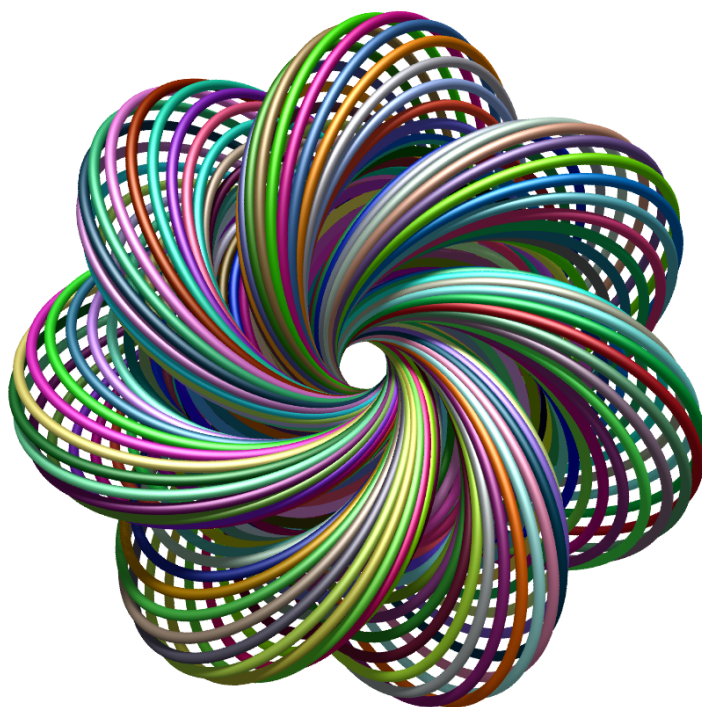


Essentials of Mathematical Methods:

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 36
- 3 ADVANCED CALCULUS 58
- 4 LINEAR ALGEBRA AND MATRIX ANALYSIS 138
- 5 BASIC FUNCTIONAL ANALYSIS 268

ii mathematical optimization methods

- 6 UNCONSTRAINED NONLINEAR OPTIMIZATION 315
- 7 CONSTRAINED NONLINEAR OPTIMIZATION 359
- 8 LINEAR OPTIMIZATION 405
- 9 CONVEX ANALYSIS AND CONVEX OPTIMIZATION 428
- 10 BASIC GAME THEORY 477

iii classical statistical methods

- 11 PROBABILITY THEORY 499
- 12 STATISTICAL DISTRIBUTIONS 603
- 13 STATISTICAL ESTIMATION THEORY 657
- 14 MULTIVARIATE STATISTICAL METHODS 716
- 15 LINEAR REGRESSION ANALYSIS 806
- 16 MONTE CARLO METHODS 900

iv dynamics modeling methods

- 17 MODELS AND ESTIMATION IN LINEAR SYSTEMS 936
- 18 STOCHASTIC PROCESS 1010

- 19 STOCHASTIC CALCULUS 1047
- 20 MARKOV CHAIN AND RANDOM WALK 1095
- 21 TIME SERIES ANALYSIS 1144

v statistical learning methods

- 22 SUPERVISED LEARNING PRINCIPLES 1223
- 23 LINEAR MODELS FOR REGRESSION 1268
- 24 LINEAR MODELS FOR CLASSIFICATION 1287
- 25 GENERATIVE MODELS 1338
- 26 K NEAREST NEIGHBORS 1360
- 27 TREE METHODS 1366
- 28 ENSEMBLE AND BOOSTING METHODS 1390
- 29 UNSUPERVISED STATISTICAL LEARNING 1420
- 30 NEURAL NETWORK AND DEEP LEARNING 1483

vi optimal control and reinforcement learning methods

- 31 CLASSICAL OPTIMAL CONTROL THEORY 1593
- 32 REINFORCEMENT LEARNING 1611

vii appendix

- A SUPPLEMENTAL MATHEMATICAL FACTS 1691
- Alphabetical Index 1721

LIST OF ALGORITHMS

1	A generic line search algorithm	324
2	Backtracking-Armijo line search algorithm	335
3	Steepest decent Backtracking-Armijo line search algorithm	338
4	Modified Newton Backtracking-Armijo line search algorithm	338
5	Quasi Newton with Wolfe line search algorithm	338
6	A generic trust-region algorithm	340
7	A linear conjugate algorithm	349
8	Iteratively reweighted least squares for p norm least square	352
9	Gauss-Newton method for nonlinear least-square algorithm	354
10	Levenberg-Marquardt method for nonlinear least-square algorithm	355
11	Newton method for root finding	356
12	Primal active-set method for strictly convex quadratic programming	378
13	First order gradient projection algorithm	381
14	The Simplex algorithm (non-degenerate system)	419
15	Primal-dual long-step path-following algorithm	424
16	A generic subgradient algorithm	463
17	Gradient projection algorithm with constant step size	470
18	Gradient projection algorithm with adaptive size	471
19	Proximal gradient algorithm	472
20	Iterative Shrinkage-Thresholding Algorithm with constant step size for L ₁ optimization	474
21	Iterative reweighed estimation for multivariate normal distribution	723
22	Alternating sparse canonical correlation analysis algorithm	743
23	EM algorithm for least square with nonconstant variance	869
24	Accept-Reject algorithm for random number generation.	904

25	Importance sampling for Monte Carlo integration.	913
26	MCMC Metropolis-Hasting algorithm	917
27	MCMC Gibbs sampling algorithm	919
28	Recursive linear least square estimation of dynamical systems.	1001
29	Recursive nonlinear least square of dynamical systems	1003
30	Kalman filtering	1007
31	Extended Kalman filter	1008
32	Coordinate descent for Lasso regression.	1279
33	Iteratively reweighted least squares for logistic regression	1293
34	Perceptron learning algorithm	1323
35	Soft margin SVM algorithm	1332
36	Multinomial Naive Bayes classification	1347
37	KNN classification and regression algorithm	1361
38	A generic decision tree generation algorithm	1374
39	ID3 classification decision tree algorithm	1379
40	A regression tree growth algorithm	1387
41	A basic bagging algorithm	1396
42	Generic Adaboost classifier algorithm	1402
43	Adaboost regressor algorithm	1405
44	A generic additive model algorithm	1407
45	Generic gradient boosting algorithm	1410
46	Gradient tree boosting algorithm	1411
47	XGBoost algorithm	1417
48	Iterative reweighted least square PCA with outliers algorithm	1432
49	Random sample consensus PCA with outliers algorithm	1432
50	Orthogonal Matching Pursuit	1434
51	K-SVD for dictionary learning.	1436
52	Online dictionary learning	1437
53	Stochastic gradient descent for matrix factorization based recommender systems.	1447

54	Isomap algorithm	1456
55	Kernel PCA algorithm	1457
56	Laplacian Eigenmap algorithm	1462
57	Diffusion map algorithm	1465
58	K-means algorithm.	1470
59	DBSCAN algorithm	1473
60	General spectral clustering algorithm	1475
61	Gaussian mixture model EM algorithm	1479
62	Full batch gradient descent algorithm	1499
63	Minibatch stochastic gradient descent algorithm	1500
64	Adam stochastic gradient descent algorithm.	1505
65	Solve forward backward stochastic differential equation via deep learning. .	1527
66	Minibatch stochastic gradient descent training of GAN.	1580
67	Minibatch stochastic gradient descent training of conditional GAN.	1587
68	Minibatch stochastic gradient descent training of Wasserstein GAN.	1589
69	The policy iteration algorithm for MDP	1618
70	Value iteration algorithm for a finite state MDP	1620
71	First-visit MC value function estimation	1627
72	MC-based reinforcement learning control	1629
73	TD(o) estimation of a value function.	1630
74	SARSA learning	1631
75	Q-learning algorithm	1632
76	TD(n) estimation of a value function.	1635
77	n -step SARSA learning	1636
78	A generic batch policy-gradient algorithm (REINFORCE)	1647
79	A basic Monte-Carlo policy-gradient algorithm	1648
80	Actor-Critic policy-gradient method	1650
81	Policy-gradient method with a value function baseline	1655
82	Neural Fitted Q Iteration (NFQ)	1660
83	Deep Q-learning with experience replay	1662

84	Asynchronous Deep Q-Learning for each thread	1667
85	Deep Q-learning with universal value function approximator	1668
86	Deep deterministic policy gradient algorithm (DDPG)	1671
87	Twin-delayed deep deterministic policy gradient (TD3)	1673
88	Trust Region Policy Optimization	1675
89	Proximal Policy Optimization	1678
90	Soft Actor-Critic (SAC) policy optimization	1681
91	Isotropic multivariate Gaussian evolution strategies for reinforcement learning	1683
92	Isotropic multivariate Gaussian parallelized evolution strategies for reinforcement learning	1683
93	Q learning with prioritized experience replay	1685
94	Deep Q-learning with hindsight experience replay	1686

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 4.9.1	Demonstration of SVD for matrices of different shapes. The dashed lines highlight the compact form SVD. 207
Figure 4.12.1	Illustration of different quadratic forms. 228
Figure 6.1.1	Demonstration of local minimizer (red, green, and blue), strict local minimizer (red and blue), and global minimizer (blue). 318
Figure 6.1.2	A complex objective function in unconstrained optimization. 319
Figure 6.1.3	Illustration of different cases in unconstrained quadratic optimization. 323
Figure 6.2.1	Drawbacks of steepest gradient descent. 326
Figure 6.2.2	Demonstration on the step choices on the iterative algorithm. (a) Large step size. (b) Small step size. 331
Figure 6.2.3	Armijo sufficient decrease condition. 335
Figure 6.3.1	Demonstration of the dogleg path as an approximation to the exact solution path in the trust-region subproblem. 343
Figure 6.4.1	Demonstration of coordinate descent procedures when A is diagonal and non-diagonal. 346
Figure 7.1.1	Demonstration of KKT condition at a local minimal for $f(x_1, x_2) = x_1^2 + x_2^2$ under constraint $x_1 + x_2 = 1$. 363
Figure 7.2.1	Demonstration of KKT condition at a local minimal for $f(x_1, x_2) = x_1^2 + x_2^2$ under constraint $x_1 + x_2 \geq 1$. 373
Figure 8.2.1	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. 409
Figure 8.2.2	Demonstration on multiple minimizers forming a convex set. 411
Figure 8.3.1	Overview of geometry approach to linear programming. 413
Figure 9.1.1	Example 2D affine hull and 3D affine hull. 430
Figure 9.1.2	Affine subspace and linear subspace. 432
Figure 9.2.1	(left) A convex set. (right) A non-convex set. 436

