

SIMULATION-BASED OPTIMIZATION: Parametric Optimization Techniques and Reinforcement Learning

ABHIJIT GOSAVI

**Department of Industrial Engineering
The State University of New York, Buffalo**



**Kluwer Academic Publishers
Boston/Dordrecht/London**

Contents

List of Figures	xvii
List of Tables	xxi
Acknowledgments	xxiii
Preface	xxv
1. BACKGROUND	1
1.1 Why this book was written	1
1.2 Simulation-based optimization and modern times	3
1.3 How this book is organized	7
2. NOTATION	9
2.1 Chapter Overview	9
2.2 Some Basic Conventions	9
2.3 Vector notation	9
2.3.1 Max norm	10
2.3.2 Euclidean norm	10
2.4 Notation for matrices	10
2.5 Notation for n -tuples	11
2.6 Notation for sets	11
2.7 Notation for Sequences	11
2.8 Notation for Transformations	11
2.9 Max, min, and arg max	12
2.10 Acronyms and Abbreviations	12
2.11 Concluding Remarks	12
3. PROBABILITY THEORY: A REFRESHER	15
3.1 Overview of this chapter	15
3.1.1 Random variables	15
3.2 Laws of Probability	16

3.2.1	Addition Law	17
3.2.2	Multiplication Law	18
3.3	Probability Distributions	21
3.3.1	Discrete random variables	21
3.3.2	Continuous random variables	22
3.4	Expected value of a random variable	23
3.5	Standard deviation of a random variable	25
3.6	Limit Theorems	27
3.7	Review Questions	28
4.	BASIC CONCEPTS UNDERLYING SIMULATION	29
4.1	Chapter Overview	29
4.2	Introduction	29
4.3	Models	30
4.4	Simulation Modeling of Random Systems	32
4.4.1	Random Number Generation	33
4.4.1.1	Uniformly Distributed Random Numbers	33
4.4.1.2	Other Distributions	36
4.4.2	Re-creation of events using random numbers	37
4.4.3	Independence of samples collected	42
4.4.4	Terminating and non-terminating systems	43
4.5	Concluding Remarks	44
4.6	Historical Remarks	44
4.7	Review Questions	45
5.	SIMULATION OPTIMIZATION: AN OVERVIEW	47
5.1	Chapter Overview	47
5.2	Stochastic parametric optimization	47
5.2.1	The role of simulation in parametric optimization	50
5.3	Stochastic control optimization	51
5.3.1	The role of simulation in control optimization	53
5.4	Historical Remarks	54
5.5	Review Questions	54
6.	RESPONSE SURFACES AND NEURAL NETS	57
6.1	Chapter Overview	57
6.2	RSM: An Overview	58
6.3	RSM: Details	59
6.3.1	Sampling	60
6.3.2	Function Fitting	60
6.3.2.1	Fitting a straight line	60
6.3.2.2	Fitting a plane	63
6.3.2.3	Fitting hyper-planes	64

6.3.2.4	Piecewise regression	65
6.3.2.5	Fitting non-linear forms	66
6.3.3	How good is the metamodel?	67
6.3.4	Optimization with a metamodel	68
6.4	Neuro-Response Surface Methods	69
6.4.1	Linear Neural Networks	69
6.4.1.1	Steps in the Widrow-Hoff Algorithm	72
6.4.1.2	Incremental Widrow-Hoff	72
6.4.1.3	Pictorial Representation of a Neuron	73
6.4.2	Non-linear Neural Networks	73
6.4.2.1	The Basic Structure of a Non-Linear Neural Network	75
6.4.2.2	The Backprop Algorithm	78
6.4.2.3	Deriving the backprop algorithm	79
6.4.2.4	Backprop with a Bias Node	82
6.4.2.5	Deriving the algorithm for the bias weight	82
6.4.2.6	Steps in Backprop	84
6.4.2.7	Incremental Backprop	86
6.4.2.8	Example D	88
6.4.2.9	Validation of the neural network	89
6.4.2.10	Optimization with a neuro-RSM model	90
6.5	Concluding Remarks	90
6.6	Bibliographic Remarks	90
6.7	Review Questions	91
7.	PARAMETRIC OPTIMIZATION	93
7.1	Chapter Overview	93
7.2	Continuous Optimization	94
7.2.1	Gradient Descent	94
7.2.1.1	Simulation and Gradient Descent	98
7.2.1.2	Simultaneous Perturbation	101
7.2.2	Non-derivative methods	104
7.3	Discrete Optimization	106
7.3.1	Ranking and Selection	107
7.3.1.1	Steps in the Rinott method	108
7.3.1.2	Steps in the Kim-Nelson method	109
7.3.2	Meta-heuristics	110
7.3.2.1	Simulated Annealing	111
7.3.2.2	The Genetic Algorithm	117
7.3.2.3	Tabu Search	119
7.3.2.4	A Learning Automata Search Technique	123
7.3.2.5	Other Meta-Heuristics	128
7.3.2.6	Ranking and selection & meta-heuristics	128
7.4	Hybrid solution spaces	128
7.5	Concluding Remarks	129

