Risk-Sensitive Reinforcement Learning via Policy Gradient Search

Other titles in Foundations and Trends® in Machine Learning

Dynamical Variational Autoencoders: A Comprehensive Review
Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber and Xavier Alameda-Pineda
ISBN: 978-1-68083-912-8

Machine Learning for Automated Theorem Proving: Learning to Solve SAT and QSATe
Sean B. Holden
ISBN: 978-1-68083-898-5

Spectral Methods for Data Science: A Statistical Perspective
Yuxin Chen, Yuejie Chi, Jianqing Fan and Cong Ma
ISBN: 978-1-68083-896-1

Tensor Regression
Jiani Liu, Ce Zhu, Zhen Long and Yipeng Liu
ISBN: 978-1-68083-886-2

Minimum-Distortion Embedding
Akshay Agrawal, Alnur Ali and Stephen Boyd
ISBN: 978-1-68083-888-6

Graph Kernels: State-of-the-Art and Future Challenges
Karsten Borgwardt, Elisabetta Ghisu, Felipe Llinares-López, Leslie O'Bray and Bastian Rieck
ISBN: 978-1-68083-770-4

Risk-Sensitive Reinforcement Learning via Policy Gradient Search

Prashanth L. A.
Department of Computer Science and Engineering,
Indian Institute of Technology Madras
prashla@cse.iitm.ac.in

Michael C. Fu
Robert H. Smith School of Business &
Institute for Systems Research,
University of Maryland, College Park
mfu@umd.edu

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

Prashanth L. A. and M. C. Fu. *Risk-Sensitive Reinforcement Learning via Policy Gradient Search*. Foundations and Trends® in Machine Learning, vol. 15, no. 5, pp. 536–692, 2022.

ISBN: 978-1-63828-027-9

Foundations and Trends® in Machine Learning
Volume 15, Issue 5, 2022

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2022, Volume 15, 6 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Contents

Risk-Sensitive Reinforcement Learning via Policy Gradient Search

Prashanth L. A.[1] and Michael C. Fu[2]

[1] *Indian Institute of Technology Madras, India; prashla@cse.iitm.ac.in*
[2] *University of Maryland, College Park, USA; mfu@umd.edu*

ABSTRACT

The objective in a traditional reinforcement learning (RL) problem is to find a policy that optimizes the expected value of a performance metric such as the infinite-horizon cumulative discounted or long-run average cost/reward. In practice, optimizing the expected value alone may not be satisfactory, in that it may be desirable to incorporate the notion of risk into the optimization problem formulation, either in the objective or as a constraint. Various risk measures have been proposed in the literature, e.g., exponential utility, variance, percentile performance, chance constraints, value at risk (quantile), conditional value-at-risk, prospect theory and its later enhancement, cumulative prospect theory.

In this monograph, we consider risk-sensitive RL in two settings: one where the goal is to find a policy that optimizes the usual expected value objective while ensuring that a risk constraint is satisfied, and the other where the risk measure is the objective. We survey some of the recent work in this area specifically where policy gradient search is the solution approach. In the first risk-sensitive RL setting, we cover popular risk measures based on variance, conditional value-at-risk, and chance constraints, and present a template for

Prashanth L. A. and Michael C. Fu (2022), "Risk-Sensitive Reinforcement Learning via Policy Gradient Search", Foundations and Trends® in Machine Learning: Vol. 15, No. 5, pp 536–692. DOI: 10.1561/2200000091.

policy gradient-based risk-sensitive RL algorithms using a Lagrangian formulation. For the setting where risk is incorporated directly into the objective function, we consider an exponential utility formulation, cumulative prospect theory, and coherent risk measures. This non-exhaustive survey aims to give a flavor of the challenges involved in solving risk-sensitive RL problems using policy gradient methods, as well as outlining some potential future research directions.

Preface

Reinforcement learning (RL) is one of the foundational pillars of artificial intelligence and machine learning. An important consideration in any optimization or control problem is the notion of risk, but its incorporation into RL has been a fairly recent development. This monograph surveys research on risk-sensitive RL that uses policy gradient search, i.e., policy optimization in a stochastic formulation, as opposed to robust optimization approaches and methods that focus on the value function.

We have tried to make the exposition completely self-contained but also organized in a manner that allows expert readers to skip background sections. In particular, those readers already familiar with Markov decision processes (MDPs), risk measures, and stochastic gradient-based search (specifically, stochastic approximation) can skip Sections 2, 3, and 4, respectively.

We have benefited from the feedback of many who read earlier drafts of the manuscript. We begin by thanking Prof. Vivek Borkar, who generously offered valuable detailed comments regarding the content, and provided material and references for the sections on the exponential cost formulation. Next, we thank Prof. Shalabh Bhatnagar for helpful discussions on the convergence analysis in the risk-as-constraint setting, and Prof. Armand Makowski for critical observations. We'd also like to thank two anonymous reviewers, whose comments and suggestions helped us improve the exposition considerably. Lastly, we thank several of our Ph.D. students — Xingyu Ren, Erfaun Noorani, Mehrdad Moharrami, Nithia Vijayan, Yi Zhou, and Mengting Chao, who read through various portions and stages of the manuscript and caught numerous typos. Any remaining errors are of course our responsibility alone.

One final note: We have chosen to include references at the end of the section in bibliographic remarks rather than cite them in the main text, so as not to interrupt the expositional flow.

1

Introduction

Markov decision processes (MDPs) provide a general framework for modeling a wide range of problems involving sequential decision making under uncertainty, which arise in many areas of applications, such as transportation, computer/communication systems, manufacturing, and supply chain management. MDPs transition from state to state probabilistically over time due to chosen actions taken by the decision maker, incurring state/action-dependent costs/rewards at each instant. The goal is to find a policy (sequence of decision rules) for choosing actions that optimizes a long-run objective function, e.g., the cumulative sum of discounted costs or the long-run average cost.

The traditional MDP setting assumes that (i) the transition dynamics (probabilities) and costs/rewards are fully specified/known, and (ii) the objective function and constraints involve standard expected value criteria. However, in a myriad of settings of practical interest, neither of these conditions holds, i.e., only ***samples*** of transitions (and costs/rewards) can be observed (e.g., in a black-box simulation model or an actual system) and/or performance measures that incorporate ***risk*** really need to be considered in the problem. In the case of the former, reinforcement learning (RL) techniques can be employed, and in the

latter setting, risk-sensitive approaches are appropriate. Although there is abundant research on both of these settings dating back decades, the work combining both aspects is more recent. Furthermore, the two settings have been predominantly pursued independently by different research communities, with RL a focus of CS/AI researchers and risk-sensitive MDPs a focus of stochastic control and operations research/management science/mathematical finance researchers.

Why risk? (Avoid merely expectations?)