- Figure 9.2.2 (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. 438
- Figure 9.2.3 An illustration of separating hyperplane theorem for two convex bodies. 440
- Figure 9.2.4 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. 443
- Figure 9.2.5 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. 443
- Figure 9.3.1 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. 445
- Figure 9.3.2 The epigraph (green area) of a convex function. 447
- Figure 9.3.3 Illustration of linear underestimator. 450
- Figure 9.5.1 Demonstration of optimality condition $\nabla f(x^*)^T(x - x^*) \geq 0, \forall x \in X$ when x^* lies on the boundary of X . 460
- Figure 11.1.1 An illustration of measurable functions. 506
- Figure 11.11.1 Visualization of central limit theorem. Samples are drawn from uniform and lognormal distributions. Sample means \bar{X}_n converge to normal distribution in distribution when n is large. 576
- Figure 12.1.1 Comparison of Laplace distribution and normal distribution. 613
- Figure 12.1.2 Density of $LN(0,1)$ and $LN(0,0.5)$. Note the positive skewness. 620
- Figure 12.1.3 Density of $NLN(0,1)$ and $NLN(0,0.5)$. Note the negative skewness. 621
- Figure 12.2.1 Distributions with left-skewness (black) and right-skewness (red). 642
- Figure 12.2.2 Distributions with zero excess Kurtosis (Normal distribution, black), positive excess kurtosis (Laplace distribution, red), and negative excess Kurtosis (Uniform distribution, blue). 643
- Figure 12.2.3 Percentile points at $\alpha = 0.1, 0.2, \dots, 0.9$ for a standard normal distribution. 645

Figure 12.2.4	QQ plot of different sample distributions against standard normal distribution, including standard normal $N(0, 1)$, shifted-scaled normal $N(50, 5)$, Student's t with degree 1, and lognormal $LN(0, 1)$. Red solid lines are the fitted linear lines. 647
Figure 13.1.1	Statistical estimation and inference scheme. 660
Figure 13.1.2	An example of biased estimator with smaller variance than unbiased estimator 665
Figure 13.1.3	Visualization of log-likelihood function for normal distributed samples. 673
Figure 13.6.1	Demonstration for rejection regions for upper-tailed one-sided hypothesis (a), two-sided hypothesis(b), and lower-tailed one-sided hypothesis (c). 700
Figure 13.7.1	QQ plot with different sample distributions, including normal, Laplace ($b = 4$), Uniform $U([-1, 1])$, Lognormal $LN(0, 1)$. Red solid lines are the fitted linear lines. 709
Figure 14.2.1	Principal components for 2D samples. 727
Figure 14.2.2	PCA eigenface analysis. 735
Figure 14.2.3	PCA eigen-digit analysis for MNIST dataset. 736
Figure 14.2.4	Demonstration of interest rate curve dynamics. 738
Figure 14.2.5	Demonstration of first three dominating PCA factor in the swap rate curve daily change. 738
Figure 14.4.1	Gaussian copula with different correlations. 760
Figure 14.4.2	Student T copula with different correlations. 764
Figure 14.4.3	Generated correlated default time via Gaussian copula with different correlations. The hazard rate for both parties is $h(t) = 0.01$. 781
Figure 14.5.1	Correlation structure for the Gaussian copula implied by the factor model. 792
Figure 14.5.2	Factor model using (a) external factors or (b) internal factors. 793
Figure 14.5.3	Scatter plot of AAPL return vs. market excess return, SMB excess return and HML excess return. 797
Figure 14.6.1	A fully connected graphical model representing the joint distribution $P(X_1, \dots, X_N)$. 800
Figure 14.6.2	Graphical model examples. 802
Figure 14.6.3	d-separation examples. 804
Figure 15.1.1	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. 809
Figure 15.3.1	Demo of heteroskedasticity in linear regression. The noises are larger at larger x values. 867

Figure 15.3.2	Demonstrations on linear regression with auto-correlated error. Observations are generated by $y_i = x_i + \epsilon_i, \epsilon_i = \rho\epsilon_{i-1} + z, z \sim N(0, 1)$. 872
Figure 15.3.3	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. 877
Figure 15.3.4	Visual scatter and box plots to identify outliers. 878
Figure 15.3.5	Different function choice for M-estimation linear regression 883
Figure 15.3.6	Common linear regression diagnosis plots 886
Figure 15.4.1	Diagnosis plots for a toy linear regression example 889
Figure 15.4.2	Diagnosis plots for the Boston Housing example 891
Figure 16.4.1	Brownian motion interpolation Demo. 925
Figure 18.1.1	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. 1014
Figure 18.3.1	Sample trajectories of Brownian motion process. 1019
Figure 18.4.1	Variance of X_t in a Brownian bridge 1029
Figure 18.5.1	A typical realized trajectory from the Poisson process with jumps at T_1, T_2 , and T_3 . 1034
Figure 19.4.1	The variance function $Var[X(t)]$ for Brownian motion (red) and OU process(black) with $a = 0.5, \sigma = 1$. 1081
Figure 19.4.2	Representative trajectories from three OU processes with different k . k has the unit of inverse year. Mean level $\mu = 50$ and volatility $\sigma = 20$. 1083
Figure 20.1.1	Example Markov chains. Arrows and numbers are transition directions and probabilities. 1098
Figure 20.1.2	Markov chain diagram for a random walk on the state space \mathbb{Z} . 1099
Figure 20.2.1	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. 1101
Figure 20.2.2	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. 1102

Figure 20.2.3	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. 1109
Figure 20.2.4	Example periodic and aperiodic Markov chains. 1111
Figure 21.1.1	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] . 1146
Figure 21.1.2	Demonstration on using STL to decompose an example CO2 concentration time series. 1151
Figure 21.2.1	Example trajectories of AR(1) models ($X_t = aX_{t-1} + Z_t$) with different choices of a . 1156
Figure 21.2.2	Example trajectories of MA(1) models (upper) and MA(2) models (lower). 1163
Figure 21.2.3	Representative simulated trajectories for AR(1) process with coefficient 1, which forms a unit-root process, and with coefficient -1. 1172
Figure 21.2.4	The ACF and PACF correlogram for a white noise process. 1179
Figure 21.2.5	The ACF and PACF correlogram for an AR(1) process with coefficient 0.8. 1179
Figure 21.2.6	The ACF and PACF correlogram for an MA(1) process. 1180
Figure 21.2.7	The ACF and PACF correlogram for an ARMA(1,1) process. 1180
Figure 21.2.8	Diagnosis plot of residuals for AR(2) model estimation. 1189
Figure 21.2.9	Diagnosis plot of residuals for AR(1) model fitted to time series generated by AR(2) ground truth model. 1190
Figure 21.4.1	Stock index SP500 daily return between 2014 and 2019. 1203
Figure 21.4.2	Simulated representative trajectories from ARCH(1) model with coefficients $a = 0.9$ and $a = 0.5$. 1205
Figure 22.1.1	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. 1225
Figure 22.1.2	Input feature data type examples. 1225
Figure 22.2.1	A simple regression problem illustrating underfitting and overfitting. Solid lines are models, and points are samples. 1229
Figure 22.2.2	The commonly observed phenomenon of overfitting and underfitting in machine learning. 1230
Figure 22.2.3	The performance of different types of models as training proceeds. 1233

Figure 22.3.1	Regression loss functions: MSE Loss, MAE Loss, Huber loss ($\delta = 0.1, 1, 3$), and Log-Cosh Loss. 1237
Figure 22.3.2	Common classification loss functions. 1239
Figure 22.3.3	Scheme for ROC curves diagram. A and B demote ROC curves of different model. 1244
Figure 22.4.1	Scheme for cross-validation error calculation procedure. 1247
Figure 22.4.2	Hyperparameter search via turning regularization parameter λ . At large λ , heavy regularization causes underfitting; at small λ , insufficient regularization causes overfitting. 1249
Figure 22.5.1	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] 1258
Figure 23.1.1	Correlation among features and the label MEDV. 1271
Figure 23.1.2	Pair plot among features and the label. 1272
Figure 23.2.1	Comparison different penalization: Lasso $L1$, elastic net, and Ridge $L2$. 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. 1281
Figure 23.2.2	Penalized regression path in a toy regression problem. 1282
Figure 23.3.1	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. 1285
Figure 23.3.2	Linear model enhancement flowchart. 1285
Figure 24.1.1	Logistic regression for classification on the Iris data set. 1290
Figure 24.1.2	Pair plot analysis results on South Africa heart disease problem. 1298
Figure 24.1.3	$L1$ regularization path for South Africa heart disease problem. 1299
Figure 24.1.4	Logistic regression result with $L1$ penalty. 1301
Figure 24.1.5	Balanced accuracy score vs. inverse regularization strength in credit card fraud detection problem. 1302
Figure 24.1.6	Logistic regression coefficients corresponding to each class. 1303

Figure 24.2.1	Geometry of decision boundary in linear Gaussian discriminant model. 1307
Figure 24.2.2	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. 1310
Figure 24.2.3	The decision boundary geometry of LDA and GDA can be understood via their decision functions 1311
Figure 24.3.1	The linear discriminant w that maximizes the separability for 2D sample points belonging to two classes. 1313
Figure 24.3.2	Fisher linear discriminate will fail to achieve class separability for complex data structures. 1317
Figure 24.4.1	Scheme of a hyperplane. 1322
Figure 24.4.2	Binary classification using the Perceptron learning algorithm. The hyperplane learned separates the two clusters. 1324
Figure 24.5.1	Left: existence of multiple separating hyperplanes in 2D binary classification problem. Right: hyperplanes with maximum margin. 1326
Figure 24.5.2	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. 1329
Figure 24.5.3	SVM classification using Gaussian kernel. The original problem cannot be separated by linear kernel. 1334
Figure 24.5.4	Comparison of classification loss functions. 1336
Figure 25.2.1	Histogram of the features group by class label (fraud vs. genuine). 1351
Figure 25.2.2	Feature density similarity and predictive performance of model 1352
Figure 26.2.1	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. 1365
Figure 27.2.1	Demonstration of decision trees. 1373
Figure 27.2.2	Different types of impurity measure. Entropy function, Gini function and classification error function. 1376
Figure 27.2.3	Different splitting strategy when variables taking more than two discrete values. 1377
Figure 27.2.4	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. 1382

Figure 27.2.5	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. 1383
Figure 27.2.6	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. 1383
Figure 27.3.1	Demonstration of a tree and input space partitioning. 1386
Figure 27.3.2	2D input space partitions cannot be represented by a regression tree. 1386
Figure 27.3.3	Regression tree demonstration in a toy example. 1388
Figure 27.3.4	Variable importance from regression tree in the Boston Housing Pricing problem. 1389
Figure 28.1.1	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. 1394
Figure 28.3.1	Illustration of adaptive boosting where sample weights are adjusted iteratively based on the classification error. 1401
Figure 29.1.1	Demonstration of SVD for matrices of two different shapes. The dashed lines highlight the compact form SVD. 1423
Figure 29.1.2	Principal components for 2D samples. 1427
Figure 29.2.1	Singular value spectrum of LSA on 20-news-group text data. 1442
Figure 29.2.2	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] 1447
Figure 29.3.1	Triangle and Tetrahedron reconstructed from distance matrix by MDS method. 1454
Figure 29.3.2	Application of MDS, based on Euclidean distance, to Swiss Roll data set cannot fully reveal of the global structure. 1455
Figure 29.3.3	Isomap analysis of MNIST dataset. 1466
Figure 29.3.4	Isomap analysis of digit '5' in MNIST dataset 1467
Figure 29.4.1	Demonstration of k-means clustering on a data set with two blobs. 1469
Figure 29.4.2	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. 1471
Figure 29.4.3	Clustering comparison between Kmeans and Kmeans++. 1472
Figure 29.4.4	DBSCAN demo with different ϵ . 1474
Figure 29.4.5	Spectral clustering demo. 1476
Figure 29.4.6	Clustering comparison between K-means and GMM. 1480
Figure 29.4.7	K-means application to image segmentation. 1481

