Risk-Constrained Reinforcement Learning with Percentile Risk Criteria

Yinlam Chow YCHOW@STANFORD.EDU
Institute for Computational & Mathematical Engineering
Stanford University
Stanford, CA 94305, USA

Mohammad Ghavamzadeh GHAVAMZA@ADOBE.COM
Adobe Research & INRIA Lille
San Jose, CA 95110, USA

Lucas Janson LJANSON@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA 94305, USA

Marco Pavone PAVONE@STANFORD.EDU
Aeronautics and Astronautics
Stanford University
Stanford, CA 94305, USA

Editor:

Abstract

In many sequential decision-making problems one is interested in minimizing an expected cumulative cost while taking into account *risk*, i.e., increased awareness of events of small probability and high consequences. Accordingly, the objective of this paper is to present efficient reinforcement learning algorithms for risk-constrained Markov decision processes (MDPs), where risk is represented via a chance constraint or a constraint on the conditional value-at-risk (CVaR) of the cumulative cost. We collectively refer to such problems as percentile risk-constrained MDPs. Specifically, we first derive a formula for computing the gradient of the Lagrangian function for percentile risk-constrained MDPs. Then, we devise policy gradient and actor-critic algorithms that (1) estimate such gradient, (2) update the policy in the descent direction, and (3) update the Lagrange multiplier in the ascent direction. For these algorithms we prove convergence to locally optimal policies. Finally, we demonstrate the effectiveness of our algorithms in an optimal stopping problem and an online marketing application.

Keywords: Markov Decision Process, Reinforcement Learning, Conditional Value-at-Risk, Chance-Constrained Optimization, Policy Gradient Algorithms, Actor-Critic Algorithms

1. Introduction

The most widely-adopted optimization criterion for Markov decision processes (MDPs) is represented by the *risk-neutral* expectation of a cumulative cost. However, in many applications one is interested in taking into account risk, i.e., increased awareness of events of small probability and high consequences. Accordingly, in *risk-sensitive* MDPs the objective is to minimize a risk-sensitive criterion such as the expected exponential utility, a variance-related measure, or percentile performance. There are several risk metrics available in the literature, and constructing a "good" risk

criterion in a manner that is both conceptually meaningful and computationally tractable remains a topic of current research.

Risk-Sensitive MDPs: One of the earliest risk metrics used for risk-sensitive MDPs is the exponential risk metric $(1/\gamma)\mathbb{E}\big[\exp(\gamma Z)\big]$, where Z represents the cumulative cost for a sequence of decisions (Howard and Matheson, 1972). In this setting, the degree of risk-aversion is controlled by the parameter γ, whose selection, however, is often challenging. This motivated the study of several different approaches. In Collins (1997), the authors considered the maximization of a strictly concave functional of the distribution of the terminal state. In Wu and Lin (1999); Boda et al. (2004); Filar et al. (1995), risk-sensitive MDPs are cast as the problem of maximizing percentile performance. Variance-related risk metrics are considered, e.g., in Sobel (1982); Filar et al. (1989). Other mean, variance, and probabilistic criteria for risk-sensitive MDPs are discussed in the survey (White, 1988).

Numerous alternative risk metrics have recently been proposed in the literature, usually with the goal of providing an "intuitive" notion of risk and/or to ensure computational tractability. *Value-at-risk* (VaR) and *conditional value-at-risk* (CVaR) represent two promising such alternatives. They both aim at quantifying costs that might be encountered in the tail of a cost distribution, but in different ways. Specifically, for continuous cost distributions, VaR_α measures risk as the maximum cost that might be incurred with respect to a given confidence level α. This risk metric is particularly useful when there is a well-defined failure state, e.g., a state that leads a robot to collide with an obstacle. A VaR_α constraint is often referred to as a chance (probability) constraint, especially in the engineering literature, and we will use this terminology in the remainder of the paper. In contrast, CVaR_α measures risk as the expected cost given that such cost is greater than or equal to VaR_α, and provides a number of theoretical and computational advantages. CVaR optimization was first developed by Rockafellar and Uryasev (Rockafellar and Uryasev, 2002, 2000) and its numerical effectiveness has been demonstrated in several portfolio optimization and option hedging problems. Risk-sensitive MDPs with a conditional value at risk metric were considered in Boda and Filar (2006); Ott (2010); Bäuerle and Ott (2011), and a mean-average-value-at-risk problem has been solved in Bäuerle and Mundt (2009) for minimizing risk in financial markets.

The aforementioned works focus on the derivation of exact solutions, and the ensuing algorithms are only applicable to relatively small problems. This has recently motivated the application of reinforcement learning (RL) methods to risk-sensitive MDPs. We will refer to such problems as risk-sensitive RL.