The focus of this monograph is not on why risk is important nor on what is the best way to incorporate it into decision making but rather on finding good risk-sensitive policies via RL policy gradient algorithms. However, to provide some motivation for incorporating risk into decision making, we briefly describe two everyday illustrative examples. The first example has to do with financial investments, where the primary objective is generally to *maximize* ***expected*** *return*. Clearly, this is not sufficient for most decision makers, who would very much like to take into consideration the "risk" of the investments, in this case taken to mean mitigating the potential downside losses. The second example is your daily commute to work. In this case, your primary objective is likely to *minimize* ***expected*** *travel time*. However, if you have an important early morning meeting, you might want to reduce the "risk" of being late by choosing an alternative that has a higher expected travel time but is unlikely to suffer a huge delay from an unexpected but rare event such as an overturned tractor-trailer. A colleague of ours avoids taking the highway to/from work for this very reason (along with safety considerations). In other words, most decision makers consider more than merely expectations. Both of these examples also serve to illustrate the more general observation that real-world decisions involve multiple objectives, where at least one of them involves the notion of risk, extending beyond the usual expected value performance measures considered in standard MDP and RL models (including commonly used metrics for analysis purposes such as expected regret in multi-armed bandit models).

Types of risk and ways to incorporate risk

As in any multi-objective optimization problem, there are many ways to incorporate risk. Again, our focus is not on advocating for one formulation over another, but to provide several different alternatives, with a solution approach for each of them. Which formulation is "better" will depend on both the problem and the problem solver(s). We illustrate this concept by revisiting our two examples.

One way to address risk in the investment problem is to minimize some measure of volatility, which could take the form of putting an upper bound on the *variance* of return. Thus, the decision problem becomes a constrained optimization of maximizing the objective of expected return *subject to a constraint* on the variance of return. This is the classic mean-variance portfolio optimization problem in finance for which Harry Markowitz was awarded the 1990 Nobel Prize in Economics.

It can be easily argued that variance is not the best measure of risk for this problem, since it also penalizes excessive upside moves, so maybe focusing on one tail (the downside risk) is more appropriate. One way to address this would be to limit the probability of a high loss to some acceptable level such as 5% or 1% or even smaller. This is known as a *chance constraint*. Conversely, one might have an upper bound on the amount of loss that might occur at a certain low probability, i.e., putting a constraint on a quantile of the loss distribution, which the financial industry defines as *value-at-risk* (VaR). A more sophisticated extension of VaR is *conditional value-at-risk* (CVaR), which also has some other nice properties that VaR does not, most notably that it is a *coherent risk measure*. Exponential utility is another way of capturing risk preferences and implicitly capturing higher moments beyond the second moment. Section 3 provides a more formal review of all of these risk concepts and metrics.

Similarly, revisiting risk in the commuting problem where the objective is to minimize travel time, a constrained optimization problem formulation would be to minimize expected travel time subject to an upper bound on the variability of travel time, or alternatively, one could instead employ a chance constraint by specifying the probability of the travel time exceeding an acceptable threshold, e.g., requiring that at least 99% of the time the travel time will be less than an hour.

Realistic problems may involve multiple constraints that need to be satisfied concurrently, such as bounds on both the variability and the probability of a rare event. In our setting, this can be easily handled, but for the sake of simplicity we will only explicitly consider the case of a single constraint, as the extension using the policy gradient approach would just involve additional Lagrange multiplier gradient estimates, but the general approach would be the same.

Finally, rather than formulating the problem with risk as a constraint, another approach is to try and include it in the objective function. Perhaps the simplest way would be as a weighted combination of the multiple objectives. While we don't address the weighed objectives formulation explicitly, it should be clear how it could also be handled as an easy special case using the techniques of this monograph. Instead, we consider more general formulations: the use of expected utility (an exponential cost formulation), which modifies the output performance measure (corresponding to investment return or travel time in the two examples), and a risk measure called *cumulative prospect theory* (CPT) that "distorts" the perceived probabilities due to the decision maker's view of the world. Demonstrating that prospect theory and CPT are able to model certain aspects of actual observed human behavior that utility theory was unable to capture was a key contribution for which (behavioral psychologist) Daniel Kahneman was awarded the 2002 Nobel Prize in Economics. Our treatment also extends the CPT formulation to a framework encompassing general coherent risk measures.

Objectives of this monograph

The main purpose of this monograph is to introduce and survey research results on policy gradient methods for reinforcement learning with risk-sensitive criteria, as well as to outline some promising avenues for future research following the risk-sensitive RL framework. We consider both constrained formulations where the traditional expected value performance measure is augmented with a risk constraint and problem formulations where the risk measure is explicitly in the objective function being optimized. Some well-known examples of risk measures to be considered as constraints, most of which were illustrated

by the two earlier examples, include variance (or higher moments), probabilities (in the form of chance constraints), quantiles or value-at-risk (VaR), and conditional value-at-risk (CVaR). As also mentioned in the examples, risk measures used explicitly as the objective function include exponential utility and some very recent work on using CPT with RL.

To be specific, the constrained risk-sensitive RL problem will be an optimization problem of the following general form:

$$\min_{\theta \in \Theta} J(\theta) \triangleq \mathbb{E}\left[D(\theta)\right] \text{ subject to } \quad G(\theta) \leq \kappa, \tag{1.1}$$

where θ denotes the policy parameter, Θ represents the policy space, $D(\theta)$ is a (stochastic) cost function, $G(\theta)$ is a risk measure, and κ denotes the acceptable risk level. In the MDP setting, the quantities may also depend on the initial state of the MDP, which is not indicated here. The most common choices for $D(\theta)$ in the MDP setting include the infinite-horizon cumulative discounted cost, total cost in a stochastic shortest path problem, and the long-run average cost. Note that we will be minimizing cost (as in the commuting example), which is more common in MDP formulations than in the RL setting, which often focuses on maximizing reward (as in the investment example). The classic "risk-neutral" formulation simply minimizes $J(\cdot)$ without the risk constraint in (1.1). Also, in contrast to the traditional setting of risk-sensitive control where J and G functions are analytically available in the MDP model, in the RL setting, J and G are unknown or cannot be calculated directly, but noisy estimates of J and G are available, e.g., samples of D could provide an unbiased estimator of J. Thus, as in the usual RL setting, traditional MDP techniques cannot be applied, whereas RL algorithms suitably adapted provide one avenue to attack such risk-sensitive MDPs, i.e., a setting when the MDP model is unknown and all the information about the system is obtained from samples resulting from the decision maker's interaction with the environment.