7.6	Bibliographic Remarks	129
7.7	Review Questions	131
8.	DYNAMIC PROGRAMMING	133
8.1	Chapter Overview	133
8.2	Stochastic processes	133
8.3	Markov processes, Markov chains and semi-Markov processes	136
8.3.1	Markov chains	139
8.3.1.1	n -step transition probabilities	140
8.3.2	Regular Markov chains	142
8.3.2.1	Limiting probabilities	143
8.3.3	Ergodicity	145
8.3.4	Semi-Markov processes	146
8.4	Markov decision problems	148
8.4.1	Elements of the Markov decision framework	151
8.5	How to solve an MDP using exhaustive enumeration	157
8.5.1	Example A	158
8.5.2	Drawbacks of exhaustive enumeration	161
8.6	Dynamic programming for average reward	161
8.6.1	Average reward Bellman equation for a policy	162
8.6.2	Policy iteration for average reward MDPs	163
8.6.2.1	Steps	163
8.6.3	Value iteration and its variants: average reward MDPs	165
8.6.4	Value iteration for average reward MDPs	165
8.6.4.1	Steps	166
8.6.5	Relative value iteration	168
8.6.5.1	Steps	168
8.6.6	A general expression for the average reward of an MDP	169
8.7	Dynamic programming and discounted reward	170
8.7.1	Discounted reward	171
8.7.2	Discounted reward MDP	171
8.7.3	Bellman equation for a policy: discounted reward	173
8.7.4	Policy iteration for discounted reward MDPs	173
8.7.4.1	Steps	174
8.7.5	Value iteration for discounted reward MDPs	175
8.7.5.1	Steps	176
8.7.6	Getting value iteration to converge faster	177
8.7.6.1	Gauss Siedel value iteration	178
8.7.6.2	Relative value iteration for discounted reward	179
8.7.6.3	Span seminorm termination	180
8.8	The Bellman equation: An intuitive perspective	181
8.9	Semi-Markov decision problems	182
8.9.1	The natural process and the decision-making process	184
8.9.2	Average reward SMDPs	186

8.9.2.1	Exhaustive enumeration for average reward SMDPs	186
8.9.2.2	Example B	187
8.9.2.3	Policy iteration for average reward SMDPs	189
8.9.2.4	Value iteration for average reward SMDPs	191
8.9.2.5	Counterexample for regular value iteration	192
8.9.2.6	Uniformization for SMDPs	193
8.9.2.7	Value iteration based on the Bellman equation	194
8.9.2.8	Extension to random time SMDPs	194
8.9.3	Discounted reward SMDPs	194
8.9.3.1	Policy iteration for discounted SMDPs	195
8.9.3.2	Value iteration for discounted reward SMDPs	195
8.9.3.3	Extension to random time SMDPs	196
8.9.3.4	Uniformization	196
8.10	Modified policy iteration	197
8.10.1	Steps for discounted reward MDPs	198
8.10.2	Steps for average reward MDPs	199
8.11	Miscellaneous topics related to MDPs and SMDPs	200
8.11.1	A parametric-optimization approach to solving MDPs	200
8.11.2	The MDP as a special case of a stochastic game	201
8.11.3	Finite Horizon MDPs	203
8.11.4	The approximating sequence method	206
8.12	Conclusions	207
8.13	Bibliographic Remarks	207
8.14	Review Questions	208
9.	REINFORCEMENT LEARNING	211
9.1	Chapter Overview	211
9.2	The Need for Reinforcement Learning	212
9.3	Generating the TPM through straightforward counting	214
9.4	Reinforcement Learning: Fundamentals	215
9.4.1	Q -factors	218
9.4.1.1	A Q -factor version of value iteration	219
9.4.2	The Robbins-Monro algorithm	220
9.4.3	The Robbins-Monro algorithm and Q -factors	221
9.4.4	Simulators, asynchronous implementations, and step sizes	222
9.5	Discounted reward Reinforcement Learning	224
9.5.1	Discounted reward RL based on value iteration	224
9.5.1.1	Steps in Q -Learning	225
9.5.1.2	Reinforcement Learning: A “Learning” Perspective	227
9.5.1.3	On-line and Off-line	229
9.5.1.4	Exploration	230
9.5.1.5	A worked-out example for Q -Learning	231
9.5.2	Discounted reward RL based on policy iteration	234