Figure 30.1.1	Scheme of an artificial neuron. 1487
Figure 30.1.2	Common activation functions in artificial neural networks. 1488
Figure 30.1.3	Scheme for an artificial neural network. 1489
Figure 30.1.4	A four-layer feed-forward neural network. 1489
Figure 30.1.5	A three-layer feed-forward neural network with three output units. 1490
Figure 30.1.6	A four-layer feed-forward neural network. 1492
Figure 30.2.1	An example saddle point at $(0, 0, 0)$, which locally minimizes the x direction but maximizes the y direction.. 1499
Figure 30.2.2	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. 1503
Figure 30.3.1	Dropout technique for a simple feedforward neural networks. The original network (left) and the neural network after some neurons being dropped out (right). 1511
Figure 30.4.1	Neural network architecture for polynomial regression. 1513
Figure 30.4.2	Polynomial regression with degree $d = 1, 3, 6, 10$. 1514
Figure 30.4.3	Visualization of first layer weight for a one-layer linear multi-class classification neural network 1515
Figure 30.4.4	Example images from the Fashion MNIST dataset. 1516
Figure 30.4.5	The confusion matrix from fashion MNIST classification results. 1517
Figure 30.4.6	Classification results for a set of randomly selected samples. 1518
Figure 30.4.7	(a) Embedding layer maps large, sparse one-hot vectors to short, dense vectors. (b) Example of low dimensional embeddings that capture semantic meanings. 1519
Figure 30.4.8	(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. 1520
Figure 30.4.9	A feed-forward neural network architecture for sentiment analysis. 1524
Figure 30.5.1	Comparison of receptive fields in fully-connected layer and local-connected layer in CNN. Credit 1528
Figure 30.5.2	Demo for one kernel 'convoluting' with an input image. 1529
Figure 30.5.3	Pooling layer demo. 1530
Figure 30.5.4	A typical CNN architecture for image classification tasks. 1531
Figure 30.5.5	Scheme for LeNet [24] 1531
Figure 30.5.6	Architecture of AlexNet. 1532

Figure 30.5.7	A typical VGG architecture: VGG-19 scheme. 1532
Figure 30.5.8	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] 1533
Figure 30.5.9	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] 1534
Figure 30.5.10	Scheme for a residual block. 1534
Figure 30.5.11	Scheme of a 18 layer ResNet 1535
Figure 30.6.1	Example images from CIFAR10 image dataset. 1536
Figure 30.6.2	Visualization of convolution layer. 1538
Figure 30.6.3	Grad-CAM method applied to understand image classification. Middle and right are class-discriminative localization map superimposed onto the original image. 1539
Figure 30.6.4	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. 1541
Figure 30.6.5	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. 1541
Figure 30.6.6	Denoising autoencoders applied to remove the noise in the MNIST dataset. 1542
Figure 30.6.7	Demonstration of neural style transfer with Van Gogh painting style. 1544
Figure 30.6.8	Demonstration of neural style transfer with Picasso painting style. 1545
Figure 30.6.9	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. 1547
Figure 30.6.10	Demonstration of a CNN filter applying to a sentence to produce a one-dimensional feature vector. 1548
Figure 30.6.11	CNN for sentence classification proposed in [43]. 1549
Figure 30.7.1	Scheme of recurrent units in a neural network (left). Recurrent neural network can be unrolled (right) 1551
Figure 30.7.2	Scheme for backpropogation through time in a simple RNN. Red arrows are backpropogation directions. 1552

Figure 30.7.3	Scheme of an LSTM cell. Modified from [45, p. 149].	1554
Figure 30.7.4	Scheme of a GRU cell. Modified from [45, p. 152].	1557
Figure 30.7.5	Different types of units in RNNs: (a) Vanilla RNN cell. (b) LSTM cell. (c) GRU cell. Credit	1558
Figure 30.7.6	Typical RNN connection to the output layer.	1559
Figure 30.7.7	Scheme for stacked RNN.	1560
Figure 30.7.8	Scheme for bidirectional RNN.	1560
Figure 30.8.1	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.	1562
Figure 30.8.2	RNN one-step forward and multiple forward prediction performance for Sine time series.	1563
Figure 30.8.3	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.	1564
Figure 30.8.4	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.	1565
Figure 30.8.5	A RNN architecture for MNIST image recognition.	1567
Figure 30.8.6	RNN architecture for sentiment analysis.	1568
Figure 30.8.7	A RNN for character-level word classification.	1569
Figure 30.8.8	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.	1570
Figure 30.8.9	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.	1571
Figure 30.9.1	Seq2seq modeling for language transformation. The input and output sequences might have different lengths, and not in synchrony.	1573

- Figure 30.9.2 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. [1574](#)
- Figure 30.9.3 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. [1576](#)
- Figure 30.9.4 A bidirectional RNN encoder system with attention mechanism. [\[51\]](#) [1577](#)
- Figure 30.10.1 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. [1579](#)
- Figure 30.10.2 Generated image samples from a GAN consisting of feed-forward networks. [1581](#)
- Figure 30.10.3 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 [\[56\]](#) [1584](#)
- Figure 30.10.4 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.[\[56\]](#) [1584](#)
- Figure 30.10.5 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. [1586](#)
- Figure 32.1.1 Policy iteration involves iteratively carrying out policy evaluation and policy improvement procedures. [1614](#)

- Figure 32.2.1 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 [1622](#)
- Figure 32.2.2 Scheme of the Atari game *breakout*. [1623](#)
- Figure 32.2.3 Policy evaluation and policy improvement framework in the context reinforcement learning. [1624](#)
- Figure 32.3.1 Neural network parameterization for a Gaussian policy (a) and (b) a generic stochastic policy. [1652](#)
- Figure 32.4.1 A typical feed-forward neural network used in NFQ to approximate the Q function. [1659](#)
- Figure 32.4.2 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 [1661](#)
- Figure 32.4.3 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. [\[10\]](#). [1664](#)
- Figure 32.4.4 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.[\[11\]](#) [1665](#)
- Figure 32.4.5 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. [1666](#)
- Figure 32.4.6 A comparison between a typical deep Q learning network (left) and a typical universal value function approximator network (right). [1668](#)
- Figure 32.4.7 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. [1670](#)
- Figure 32.5.1 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. [1687](#)

Figure 32.5.2 One representative trajectory on the curved surface via a control policy learned via curriculum learning on a low-dimensional manifold. [1688](#)

LIST OF TABLES

Table 13.7.1	Test on mean with known variance σ^2	708
Table 13.7.2	Test on mean with unknown variance σ^2	710
Table 13.7.3	Test on variance	710
Table 13.7.4	Test on variance comparison between two samples	711
Table 14.2.1	Eigenvectors and eigenvalues for swap rate daily change	739
Table 14.5.1	statistics on Fama-French 3 factors from July 1963 to Dec. 1991.	796
Table 14.5.2	AAPL stock return modeled by the Fama-French 3 factor model.	798
Table 20.2.1	Summary of Markov chain state property[3, p. 140]	1113
Table 21.2.1	Summary of PACF and ACF for AR, MA, and ARMA processes.	1179
Table 23.1.1	Linear regression results.	1273
Table 24.1.1	Logistic regression results on South Africa heart disease problem.	1299
Table 29.2.1	Most frequent words in the top 8 topics	1443
Table 29.2.2	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.	1445
Table 32.2.1	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 .	1634
Table A.9.1	Closed Newton-Cotes Formula	1711

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 a 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	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		LINEAR ALGEBRA AND MATRIX ANALYSIS	138
4.1		Theory for system of linear equations	143
	4.1.1	Overview	143
	4.1.2	Homogeneous systems	143
	4.1.3	Non-homogeneous systems	144
	4.1.4	Overdetermined vs. underdetermined systems	145
	4.1.5	Solution methods	146
	4.1.6	Error bounds in numerical solutions	151
	4.1.6.1	Condition number	151
	4.1.6.2	Error bounds	151
4.2		Vector space theory	153
	4.2.1	Vector space	153
	4.2.2	Subspace	154
	4.2.3	Sum and direct sum	155
	4.2.4	Basis and dimensions	157
	4.2.5	Complex vector space vs. real vector space	159
4.3		Linear maps & linear operators	161
	4.3.1	Basic concepts of linear maps	161
	4.3.2	Fundamental theorem of linear maps	162
	4.3.3	Isomorphism	164
	4.3.4	Coordinate map properties	166
	4.3.5	Change of basis and similarity	167
	4.3.5.1	Change of basis for coordinate vector	167
	4.3.5.2	Change of basis for linear maps	167
	4.3.6	Linear maps and matrices	167
	4.3.6.1	Similarity	169

4.4	Fundamental theorems of ranks and linear algebra	170
4.4.1	Basics of ranks	170
4.4.2	Fundamental theorem of ranks	171
4.4.3	Fundamental theorem of linear algebra	172
4.5	Complementary subspaces and projections	174
4.5.1	General complementary subspaces	174
4.5.2	Orthogonal complementary spaces and projections	176
4.5.3	Decomposition of orthogonal projectors	180
4.6	Orthonormal basis and projections	183
4.6.1	Gram-Schmidt Procedure	183
4.6.2	Orthogonal-triangular decomposition	183
4.6.3	Orthonormal basis for linear operators	184
4.6.4	Riesz representation theorem	185
4.7	Eigenvectors and eigenvalues of Matrices: general theory	186
4.7.1	Existence and properties of eigenvalues	186
4.7.2	Properties of eigenvectors	188
4.7.3	Right and left eigenvectors	189
4.7.4	Diagonalizable matrices	190
4.8	Eigenvalue and eigenvectors of matrices: case studies	193
4.8.1	Real diagonalizable matrix	193
4.8.2	Real symmetric matrix	194
4.8.2.1	Spectral properties	194
4.8.2.2	Rayleigh quotients	196
4.8.2.3	Pointcare inequality	199
4.8.3	Hermitian matrix	200
4.8.4	Matrix congruence	202
4.8.5	Complex symmetric matrix	203
4.8.6	Unitary, orthonormal & rotation matrix	203
4.9	Singular Value Decomposition theory	205
4.9.1	SVD fundamentals	205
4.9.2	SVD and matrix norm	207
4.9.3	SVD vs. eigendecomposition	208
4.9.4	SVD low rank approximation	209
4.9.4.1	Frobenius norm low rank approximation	209
4.9.4.2	Two-norm low rank approximation	211
4.10	Generalized eigenvectors and Jordan normal forms	213
4.10.1	Generalized eigenvectors	213
4.10.2	Upper triangle matrix and nilpotent matrix	216
4.10.3	Jordan normal forms	218
4.11	Matrix factorization	222
4.11.1	Orthogonal-triangular decomposition	222

