Xi-Ren Cao

Stochastic Learning and Optimization

- A Sensitivity-Based Approach

With 119 Figures, 27 Tables, and 212 Problems

Springer

To the memory of my parents
Cao Yun Jiu and Guo Wen Ying

Preface

Performance optimization is very important in the design and operation of modern engineering systems in many areas, including communications (Internet and wireless), manufacturing, robotics, and logistics. Most engineering systems are too complicated to be modelled, or the system parameters cannot be easily identified. Therefore, learning techniques have to be utilized.

A Brief Description of Learning and Optimization

Learning and optimization of stochastic systems is a multi-disciplinary area that has attracted wide attention from researchers in many disciplines including control systems, operations research, and computer science. Areas such as perturbation analysis (PA) in discrete event dynamic systems (DEDSs), Markov decision processes (MDPs) in operations research, reinforcement learning (RL) in computer science, neuro-dynamic programming (NDP), identification, and adaptive control (I&AC) in control systems, share a common goal: to make the "best decision" to optimize a system's performance.

Different areas take different perspectives and have different formulations for the problems with the same goal. *This book provides an overview of these different areas, PA, MDPs, RL, and I&AC, with a unified framework based on a sensitivity point of view. It also introduces new approaches and proposes new research topics and directions with this sensitivity-based framework.*

Roughly speaking, with RL, we learn how to make decisions to improve a system's performance by observing and analyzing the system's current behavior; the structure and the parameters of the system may not be known and even may not need to be estimated. PA estimates the derivatives of a system's performance with respect to the system's parameters by observing and analyzing the system's behavior. Optimization is achieved by combining performance derivative estimation and other optimization techniques such as stochastic approximation. MDPs provide a theoretical foundation for performance optimization of systems with a Markov model [21, 216]. In adaptive control, the system behavior is described by differential or difference equations; when the system parameters are unknown, they have to be identified using the observed data. Adaptive control together with identification achieves the same goal as learning and optimization.

The goal of these research areas is the same: to find a policy that optimizes a system's performance, by using the information "learned" by observing or analyzing the system's behavior. Given a system's status or history, a policy determines an action to be applied to the system, which controls the system evolution. In some cases, policies depend on continuous parameters and the policy space is continuous; in other cases, the policy space is discrete and usually contains a huge number of policies.

A Sensitivity-Based View

Recent research indicates that the various disciplines in learning and optimization can be explained from a unified point of view based on the performance sensitivities in the policy space [56]. The fundamental elements of learning and optimization are two types of performance sensitivity formulas, one for performance derivatives at any policy in the policy space, the other, for performance differences between any two policies in the policy space. With these two types of sensitivity formulas, existing results in the various areas and their relations can be derived or explained in a simple and intuitive way, new approaches can be introduced, and the average, discounted, and other performance criteria can be treated in the same way.

The unified framework is based on a few simple and fundamental facts: Naturally, by observing and analyzing a system's behavior under one policy, we cannot know the system's performance under other policies, if no structural information about the system is known; and we can only compare the performance of two policies at a time. The question is, with these fundamental limitations, how can we achieve our goal of performance optimization by using as little information about the system structure as possible and with as little computation effort as possible?

Thinking along this direction, we find that two things can be done: First, if the policy space is continuous, *with some knowledge about the system structure (e.g., queueing or Markov) and by the PA principles, we may estimate the performance derivatives at a policy along any direction in the policy space by observing/analyzing the behavior of the system under this policy [70, 62, 69].* This leads to the performance derivative formula; all the quantities required to calculate the derivative along a given direction can be obtained by analyzing the sample paths of the current policy. The performance derivative formula forms the basis for PA and the "policy gradient" approach that was proposed recently in the RL research community.