We propose to solve the constrained optimization problem (1.1) by performing gradient descent search on the Lagrangian objective function. As depicted in Figure 1.1, the risk-sensitive policy gradient algorithm requires estimators $\widehat{\nabla} J(\theta)$, $\widehat{\nabla} G(\theta)$ and $\widehat{G}(\theta)$ of $\nabla J(\theta)$, $\nabla G(\theta)$ and $G(\theta)$, respectively. Then, two-timescale gradient-based search algorithms

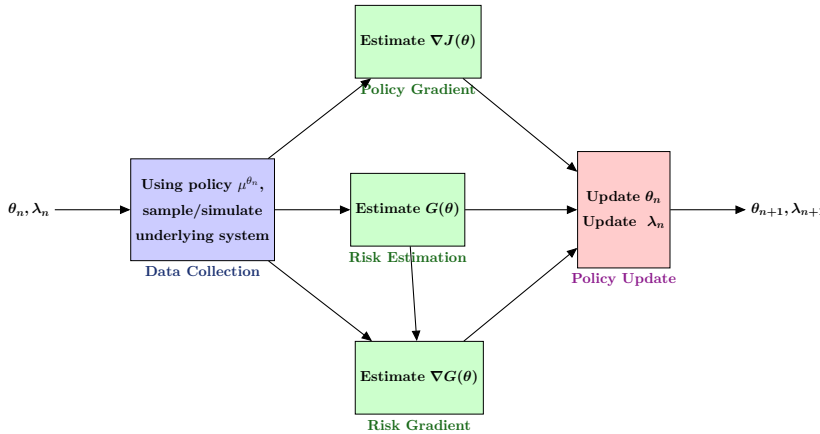

Figure 1.1: Schematic of risk-sensitive policy gradient algorithm for constrained optimization (underlying system could be a simulation model or a real system).

taking the following form will be developed (where λ is the Lagrange multiplier to be optimized along with the policy parameter θ):

$$\lambda_{n+1} = \left[\lambda_n + \zeta_1(n)\left(\widehat{G}(\theta_n) - \kappa\right)\right]^+,$$
$$\theta_{n+1} = \Gamma\left[\theta_n - \zeta_2(n)\left(\widehat{\nabla} J(\theta_n) + \lambda_n \widehat{\nabla} G(\theta_n)\right)\right],$$

where $[x]^+ = \max(0, x)$, Γ is a projection into Θ, and $\{\zeta_1(n), \zeta_2(n)\}$ are step-size sequences selected such that the θ update is on the faster timescale and the λ update is on the slower timescale (see Section 5.2 for details).

In addition to the risk-constrained problem (1.1), we also consider a risk-sensitive problem where the risk measure is explicitly incorporated into the objective function, i.e., the following optimization problem:

$$\min_{\theta \in \Theta} G(\theta), \tag{1.2}$$

where G is a risk objective function involving exponential utility, CPT, or a coherent risk measure. For solving the problem (1.2), we propose a policy gradient algorithm that incorporates the following iterative update:

$$\theta_{n+1} = \Gamma\left[\theta_n - \zeta(n)\widehat{\nabla} G(\theta_n)\right],$$

where $\{\zeta(n)\}$ is a step-size sequence, $\widehat{\nabla} G(\theta_n)$ is an estimate of $\nabla G(\theta_n)$, and Γ is a projection operator that keeps the iterate θ_n bounded within the set Θ as in the case of the risk-constrained policy gradient algorithm above (see Section 5.1 for details).

Challenges in risk-sensitive RL

Risk-sensitive RL is generally more challenging than its risk-neutral counterpart. For instance, for a discounted-cost MDP, there exists a Bellman equation for the variance of the return, but the underlying Bellman operator is not necessarily monotone, so that policy iteration is no longer guaranteed to lead to an optimal policy. Moreover, finding a globally mean-variance optimal policy in a discounted-cost MDP is NP-hard, even in the classic MDP setting where the transition model is known. Average-cost MDP problems also are generally NP-hard, e.g., consider a risk measure that is not the plain variance of the average cost and instead is a variance of a quantity that measures the deviation of the single-stage cost from the average cost. Finally, in comparison to variance/CVaR, CPT is a non-coherent and non-convex measure, ruling out the usual Bellman equation-based dynamic programming (DP) approaches when optimizing the MDP CPT-value.

The computational complexity results summarized in the previous paragraph imply that finding guaranteed global optima of risk-sensitive MDP formulations described by (1.1) or (1.2) is not computationally practical, motivating the need for algorithms that approximately solve such MDP formulations. In this monograph, we focus on policy gradient-type learning algorithms where the policies are parameterized in a continuous space, and an iterative search for a better policy occurs through a gradient-descent update. Actor-critic methods are a popular subclass of policy gradient methods and were among the earliest to be investigated in RL. They are comprised of an *Actor* that improves the current policy via gradient descent (as in policy gradient schemes) and a *Critic* that incorporates feature-based representations to approximate

the value function. The latter approximation is necessary to handle the curse of dimensionality. Regular policy gradient schemes usually rely on Monte Carlo methods for policy evaluation, an approach that suffers from high variance as compared to actor-critic schemes. On the other hand, function approximation introduces a bias in the policy evaluation. A policy gradient/actor-critic scheme with provable convergence to a locally risk-optimal policy would require careful synthesis of techniques from stochastic approximation, stochastic gradient estimation approaches, and importance sampling.

Several of the constituent solution pieces require significant research for various risk measures. For example, consider the "policy evaluation" part of the overall algorithm in a risk-sensitive MDP, which requires estimating $J(\theta)$ and $G(\theta)$, given samples obtained by simulating the MDP with policy θ. If $J(\theta)$ is one of the usual MDP optimization objectives such as discounted total cost, long-run average cost, or total cost (in a finite-horizon MDP), then estimating $J(\theta)$ can be performed using one of the existing algorithms. Temporal difference (TD) learning is a well-known algorithm that can learn the objective value along a sample path for a given θ. However, estimating $G(\theta)$ using TD-type learning algorithms is infeasible in many cases. For instance, consider variance as the risk measure in a discounted-cost MDP. In this case, even though there is a Bellman equation, the operator underlying this equation is not monotone, ruling out a TD-type learning algorithm. More recently, CVaR-constrained MDPs have been considered, though a variance-reduced CVaR estimation algorithm is still needed. In other words, there is no algorithm in an RL context that incorporates a variance reduction technique such as importance sampling and is provably convergent. Note that variance reduction is necessary, because CVaR is based on the tail of the distribution.

Going beyond the prediction problem, designing policy gradient algorithms is challenging for a risk-sensitive MDP, as it requires estimating the (policy) gradient of the risk measure considered, a nontrivial task in the RL context. For instance, in a discounted-cost MDP context, the policy gradient theorem variant that accounts for the variance of the cumulative discounted cost does not lend itself to an RL algorithm. An alternative is to apply a finite differences method such as simultaneous

perturbation stochastic approximation (SPSA), which amounts to treating the MDP as a black box, and such an approach would ignore the underlying Markovian structure of the problem, which is the case with the existing policy gradient algorithms for optimizing the CPT-value in any of the MDP settings.