4.11.2	LU decomposition	223
4.11.3	Cholesky decomposition	223
4.12	Positive definite matrices and quadratic forms	225
4.12.1	Quadratic forms	225
4.12.2	Real symmetric non-negative definite matrix	226
4.12.2.1	Characterization	226
4.12.2.2	Decomposition and transformation	229
4.12.2.3	Matrix square root	230
4.12.2.4	Maximization of quadratic forms	231
4.12.2.5	Gramian matrix	233
4.12.3	Completing the square	234
4.13	Matrix norm and spectral estimation	235
4.13.1	Basics	235
4.13.2	Singularity from matrix norm and spectral radius	236
4.13.3	Gerschgorin theorem	237
4.13.4	Irreducible matrix and stronger results	238
4.14	Pseudoinverse of matrix	239
4.14.1	Pseudoinverse for full rank system	239
4.14.2	Pseudoinverse for general matrix	241
4.14.3	Application in linear systems	243
4.15	Multilinear forms	246
4.15.1	Bilinear forms	246
4.15.2	Multilinear forms	247
4.16	Determinant	250
4.16.1	Basic properties	250
4.16.2	Vandermonde matrix and determinant	256
4.17	Numerical iteration analysis	258
4.17.1	Numerical linear equation solution	258
4.17.1.1	Goals and general principles	258
4.17.1.2	Jacobi algorithm	258
4.17.1.3	Gauss Seidel algorithm	259
4.17.2	Power method for eigen-decomposition	259
4.18	Appendix: supplemental results for polynomials	262
4.18.1	Basics	262
4.18.2	Factorization of polynomial over \mathbb{C}	263
4.18.3	Factorization of polynomial over \mathbb{R}	264
4.19	Notes on bibliography	266
5	BASIC FUNCTIONAL ANALYSIS	268
5.1	Normed vector space	270
5.1.1	Basic properties	270
5.1.2	Equivalence of norms	272

5.2	Contraction mapping and fixed point theorems	274
5.2.1	Complete normed space (Banach space)	274
5.2.2	Contraction mapping	275
5.2.3	Banach fixed point theorem	276
5.2.4	Applications in root finding	277
5.2.5	Application to numerical linear equations	277
5.2.6	Applications to integral and differential equations	278
5.3	Inner product space and Hilbert space	281
5.3.1	Inner product space (pre-Hilbert space) and Hilbert space	281
5.3.1.1	Foundations	281
5.3.2	Hilbert spaces	283
5.3.2.1	Basics	283
5.3.3	Orthogonal decomposition of Hilbert spaces	284
5.3.3.1	Orthogonality	284
5.3.4	Projection and orthogonal decomposition	285
5.4	Approximations in Hilbert space	288
5.4.1	Approximation via projection	288
5.4.2	Application examples	289
5.4.2.1	Orthogonal projection and normal equations in \mathbb{R}^n	289
5.4.2.2	Approximation by continuous polynomials	291
5.4.2.3	Legendre polynomial via Gram-Schmidt process	292
5.5	Orthonormal systems	293
5.5.1	Basic definitions	293
5.5.2	Gram-Schmidt process	293
5.5.3	Properties of orthonormal systems	293
5.5.4	Orthonormal expansion in Hilbert space	295
5.5.5	Complete orthonormal system	296
5.5.5.1	Weierstrass approximation theorem for polynomials	298
5.5.5.2	Examples of complete orthonormal function set	298
5.6	Theory for trigonometric Fourier Series	300
5.6.1	Basic definitions	300
5.6.2	Completeness of Fourier series	302
5.6.3	Complex representation	303
5.7	Fourier transform	305
5.7.1	Definitions and basic concepts	305
5.7.2	Convolution theorem	307
5.7.3	Fourier transform and Fourier series	308
5.7.4	Discrete Fourier transform	309
5.7.4.1	Properties	309
5.8	Notes on bibliography	312

ii mathematical optimization methods

6	UNCONSTRAINED NONLINEAR OPTIMIZATION	315
6.1	Optimality conditions	317
6.1.1	Optimality concepts	317
6.1.2	Necessary and sufficient conditions	319
6.1.3	Special case: unconstrained quadratic programming	322
6.2	Line search method	324
6.2.1	A generic algorithm	324
6.2.2	Theory and computation of descent directions	325
6.2.2.1	Gradient descent direction and properties	325
6.2.2.2	Curvature-modified descent direction	326
6.2.2.3	Quasi-Newton method	328
6.2.2.4	Subspace optimization in quadratic forms	330
6.2.3	Theory and computation of step length	331
6.2.3.1	Overview	331
6.2.3.2	Lipschitz bounded convex functions	331
6.2.3.3	Backtracking-Armijo step size search	334
6.2.3.4	Wolfe condition	336
6.2.4	Complete algorithms	337
6.3	Trust region method	339
6.3.1	Motivation and the framework	339
6.3.2	Cauchy point method	340
6.3.3	Exact solution method	342
6.3.4	Approximate method	342
6.4	Conjugate gradient method	345
6.4.1	Motivating problems	345
6.4.2	Theory conjugate direction	346
6.4.3	Linear conjugate gradient algorithm	347
6.5	Least square problems	350
6.5.1	Linear least square theory and algorithm	350
6.5.1.1	Linear least square problems	350
6.5.1.2	SVD methods	351
6.5.1.3	Extension to L^p norm optimization	351
6.5.2	nonlinear least square problem	352
6.5.3	Line search Gauss-Newton method	353
6.5.4	Trust region method	355
6.5.5	Application: roots for nonlinear equation	355
6.6	Notes on bibliography	357
7	CONSTRAINED NONLINEAR OPTIMIZATION	359
7.1	Quadratic optimization I: equality constraints	361
7.1.1	Problem formulation	361

7.1.2	Optimality condition	361
7.1.2.1	General case	361
7.1.2.2	Positive semi-definitive quadratic programming	364
7.1.3	Solving KKT systems	365
7.1.3.1	Factorization approach	365
7.1.3.2	Range space approach	366
7.1.4	Linear least square with linear constraints	366
7.1.4.1	Least norm problem	366
7.1.5	Application: Markovitz Portfolio Optimization Model	368
7.2	Quadratic optimization II: inequality constraints	371
7.2.1	Problem formulation	371
7.2.2	Optimality conditions	371
7.2.2.1	Pure inequality case	371
7.2.2.2	General constrained optimization	373
7.2.2.3	Positive semi-definitive quadratic programming	374
7.2.3	Primal active-set method	375
7.2.4	Gradient projection method	380
7.2.5	Dual convex quadratic programming	381
7.3	General equality constrained optimization	383
7.3.1	Feasible path and optimality	383
7.3.2	Constraint qualification and Lagrange theory	384
7.3.3	Second order condition	387
7.4	General inequality constrained optimization	391
7.4.1	Feasible path and optimality	391
7.4.2	Constraint qualifications and KKT conditions	392
7.4.3	Second order conditions	396
7.5	Envelope theorem and sensitive analysis	400
7.6	Notes on bibliography	403
8	LINEAR OPTIMIZATION	405
8.1	Equality constrained linear programming	406
8.2	Inequality constrained linear programming	408
8.2.1	Linear optimization with inequality constraints	408
8.2.2	Geometry of linear programming	408
8.2.3	Optimality property and condition	410
8.2.4	Standard form of linear programming	411
8.2.5	Application examples	412
8.3	Linear programming geometry and simplex algorithm	413
8.3.1	Geometrical approach to linear programming	413
8.3.1.1	Overview	413
8.3.1.2	Vertex and optimality	413
8.3.1.3	Descent direction at a vertex	417

8.3.1.4	Stepping along a descent direction	418
8.3.2	The simplex algorithm	419
8.4	Interior point method	420
8.4.1	Optimality condition	420
8.4.2	Newton step and perturbed system	422
8.4.3	Algorithms	424
8.5	Notes on bibliography	426
9	CONVEX ANALYSIS AND CONVEX OPTIMIZATION	428
9.1	Affine sets	430
9.1.1	Basic concepts	430
9.1.2	Affine independence and dimensions	432
9.2	Convex sets and properties	436
9.2.1	Concepts of convex sets	436
9.2.2	Projection theorems	438
9.2.3	Separation theorems	439
9.2.3.1	Separating hyperplane theorem	439
9.2.3.2	Farka's lemma	441
9.3	Convex functions	445
9.3.1	Basic concepts	445
9.3.2	Connection to convex set	447
9.3.3	Strongly convex functions	447
9.3.4	Operations preserve convexity	448
9.3.5	Convexity and derivatives	449
9.3.6	Subgradient	451
9.4	Duality theory	453
9.5	Convex optimization and optimality conditions	457
9.5.1	Local optimality vs. global optimality	457
9.5.2	Unconstrained optimization optimality conditions	458
9.5.3	Constrained optimization optimality conditions	458
9.6	Subgradient methods	463
9.6.1	A generic algorithm for unconstrained problem	463
9.6.2	Convergence under Lipschitz smoothness	465
9.6.3	Projected gradient methods	468
9.6.3.1	Foundations	468
9.6.3.2	Algorithms	470
9.6.4	Proximal gradient methods	471
9.6.4.1	Foundations	471
9.6.4.2	Algorithms	472
9.6.4.3	Case study: sparsity regularization problem	473
9.7	Notes on bibliography	475
10	BASIC GAME THEORY	477

10.1	Static normal form game	478
10.1.1	Normal form game concepts	478
10.1.2	Pure strategy and equilibrium	478
10.1.2.1	Solution concepts	478
10.1.3	Mixed strategy and equilibrium	482
10.1.4	Pareto optimality	484
10.2	Zero-sum matrix game	485
10.2.1	Fundamentals	485
10.2.2	Optimal strategy and Nash equilibrium	486
10.2.3	Saddle points as solutions	488
10.2.4	Maxmin strategies and Nash equilibrium	489
10.2.5	Linear programming approach to optimal strategy	492
10.3	Notes on bibliography	496

iii classical statistical methods

11	PROBABILITY THEORY	499
11.1	σ algebra	501
11.1.1	σ algebra concepts	501
11.1.2	Generation of sigma algebra	501
11.1.3	Partition of sample space	502
11.1.4	Filtration & information	502
11.1.5	Borel σ algebra	503
11.1.6	Measurable set and measurable space	504
11.2	Probability space	507
11.2.1	Event, sample point and sample space	507
11.2.2	Probability space	507
11.2.3	Properties of probability measure	509
11.2.4	Conditional probability	510
11.2.4.1	Basics	510
11.2.4.2	Bayes' theorem	511
11.2.4.3	Independence of events and sigma algebra	512
11.3	Measurable map and random variable	515
11.3.1	Random variable	515
11.3.2	σ algebra of random variables	516
11.3.3	Independence of random variables	517
11.4	Distributions of random variables	519
11.4.1	Basic concepts	519
11.4.1.1	Probability mass function	519
11.4.1.2	Distributions on \mathbb{R}^n	519
11.4.1.3	Probability density function	520
11.4.1.4	Conditional distributions	521

11.4.1.5	Bayes law	522
11.4.2	Independence	523
11.4.3	Conditional independence	525
11.4.4	Transformations	525
11.4.4.1	Transformation for univariate distribution	525
11.4.4.2	Location-scale transformation	526
11.4.4.3	Transformation for multivariate distribution	527
11.5	Expectation, variance, and covariance	531
11.5.1	Expectation	531
11.5.2	Expectation in the Lebesgue framework	532
11.5.3	Properties of expectation	534
11.5.4	Variance and covariance	534
11.5.5	Conditional variance	535
11.5.6	Delta method	536
11.6	Moment generating functions and characteristic functions	538
11.6.1	Moment generating function	538
11.6.2	Characteristic function	541
11.6.3	Joint moment generating functions for random vectors	542
11.6.4	Cumulants	543
11.7	Conditional expectation	545
11.7.1	General intuitions	545
11.7.2	Formal definitions	545
11.7.3	Different versions of conditional expectation	547
11.7.3.1	Conditioning on an event	547
11.7.3.2	Conditioning on a discrete random variable as a new random variable	547
11.7.3.3	Condition on random variable vs. event vs σ algebra	548
11.7.4	Properties	548
11.7.4.1	Linearity	548
11.7.4.2	Taking out what is known	549
11.7.4.3	Law of total expectation	549
11.7.4.4	Law of iterated expectations	551
11.7.4.5	Conditioning on independent random variable/ σ algebra	551
11.7.4.6	Least Square minimizing property	552
11.8	The Hilbert space of random variables	553
11.8.1	Definitions	553
11.8.2	Subspaces, projections, and approximations	553
11.8.3	Connection to conditional expectation	558
11.9	Probability inequalities	561
11.9.1	Chebychev inequalities	561