Second, if the policy space is discrete, the performance difference formula forms the basis for optimization. The difference formula compares the performance of the system under two policies. However, unlike the derivative formula, the difference formula involves quantities for both policies and it is not possible to know the performance of another policy, or the difference in the performance of a system under two policies, by observing or analyzing only the behavior of the system under one policy. Fortunately, *by the particular factorized form of the performance difference formula, under some* *structural conditions, we can always use the information learned from observing/analyzing the system behavior under a policy to find another policy under which the performance of the system is better, if such better policies exist.* This leads to policy iteration: learn from a policy to find another better policy, and learn from this better policy to find an even better policy, and so on. Thus, the performance difference formula forms the basis for policy-iteration type approaches to performance optimization. We will show that the results in I&AC can also be derived using this principle, which also provides a learning-based perspective to the area.

The fundamental quantity in the two sensitivity formulas (and thus in the two types of optimization approaches, the gradient-based and the policy-iteration-based) is the performance potential, which has a clear physical meaning: It measures the "potential" contribution of a state to the system performance. The difference of the potentials of two states measures the effect of changing from one state to the other on the system performance. Such a change from one state to the other is called a perturbation in PA (or simply called a "jump" on a sample path). In RL, many efficient algorithms (e.g., TD(λ) and Q-learning) have been developed for estimating the potential and its variant Q-factor and their values for the optimal policies.

The physical interpretation of the potentials leads to the fundamental principle in PA: The effect of any change in a system's structure or parameters can be decomposed into the effects of many jumps among states (or many perturbations). With this principle, we can use the potentials as building blocks to construct new sensitivity formulas by first principles for many problems that do not fit into the standard formulation in the existing literature [59]. Since, as explained, such sensitivity formulas serve as the basis for learning and optimization, the sensitivity construction approach opens up a new direction: New learning and optimization schemes can be developed based on these new sensitivity formulas, and special system features can be utilized.

One of the approaches developed based on sensitivity construction is called the *event-based optimization* where actions can be taken only when some events happen. This approach utilizes the special feature of a system captured by events. Policy depends on events, rather than on states. An event is defined as a set of state transitions and, therefore, an event occurring in the present contains some information about the next state, i.e., the future. In many modern engineering systems in information technology, such information is accessible before actions are taken, and the standard Markov model does not capture this special property. Thus, in some cases, event-based policies may perform better than state-based ones. Furthermore, the number of events usually scales to the system size, which is much smaller than that of the states, which grows exponentially with the size of the system. Thus, under some conditions, this approach may provide a possibility to overcome or to alleviate a computational difficulty: the curse of dimensionality. In addition, many existing approaches, such as partially observed MDPs (POMDPs), state and time aggregation, hierarchical control (hybrid systems), options, and singular perturbation, can be treated as special cases of the event-based optimization by defining different events to capture the special features of these different problems.

The Unique Features of This Book

Compared with other books in the area of learning and optimization, this book is unique in the following aspects.

1. The book covers various disciplines in learning and optimization, including PA, MDPs, RL, and I&AC, with a unified framework based on a sensitivity perspective in the policy space. Many results can be explained with the two types of fundamental sensitivity formulas in a simple way.
2. We emphasize physical interpretations rather than mathematics. With the intuitive physical explanations, we propose to construct new sensitivity formulas with performance potentials as building blocks. The physical intuition may provide insights that complement to other existing approaches.
3. With the unified framework and the construction approach, we introduce the recently-developed event-based optimization approach; this approach opens up a research direction in overcoming/alleviating the curse of dimensionality issue by utilizing the system's special features.
4. The performance difference-based approach is applied to all the MDP problems, including ergodic and multi-chain systems, average and discounted performance criteria, and even bias optimality and nth-bias optimality. It is shown that the nth-bias optimal policies eventually lead to the Blackwell optimal policies. This approach provides a simple, intuitively clear, and comprehensive presentation of all these problems in MDPs in a unified way. This presentation of MDPs is unique in existing books.

The Contents of This Book

Chapter 1 presents an introduction, which consists of an overview of the different disciplines in learning and optimization and a discussion of the event-based approach. This chapter serves as a road map for the book. The rest of the book consists of three parts. Part I, consisting of Chapters 2 to 7, describes how the sensitivity point of view in policy spaces leads to the main concepts and results in PA, MDPs, RL, and I&AC. Part II, consisting of Chapters 8 and 9, presents the recent developments in event-based learning and optimization with this sensitivity point of view. Part III consists of three appendices that provide the mathematical background required for this book.