9.5.2.1	<i>Q</i> -factor version of regular policy iteration	235
9.5.2.2	Steps in the <i>Q</i> -factor version of regular policy iteration	235
9.5.2.3	Steps in <i>Q-P</i> -Learning	237
9.6	Average reward Reinforcement Learning	238
9.6.1	Discounted RL for average reward MDPs	238
9.6.2	Average reward RL based on value iteration	238
9.6.2.1	Steps in Relative <i>Q</i> -Learning	239
9.6.2.2	Calculating the average reward of a policy in a simulator	240
9.6.3	Other algorithms for average reward MDPs	241
9.6.3.1	Steps in <i>R</i> -Learning	241
9.6.3.2	Steps in SMART for MDPs	242
9.6.4	An RL algorithm based on policy iteration	244
9.6.4.1	Steps in <i>Q-P</i> -Learning for average reward	244
9.7	Semi-Markov decision problems and RL	245
9.7.1	Discounted Reward	245
9.7.1.1	Steps in <i>Q</i> -Learning for discounted reward DTMDPs	245
9.7.1.2	Steps in <i>Q-P</i> -Learning for discounted reward DTMDPs	246
9.7.2	Average reward	247
9.7.2.1	Steps in SMART for SMDPs	248
9.7.2.2	Steps in <i>Q-P</i> -Learning for SMDPs	250
9.8	RL Algorithms and their DP counterparts	252
9.9	Actor-Critic Algorithms	252
9.10	Model-building algorithms	253
9.10.1	<i>H</i> -Learning for discounted reward	254
9.10.2	<i>H</i> -Learning for average reward	255
9.10.3	Model-building <i>Q</i> -Learning	257
9.10.4	Model-building relative <i>Q</i> -Learning	258
9.11	Finite Horizon Problems	259
9.12	Function approximation	260
9.12.1	Function approximation with state aggregation	260
9.12.2	Function approximation with function fitting	262
9.12.2.1	Difficulties	262
9.12.2.2	Steps in <i>Q</i> -Learning coupled with neural networks	264
9.12.3	Function approximation with interpolation methods	265
9.12.4	Linear and non-linear functions	269
9.12.5	A robust strategy	269
9.12.6	Function approximation: Model-building algorithms	270
9.13	Conclusions	270
9.14	Bibliographic Remarks	271
9.14.1	Early works	271
9.14.2	Neuro-Dynamic Programming	271
9.14.3	RL algorithms based on <i>Q</i> -factors	271
9.14.4	Actor-critic Algorithms	272
9.14.5	Model-building algorithms	272

9.14.6	Function Approximation	273
9.14.7	Some other references	273
9.14.8	Further reading	273
9.15	Review Questions	273
10.	MARKOV CHAIN AUTOMATA THEORY	277
10.1	Chapter Overview	277
10.2	The MCAT framework	278
10.2.1	The working mechanism of MCAT	278
10.2.2	Step-by-step details of an MCAT algorithm	280
10.2.3	An illustrative 3-state example	282
10.2.4	What if there are more than two actions?	284
10.3	Concluding Remarks	285
10.4	Bibliographic Remarks	285
10.5	Review Questions	285
11.	CONVERGENCE: BACKGROUND MATERIAL	287
11.1	Chapter Overview	287
11.2	Vectors and Vector Spaces	288
11.3	Norms	290
11.3.1	Properties of Norms	291
11.4	Normed Vector Spaces	291
11.5	Functions and Mappings	291
11.5.1	Domain and Range of a function	291
11.5.2	The notation for transformations	293
11.6	Mathematical Induction	294
11.7	Sequences	297
11.7.1	Convergent Sequences	298
11.7.2	Increasing and decreasing sequences	300
11.7.3	Boundedness	300
11.8	Sequences in \mathcal{R}^n	306
11.9	Cauchy sequences in \mathcal{R}^n	307
11.10	Contraction mappings in \mathcal{R}^n	308
11.11	Bibliographic Remarks	315
11.12	Review Questions	315
12.	CONVERGENCE: PARAMETRIC OPTIMIZATION	317
12.1	Chapter Overview	317
12.2	Some Definitions and a result	317
12.2.1	Continuous Functions	318
12.2.2	Partial derivatives	319
12.2.3	A continuously differentiable function	319
12.2.4	Stationary points, local optima, and global optima	319