11.9.2	Jensen's inequality	562
11.9.3	Holder's, Minkowski, and Cauchy-Schwarz inequalities	563
11.9.4	Popoviciu's inequality for variance	565
11.10	Convergence of random variables	567
11.10.1	Different levels of equivalence among random variables	567
11.10.2	Convergence almost surely	567
11.10.3	Convergence in probability	568
11.10.4	Mean square convergence	570
11.10.5	Convergence in distribution	570
11.11	Law of Large Number and Central Limit theorem	573
11.11.1	Law of Large Numbers	573
11.11.2	Central limit theorem	575
11.12	Finite sampling models	578
11.12.1	Counting principles	578
11.12.2	Matching problem	581
11.12.3	Birthday problem	583
11.12.4	Coupon collection problem	584
11.12.5	Balls into bins model	585
11.13	Order statistics	588
11.14	Information theory	592
11.14.1	Concept of entropy	592
11.14.2	Entropy maximizing distributions	593
11.14.3	KL divergence	597
11.14.4	Conditional entropy and mutual information	598
11.14.5	Cross-entropy	599
11.15	Notes on bibliography	601
12	STATISTICAL DISTRIBUTIONS	603
12.1	Common distributions and properties	604
12.1.1	Overview	604
12.1.2	Bernoulli distribution	604
12.1.3	Poisson distribution	604
12.1.4	Geometric distribution	606
12.1.5	Binomial distribution	607
12.1.6	Normal distribution	609
12.1.7	Half-normal distribution	612
12.1.8	Laplace distribution	612
12.1.9	Multivariate Gaussian/normal distribution	613
12.1.9.1	Basic definitions	613
12.1.9.2	Affine transformation and its consequences	615
12.1.9.3	Marginal and conditional distribution	616
12.1.9.4	Box Muller transformation	618

12.1.10	Lognormal distribution	619
12.1.10.1	Univariate lognormal distribution	619
12.1.10.2	Extension to univariate lognormal distribution	620
12.1.10.3	Multivariate lognormal distribution	622
12.1.11	Exponential distribution	623
12.1.12	Gamma distribution	624
12.1.13	Hypergeometric distribution	626
12.1.14	Beta distribution	626
12.1.15	Multinomial distribution	629
12.1.16	Dirichlet distribution	630
12.1.17	χ^2 -distribution	631
12.1.17.1	Basic properties	631
12.1.17.2	Noncentral chi-squared distribution	633
12.1.18	Wishart distribution	633
12.1.19	t -distribution	634
12.1.19.1	Standard t distribution	634
12.1.19.2	classical t distribution	635
12.1.19.3	Multivariate t distribution	635
12.1.20	F -distribution	636
12.1.21	Empirical distributions	637
12.1.22	Heavy-tailed distributions	638
12.1.22.1	Basic characterization	638
12.1.22.2	Pareto and power distribution	638
12.1.22.3	Student t distribution family	639
12.1.22.4	Gaussian mixture distributions	640
12.2	Characterizing distributions	642
12.2.1	Skewness and kurtosis	642
12.2.2	Percentiles and quantiles	644
12.2.2.1	Basics	644
12.2.2.2	Cornish-Fisher expansion	646
12.3	Moment matching approximation methods	648
12.4	Gaussian quadratic forms	650
12.4.0.1	Quadratic forms and chi-square distribution	650
12.4.0.2	Applications	653
12.5	Notes on bibliography	656
13	STATISTICAL ESTIMATION THEORY	657
13.1	Parameter estimators	660
13.1.1	Overview	660
13.1.2	Statistic and Estimator	661
13.1.2.1	Statistic	661
13.1.2.2	Estimator properties	662

13.1.2.3	Variance-bias decomposition	664
13.1.2.4	Consistence	666
13.1.2.5	Efficiency	668
13.1.2.6	Robust statistics	669
13.1.3	Method of moments	670
13.1.4	Maximum likelihood estimation	671
13.1.4.1	Basic concepts	671
13.1.4.2	MLE examples	673
13.1.4.3	Bias and consistence of MLE	676
13.2	Information and efficiency	679
13.2.1	Fisher information	679
13.2.2	Cramer-Rao lower bound	683
13.2.2.1	Preliminary: information inequality	683
13.2.2.2	Cramer-Rao lower bound: univariate case	684
13.2.2.3	Cramer-Rao lower bound: multivariate case	685
13.2.3	Efficient estimators	687
13.2.4	Asymptotic normality and efficiency of MLE	688
13.3	Sufficiency and data reduction	689
13.3.1	Sufficient estimators	689
13.3.2	Factorization theorem	690
13.4	Bayesian estimation theory	693
13.4.1	Overview	693
13.4.2	Basics	693
13.5	Bootstrap method	696
13.6	Hypothesis testing theory	698
13.6.1	Basics	698
13.6.2	Characterizing errors and power	701
13.6.3	Power of a statistical test	702
13.6.4	Common statistical tests	704
13.6.4.1	Chi-square goodness-of-fit test	704
13.6.4.2	Chi-square test for statistical independence	706
13.6.4.3	Kolmogorov-Smirnov goodness-of-fit test	707
13.7	Hypothesis testing on normal distributions	708
13.7.1	Normality test	708
13.7.2	Sample mean with known variance	708
13.7.3	Sample mean with unknown variance	710
13.7.4	Variance test	710
13.7.5	Variance comparison test	711
13.7.6	Person correlation t test	711
13.7.7	Two sample tests	712
13.7.7.1	Two-sample z test	712

13.7.7.2	Two-sample t test	712
13.7.7.3	Paired Data	713
13.7.8	Interval estimation for normal distribution	713
13.8	Notes on bibliography	715
14	MULTIVARIATE STATISTICAL METHODS	716
14.1	Multivariate data and distribution	718
14.1.1	Sample statistics	718
14.1.2	Multivariate Gaussian distribution	719
14.1.3	Estimation methods	721
14.1.3.1	Maximum likelihood estimation	721
14.1.3.2	Weighted estimation	723
14.2	Principal component analysis (PCA)	725
14.2.1	Statistical fundamentals of PCA	725
14.2.1.1	PCA for random vectors	725
14.2.1.2	Sample principal components	726
14.2.2	Geometric fundamentals of PCA	729
14.2.2.1	Optimization approach	729
14.2.2.2	Properties	731
14.2.3	Probabilistic PCA	732
14.2.4	Applications	734
14.2.4.1	Eigenfaces and eigendigits	734
14.2.4.2	Interest rate curve dynamics modeling	736
14.3	Canonical correlation analysis	740
14.3.1	Basics	740
14.3.2	Sparse CCA	742
14.4	Copulas and dependence modeling	744
14.4.1	Definitions and properties	744
14.4.2	Copulas and distributions	747
14.4.2.1	Fundamentals	747
14.4.2.2	Survival copula	754
14.4.2.3	Partial differential and conditional distribution	755
14.4.3	Common copula functions	759
14.4.3.1	Gaussian copula	759
14.4.3.2	t copula	764
14.4.3.3	Common copula functions: other copula	764
14.4.4	Dependence and copula	765
14.4.4.1	Linear correlations	765
14.4.4.2	Rank correlations	767
14.4.4.3	Tail dependence	773
14.4.5	Estimating copula function	775
14.4.5.1	Empirical copula method	775

14.4.5.2	Maximum likelihood method	776
14.4.6	Applications of copula	777
14.4.6.1	Generating correlated uniform random number	777
14.4.6.2	Generating general correlated random number	779
14.4.6.3	Multivariate distribution approximation with Gaussian copula	783
14.5	Covariance structure and factor analysis	784
14.5.1	The orthogonal factor model	784
14.5.1.1	Motivation and factor models	784
14.5.1.2	Covariance structure implied by factor model	784
14.5.2	Parameter estimation	786
14.5.2.1	Data collection and preparation	786
14.5.2.2	PCA method	787
14.5.2.3	Maximum likelihood method	788
14.5.3	Factor score estimation	788
14.5.4	Application I: Joint default modeling	789
14.5.4.1	Single factor model	789
14.5.4.2	Multiple factor model	792
14.5.5	Application II: factor models for stock return	793
14.5.5.1	Overview	793
14.5.5.2	The Fama-French 3 factor model	795
14.6	Graphical models	799
14.6.1	Fundamentals	799
14.7	Notes on Bibliography	805
15	LINEAR REGRESSION ANALYSIS	806
15.1	Linear regression analysis: basics	808
15.1.1	Linear regression models	808
15.1.2	Ordinary least square (OLS): fundamentals	811
15.1.2.1	Review on orthogonal projections	811
15.1.2.2	OLS results	812
15.1.2.3	OLS results with demeaned data	817
15.1.2.4	Gauss-Markov theorem	820
15.1.2.5	Variance decomposition	821
15.1.2.6	Residual and variance estimation	824
15.1.3	Ordinary least square (OLS): Additional topics	825
15.1.3.1	Orthogonal input and successive regression	825
15.1.3.2	Frisch-Waugh-Lovell(FWL) theorem and partial regression	827
15.1.3.3	Forecasting analysis with normality assumption	828
15.1.4	Hypothesis testing and analysis of variance	830
15.1.4.1	Distribution of coefficients	831

15.1.4.2	t test and normality test of single coefficients	833
15.1.4.3	F lack-of-fit test	835
15.1.4.4	χ^2 test for variance	837
15.1.5	Maximum likelihood method with normality assumption	838
15.1.6	Asymptotic properties of least square solutions	840
15.1.6.1	Asymptotic properties of standard OLS	840
15.1.6.2	Asymptotic efficiency of standard OLS	842
15.1.7	Partial and multiple correlation	842
15.1.7.1	Multiple correlation coefficient, R^2	842
15.1.7.2	Partial correlation coefficient	845
15.1.8	Generalized linear regression (GLR)	846
15.1.8.1	Linear regression with structural error	846
15.1.8.2	Generalized least square solution	847
15.1.8.3	Gauss-Markov theorem for GLR	849
15.1.8.4	Feasible GLS	850
15.1.9	Linear structure in joint distributions	850
15.2	Model specification and selection	853
15.2.1	Model order mis-specification	853
15.2.1.1	Omission of relevant regressors	853
15.2.1.2	Inclusion of irrelevant regressors	854
15.2.2	Model selection methods	856
15.2.2.1	Adjusted R square method	856
15.2.2.2	F test method	856
15.2.2.3	Information criterion methods	858
15.2.2.4	Bayesian information criterion (BIC)	859
15.2.3	Test for structure change	860
15.3	Linear regression analysis: diagnostics & solutions	862
15.3.1	Multi-collinearity	862
15.3.1.1	Detection and characterization	862
15.3.1.2	Regressor linear regression and variance inflation factor	862
15.3.1.3	Principal component linear regression (PCLR)	865
15.3.2	Rank deficiency and rigid regression	865
15.3.3	Heteroskedasticity	867
15.3.3.1	Test for heteroskedasticity	867
15.3.3.2	Heteroskedasticity robust estimator	868
15.3.3.3	Feasible weighted least square	868
15.3.4	Residual normality test	870
15.3.4.1	Jarque-Bera test	870
15.3.4.2	D'Agostino's K^2 test	870
15.3.5	Autocorrelation of errors	871