Outline of the remaining sections

Section 2 provides an overview of MDPs and outlines the standard formulations for discounted-cost and average-cost MDPs and stochastic shortest path total-cost MDP problems. Examples and basic theoretical results are included for the benefit of readers less familiar with MDPs. Section 3 introduces all of the risk measures used in the monograph, namely exponential cost, variance, CVaR, coherent risk measures, chance constraints, and CPT. Section 4 provides an introduction to temporal difference learning and two gradient estimation techniques, namely simultaneous perturbation (stochastic approximation) and the likelihood ratio method. Section 5 presents two templates for risk-sensitive policy gradient algorithms, one for the setting where the risk measure is the objective, and the other for the setting where the risk measure is featured in the constraint. This chapter also presents a convergence analysis of the template algorithms for both settings. Section 6 develops policy gradient algorithms for four special cases of risk-sensitive MDPs for the constrained optimization problem posed in (1.1), with variance, CVaR, and a chance constraint used as the risk measure constraint. Section 7 develops policy gradient algorithms for three risk-sensitive MDP formulations in the unconstrained optimization setting of (1.2) with risk explicitly as the objective: exponential cost, CPT, and coherent risk measures. Finally, Section 8 provides concluding remarks and identifies some interesting future research directions.

A brief note on notation

Throughout the monograph, the functions J, G, and D may show one, two, or no arguments, depending on the context. Specifically, the two possible arguments would be θ, the policy parameter, as in (1.1) or (1.2),

or a state of the MDP (e.g., x_0, x, i, j), as described in Section 2. This is particularly relevant to Sections 5, 6, and 7. The same "convention" is used for other analogous counterparts such as the variance and squared versions of these quantities. On the other hand, dependence on an entire MDP policy μ is represented as subscript, e.g., J_μ. Gradients represented by ∇ are assumed to be with respect to θ unless otherwise indicated, e.g., ∇_λ denoting a gradient with respect to the Lagrange multiplier λ. Finally, all vectors will be assumed to be column vectors, and superscript "$\intercal$" will be used to denote the matrix/vector transpose operation.

Bibliographic remarks

MDPs have a long history dating back to the work of Richard E. Bellman. For a rigorous introduction, the reader is referred to the books by Puterman (1994) and Bertsekas (2007), and for reinforcement learning, the books by Bertsekas and Tsitsiklis (1996), Sutton and Barto (2018), and Szepesvári (2011). Material in this book drawn from our own research includes Prashanth and Ghavamzadeh (2013), Prashanth and Ghavamzadeh (2016), Prashanth (2014), Prashanth *et al.* (2016), and Gopalan *et al.* (2017). Cumulative prospect theory (CPT) was introduced by Tversky and Kahneman (1992) as a successor to prospect theory, which was one of the central contributions cited for Daniel Kahneman receiving the Nobel Memorial Prize in Economic Sciences in 2002.

References for the various risk measures include the following: mean-variance tradeoff (Markowitz, 1952), exponential utility (Arrow, 1971; Howard and Matheson, 1972), the percentile performance (Filar *et al.*, 1995), the use of chance constraints (Prekopa, 2003), stochastic dominance constraints (Dentcheva and Ruszczynski, 2003), value at risk (VaR), and conditional value-at-risk (CVaR) (Rockafellar and Uryasev, 2000; Ruszczyński, 2010; Shen *et al.*, 2013). The concept of a coherent risk measure was introduced by Artzner *et al.* (1999), see also Föllmer and Schied (2004), with the extension to multi-period settings treated in Riedel (2004), Ruszczyński and Shapiro (2006), Ruszczyński (2010), Cavus and Ruszczynski (2014), Tallec (2007), and Choi (2009).

The large body of literature utilizing the exponential utility formulation includes the classic formulation by Howard and Matheson (1972); related work includes Whittle (1990), Browne (1995), Fleming and McEneaney (1995), Hernández-Hernández and Marcus (1996), Marcus *et al.* (1997), Fernández-Gaucherand and Marcus (1997), Hernández-Hernández and Marcus (1999), Coraluppi and Marcus (1999a), Coraluppi and Marcus (1999b), Coraluppi and Marcus (2000), Borkar and Meyn (2002), and Bäuerle and Rieder (2014). For a survey of risk-sensitive RL under the exponential utility formulation, the reader is referred to Borkar (2010).

Another approach to risk/uncertainty is the robust optimization approach. In the setting of Markov decision processes, Iyengar (2005) is an early seminal work in this area, where a robust optimal policy is defined relative to uncertainty in the underlying transition probabilities. We do not pursue the robust approach in this monograph.

The existence of a Bellman equation for the variance of the return, where the underlying Bellman operator is not necessarily monotone, can be found in Sobel (1982). The result that finding a globally mean-variance optimal policy in a discounted-cost MDP is NP-hard can be found in Mannor and Tsitsiklis (2013). The use of variance of a quantity that measures the deviation of the single-stage cost from the average cost can be found in Filar *et al.* (1989). The result that solving an average-cost MDP under this notion of variance is NP-hard is shown in Filar *et al.* (1989).

Actor-critic methods investigated in RL are found in Barto *et al.* (1983) and Sutton (1984). Temporal difference (TD) learning can be found in Sutton (1988). More recently, CVaR-constrained MDPs have been considered in Borkar and Jain (2010), Prashanth (2014), and Tamar *et al.* (2014a), though a variance-reduced CVaR estimation algorithm is still needed.

The application of simultaneous perturbation stochastic approximation (SPSA) to policy gradient search for mean-variance optimization in discounted-cost MDPs is considered in Prashanth and Ghavamzadeh (2016) and for optimizing CPT-value in Prashanth *et al.* (2016).

Prospect theory (PT) was introduced in Kahneman and Tversky (1979), and cumulative prospect theory (CPT) in Tversky and Kahne-

man (1992), with experiments on humans reported in Starmer (2000) and Tversky and Kahneman (1992). More work adopting this approach includes Lin (2013), Lin and Marcus (2013b), Lin and Marcus (2013a), and Lin *et al.* (2018); see also Cavus and Ruszczynski (2014).

Variance as a risk measure in a discounted-cost and average-cost MDP, respectively, are based on Prashanth and Ghavamzadeh (2013) and Prashanth and Ghavamzadeh (2016). CVaR as a risk measure is based on Prashanth (2014). CPT as the risk measure is based on Prashanth *et al.* (2016) and Jie *et al.* (2018).

