

Contents

<i>Preface</i>	<i>page</i>	ix
1 Introduction		1
1.1 Data Analysis and Optimization		1
1.2 Least Squares		4
1.3 Matrix Factorization Problems		5
1.4 Support Vector Machines		6
1.5 Logistic Regression		9
1.6 Deep Learning		11
1.7 Emphasis		13
2 Foundations of Smooth Optimization		15
2.1 A Taxonomy of Solutions to Optimization Problems		15
2.2 Taylor's Theorem		16
2.3 Characterizing Minima of Smooth Functions		18
2.4 Convex Sets and Functions		20
2.5 Strongly Convex Functions		22
3 Descent Methods		26
3.1 Descent Directions		27
3.2 Steepest-Descent Method		28
3.2.1 General Case		28
3.2.2 Convex Case		29
3.2.3 Strongly Convex Case		30
3.2.4 Comparison between Rates		32
3.3 Descent Methods: Convergence		33
3.4 Line-Search Methods: Choosing the Direction		36
3.5 Line-Search Methods: Choosing the Steplength		38

vi	Contents	
3.6	Convergence to Approximate Second-Order Necessary Points	42
3.7	Mirror Descent	44
3.8	The KL and PL Properties	51
4	Gradient Methods Using Momentum	55
4.1	Motivation from Differential Equations	56
4.2	Nesterov's Method: Convex Quadratics	58
4.3	Convergence for Strongly Convex Functions	62
4.4	Convergence for Weakly Convex Functions	66
4.5	Conjugate Gradient Methods	68
4.6	Lower Bounds on Convergence Rates	70
5	Stochastic Gradient	75
5.1	Examples and Motivation	76
5.1.1	Noisy Gradients	76
5.1.2	Incremental Gradient Method	77
5.1.3	Classification and the Perceptron	77
5.1.4	Empirical Risk Minimization	78
5.2	Randomness and Steplength: Insights	80
5.2.1	Example: Computing a Mean	80
5.2.2	The Randomized Kaczmarz Method	82
5.3	Key Assumptions for Convergence Analysis	85
5.3.1	Case 1: Bounded Gradients: $L_g = 0$	86
5.3.2	Case 2: Randomized Kaczmarz: $B = 0, L_g > 0$	86
5.3.3	Case 3: Additive Gaussian Noise	86
5.3.4	Case 4: Incremental Gradient	87
5.4	Convergence Analysis	87
5.4.1	Case 1: $L_g = 0$	89
5.4.2	Case 2: $B = 0$	90
5.4.3	Case 3: B and L_g Both Nonzero	92
5.5	Implementation Aspects	93
5.5.1	Epochs	93
5.5.2	Minibatching	94
5.5.3	Acceleration Using Momentum	94
6	Coordinate Descent	100
6.1	Coordinate Descent in Machine Learning	101
6.2	Coordinate Descent for Smooth Convex Functions	103
6.2.1	Lipschitz Constants	104
6.2.2	Randomized CD: Sampling with Replacement	105
6.2.3	Cyclic CD	110

Contents	vii
6.2.4 Random Permutations CD: Sampling without Replacement	112
6.3 Block-Coordinate Descent	113
7 First-Order Methods for Constrained Optimization	118
7.1 Optimality Conditions	118
7.2 Euclidean Projection	120
7.3 The Projected Gradient Algorithm	122
7.3.1 General Case: A Short-Step Approach	123
7.3.2 General Case: Backtracking	124
7.3.3 Smooth Strongly Convex Case	125
7.3.4 Momentum Variants	126
7.3.5 Alternative Search Directions	126
7.4 The Conditional Gradient (Frank–Wolfe) Method	127
8 Nonsmooth Functions and Subgradients	132
8.1 Subgradients and Subdifferentials	134
8.2 The Subdifferential and Directional Derivatives	137
8.3 Calculus of Subdifferentials	141
8.4 Convex Sets and Convex Constrained Optimization	144
8.5 Optimality Conditions for Composite Nonsmooth Functions	146
8.6 Proximal Operators and the Moreau Envelope	148
9 Nonsmooth Optimization Methods	153
9.1 Subgradient Descent	155
9.2 The Subgradient Method	156
9.2.1 Steplengths	158
9.3 Proximal-Gradient Algorithms for Regularized Optimization	160
9.3.1 Convergence Rate for Convex f	162
9.4 Proximal Coordinate Descent for Structured Nonsmooth Functions	164
9.5 Proximal Point Method	167
10 Duality and Algorithms	170
10.1 Quadratic Penalty Function	170
10.2 Lagrangians and Duality	172
10.3 First-Order Optimality Conditions	174
10.4 Strong Duality	178
10.5 Dual Algorithms	179
10.5.1 Dual Subgradient	179
10.5.2 Augmented Lagrangian Method	180

10.5.3	Alternating Direction Method of Multipliers	181
10.6	Some Applications of Dual Algorithms	182
10.6.1	Consensus Optimization	182
10.6.2	Utility Maximization	184
10.6.3	Linear and Quadratic Programming	185
11	Differentiation and Adjoints	188
11.1	The Chain Rule for a Nested Composition of Vector Functions	188
11.2	The Method of Adjoints	190
11.3	Adjoints in Deep Learning	191
11.4	Automatic Differentiation	192
11.5	Derivations via the Lagrangian and Implicit Function Theorem	195
11.5.1	A Constrained Optimization Formulation of the Progressive Function	195
11.5.2	A General Perspective on Unconstrained and Constrained Formulations	197
11.5.3	Extension: Control	197
	Appendix	200
A.1	Definitions and Basic Concepts	200
A.2	Convergence Rates and Iteration Complexity	203
A.3	Algorithm 3.1 Is an Effective Line-Search Technique	204
A.4	Linear Programming Duality, Theorems of the Alternative	205
A.5	Limiting Feasible Directions	208
A.6	Separation Results	209
A.7	Bounds for Degenerate Quadratic Functions	213
	<i>Bibliography</i>	216
	<i>Index</i>	223