Simulation-Based Optimization: The Second Edition

A. Gosavi[a]

This book is situated at the intersection of dynamic programming, **reinforcement learning** (RL), and **simulation-based optimization**, and it is deeply influenced by the foundational works that shaped these fields. The material presented here is primarily based on **peer-reviewed literature**; however, all content has been carefully studied and synthesized over several years prior to inclusion in this volume. Most numerical experiments are also drawn from the literature, with the exception of illustrative toy problems developed for pedagogical purposes. In the first edition (2003), computer programs were included in an appendix; in the present (2nd) edition, these have been made available online, thereby freeing substantial space for additional material.

The rigorous treatment of **Markov decision processes** (MDPs) owes much to the classical framework described in a structured manner in the highly cited book of Puterman (1994). The analytical depth of Bertsekas and Tsitsiklis (1996) has been particularly influential in clarifying the connection between dynamic programming and reinforcement learning through a key stochastic approximation method known as the Robbins–Monro algorithm (Robbins and Monro, 1951). The treatment of static (parametric) optimization draws on a broad range of contributions from the literature.

Reinforcement learning has its early roots in the actor–critic architecture of Barto et al. (1983), followed by the seminal contribution of Watkins (1989) in the form of Q-learning. Early foundations of control-based learning approaches were also laid in related research in approximate dynamic programming (Werbös, 1990, 1992) and in automata theory (Wheeler and Narenda, 1986). In the broader RL community, the text by Sutton and Barto (1998) has served as a central reference, reflecting the foundational role of these authors in the development of the field. Asynchronous stochastic approximation is analyzed in the book using arguments from Borkar (1998), and the two timescale updating conditions used in the book are from Borkar (1997). The account on neural networks in this book draws on Mitchell (1997).

Building on these foundations, the book develops several new contributions across both editions. These include R-SMART (Gosavi, 2004b, 2011), a two-timescale algorithm for stable solutions of average-reward MDPs and semi-MDPs (SMDPs) in the infinite-horizon setting, as well as the Q-Policy Learning algorithm (Gosavi, 2004a). The automata-theoretic approach to average reward SMDPs in the infinite-horizon setting is based on Gosavi et al. (2004), and the deterministic bound on the Q-value of Q-learning is taken from Gosavi (2006) .

On the static side of simulation-based optimization, the book covers a range of algorithms including simultaneous perturbation (Spall, 1992, 2003), stochastic adaptive search (Zabinsky, 2003; Alrefaei and Andradóttir, 1999), learning automata (Thathachar and Sastry, 1987), nested partitions (Shi and Olafsson, 2008), and the

CONTACT: A. Gosavi. Email: gosavia@mst.edu; Missouri University of Science and Technology

stochastic ruler method (Alrefaei and Andradóttir, 2001).

On the applied side, the book emphasizes computational methods grounded in practice. It includes detailed discussions on algorithmic parameter selection, the backpropagation algorithm of neural networks, step-by-step recipes for a large number of algorithms, and, at the end, case studies involving simultaneous perturbation and simulated annealing for airline optimization (revenue management) problems and reinforcement learning for maintenance optimization problems and airline revenue management problems.

Taken together, the two broad parts of the book—one focused on static optimization and the other on dynamic optimization—reflect an effort to unify algorithmic design with real-world applicability for optimizing stochastic discrete-event systems.

The following provides a brief overview of the chapters:

- **Chapter 1: Background** – Introduces the foundational concepts underlying the book.
- **Chapter 2: Discrete-Event Simulation** – Presents fundamental concepts in discrete-event simulation for readers unfamiliar with simulation models of stochastic systems.
- **Chapter 3: Overview of the Book** – Provides a unified framework for simulation-based optimization, distinguishing between static and dynamic (control) problems in discrete-event systems.
- **Chapter 4: Response Surfaces and Neural Networks** – Introduces function approximation methods based on regression and neural networks (linear and nonlinear, including backpropagation).
- **Chapter 5: Parametric (Static) Optimization Techniques** – Covers a broad range of techniques including meta-heuristics and simultaneous perturbation.
- **Chapter 6: Dynamic Programming** – Focuses on value, policy, and modified policy iteration algorithms for discounted and average-reward MDPs with known transition models, along with backward recursion for finite-horizon problems and linear programming formulations for infinite-horizon problems.
- **Chapter 7: Reinforcement Learning** – Covers Q-learning, Q-Policy learning, SARSA, approximate policy iteration, and related algorithms for MDPs and SMDPs when transition probabilities are unknown.
- **Chapter 8: Stochastic Search Techniques** – Discusses actor–critic techniques and automata-theoretic approaches for MDPs and SMDPs.
- **Chapter 9: Background Material for Mathematical Convergence Analysis** — Develops foundational material on convergence, including Cauchy sequences, Banach fixed-point theory, and ordinary differential equations for analyzing stochastic approximation schemes.
- **Chapter 10: Convergence Analysis of Static Optimization Techniques** – Develops asymptotic convergence theory for static optimization techniques, primarily using arguments based on ergodic Markov chains.
- **Chapter 11: Convergence Analysis of Control Optimization Techniques** – Provides convergence proofs for dynamic programming and reinforcement learning algorithms, including analysis of two timescale schemes.
- **Chapter 12: Case Studies** – Presents large-scale applications of the simulation-based optimization methodology.

The book represents a selected collection of topics, primarily focussed on *model-free* methods of simulation-based optimization. For model-based static optimization, the

reader is referred to the excellent textbook of Fu and Hu (1997).

References

Alrefaei, M. and Andradóttir, S. (1999). A simulated annealing algorithm with constant temperature for discrete stochastic optimization. *Management Science*, 45(5):748–764.

Alrefaei, M. and Andradóttir, S. (2001). A modification of the stochastic ruler method for discrete stochastic optimization. *European Journal of Operational Research*, 133:160–182.

Barto, A., Sutton, R., and Anderson, C. (1983). Neuronlike elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 13:835–846.

Bertsekas, D. and Tsitsiklis, J. (1996). *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, USA.

Borkar, V. (1997). Stochastic approximation with two-time scales. *Systems and Control Letters*, 29:291–294.

Borkar, V. (1998). Asynchronous stochastic approximation. *SIAM J. Control Optim.*, 36 No 3:840–851.

Fu, M. and Hu, J. (1997). *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*. Kluwer, Norwell, MA, USA.

Gosavi, A. (2004a). A reinforcement learning algorithm based on policy iteration for average reward: Empirical results with yield management and convergence analysis. *Machine Learning*, 55:5–29.

Gosavi, A. (2004b). Reinforcement learning for long-run average cost. *European Journal of Operational Research*, 155:654–674.

Gosavi, A. (2006). Boundedness of iterates in Q-learning. *Systems and Control Letters*, 55:347–349.

Gosavi, A. (2011). Target-sensitive control of Markov and semi-Markov processes. *International Journal of Control, Automation, and Systems*, 9(5):1–11.

Gosavi, A., Das, T. K., and Sarkar, S. (2004). A simulation-based learning automata framework for solving semi-Markov decision problems. *IIE Transactions*, 36:557–567.

Mitchell, T. M. (1997). *Machine Learning*. McGraw Hill, Boston, MA, USA.

Puterman, M. L. (1994). *Markov Decision Processes*. Wiley Interscience, New York, NY, USA.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statist.*, 22:400–407.

Shi, L. and Olafsson, S. (2008). *Nested Partitions Method, Theory and Applications*. Springer.

Spall, J. (1992). Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation. *IEEE Transactions on Automatic Control*, 37:332–341.

Spall, J. (2003). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley, NY.

Sutton, R. and Barto, A. (1998). *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, USA.

Thathachar, M. and Sastry, P. (1987). Learning optimal discriminant functions through a cooperative game of automata. *IEEE Transactions on Systems, Man,*

and Cybernetics, 17:73–85.

Watkins, C. (1989). *Learning from Delayed Rewards*. PhD thesis, Kings College, Cambridge, England.

Werbös, P. (1990). A menu of designs for reinforcement learning over time. In *Neural Networks for Control*, pages 67–95. MIT Press, MA.

Werbös, P. (1992). Approximate dynamic programming for real-time control and neural modeling. In *Handbook of intelligent control; ed.D.A. White and D.A. Sofge*. Van Nostrand Reinhold, NY.

Wheeler, R. M. and Narenda, K. S. (1986). Decentralized learning in finite Markov chains. *IEEE Transactions on Automatic Control*, 31(6):373–376.

Zabinsky, Z. (2003). *Stochastic Adaptive Search for Global Optimization*. Springer, NY.