A sampling but nowhere near exhaustive list of other risk-sensitive RL work includes the following. In Tamar *et al.* (2012), variance as risk is considered in a stochastic shortest path context, and a policy gradient algorithm using the likelihood ratio method is provided. In Mihatsch and Neuneier (2002), a modified temporal differences algorithm is proposed and connected to the exponential utility approach. A general policy gradient algorithm that handles a class of risk measures that includes CVaR is presented in Tamar *et al.* (2015b). An early work that considers a constrained MDP setting similar to that in (1.1) is Borkar (2005), where the objective is average cost and the constraint is also an average-cost function different from the objective function. A modification of this formulation to a discounted-cost MDP, incorporating function approximation, was treated in Bhatnagar (2010). CVaR optimization in a constrained MDP setup was also explored in Borkar and Jain (2010), but the algorithm proposed there requires that the single-stage cost be separable. Optimization of risk measures that include CVaR in an unconstrained MDP setting using RL algorithms with function approximation can be found in Jiang and Powell (2017).

References

Abdellaoui, M. (2000). "Parameter-free elicitation of utility and probability weighting functions". *Management Science*. 46(11): 1497–1512.

Abounadi, J., D. P. Bertsekas, and V. S. Borkar. (2001). "Learning algorithms for Markov decision processes with average cost". *SIAM Journal on Control and Optimization*. 40(3): 681–698.

Abounadi, J., D. P. Bertsekas, and V. S. Borkar. (2002). "Stochastic approximation for nonexpansive maps: Application to Q-learning algorithms". *SIAM Journal on Control and Optimization*. 41(1): 1–22.

Acerbi, C. (2002). "Spectral measures of risk: A coherent representation of subjective risk aversion". *Journal of Banking & Finance*. 26(7): 1505–1518.

Aleksandrov, V., V. Sysoyev, and V. Shemeneva. (1968). "Stochastic optimization". *Engineering Cybernetics,* 5: 11–16.

Altman, E. (1999). *Constrained Markov Decision Processes*. Vol. 7. CRC Press.

Arapostathis, A., V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus. (1993). "Discrete-time controlled Markov processes with average cost criterion: A survey". *SIAM Journal on Control and Optimization*. 31: 282–344.

Arrow, K. J. (1971). *Essays in the Theory of Risk Bearing*. Chicago, IL: Markham.

Artzner, P., F. Delbaen, J. Eber, and D. Heath. (1999). "Coherent measures of risk". *Mathematical Finance*. 9(3): 203–228.

Asmussen, S. and P. Glynn. (2007). *Stochastic Simulation: Algorithms and Analysis*. New York: Springer.

Balbás, A., J. Garrido, and S. Mayoral. (2009). "Properties of distortion risk measures". *Methodology and Computing in Applied Probability*. 11(3): 385–399.

Barakat, A., P. Bianchi, W. Hachem, and S. Schechtman. (2021). "Stochastic optimization with momentum: Convergence, fluctuations, and traps avoidance". *Electronic Journal of Statistics*. 15(2): 3892–3947.

Barberis, N. C. (2013). "Thirty years of prospect theory in economics: A review and assessment". English. *Journal of Economic Perspectives*: 173–196.

Bardou, O., N. Frikha, and G. Pages. (2009). "Computing VaR and CVaR using stochastic approximation and adaptive unconstrained importance sampling". *Monte Carlo Methods and Applications*. 15(3): 173–210.

Bartlett, P. L. and J. Baxter. (2011). "Infinite-horizon policy-gradient estimation". *arXiv preprint arXiv:1106.0665*.

Barto, A., R. S. Sutton, and C. Anderson. (1983). "Neuron-like elements that can solve difficult learning control problems". *IEEE Transaction on Systems, Man and Cybernetics*. 13: 835–846.

Basu, A., T. Bhattacharyya, and V. S. Borkar. (2008). "A learning algorithm for risk-sensitive cost". *Mathematics of Operations Research*. 33(4): 880–898.

Bäuerle, N. and U. Rieder. (2014). "More risk-sensitive Markov decision processes". *Mathematics of Operations Research*. 39(1): 105–120.

Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press.

Bertsekas, D. P. (2007). *Dynamic Programming and Optimal Control, Vols. 1 & 2*. 3rd. Athena Scientific.

Bertsekas, D. P. (2012). *Dynamic Programming and Optimal Control, Vol. II, 4th edition*. Athena Scientific.

Bertsekas, D. P. and J. N. Tsitsiklis. (1996). *Neuro-Dynamic Programming*. Athena Scientific.

Bhandari, J., D. Russo, and R. Singal. (2018). "A Finite Time Analysis of Temporal Difference Learning With Linear Function Approximation". In: *Conference On Learning Theory*. 1691–1692.

Bhat, S. P. and L. A. Prashanth. (2019). "Concentration of risk measures: A Wasserstein distance approach". In: *Advances in Neural Information Processing Systems*. 11739–11748.

Bhatnagar, S. (2010). "An actor–critic algorithm with function approximation for discounted cost constrained Markov decision processes". *Systems & Control Letters*. 59(12): 760–766.

Bhatnagar, S., V. S. Borkar, and M. Akarapu. (2006). "A Simulation-Based Algorithm for Ergodic Control of Markov Chains Conditioned on Rare Events". *Journal of Machine Learning Research*. 7(70): 1937–1962.

Bhatnagar, S., H. L. Prasad, and L. Prashanth. (2013). *Stochastic Recursive Algorithms for Optimization*. Vol. 434. Springer.

Bhatnagar, S., R. Sutton, M. Ghavamzadeh, and M. Lee. (2009). "Natural actor-critic algorithms". *Automatica*. 45(11): 2471–2482.

Borkar, V. S. (2001). "A sensitivity formula for risk-sensitive cost and the actor–critic algorithm". *Systems & Control Letters*. 44(5): 339–346.

Borkar, V. S. (2002). "Q-learning for risk-sensitive control". *Mathematics of operations research*. 27(2): 294–311.

Borkar, V. S. (2005). "An actor-critic algorithm for constrained Markov decision processes". *Systems & Control Letters*. 54(3): 207–213.

Borkar, V. S. (2008). *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press.

Borkar, V. S. (2010). "Learning algorithms for risk-sensitive control". In: *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems–MTNS*. Vol. 5. No. 9.

Borkar, V. S. and R. Jain. (2010). "Risk-constrained Markov decision processes". In: *IEEE Conference on Decision and Control*. 2664–2669.

Borkar, V. S. and S. P. Meyn. (2000). "The ODE method for convergence of stochastic approximation and reinforcement learning". *SIAM Journal on Control and Optimization*. 38(2): 447–469.

Borkar, V. S. and S. P. Meyn. (2002). "Risk-sensitive optimal control for Markov decision processes with monotone cost". *Mathematics of Operations Research*. 27(1): 192–209.

Borkar, V. S. (1997). "Stochastic approximation with two time scales". *Systems & Control Letters*. 29(5): 291–294.