Part I starts with PA in Chapter 2. We derive the performance derivative formulas, by using performance potentials or realization factors as building blocks, for Markov systems and queueing systems. The sample-path-based sensitivity point of view in PA is the core of the unified approach of this book. In Chapter 3, we discuss performance potentials and develop sample-path-based algorithms for estimating potentials and performance derivatives, as well as for performance optimization with the potentials. In Chapter 4, we show how policy iteration for both uni-chain and multi-chain MDPs can be easily derived from the performance difference formulas; this approach applies in the same way to both average and discounted criteria, as well as bias optimality, etc. We also define and solve the nth-bias optimality problem with the same approach. On-line policy iteration algorithms are developed in Chapter 5 with the potentials estimated from the sample paths. Chapter 6 presents basic results of RL, which is essentially a combination of stochastic approximation and the sample-path-based estimation of the potentials and their variants Q-factors. In Chapter 7, we show that the on-line policy iteration approach can be applied to I&AC problems, including linear systems and some non-linear systems.

In Part II, Chapter 8 presents the event-based optimization approach. This approach provides a possible way to address the difficult issue of the curse of dimensionality by utilizing particular system structures; in some cases event-based policies may perform better than state-based ones. The construction of sensitivity formulas with performance potentials as building blocks for general problems is presented in Chapter 9.

How to Use This Book

This book provides, in a unified way, good introductory materials for graduate students and engineers who wish to have an overview of learning and optimization theory, the related methodologies in different disciplines, including PA, MDPs, RL, I&AC, and stochastic approximation, and their relations. The new perspective presented in this book is helpful in finding new research topics. Thus, the book is useful to researchers in these areas who wish to find some motivation and to promote inter-disciplinary collaborations. In addition, engineers, in particular those in information technology, may find the ideas and methodologies introduced in this book useful in their practical applications.

The chapters and sections marked with asterisks "$*$" are supplementary reading material and can be omitted by first-time readers. Each chapter contains a considerable number of problems that may help students to enhance their understanding of the main contents. Some of the problems are summaries of past research topics and might be difficult. These are also marked with asterisks. Solutions to the problems are available upon request and can be found on my website http://www.ee.ust.hk/∼eecao.

An earlier version of this book was used as the textbook for graduate courses at the Hong Kong University of Science and Technology and Tsinghua University in Beijing, China. A suggested time table for a course in a fourteen-week term (three hours per week) is as follows.

Chapters	Sections		Hours	Weeks
A-C	A.1 - C.2		3	1
1	1.1 - 1.4		2	2/3
2	2.1		4	4/3
	2.2		1	1/3
	2.4		3	1
3	3.1 - 3.3		3	1
Review			1	1/3
4	4.1		3	1
	4.2		3	1
5	5.1 - 5.2		3	1
6	6.1 - 6.4		4	4/3
7	7.1 - 7.3		2	2/3
8	8.1 - 8.5		5	5/3
9	9.1 - 9.2		1	1/3
Review			1	1/3
Examination			3	1
		Total	42	14

The following are some suggestions and comments about the contents covered in each chapter:

0. The contents covered in the appendices are the prerequisite of the course. Three hours are not enough to cover the details in the three appendices. In a brief review, we may focus more on probability and the theory on Markov chains, which are closely related to the main concepts presented in the book. Appendix B is mainly related to Chapter 4, and Appendix C is mainly related to Section 2.4. Some results can be reviewed when the main texts are taught.
1. For students with a background in control, Section 1.1 on policies can be taught fast. Sections 1.2-1.3 are intended to give an overview of the different disciplines, and they should be revisited after studying Part I to get a better picture.
2. The main part of Chapter 2 is Sections 2.1 and 2.4.
3. Section 3.2 is relatively new in the literature.
4. Sections 4.1 and 4.2 cover the main ideas and methodologies. If time permits, we may cover the main results in Section 4.3 without going through the proofs.
5. The proof in Section 5.2.3 is interesting, but it is a bit technical and requires some careful thinking.
6. In Chapter 6, we emphasize on the intuitions behind the development of recursive algorithms, with the principles in stochastic approximation, and we do not intend to provide proofs for these algorithms. The algorithms for performance derivative estimates are new research topics in recent years.

