Data augmentation	1591
32.3.4.6	Label smoothing	1591
32.4	Feed-forward neural network examples	1592
32.4.1	Linear regression and classification	1592
32.4.2	Image classification	1594
32.4.3	Word embedding	1597
32.4.4	Sentiment analysis	1601
32.4.5	Approximating numerical partial differential equations	1603
32.5	Convolutional neural networks (CNN)	1607
32.5.1	Foundations	1607
32.5.2	CNN classical architectures	1610
32.5.2.1	LeNet	1610
32.5.2.2	AlexNet	1610
32.5.2.3	VGG	1611
32.5.2.4	ResNet	1612
32.6	CNN application examples	1615
32.6.1	Image classification	1615
32.6.2	Visualizing CNN	1616
32.6.2.1	Visualizing filters	1616
32.6.2.2	Visualizing classification activation map	1617
32.6.3	Autoencoders and denoising	1619
32.6.3.1	Autoencoders	1619
32.6.3.2	Denoising autoencoder	1620
32.6.4	Neural style transfer	1621
32.6.5	Visual based deep reinforcement learning	1624
32.6.6	Sentence classification	1626
32.7	Recurrent neural networks (RNN)	1629
32.7.1	Recurrent units	1629
32.7.1.1	Simple recurrent unit (SRU)	1629
32.7.1.2	Simple RNN and its approximation capability	1630
32.7.1.3	Backpropagation through time (BPTT)	1630
32.7.2	Recurrent unit variants	1632
32.7.2.1	Long short term memory (LSTM)	1632

32.7.2.2	Gated Recurrent Unit (GRU)	1635
32.7.3	Common RNN architectures	1636
32.8	RNN application examples	1640
32.8.1	Time series prediction	1640
32.8.1.1	Simple RNN prediction	1640
32.8.1.2	Deep autoregressive (DeepAR) model	1643
32.8.1.3	Deep factor model	1645
32.8.2	MNIST classification with sequential observation	1646
32.8.3	Sentiment classification	1646
32.8.4	Character-level language modeling	1647
32.8.4.1	Word classification	1647
32.8.4.2	Text generation	1648
32.9	Sequence-to-sequence modeling	1651
32.9.1	Encoder decoder model	1651
32.9.2	Attention mechanism	1653
32.10	Generative adversarial network (GAN)	1657
32.10.1	Canonical GAN	1657
32.10.1.1	Basics	1657
32.10.1.2	An example	1659
32.10.1.3	Understand training difficulties in GAN	1660
32.10.1.4	Deep Convolutional GAN (DCGAN)	1662
32.10.2	Conditional GAN	1664
32.10.3	Wasserstein GAN (WGAN)	1665
32.11	Notes on Bibliography	1669

vi optimal control and reinforcement learning methods

33	CLASSICAL OPTIMAL CONTROL THEORY	1675
33.1	Basic problem	1676
33.2	Controllability & observability	1677
33.3	Dynamic programming principle	1678
33.3.1	Principle of optimality	1678
33.3.2	The Hamilton-Jacobi-Bellman equation (finite horizon)	1678
33.3.3	The Hamilton-Jacobi-Bellman equation (infinite horizon)	1679
33.4	Deterministic linear quadratic control	1682
33.4.1	Linear quadratic control (finite horizon)	1682
33.4.2	Linear quadratic control(infinite horizon)	1683
33.5	Continuous-time stochastic optimal control	1685
33.5.1	HJB equation for general nonlinear systems	1685
33.5.2	Linear Gaussian quadratic system	1686
33.6	Stochastic dynamic programming	1687

33.6.1	Discrete-time Stochastic dynamic programming: finite horizon	1687
33.6.2	Discrete-time stochastic dynamic programming: infinite horizon	1689
33.6.2.1	Fundamentals	1689
33.6.2.2	Convergence analysis	1690
33.7	Notes on bibliography	1692
34	REINFORCEMENT LEARNING	1694
34.1	Preliminaries	1697
34.1.1	Notations	1697
34.1.2	Finite state Markov decision process	1697
34.1.3	Policy iteration and value iteration	1699
34.1.3.1	Policy iteration	1699
34.1.3.2	Value iteration	1703
34.2	Reinforcement learning theory	1707
34.2.1	Overview	1707
34.2.2	State-action Value function (Q function)	1709
34.2.3	Monte-Carlo method	1711
34.2.3.1	On-policy value estimation	1711
34.2.3.2	Off-policy value estimation	1713
34.2.3.3	MC-based reinforcement learning control	1713
34.2.4	TD(o) learning	1714
34.2.4.1	TD(o) for value estimation	1714
34.2.4.2	On-policy reinforcement learning control	1715
34.2.4.3	Off-policy reinforcement learning control	1716
34.2.5	TD(n) learning	1718
34.2.5.1	Motivation and concepts	1718
34.2.5.2	TD(n) for value estimation	1719
34.2.5.3	TD(n) for reinforcement learning control	1721
34.2.6	Standing challenges in reinforcement learning	1721
34.2.6.1	Curse of dimensionality	1721
34.2.6.2	Sample efficiency	1722
34.2.6.3	Exploration-exploitation dilemma	1722
34.2.6.4	Deadly triad	1722
34.3	Policy gradient learning	1723
34.3.1	Stochastic policy gradient fundamentals	1723
34.3.1.1	Preliminaries: derivative and expectation	1723
34.3.1.2	Theoretical framework based on finite-horizon trajectories	1724
34.3.1.3	Theoretical framework based on distributions*	1727
34.3.1.4	Estimate policy gradient and basic algorithms	1731