Bottou, L., F. E. Curtis, and J. Nocedal. (2018). "Optimization methods for large-scale machine learning". *Siam Review*. 60(2): 223–311.

Brandiere, O. and M. Duflo. (1996). "Les algorithmes stochastiques contournent-ils les pieges?" In: *Annales de l'IHP Probabilités et Statistiques*. Vol. 32. No. 3. 395–427.

Brown, D. B. (2007). "Large deviations bounds for estimating conditional value-at-risk". *Operations Research Letters*. 35(6): 722–730.

Browne, S. (1995). "Optimal Investment Policies for a Firm With a Random Risk Process: Exponential Utility and Minimizing the Probability of Ruin". *Mathematics of Operations Research*. 20(4): 937–958. DOI: 10.1287/moor.20.4.937.

Camerer, C. F. (1989). "An experimental test of several generalized utility theories". *Journal of Risk and Uncertainty*. 2(1): 61–104.

Camerer, C. F. (1992). "Recent tests of generalizations of expected utility theory". In: *Utility Theories: Measurements and Applications*. Springer. 207–251.

Camerer, C. F. and T.-H. Ho. (1994). "Violations of the betweenness axiom and nonlinearity in probability". *Journal of Risk and Uncertainty*. 8(2): 167–196.

Cavus, O. and A. Ruszczynski. (2014). "Risk-averse control of undiscounted transient Markov models". *SIAM Journal on Control and Optimization*. 52(6): 3935–3966.

Chang, H. S., M. C. Fu, J. Hu, and S. I. Marcus. (2007). *Simulation-based Algorithms for Markov Decision Processes*. Springer.

Chang, H. S., J. Hu, M. C. Fu, and S. I. Marcus. (2013). *Simulation-based Algorithms for Markov Decision Processes*. Springer.

Charnes, A., W. W. Cooper, and G. H. Symonds. (1958). "Cost horizons and certainty equivalents: An approach to stochastic programming of heating oil". *Management Science*. 4: 253–263.

Choi, S. (2009). "Risk-averse Newsvendor Models". *PhD thesis*. Rutgers.

Chow, Y., M. Ghavamzadeh, L. Janson, and M. Pavone. (2017). "Risk-constrained reinforcement learning with percentile risk criteria". *The Journal of Machine Learning Research*. 18(1): 6070–6120.

Conlisk, J. (1989). "Three variants on the Allais example". *The American Economic Review*: 392–407.

Coraluppi, S. P. and S. I. Marcus. (1999a). "Risk-Sensitive and Minimax Control of Discrete-Time, Finite-State Markov Decision Processes". *Automatica*. 35: 301–309.

Coraluppi, S. P. and S. I. Marcus. (1999b). "Risk-Sensitive, Minimax, and Mixed Risk Neutral/Minimax Control of Markov Decision Processes". In: *Stochastic Analysis, Control, Optimization and Applications: A Volume in Honor of W. H. Fleming*. Boston: Birkhauser. 21–40.

Coraluppi, S. P. and S. I. Marcus. (2000). "Mixed risk-neutral/minimax control of discrete-time, finite-state Markov decision processes". *IEEE Transactions on Automatic Control*. 45: 528–532.

Dalal, G., B. Szörényi, and G. Thoppe. (2020). "A tale of two-timescale reinforcement learning with the tightest finite-time bound". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 04. 3701–3708.

Dalal, G., B. Szörényi, G. Thoppe, and S. Mannor. (2018). "Finite sample analyses for TD(0) with function approximation". In: *Thirty-Second AAAI Conference on Artificial Intelligence*.

Dentcheva, D. and A. Ruszczynski. (2003). "Optimization with stochastic dominance constraints". *SIAM Journal on Optimization*. 14(2): 548–566.

Derman, C. (1970). *Finite State Markovian Decision Processes*. Academic Press.

Fernández-Gaucherand, E. and S. I. Marcus. (1997). "Risk-sensitive optimal control of hidden Markov models: Structural results". *IEEE Transactions on Automatic Control*. 42: 1418–1422.

Filar, J., L. Kallenberg, and H. Lee. (1989). "Variance-penalized Markov decision processes". *Mathematics of Operations Research*. 14(1): 147–161.

Filar, J., D. Krass, and K. Ross. (1995). "Percentile performance criteria for limiting average Markov decision processes". *IEEE Transaction of Automatic Control*. 40(1): 2–10.

Fleming, W. H. and W. M. McEneaney. (1995). "Risk-sensitive control on an infinite time horizon". *SIAM Journal on Control and Optimization*. 33(6): 1881–1915.

Föllmer, H. and A. Schied. (2002). "Convex measures of risk and trading constraints". *Finance and stochastics*. 6(4): 429–447.

Föllmer, H. and A. Schied. (2004). *Stochastic Finance: An Introduction in Discrete Time*. de Gruyter.

Föllmer, H. and A. Schied. (2016). *Stochastic finance*. de Gruyter.

Fu, M. C. (2006). "Gradient Estimation". In: *Handbooks in Operations Research and Management Science: Simulation*. Ed. by S. G. Henderson and B. L. Nelson. Elsevier. Chap. 19. 575–616.

Fu, M. C. (2015). "Stochastic Gradient Estimation". In: *Handbook on Simulation Optimization*. Ed. by M. C. Fu. Springer. Chap. 5.

Ge, R., F. Huang, C. Jin, and Y. Yuan. (2015). "Escaping from saddle points—online stochastic gradient for tensor decomposition". In: *Conference on Learning Theory*. PMLR. 797–842.

Ghadimi, S. and G. Lan. (2013). "Stochastic first-and zeroth-order methods for nonconvex stochastic programming". *SIAM Journal on Optimization*. 23(4): 2341–2368.

Glynn, P. W. (1987). "Likelilood ratio gradient estimation: an overview". In: *Proceedings of the 19th conference on Winter simulation*. ACM. 366–375.

Gonzalez, R. and G. Wu. (1999). "On the shape of the probability weighting function". *Cognitive psychology*. 38(1): 129–166.

Gopalan, A., L. A. Prashanth, M. C. Fu, and S. I. Marcus. (2017). "Weighted bandits or: How bandits learn distorted values that are not expected". In: *AAAI Conference on Artificial Intelligence*. 1941–1947.

Gosavi, A. (2003). *Simulation-Based Optimization: Parametric Optimization Techniques and Reinforcement Learning*. Kluwer.

Gower, R. M., M. Schmidt, F. Bach, and P. Richtárik. (2020). "Variance-reduced methods for machine learning". *Proceedings of the IEEE*. 108(11): 1968–1983.

Harless, D. W. (1992). "Predictions about indifference curves inside the unit triangle: A test of variants of expected utility theory". *Journal of Economic Behavior & Organization*. 18(3): 391–414.