7. In Chapter 7, it is easy to convince students that a control system can be modelled as an MDP. The extension of MDP from a discrete state space to a continuous state space is of no conceptual difficulty. We may cover only the LQ problem as an example.
8. Section 8.1 provides a nice overview of the event-based optimization approach. If we wish to avoid studying the tedious mathematical formulation, we may study the two examples to obtain a clear picture of the approach.
9. Section 9.2 provides the basic ideas for the construction of the performance difference formula. Other sections illustrate the flexibility of this approach and are for additional reading.

The following is a suggested time table for a course in a nine-week term (three hours per week).

Chapters	Sections		Hours	Weeks
A-C	A.1 - C.2		1.5	1/2
1	1.1 - 1.4		1.5	1/2
2	2.1		4	4/3
3	3.1 - 3.3		2	2/3
4	4.1		3	1
	4.2.1		2	2/3
5	5.1 - 5.2		2	2/3
6	6.1 - 6.4		3	1
7	7.1 - 7.3		2	2/3
8	8.1 - 8.5		3	1
9	9.1 - 9.2		1	1/3
Examinations			2	2/3
		Total	27	9

The additional suggestions are as follows.

1. We do not have time to cover PA of queueing systems in Section 2.4, PA-based optimization of queueing systems in Section 3.3, etc.
2. In Chapter 4, we may only briefly introduce the concept of the nth bias and the problem of the nth-bias optimality.
3. Event-based optimization can be introduced via examples.

Acknowledgements

A large portion of the book is based on my research on learning and optimization conducted over the past two to three decades, starting early 1980s when I was at Harvard University and working on PA. Many people have helped me in my research in various ways during this period.

I would like to express my sincere thanks to Prof. Yu-Chi Ho for his continuing support and encouragement; his insights and inspiration have made a significant impact on my research. I would like to thank the following people for joint works and/or insightful discussions in various periods on topics that are related to the materials covered in this book: K. J. Åström, T. Başar, A. G. Barto, D. P. Bertsekas, R. W. Brockett, C. G. Cassandras, H. F. Chen, A. Ephremides, H. T. Fang, E. A. Feinberg, M. C. Fu, P. Glasserman, W. B. Gong, X. P. Guo, B. Heidergott, P. V. Kokotovic, F. L. Lewis, L. Ljung, D. J. Ma, S. I. Marcus, S. P. Meyn, G. Ch. Pflug, L. Qiu, Z. Y. Ren, J. Si, R. Suri, J. N. Tsitsiklis, B. Van Roy, P. Varaiya, A. F. Veinott, Y. Wardi, Y. W. Wan, and J. Y. Zhang. I also wish to thank those people who have carefully read parts of the early draft of this book and provided useful comments regarding the presentations of the book and corrected typos: F. Cao, H. F. Chen, T. W. Chen, X. P. Guo, Q. L. Li, Y. J. Li, D. Y. Shi, L. Xia, Y. K. Xu, and J. Y. Zhang. In addition, I particularly appreciate the tedious work of Y. K. Xu and J. Y. Zhang in making Latex files for many pictures in the book. I also thank V. Unkefer for her technical editing of most part of this book and thank J. Q. Shen for drawing the figure for the book cover. Of course, all errors remain my responsibility. My sincere thanks also go to Harvard University, Digital Equipment Corporation, U.S.A., and the Hong Kong University of Science and Technology for providing me with financial support as well as excellent research environment during the past years.

Finally, I wish to express my sincere appreciation to my wife, Mindy Wang Cao, for her continuing support and understanding under all circumstance in the past years.

Hong Kong
April 2007

Xi-Ren Cao
The Hong Kong University
of Science and Technology
eecao@ust.hk