15.3.5.1	Motivation and general remarks	871
15.3.5.2	Test of autocorrelation of errors	872
15.3.5.3	Models with known autocorrelation	874
15.3.5.4	Transformation to generalized linear regression	875
15.3.6	Outliers analysis and robust linear regression	877
15.3.6.1	Outliers and influential points	877
15.3.6.2	Outlier impact analysis	878
15.3.6.3	Robust M-estimation linear regression	881
15.3.7	Visual diagnosis	884
15.4	Linear regression case studies	887
15.4.1	Standard linear regression	887
15.4.2	Boston Housing example	889
15.5	Multivariate multiple linear regression (MMLR)	892
15.5.1	Canonical MMLR	892
15.5.1.1	Motivation and model	892
15.5.1.2	Ordinary least square solution	893
15.5.2	Reduced rank regression	894
15.6	Notes on Bibliography	899
16	MONTE CARLO METHODS	900
16.1	Generating random variables	902
16.1.1	Inverse transform method	902
16.1.2	Box-Muller method for standard normal random variable	904
16.1.3	Acceptance-rejection method	904
16.1.4	Composition approach	905
16.1.5	Generate dependent continuous random variables	908
16.1.5.1	Multivariate normal and lognormal distribution	908
16.1.5.2	Multivariate student t distribution	908
16.1.5.3	General joint distribution	909
16.1.6	Generate discrete random variables	909
16.1.6.1	Generate single discrete random variables	909
16.1.6.2	Generate correlated discrete random variables	910
16.2	Monte Carlo integration	912
16.2.1	Naive approach	912
16.2.2	Importance sampling	913
16.3	Markov chain Monte Carlo	916
16.3.1	Basics	916
16.3.1.1	Markov chain Monte Carlo (MCMC)	916
16.3.2	Metropolis-Hasting algorithm	917
16.3.3	Gibbs sampling	919
16.4	Monte Carlo for random processes	920
16.4.1	Simulating stochastic differential equations	920

16.4.1.1	Simulating Brownian motion	920
16.4.1.2	Simulating linear arithmetic SDE	920
16.4.1.3	Simulating linear geometric SDE	921
16.4.1.4	Simulation mean-reversion(OU) process	921
16.4.2	Stochastic interpolation	922
16.4.2.1	Interpolating Gaussian processes	922
16.4.2.2	Interpolating one Dimensional Brownian motions	923
16.4.2.3	Interpolating multi-dimensional Brownian motions	925
16.5	Monte Carlo variance reduction	928
16.5.1	Antithetic sampling	928
16.5.1.1	Basic principles	928
16.5.1.2	Methods and analysis	929
16.5.2	Control variates	931
16.5.2.1	Basic principles	931
16.5.2.2	Multiple control variates	933
16.6	Notes on bibliography	934

iv dynamics modeling methods

17	MODELS AND ESTIMATION IN LINEAR SYSTEMS	936
17.1	Difference equation	937
17.1.1	Introduction	937
17.1.2	Solution structure of linear difference equations	938
17.1.3	Solution to non-homogeneous equation	940
17.1.4	Linear equations with constant coefficients	941
17.1.4.1	Basic case	941
17.1.4.2	General case	944
17.2	Differential equations	947
17.2.1	Linear differential equations	947
17.2.1.1	Concepts	947
17.2.1.2	Wronskian and linear independence	948
17.2.1.3	General solution theory	951
17.2.1.4	Existence & uniqueness of solution	954
17.2.2	Linear homogeneous differential equations with constant coefficients	955
17.2.2.1	The key identity	955
17.2.2.2	The case of real roots	957
17.2.2.3	The case of complex roots	958
17.2.2.4	The complete solution set	958
17.2.3	Solution to non-homogeneous ODEs	961
17.2.3.1	General principles	961
17.2.3.2	Key identity approach	962

17.2.4	First order linear differential equation	968
17.3	Linear system	970
17.3.1	Solution space for linear homogeneous system	970
17.3.2	Linear independence and the Wronskian	972
17.3.3	The fundamental system and solution method	974
17.3.4	The non-homogeneous linear equation	975
17.3.5	Conversion of linear differential/difference equation to linear systems	979
17.3.6	Solution method for discrete system	980
17.4	Linear system with constant coefficients	981
17.4.1	General solutions	981
17.4.2	System eigenvector method: continuous-time system	981
17.4.2.1	Diagonalizable system	981
17.4.2.2	two-by-two non-diagonalizable system	988
17.4.2.3	Non-diagonalizable system	989
17.4.3	Equilibrium point	991
17.4.3.1	Discrete-time system	991
17.4.3.2	Continuous-time system	992
17.4.4	Stability	993
17.4.5	Complex eigenvalues/eigenvectors	995
17.4.6	Boundedness of linear systems	995
17.4.7	One dimensional Nonlinear dynamical system analysis	996
17.5	Least square estimation of constant vectors	998
17.5.1	linear static estimation from single measurement with no prior information	998
17.5.2	linear static estimation from single measurement with prior information	999
17.5.3	Batch and recursive least square estimation with multiple measurements	1000
17.5.4	Nonlinear least square estimation	1003
17.6	Kalman filter	1004
17.6.1	Preliminary: error propagation in linear systems	1004
17.6.1.1	Discrete-time system	1004
17.6.1.2	Continuous-time system	1004
17.6.2	Batch estimation	1005
17.6.3	From batch estimation to Kalman filter	1006
17.6.4	Extended Kalman filter for nonlinear system	1007
17.7	Notes on bibliography	1009
18	STOCHASTIC PROCESS	1010
18.1	Stochastic process	1012
18.1.1	Basic definition and concepts	1012

18.1.2	Stationarity	1014
18.2	Gaussian process	1017
18.2.1	Basic Gaussian process	1017
18.2.2	Stationarity	1018
18.3	Brownian motion (Wiener process)	1019
18.3.1	Definition and elementary properties	1019
18.3.2	Multi-dimensional Brownian motion	1021
18.3.3	Asymptotic behaviors	1022
18.3.4	The reflection principle	1023
18.3.5	Quadratic variation	1024
18.3.6	Discrete-time approximations and simulation	1026
18.4	Brownian motion variants	1027
18.4.1	Gaussian process generated by Brownian motion	1027
18.4.2	Brownian bridge	1028
18.4.2.1	Constructions	1028
18.4.2.2	Applications	1032
18.4.3	Geometric Brownian motion	1032
18.5	Poisson process	1034
18.5.1	Basics	1034
18.5.2	Arrival and Inter-arrival Times	1035
18.6	Martingale theory	1037
18.6.1	Preliminaries: Filtration and adapted process	1037
18.6.1.1	Basic concepts in filtration	1037
18.6.1.2	Filtration for Brownian motion	1039
18.6.2	Basics of martingales	1039
18.6.3	Martingale transformation	1042
18.7	Stopping time	1043
18.7.1	Stopping time examples	1043
18.7.1.1	First passage time	1043
18.7.1.2	Trivial stopping time	1043
18.7.1.3	Counter example: last exit time	1044
18.7.2	Wald's equation	1044
18.7.3	Optional stopping	1045
18.7.4	martingale method for first hitting time	1045
18.8	Notes on bibliography	1046
19	STOCHASTIC CALCULUS	1047
19.1	Ito stochastic integral	1048
19.1.1	Motivation	1048
19.1.2	Construction of Ito integral	1048
19.1.2.1	Ito integral of a simple process	1048
19.1.2.2	Ito integral of a general process	1051

19.1.3	Ito integral with deterministic integrands (Wiener integral)	1055
19.1.3.1	Basics	1055
19.1.3.2	Integration by parts	1058
19.2	Stochastic differential equations	1061
19.2.1	Ito Stochastic differential equations	1061
19.2.2	Ito's lemma	1062
19.2.3	Useful results of Ito's lemma	1065
19.2.3.1	Product rule and quotient rule	1065
19.2.3.2	Logarithm and exponential	1066
19.2.3.3	Ito integral by parts	1066
19.2.3.4	Differentiate integrals of Ito process	1067
19.3	Linear SDE	1069
19.3.1	State-independent linear arithmetic SDE	1069
19.3.2	State-independent linear geometric SDE	1069
19.3.3	Multiple dimension extension	1071
19.3.4	Exact SDE	1073
19.3.5	Calculation mean and variance from SDE	1074
19.3.6	Integrals of Ito SDE	1075
19.4	Ornstein-Uhlenbeck(OU) process	1080
19.4.1	OU process	1080
19.4.1.1	Constant coefficient OU process	1080
19.4.1.2	Time-dependent coefficient OU process	1084
19.4.1.3	Integral of OU process	1086
19.4.2	Exponential OU process	1089
19.4.3	Parameter estimation for OU process	1091
19.4.4	Multiple factor extension	1091
19.5	Notes on bibliography	1094
20	MARKOV CHAIN AND RANDOM WALK	1095
20.1	Discrete-time Markov chain	1097
20.1.1	The model	1097
20.1.2	Evolution of discrete chain	1099
20.2	Classification of states	1101
20.2.1	accessibility and communicating classes	1101
20.2.2	Transient and recurrent states and classes	1103
20.2.2.1	Transient and recurrent states	1103
20.2.2.2	From states to classes	1107
20.2.2.3	Qualitative classification of recurrent and transient classes	1108
20.2.3	Periodicity	1109
20.2.4	Positive and null recurrent	1111
20.2.5	Summary	1113

20.3	Absorption analysis	1114
20.3.1	Matrix structure for adsorption analysis	1114
20.3.2	Absorbing Markov chains	1116
20.3.3	Hitting and return analysis	1118
20.3.4	Examples	1120
20.3.4.1	Consecutive coin toss game	1120
20.4	Limiting behavior & distributions	1122
20.4.1	Preliminary: eigenvalue propoties of stochastic matrices	1122
20.4.1.1	Preliminary: Frobenius-Perron matrix theory	1122
20.4.1.2	More general situations	1124
20.4.2	Limiting theorem	1126
20.4.2.1	Limiting distribution	1126
20.4.2.2	Extensions via Long run return analysis	1130
20.4.3	Application: PageRank algorithm	1132
20.5	Detailed balance and spectral properties	1134
20.6	Random walk	1136
20.6.1	Basic concepts and properties	1136
20.6.2	Persistent random walk	1136
20.6.3	Asymptotic properties	1138
20.6.4	Gambler's ruin problems	1138
20.7	Notes on bibliography	1143
21	TIME SERIES ANALYSIS	1144
21.1	Overview of time series analysis	1145
21.1.1	Introduction to time series	1145
21.1.2	Stationarity	1146
21.1.2.1	Stationarity concept	1146
21.1.2.2	Rolling analysis	1148
21.1.3	Remove trend and seasonality	1149
21.2	Linear stationary process theory	1152
21.2.1	Preliminaries: the lag operator and polynomial	1152
21.2.2	Linear process	1153
21.2.3	Autoregressive (AR) process	1155
21.2.3.1	Basics	1155
21.2.3.2	Stationarity and invertibility condition	1158
21.2.3.3	Forecasting	1159
21.2.4	Moving average (MA) process	1162
21.2.4.1	Basics	1162
21.2.4.2	Stationarity and invertibility	1165
21.2.4.3	Forecasting	1166
21.2.5	ARMA process	1169
21.2.5.1	Basic properties	1169