Heidergott, B. and F. Vázquez-Abad. (2000). "Measure-valued differentiation for stochastic processes: the finite horizon case". *Tech. rep.* No. Report 2000-033. EURANDOM.

Hernández-Hernández, D. and S. I. Marcus. (1996). "Risk sensitive control of Markov processes in countable state space". *Systems & Control Letters*. 29(3): 147–155.

Hernández-Hernández, D. and S. I. Marcus. (1999). "Existence of risk sensitive optimal stationary policies for controlled Markov processes". *Applied Mathematics and Optimization*. 40: 273–285.

Howard, R. A. and J. E. Matheson. (1972). "Risk-sensitive Markov decision processes". *Management Science*. 18: 356–369.

Iyengar, G. N. (2005). "Robust dynamic programming". *Mathematics of Operations Research*. 30(2): 257–280. URL: http:https://doi.org/10.1287/moor.1040.0129.

Jiang, D. R. and W. B. Powell. (2017). "Risk-averse approximate dynamic programming with quantile-based risk measures". *Mathematics of Operations Research*. 43(2): 554–579.

Jie, C., L. A. Prashanth, M. C. Fu, S. I. Marcus, and C. Szepesvári. (2018). "Stochastic optimization in a cumulative prospect theory framework". *IEEE Transactions on Automatic Control*. 63(9): 2867–2882.

Jin, C., R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. (2017). "How to escape saddle points efficiently". In: *International Conference on Machine Learning*. PMLR. 1724–1732.

Kahneman, D. and A. Tversky. (1979). "Prospect theory: An analysis of decision under risk". *Econometrica*: 263–291.

Kiefer, J. and J. Wolfowitz. (1952). "Stochastic estimation of the maximum of a regression function". *Annals of Mathematical Statistics*. 23: 462–266.

Kolla, R. K., L. A. Prashanth, S. P. Bhat, and K. P. Jagannathan. (2019). "Concentration bounds for empirical conditional value-at-risk: The unbounded case". *Operations Research Letters*. 47(1): 16–20.

Konda, V. R. and V. S. Borkar. (1999). "Actor-Critic–Type Learning Algorithms for Markov Decision Processes". *SIAM Journal on control and Optimization*. 38(1): 94–123.

Konda, V. R. and J. N. Tsitsiklis. (2004). "Convergence rate of linear two-time-scale stochastic approximation". *The Annals of Applied Probability*. 14(2): 796–819.

Kushner, H. and D. Clark. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag. 191–196.

Lin, K. (2013). "Stochastic Systems with Cumulative Prospect Theory". *PhD thesis*. University of Maryland, College Park.

Lin, K., C. Jie, and S. I. Marcus. (2018). "Probabilistically distorted risk-sensitive infinite-horizon dynamic programming". *Automatica*.

Lin, K. and S. I. Marcus. (2013a). "Cumulative weighting optimization: The discrete case". In: *Proceedings of the 2013 Winter Simulation Conference*. Washington, D.C.: Institute of Electrical and Electronic Engineers, Inc.

Lin, K. and S. I. Marcus. (2013b). "Dynamic programming with non-convex risk-sensitive measures". In: *Proceedings of the 2013 American Control Conference*. Washington, D.C.

Mannor, S. and J. N. Tsitsiklis. (2013). "Algorithmic aspects of mean–variance optimization in Markov decision processes". *European Journal of Operational Research*. 231(3): 645–653.

Marbach, P. and J. N. Tsitsiklis. (2001). "Simulation-based optimization of Markov reward processes". *IEEE Transactions on Automatic Control*. 46(2): 191–209.

Marcus, S. I., E. Fernández-Gaucherand, S. C. D. Hernández-Hernández, and P. Fard. (1997). "Risk sensitive Markov decision processes". In: *Systems and Control in the Twenty-First Century*. Ed. by C. I. Byrnes. Boston: Birkhauser. 263–279.

Markowitz, H. (1952). "Portfolio selection". *The Journal of Finance*. 7(1): 77–91.

Mas-Colell, A., M. Whinston, and J. Green. (1995). *Microeconomic theory*. Oxford University Press.

Mihatsch, O. and R. Neuneier. (2002). "Risk-sensitive reinforcement learning". *Machine Learning*. 49(2): 267–290.

Miller, L. B. and H. Wagner. (1965). "Chance-constrained programming with joint constraints". *Operations Research*. 13: 199–213.

Moharrami, M., Y. Murthy, A. Roy, and R. Srikant. (2022). "A Policy Gradient Algorithm for the Risk-Sensitive Exponential Cost MDP". arXiv: 2202.04157 [eess.SY].

Mokkadem, A. and M. Pelletier. (2006). "Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms". *The Annals of Applied Probability*. 16(3): 1671–1702.

Nemirovski, A. and A. Shapiro. (2007). "Convex Approximations of Chance Constrained Programs". *SIAM Journal on Optimization*. 17(4): 969–996.

Papini, M., D. Binaghi, G. Canonaco, M. Pirotta, and M. Restelli. (2018). "Stochastic variance-reduced policy gradient". In: *International Conference on Machine Learning*. Vol. 80. *Proceedings of Machine Learning Research*. PMLR. 4026–4035.

Pemantle, R. (1990). "Nonconvergence to unstable points in urn models and stochastic approximations". *The Annals of Probability*. 18(2): 698–712.

Pflug, G. C. (1989). "Sampling derivatives of probabilities". *Computing*. 42: 315–328.

Pflug, G. C. (1996). *Optimization of Stochastic Models*. Kluwer Academic.

Polyak, B. T. and A. B. Juditsky. (1992). "Acceleration of stochastic approximation by averaging". *SIAM Journal on Control and Optimization*. 30(4): 838–855.

Prashanth, L. A. (2014). "Policy gradients for CVaR-constrained MDPs". In: *Algorithmic Learning Theory (ALT)*. 155–169.

Prashanth, L. A., S. Bhatnagar, M. C. Fu, and S. I. Marcus. (2018). "Adaptive system optimization using random directions stochastic approximation". *IEEE Transactions on Automatic Control*. 62(5): 2223–2238.

Prashanth, L. A. and M. Ghavamzadeh. (2013). "Actor-critic algorithms for risk-sensitive MDPs". In: *Advances in Neural Information Processing Systems (NIPS)*. 252–260.

Prashanth, L. A. and M. Ghavamzadeh. (2016). "Variance-constrained actor-critic algorithms for discounted and average reward MDPs". *Machine Learning*. 105(3): 367–417.

Prashanth, L. A., K. Jagannathan, and R. K. Kolla. (2020). "Concentration bounds for CVaR estimation: The cases of light-tailed and heavy-tailed distributions". In: *International Conference on Machine Learning*. Vol. 119. PMLR. 5577–5586.