12.2.5	Taylor's theorem	320
12.3	Convergence of gradient-descent approaches	323
12.4	Perturbation Estimates	327
12.4.1	Finite Difference Estimates	327
12.4.2	Notation	328
12.4.3	Simultaneous Perturbation Estimates	328
12.5	Convergence of Simulated Annealing	333
12.6	Concluding Remarks	341
12.7	Bibliographic Remarks	341
12.8	Review Questions	341
13.	CONVERGENCE: CONTROL OPTIMIZATION	343
13.1	Chapter Overview	343
13.2	Dynamic programming transformations	344
13.3	Some definitions	345
13.4	Monotonicity of T , $T_{\hat{\mu}}$, L , and $L_{\hat{\mu}}$	346
13.5	Some results for average & discounted MDPs	347
13.6	Discounted reward and classical dynamic programming	349
13.6.1	Bellman Equation for Discounted Reward	349
13.6.2	Policy Iteration	356
13.6.3	Value iteration for discounted reward MDPs	359
13.7	Average reward and classical dynamic programming	364
13.7.1	Bellman equation for average reward	365
13.7.2	Policy iteration for average reward MDPs	368
13.7.3	Value Iteration for average reward MDPs	372
13.8	Convergence of DP schemes for SMDPs	379
13.9	Convergence of Reinforcement Learning Schemes	379
13.10	Background Material for RL Convergence	380
13.10.1	Non-Expansive Mappings	380
13.10.2	Lipschitz Continuity	380
13.10.3	Convergence of a sequence with probability 1	381
13.11	Key Results for RL convergence	381
13.11.1	Synchronous Convergence	382
13.11.2	Asynchronous Convergence	383
13.12	Convergence of RL based on value iteration	392
13.12.1	Convergence of Q -Learning	392
13.12.2	Convergence of Relative Q -Learning	397
13.12.3	Finite Convergence of Q -Learning	397
13.13	Convergence of Q - P -Learning for MDPs	400
13.13.1	Discounted reward	400
13.13.2	Average Reward	401
13.14	SMDPs	402

13.14.1 Value iteration for average reward	402
13.14.2 Policy iteration for average reward	402
13.15 Convergence of Actor-Critic Algorithms	404
13.16 Function approximation and convergence analysis	405
13.17 Bibliographic Remarks	406
13.17.1 DP theory	406
13.17.2 RL theory	406
13.18 Review Questions	407
14. CASE STUDIES	409
14.1 Chapter Overview	409
14.2 A Classical Inventory Control Problem	410
14.3 Airline Yield Management	412
14.4 Preventive Maintenance	416
14.5 Transfer Line Buffer Optimization	420
14.6 Inventory Control in a Supply Chain	423
14.7 AGV Routing	424
14.8 Quality Control	426
14.9 Elevator Scheduling	427
14.10 Simulation optimization: A comparative perspective	429
14.11 Concluding Remarks	430
14.12 Review Questions	430
15. CODES	433
15.1 Introduction	433
15.2 C programming	434
15.3 Code Organization	436
15.4 Random Number Generators	437
15.5 Simultaneous Perturbation	439
15.6 Dynamic Programming Codes	441
15.6.1 Policy Iteration for average reward MDPs	442
15.6.2 Relative Iteration for average reward MDPs	447
15.6.3 Policy Iteration for discounted reward MDPs	450
15.6.4 Value Iteration for discounted reward MDPs	453
15.6.5 Policy Iteration for average reward SMDPs	460
15.7 Codes for Neural Networks	464
15.7.1 Neuron	465
15.7.2 Backprop Algorithm — Batch Mode	470
15.8 Reinforcement Learning Codes	478
15.8.1 Codes for Q -Learning	478
15.8.2 Codes for Relative Q -Learning	486
15.8.3 Codes for Relaxed-SMART	495

15.9 Codes for the Preventive Maintenance Case Study	506
15.9.1 Learning Codes	507
15.9.2 Fixed Policy Codes	521
15.10 MATLAB Codes	531
15.11 Concluding Remarks	535
15.12 Review Questions	535
16. CONCLUDING REMARKS	537
References	539
Index	551