21.2.6	Unit root AR process	1171
21.2.6.1	Unit root process	1171
21.2.6.2	Trend stationarity vs. unit root process	1173
21.2.6.3	Unit root test	1173
21.2.6.4	Forecasting	1174
21.2.7	Correlation analysis	1174
21.2.7.1	Autocorrelation statistical analysis	1174
21.2.7.2	Partial autocorrelation function theory	1175
21.2.7.3	Correlogram analysis example	1178
21.2.8	Model analysis and calibration	1180
21.2.8.1	Order selection	1180
21.2.8.2	Yule-Walker equations and related methods	1181
21.2.8.3	Linear regression approach	1184
21.2.8.4	Maximum likelihood estimation	1186
21.2.8.5	Example: a toy example	1187
21.2.9	Wold Representation theorem	1190
21.3	Extensions to multivariate time series	1193
21.3.1	Introduction	1193
21.3.2	Vector autoregressive models	1194
21.3.2.1	VAR(1) model	1194
21.3.2.2	VAR(2) model	1196
21.3.2.3	VAR(p) model	1198
21.3.3	Vector moving-average model	1200
21.4	Autoregressive conditional heteroscedastic model	1203
21.4.1	ARCH models	1203
21.4.1.1	The motivation and the model	1203
21.4.1.2	Statistical properties	1205
21.4.1.3	Variance forecasting	1211
21.4.1.4	Detect ARCH effect	1214
21.4.1.5	Parameter estimation	1215
21.4.2	GARCH models	1215
21.4.2.1	The model	1215
21.4.2.2	Connecting GARCH to ARCH	1218
21.4.2.3	Variance forecasting	1218
21.5	Notes on Bibliography	1221

v statistical learning methods

22	SUPERVISED LEARNING PRINCIPLES	1223
22.1	The supervised learning problem	1224
22.1.1	Concepts	1224
22.1.2	Framework	1226

22.2	Variance bias trade-off	1228
22.2.1	Underfitting and Overfitting	1228
22.2.2	Variance and bias trade-off	1230
22.2.3	Examples	1233
22.2.3.1	Linear regression	1233
22.2.4	No free lunch theorem	1234
22.3	Model loss and evaluation	1236
22.3.1	Common loss functions	1236
22.3.2	Model evaluation metrics	1240
22.3.2.1	Regression metrics	1240
22.3.2.2	Classification metrics	1241
22.3.2.3	ROC and PRC metrics	1243
22.3.2.4	Metrics for imbalanced data	1245
22.4	Model selection methods	1246
22.4.1	The training-validation-testing idea	1246
22.4.2	Cross-validation	1246
22.5	Data and feature engineering	1251
22.5.1	Data preprocessing	1251
22.5.1.1	Data standardization	1251
22.5.1.2	Data normalization	1252
22.5.1.3	Handle categorical data	1252
22.5.1.4	Handle missing values	1253
22.5.1.5	Dimensional reduction	1253
22.5.1.6	Centering kernel matrix	1253
22.5.2	Feature engineering I: basic routines	1254
22.5.2.1	Nonlinear transformation	1254
22.5.2.2	Polynomial features	1254
22.5.2.3	Binning	1254
22.5.3	Feature engineering II: feature selection	1255
22.5.3.1	Filtering methods	1255
22.5.3.2	Recursive elimination methods	1256
22.5.3.3	Regularization methods	1256
22.5.4	Feature engineering III: feature extraction	1256
22.5.4.1	Text analytics	1256
22.5.4.2	Image	1257
22.5.4.3	Time series	1258
22.5.5	Imbalanced data	1259
22.5.5.1	Motivations	1259
22.5.5.2	Data resampling: undersampling	1259
22.5.5.3	Data resampling: upsampling	1259
22.5.5.4	Choice of loss functions, algorithms, and metrics	1260

22.6	Kernel methods	1261
22.6.1	Basic concepts of kernels and feature maps	1261
22.6.2	Mercer's theorem	1261
22.6.3	Common kernels	1263
22.6.4	Kernel trick and elementary algorithms using kernels	1265
22.6.5	Elementary algorithms	1265
22.7	Note on bibliography	1267
23	LINEAR MODELS FOR REGRESSION	1268
23.1	Standard linear regression	1269
23.1.1	Ordinary linear regression	1269
23.1.2	Application examples	1270
23.1.2.1	Boston housing prices	1270
23.2	Penalized linear regression	1274
23.2.1	Ridge regression	1274
23.2.1.1	Basics	1274
23.2.1.2	Dual form of ridge regression	1277
23.2.2	Lasso regression	1277
23.2.3	Elastic net	1280
23.2.4	Shrinkage Comparison	1280
23.2.5	Effective degree of freedom	1283
23.3	Basis function extension	1284
23.4	Note on bibliography	1286
24	LINEAR MODELS FOR CLASSIFICATION	1287
24.1	Logistic regression	1289
24.1.1	Logistic regression model	1289
24.1.2	Parameter estimation via maximum likelihood estimation	1290
24.1.3	Logistic regression with regularization	1294
24.1.4	Feature augmentation strategies	1295
24.1.5	Multinomial logistic regression	1296
24.1.6	Application examples	1297
24.1.6.1	South Africa heart disease	1297
24.1.6.2	Credit card fraud detection	1300
24.1.6.3	MNIST	1302
24.2	Gaussian discriminate analysis	1304
24.2.1	Linear Gaussian discriminant model	1304
24.2.1.1	The model	1304
24.2.1.2	Model parameter estimation	1305
24.2.1.3	Geometry of decision boundary	1305
24.2.2	Quadratic Gaussian discriminant model	1307
24.2.2.1	The model	1307
24.2.2.2	Model parameter estimation	1309

24.2.3	Application examples	1309
24.2.3.1	A toy example	1309
24.3	Fisher Linear discriminate analysis (Fisher LDA)	1312
24.3.1	One dimensional linear discriminant	1312
24.3.1.1	Basics	1312
24.3.1.2	Application in classification	1314
24.3.1.3	Possible issues	1316
24.3.2	Multi-dimensional linear discriminate	1317
24.3.2.1	Basics	1317
24.3.2.2	Application in classification	1318
24.3.3	Supervised dimensional reduction via Fisher LDA	1320
24.4	Separating hyperplane and Perceptron learning algorithm	1322
24.4.1	Basic geometry of hyperplanes	1322
24.4.2	The Perceptron learning algorithm	1323
24.5	Support vector machine classifier	1325
24.5.1	Motivation and formulation	1325
24.5.2	Optimality condition and dual form	1326
24.5.3	Soft margin SVM	1328
24.5.3.1	Basics	1328
24.5.3.2	Optimality condition for soft margin SVM	1329
24.5.3.3	Algorithm	1331
24.5.4	SVM with kernels	1333
24.5.5	A unified perspective from loss functions	1334
24.6	Note on bibliography	1337
25	GENERATIVE MODELS	1338
25.1	Naive Bayes classifier (NBC)	1342
25.1.1	Overview	1342
25.1.2	Binomial NBC	1342
25.1.3	Multinomial NBC	1344
25.1.4	Gaussian NBC	1347
25.1.5	Discussion	1349
25.2	Application	1350
25.2.1	Classifying documents using bag of words	1350
25.2.2	Credit card fraud prediction	1350
25.3	Supporting mathematical results	1353
25.3.1	Beta-binomial model	1353
25.3.1.1	The model	1353
25.3.1.2	Parameter inference	1354
25.3.2	Dirichlet-multinomial model	1356
25.3.2.1	The model	1356
25.3.2.2	Parameter inference	1359

26	K NEAREST NEIGHBORS	1360
26.1	Principles	1361
26.1.1	The algorithm	1361
26.1.2	Metrics and features	1362
26.2	Application examples	1364
27	TREE METHODS	1366
27.1	Preliminaries: entropy concepts	1367
27.1.1	Concept of entropy	1367
27.1.2	Conditional entropy and mutual information	1368
27.2	Classification tree	1372
27.2.1	Basic concepts of decision tree learning	1372
27.2.2	A generic tree-growth algorithm	1373
27.2.3	Splitting criterion	1374
27.2.4	Tree pruning	1378
27.2.5	Practical algorithms	1378
27.2.6	Examples	1381
27.2.6.1	Tree structures in Iris data classification	1381
27.3	Regression tree	1384
27.3.1	Basics	1384
27.3.2	Practical algorithms	1387
27.3.3	Examples	1387
27.3.3.1	A toy example	1387
27.3.3.2	Boston Housing prices	1388
28	ENSEMBLE AND BOOSTING METHODS	1390
28.1	Motivation and overview	1393
28.2	Bagging Methods	1395
28.2.1	A basic bagging method	1395
28.2.2	Tree bagging	1396
28.2.3	Random Forest	1398
28.3	Adaboost	1401
28.3.1	Adaboost classifier	1401
28.3.2	Adaboost regressor	1405
28.3.3	Additive model framework	1406
28.3.3.1	Generic additive model algorithm	1406
28.3.3.2	Adaboost as a special additive model	1407
28.4	Gradient boosting machines	1409
28.4.1	Fundamental	1409
28.4.2	Gradient boosting tree	1410
28.5	XGBoost	1415
28.6	Notes on Bibliography	1419

29	UNSUPERVISED STATISTICAL LEARNING	1420
29.1	Singular value decomposition (SVD) and matrix factorization	1421
29.1.1	SVD theory	1421
29.1.1.1	SVD fundamentals	1421
29.1.1.2	SVD and matrix norm	1423
29.1.1.3	SVD low rank approximation	1424
29.1.2	Principal component analysis (PCA)	1426
29.1.2.1	Statistical perspective of PCA	1426
29.1.2.2	Geometric fundamentals of PCA	1429
29.1.2.3	Robust PCA with outliers	1431
29.1.3	Sparse coding and dictionary learning	1433
29.1.3.1	Sparse coding	1433
29.1.3.2	Dictionary learning	1434
29.1.3.3	Online dictionary learning	1436
29.1.4	Non-negative matrix factorization	1438
29.2	Advanced applications of matrix factorization methods	1440
29.2.1	Latent semantic analysis	1440
29.2.2	Collaborative filtering in recommender systems	1443
29.2.3	Co-occurrence based word embedding	1448
29.3	Manifold learning	1450
29.3.1	Overview	1450
29.3.2	Preliminary: multidimensional scaling (MDS)	1450
29.3.2.1	Motivation	1450
29.3.2.2	Solution to classical MDS	1451
29.3.3	Isomap	1455
29.3.4	Kernel PCA	1456
29.3.5	Laplacian eigenmap	1458
29.3.5.1	Preliminary: graph Laplacian	1458
29.3.5.2	Laplacian eigenmap	1460
29.3.6	Diffusion map	1463
29.3.7	Application examples	1466
29.3.7.1	MNIST	1466
29.4	Clustering	1468
29.4.1	Overview	1468
29.4.2	K-means	1468
29.4.2.1	Canonical K-means	1468
29.4.2.2	K means++	1471
29.4.2.3	Kernel K means	1472
29.4.3	Density-based spatial clustering of applications with noise (DB-SCAN)	1473
29.4.4	Spectral clustering	1475