34.3.1.5	Bootstrap and Actor-Critic methods	1733
34.3.1.6	Common stochastic policies and their representations	1735
34.3.2	Advanced methods for policy gradient estimation	1737
34.3.2.1	Stochastic policy gradient with baseline	1737
34.3.2.2	Generalized advantage estimation	1740
34.3.2.3	Summary of stochastic gradient descent forms	1741
34.3.3	Deterministic policy gradient	1742
34.4	Algorithms zoo	1744
34.4.1	Neural Fitted Q Iteration (NFQ)	1744
34.4.2	Canonical deep Q learning	1745
34.4.3	DQN variants	1747
34.4.3.1	Overview	1747
34.4.3.2	Double Q learning	1748
34.4.3.3	Dueling network	1748
34.4.3.4	Deep Recurrent Q network (DRQN)	1749
34.4.3.5	Asynchronous Methods	1749
34.4.4	Universal value function approximator	1751
34.4.5	Deep deterministic policy gradient (DDPG) algorithm	1754
34.4.6	Twin-delayed deep deterministic policy gradient (TD3)	1756
34.4.7	Trust Region Policy Optimization (TRPO)	1759
34.4.7.1	TRPO	1759
34.4.7.2	Evaluating Hessian of KL-divergence	1761
34.4.8	Proximal Policy Optimization (PPO)	1763
34.4.9	Soft Actor-Critic(SAC)	1764
34.4.9.1	Entropy regulated reinforcement learning	1764
34.4.9.2	The SAC algorithm	1766
34.4.10	Evolution strategies	1767
34.5	Advanced training strategies	1771
34.5.1	Priority experience replay	1771
34.5.2	Hindsight experience generation	1772
34.5.3	Reverse goal generation	1773
34.5.4	Reverse goal generation on low-dimensional manifolds	1774
34.5.4.1	Key idea	1774
34.5.4.2	Example: navigation on a curved surface	1774
34.6	Notes on bibliography	1776

vii appendix

A	SUPPLEMENTAL MATHEMATICAL FACTS	1780
A.1	Basic logic for proof	1782
A.2	Some common limits	1783

A.3	Common series summation	1785
A.4	Some common spaces	1786
A.4.1	Notations on continuously differentiable functions	1786
A.5	Different modes of continuity	1788
A.5.1	continuity vs. uniform continuity	1789
A.6	Exchanges of limits	1790
A.6.1	Overall remark	1790
A.6.2	exchange limits with infinite summations	1790
A.6.3	Exchange limits with integration and differentiation	1790
A.6.4	Exchange differentiation with integration	1791
A.6.5	Exchange limit and function evaluations	1792
A.7	Useful inequalities	1793
A.7.1	Gronwall's inequality	1793
A.7.2	Inequality for norms	1793
A.7.3	Young's inequality for product	1794
A.8	Useful properties of matrix	1795
A.8.1	Matrix derivatives	1795
A.8.2	Matrix inversion lemma	1795
A.8.3	Block matrix	1796
A.8.4	Matrix trace	1797
A.8.5	Matrix elementary operator	1798
A.8.6	Matrix determinant	1800
A.9	Numerical integration	1801
A.9.1	Gaussian quadrature	1802
A.10	Vector calculus	1803
A.11	Numerical linear algebra computation complexity	1804
A.12	Distributions	1805
A.13	Common integrals	1806
A.14	Nonlinear root finding	1807
A.14.1	Bisection method	1807
A.14.2	Newton method	1807
A.14.3	Secant method	1807
A.15	Interpolation	1809
A.15.1	cubic interpolation	1809
Alphabetical Index		1811

Part I

MATHEMATICAL FOUNDATIONS