Prashanth, L. A., C. Jie, M. C. Fu, S. I. Marcus, and C. Szepesvári. (2016). "Cumulative prospect theory meets reinforcement learning: prediction and control". In: *International Conference on Machine Learning*. 1406–1415.

Prashanth, L. A., N. Korda, and R. Munos. (2021). "Concentration bounds for temporal difference learning with linear function approximation: The case of batch data and uniform sampling". *Mach. Learn.* 110(3): 559–618.

Prekopa, A. (2003). "Probabilistic programming". In: *Stochastic Programming*. Ed. by A. Ruszczyński and Shapiro. Elsevier, Amsterdam.

Prékopa, A. (1970). "On probabilistic constrained programming". In: *Proceedings of the Princeton Symposium on Mathematical Programming*. Princeton University Press, Princeton, NJ. 113–138.

Prelec, D. (1998). "The probability weighting function". *Econometrica*: 497–527.

Puterman, M. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Reiman, M. and A. Weiss. (1989). "Sensitivity analysis for simulations via likelihood ratios". *Operations Research*. 37: 830–844.

Riedel, F. (2004). "Dynamic coherent risk measures". *Stochastic Processes and Their Applications*. 112: 185–200.

Robbins, H. and S. Monro. (1951). "A stochastic approximation method". *The Annals of Mathematical Statistics*: 400–407.

Rockafellar, R. T. and S. Uryasev. (2000). "Optimization of conditional value-at-risk". *Journal of Risk*. 2: 21–42.

Ross, S. M. (1983). *Introduction to Stochastic Dynamic Programming*. Academic Press.

Rubinstein, R. Y. (1989). "Sensitivity analysis of computer simulation models via the score efficient". *Operations Research*. 37: 72–81.

Ruppert, D. (1991). "Stochastic approximation". *Handbook of Sequential Analysis*: 503–529.

Ruszczyński, A. (2010). "Risk-averse dynamic programming for Markov decision processes". *Mathematical Programming*. 125: 235–261.

Ruszczyński, A. and A. Shapiro. (2006). "Conditional risk mappings". *Mathematics of Operations Research*. 31(3): 544–561.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*. Vol. 162. John Wiley & Sons.

Shapiro, A., D. Dentcheva, and A. Ruszczyński. (2014). *Lectures on Stochastic Programming: Modeling and Theory*. SIAM.

Shen, Y., W. Stannat, and K. Obermayer. (2013). "Risk-sensitive Markov control processes". *SIAM Journal on Control and Optimization*. 51(5): 3652–3672.

Shen, Z., A. Ribeiro, H. Hassani, H. Qian, and C. Mi. (2019). "Hessian aided policy gradient". In: *International Conference on Machine Learning*. PMLR. 5729–5738.

Sion, M. (1958). "On general minimax theorems". *Pacific J. Math*. 8(1): 171–176.

Sobel, M. (1982). "The variance of discounted Markov decision processes". *Journal of Applied Probability*: 794–802.

Sopher, B. and G. Gigliotti. (1993). "A test of generalized expected utility theory". *Theory and Decision*. 35(1): 75–106.

Spall, J. C. (1992). "Multivariate stochastic approximation using simultaneous perturbation gradient approximation". *IEEE Transactions on Automatic Control*. 37: 332–341.

Srikant, R. and L. Ying. (2019). "Finite-Time Error Bounds For Linear Stochastic Approximation and TD Learning". In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Vol. 99. *Proceedings of Machine Learning Research*. PMLR. 2803–2830.

Starmer, C. (2000). "Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk". *Journal of Economic Literature*: 332–382.

Sutton, R. S. (1984). "Temporal Credit Assignment in Reinforcement Learning". *PhD thesis*. University of Massachusetts Amherst.

Sutton, R. S. (1988). "Learning to predict by the methods of temporal differences". *Machine Learning*. 3(1): 9–44.

Sutton, R. S., D. A. McAllester, S. P. Singh, and Y. Mansour. (1999). "Policy gradient methods for reinforcement learning with function approximation." In: *NIPS*. Vol. 99. 1057–1063.

Sutton, R. S. and A. G. Barto. (2018). *Reinforcement Learning: An Introduction*. 2nd ed. Cambridge, MA: MIT Press.

Szepesvári, C. (2011). "Reinforcement learning algorithms for MDPs". *Wiley Encyclopedia of Operations Research and Management Science*.

Tallec, Y. L. (2007). "Robust, Risk-sensitive, and Data-driven Control of Markov Decision Processes". *PhD thesis*. MIT.

Tamar, A., D. D. Castro, and S. Mannor. (2012). "Policy gradients with variance related risk criteria". In: *Proceedings of the Twenty-Ninth International Conference on Machine Learning*. 387–396.

Tamar, A., Y. Chow, M. Ghavamzadeh, and S. Mannor. (2015a). "Policy gradient for coherent risk measures". In: *Advances in Neural Information Processing Systems*. Vol. 28. 1468–1476.

Tamar, A., Y. Chow, M. Ghavamzadeh, and S. Mannor. (2015b). "Policy gradient for coherent risk measures". *CoRR*. abs/1502.03919. arXiv: 1502.03919.

Tamar, A., D. Di Castro, and S. Mannor. (2013). "Temporal difference methods for the variance of the reward to go". In: *International Conference on Machine Learning*. 495–503.

Tamar, A., Y. Glassner, and S. Mannor. (2014a). "Optimizing the CVaR via sampling". *arXiv preprint arXiv:1404.3862*.

Tamar, A., Y. Glassner, and S. Mannor. (2014b). "Policy gradients beyond expectations: Conditional Value-at-Risk". *arXiv preprint arXiv:1404.3862*.

Thomas, P. and E. Learned-Miller. (2019). "Concentration Inequalities for Conditional Value at Risk". In: *International Conference on Machine Learning*. 6225–6233.

Tsitsiklis, J. N. and B. Van Roy. (1997). "An analysis of temporal-difference learning with function approximation". *IEEE Transactions on Automatic Control*. 42(5): 674–690.

Tsitsiklis, J. N. and B. Van Roy. (1999). "Average cost temporal-difference learning". *Automatica*. 35(11): 1799–1808.

Tversky, A. and D. Kahneman. (1992). "Advances in prospect theory: Cumulative representation of uncertainty". *Journal of Risk and Uncertainty*. 5(4): 297–323.

Wang, Y. and F. Gao. (2010). "Deviation inequalities for an estimator of the conditional value-at-risk". *Operations Research Letters*. 38(3): 236–239.

Whittle, P. (1990). *Risk-sensitive Optimal Control*. *Wiley-Interscience series in systems and optimization*. Wiley. ISBN: 9780471926221.

Zhang, K., A. Koppel, H. Zhu, and T. Basar. (2020). "Global convergence of policy gradient methods to (almost) locally optimal policies". *SIAM Journal on Control and Optimization*. 58(6): 3586–3612.