29.4.5	Gaussian mixture models (GMM)	1476
29.4.5.1	Preliminaries: Expectation Maximization (EM) algorithm	1476
29.4.5.2	The GMM model and algorithm	1478
29.4.6	Application examples	1480
29.4.6.1	Image segmentation	1480
29.5	Notes on Bibliography	1482
30	NEURAL NETWORK AND DEEP LEARNING	1483
30.1	Neural network foundations	1485
30.1.1	From machine learning to deep learning	1485
30.1.2	Neurons and neural networks	1486
30.1.2.1	Artificial neurons	1486
30.1.2.2	Artificial neural networks	1488
30.1.3	Universal approximation	1491
30.1.4	Training via backpropagation	1492
30.2	Optimization algorithms	1498
30.2.1	Motivation	1498
30.2.2	Full Batch gradient descent	1499
30.2.3	Minibatch stochastic gradient descent	1500
30.2.4	Adaptive gradient method	1501
30.2.4.1	Adaptive gradient (AdaGrad)	1501
30.2.4.2	RMSProp & AdaDelta	1501
30.2.5	Momentum method	1503
30.2.6	Combined together: adaptive momentum (Adam)	1504
30.3	Training and regularization techniques	1506
30.3.1	Choices of activation functions	1506
30.3.2	Weight initialization	1506
30.3.2.1	Motivation	1506
30.3.2.2	Xvaier initialization	1507
30.3.2.3	He initialization	1508
30.3.3	Data normalization	1508
30.3.3.1	Initial data standardization	1508
30.3.3.2	Batch normalization	1509
30.3.4	Regularization	1510
30.3.4.1	L_p regularization	1510
30.3.4.2	Weight decay	1510
30.3.4.3	Early stopping	1511
30.3.4.4	Dropout	1511
30.3.4.5	Data augmentation	1512
30.3.4.6	Label smoothing	1512
30.4	Feed-forward neural network examples	1513

30.4.1	Linear regression and classification	1513
30.4.2	Image classification	1515
30.4.3	Word embedding	1518
30.4.4	Sentiment analysis	1522
30.4.5	Approximating numerical partial differential equations	1524
30.5	Convolutional neural networks (CNN)	1528
30.5.1	Foundations	1528
30.5.2	CNN classical architectures	1531
30.5.2.1	LeNet	1531
30.5.2.2	AlexNet	1531
30.5.2.3	VGG	1532
30.5.2.4	ResNet	1533
30.6	CNN application examples	1536
30.6.1	Image classification	1536
30.6.2	Visualizing CNN	1537
30.6.2.1	Visualizing filters	1537
30.6.2.2	Visualizing classification activation map	1538
30.6.3	Autoencoders and denoising	1540
30.6.3.1	Autoencoders	1540
30.6.3.2	Denoising autoencoder	1541
30.6.4	Neural style transfer	1542
30.6.5	Visual based deep reinforcement learning	1545
30.6.6	Sentence classification	1547
30.7	Recurrent neural networks (RNN)	1550
30.7.1	Recurrent units	1550
30.7.1.1	Simple recurrent unit (SRU)	1550
30.7.1.2	Simple RNN and its approximation capability	1551
30.7.1.3	Backpropagation through time (BPTT)	1551
30.7.2	Recurrent unit variants	1553
30.7.2.1	Long short term memory (LSTM)	1553
30.7.2.2	Gated Recurrent Unit (GRU)	1556
30.7.3	Common RNN architectures	1557
30.8	RNN application examples	1561
30.8.1	Time series prediction	1561
30.8.1.1	Simple RNN prediction	1561
30.8.1.2	Deep autoregressive (DeepAR) model	1564
30.8.1.3	Deep factor model	1566
30.8.2	MNIST classification with sequential observation	1567
30.8.3	Sentiment classification	1567
30.8.4	Character-level language modeling	1568
30.8.4.1	Word classification	1568

30.8.4.2	Text generation	1569
30.9	Sequence-to-sequence modeling	1572
30.9.1	Encoder decoder model	1572
30.9.2	Attention mechanism	1574
30.10	Generative adversarial network (GAN)	1578
30.10.1	Canonical GAN	1578
30.10.1.1	Basics	1578
30.10.1.2	An example	1580
30.10.1.3	Understand training difficulties in GAN	1581
30.10.1.4	Deep Convolutional GAN (DCGAN)	1583
30.10.2	Conditional GAN	1585
30.10.3	Wasserstein GAN (WGAN)	1587
30.11	Notes on Bibliography	1591

vi optimal control and reinforcement learning methods

31	CLASSICAL OPTIMAL CONTROL THEORY	1593
31.1	Basic problem	1594
31.2	Controllability & observability	1595
31.3	Dynamic programming principle	1596
31.3.1	Principle of optimality	1596
31.3.2	The Hamilton-Jacobi-Bellman equation (finite horizon)	1596
31.3.3	The Hamilton-Jacobi-Bellman equation (infinite horizon)	1597
31.4	Deterministic linear quadratic control	1600
31.4.1	Linear quadratic control (finite horizon)	1600
31.4.2	Linear quadratic control(infinite horizon)	1601
31.5	Continuous-time stochastic optimal control	1603
31.5.1	HJB equation for general nonlinear systems	1603
31.5.2	Linear Gaussian quadratic system	1604
31.6	Stochastic dynamic programming	1605
31.6.1	Discrete-time Stochastic dynamic programming: finite horizon	1605
31.6.2	Discrete-time stochastic dynamic programming: infinite horizon	1607
31.6.2.1	Fundamentals	1607
31.6.2.2	Convergence analysis	1608
31.7	Notes on bibliography	1610
32	REINFORCEMENT LEARNING	1611
32.1	Preliminaries	1612
32.1.1	Notations	1612
32.1.2	Finite state Markov decision process	1612
32.1.3	Policy iteration and value iteration	1614

32.1.3.1	Policy iteration	1614
32.1.3.2	Value iteration	1618
32.2	Reinforcement learning theory	1622
32.2.1	Overview	1622
32.2.2	State-action Value function (Q function)	1624
32.2.3	Monte-Carlo method	1626
32.2.3.1	On-policy value estimation	1626
32.2.3.2	Off-policy value estimation	1628
32.2.3.3	MC-based reinforcement learning control	1628
32.2.4	TD(0) learning	1629
32.2.4.1	TD(0) for value estimation	1629
32.2.4.2	On-policy reinforcement learning control	1630
32.2.4.3	Off-policy reinforcement learning control	1631
32.2.5	TD(n) learning	1633
32.2.5.1	Motivation and concepts	1633
32.2.5.2	TD(n) for value estimation	1634
32.2.5.3	TD(n) for reinforcement learning control	1636
32.2.6	Standing challenges in reinforcement learning	1636
32.2.6.1	Curse of dimensionality	1636
32.2.6.2	Sample efficiency	1637
32.2.6.3	Exploration-exploitation dilemma	1637
32.2.6.4	Deadly triad	1637
32.3	Policy gradient learning	1638
32.3.1	Stochastic policy gradient fundamentals	1638
32.3.1.1	Preliminaries: derivative and expectation	1638
32.3.1.2	Theoretical framework based on finite-horizon trajectories	1639
32.3.1.3	Theoretical framework based on distributions*	1642
32.3.1.4	Estimate policy gradient and basic algorithms	1646
32.3.1.5	Bootstrap and Actor-Critic methods	1648
32.3.1.6	Common stochastic policies and their representations	1650
32.3.2	Advanced methods for policy gradient estimation	1652
32.3.2.1	Stochastic policy gradient with baseline	1652
32.3.2.2	Generalized advantage estimation	1655
32.3.2.3	Summary of stochastic gradient descent forms	1656
32.3.3	Deterministic policy gradient	1657
32.4	Algorithms zoo	1659
32.4.1	Neural Fitted Q Iteration (NFQ)	1659
32.4.2	Canonical deep Q learning	1660
32.4.3	DQN variants	1662

32.4.3.1	Overview	1662
32.4.3.2	Double Q learning	1663
32.4.3.3	Dueling network	1663
32.4.3.4	Deep Recurrent Q network (DRQN)	1664
32.4.3.5	Asynchronous Methods	1664
32.4.4	Universal value function approximator	1666
32.4.5	Deep deterministic policy gradient (DDPG) algorithm	1669
32.4.6	Twin-delayed deep deterministic policy gradient (TD3)	1671
32.4.7	Trust Region Policy Optimization (TRPO)	1674
32.4.7.1	TRPO	1674
32.4.7.2	Evaluating Hessian of KL-divergence	1676
32.4.8	Proximal Policy Optimization (PPO)	1677
32.4.9	Soft Actor-Critic(SAC)	1679
32.4.9.1	Entropy regulated reinforcement learning	1679
32.4.9.2	The SAC algorithm	1680
32.4.10	Evolution strategies	1681
32.5	Advanced training strategies	1684
32.5.1	Priority experience replay	1684
32.5.2	Hindsight experience generation	1685
32.5.3	Reverse goal generation	1686
32.5.4	Reverse goal generation on low-dimensional manifolds	1687
32.5.4.1	Key idea	1687
32.5.4.2	Example: navigation on a curved surface	1688
32.6	Notes on bibliography	1689

vii appendix

A	SUPPLEMENTAL MATHEMATICAL FACTS	1691
A.1	Basic logic for proof	1692
A.2	Some common limits	1693
A.3	Common series summation	1695
A.4	Some common spaces	1696
A.4.1	Notations on continuously differentiable functions	1697
A.5	Different modes of continuity	1698
A.5.1	continuity vs. uniform continuity	1699
A.6	Exchanges of limits	1700
A.6.1	Overall remark	1700
A.6.2	exchange limits with infinite summations	1700
A.6.3	Exchange limits with integration and differentiation	1700
A.6.4	Exchange differentiation with integration	1701
A.6.5	Exchange limit and function evaluations	1702
A.7	Useful inequalities	1703

A.7.1	Gronwall's inequality	1703
A.7.2	Inequality for norms	1703
A.7.3	Young's inequality for product	1704
A.8	Useful properties of matrix	1705
A.8.1	Matrix derivatives	1705
A.8.2	Matrix inversion lemma	1705
A.8.3	Block matrix	1706
A.8.4	Matrix trace	1707
A.8.5	Matrix elementary operator	1708
A.8.6	Matrix determinant	1710
A.9	Numerical integration	1711
A.9.1	Gaussian quadrature	1712
A.10	Vector calculus	1714
A.11	Numerical linear algebra computation complexity	1715
A.12	Distributions	1716
A.13	Common integrals	1717
A.14	Nonlinear root finding	1718
A.14.1	Bisection method	1718
A.14.2	Newton method	1718
A.14.3	Secant method	1718
A.15	Interpolation	1720
A.15.1	cubic interpolation	1720
Alphabetical Index		1721

Part I

MATHEMATICAL FOUNDATIONS