Risk-Sensitive RL: To address large-scale problems, it is natural to apply reinforcement learning (RL) techniques to risk-sensitive MDPs. Reinforcement learning (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998) can be viewed as a class of sampling-based methods for solving MDPs. Popular reinforcement learning techniques include policy gradient (Williams, 1992; Marbach, 1998; Baxter and Bartlett, 2001) and actor-critic methods (Sutton et al., 2000; Konda and Tsitsiklis, 2000; Peters et al., 2005; Borkar, 2005; Bhatnagar et al., 2009; Bhatnagar and Lakshmanan, 2012), whereby policies are parameterized in terms of a parameter vector and policy search is performed via gradient flow approaches. One effective way to estimate gradients in RL problems is by simultaneous perturbation stochastic approximation (SPSA) (Spall, 1992). Risk-sensitive RL with expected exponential utility has been considered in Borkar (2001, 2002). More recently, the works in Tamar et al. (2012); Prashanth and Ghavamzadeh (2013) present RL algorithms for several variance-related risk measures, the works in Morimura et al. (2010); Tamar et al. (2015); Petrik and Subramanian (2012) consider CVaR-based formulations, and the works in Tallec (2007); Shapiro et al. (2013) consider nested CVaR-based formulations.

Risk-Constrained RL and Paper Contributions: Despite the rather large literature on risk-sensitive MDPs and RL, *risk-constrained* formulations have largely gone unaddressed, with only a few exceptions, e.g., Chow and Pavone (2013); Borkar and Jain (2014). Yet constrained formulations naturally arise in several domains, including engineering, finance, and logistics, and provide a principled approach to address multi-objective problems. The objective of this paper is to fill this gap, by devising policy gradient and actor-critic algorithms for risk-constrained MDPs where risk is represented via a constraint on the conditional value-at-risk (CVaR) of the cumulative cost, or as a chance constraint. Specifically, the contribution of this paper is fourfold.

1. We formulate two risk-constrained MDP problems. The first one involves a CVaR constraint and the second one involves a chance constraint. For the CVaR-constrained optimization problem, we consider both discrete and continuous cost distributions. By re-writing the problems using a Lagrangian formulation, we derive for both problems a Bellman optimality condition with respect to an augmented MDP.

2. We devise a trajectory-based policy gradient algorithm for both CVaR-constrained and chance-constrained MDPs. The key novelty of this algorithm lies in an unbiased gradient estimation procedure under Monte Carlo sampling. Using an ordinary differential equation (ODE) approach, we establish convergence of the algorithm to locally optimal policies.

3. Using the aforementioned Bellman optimality condition, we derive several actor-critic algorithms to optimize policy and value function approximation parameters in an online fashion. As for the trajectory-based policy gradient algorithm, we show that the proposed actor-critic algorithms converge to locally optimal solutions.

4. We demonstrate the effectiveness of our algorithms in an optimal stopping problem as well as in a realistic personalized ad recommendation problem (see Derfer et al. 2007 for more details). For the latter problem, we empirically show that our CVaR-constrained RL algorithms successfully guarantee that the worst-case revenue is lower-bounded by the pre-specified company yearly target.

The rest of the paper is structured as follows. In Section 2 we introduce our notation and rigorously state the problem we wish to address, namely risk-constrained RL. The next two sections provide various RL methods to approximately compute (locally) optimal policies for CVaR constrained MDPs. A trajectory-based policy gradient algorithm is presented in Section 3 and its convergence analysis is provided in Appendix A (Appendix A.1 provides the gradient estimates of the CVaR parameter, the policy parameter, and the Lagrange multiplier, and Appendix A.2 gives their convergence proofs). Actor-critic algorithms are presented in Section 4 and their convergence analysis is provided in Appendix B (Appendix B.1 derives the gradient of the Lagrange multiplier as a function of the state-action value function, Appendix B.2.1 analyzes the convergence of the critic, and Appendix B.2.2 provides the multi-timescale convergence results of the CVaR parameter, the policy parameter, and the Lagrange multiplier). Section 5 generalizes the above policy gradient and actor-critic methods to the chance-constrained case. Empirical evaluation of our algorithms is the subject of Section 6. Finally, we conclude the paper in Section 7, where we also provide directions for future work.

This paper generalizes earlier results by the authors presented in Chow and Ghavamzadeh (2014).

2. Preliminaries

We begin by defining some notation that is used throughout the paper, as well as defining the problem addressed herein and stating some basic assumptions.

2.1 Notation

We consider decision-making problems modeled as a finite MDP (an MDP with finite state and action spaces). A finite MDP is a tuple $(\mathcal{X}, \mathcal{A}, C, D, P, P_0)$ where $\mathcal{X} = \{1, \ldots, n, x_{\text{Tar}}\}$ and $\mathcal{A} = \{1, \ldots, m\}$ are the state and action spaces, x_{Tar} is a recurrent target state, and for a state x and an action a, $C(x, a)$ is a cost function with $|C(x, a)| \leq C_{\max}$, $D(x, a)$ is a constraint cost function with $|D(x, a)| \leq D_{\max}$ [1], $P(\cdot|x, a)$ is the transition probability distribution, and $P_0(\cdot)$ is the initial state distribution. For simplicity, in this paper we assume $P_0 = \mathbf{1}\{x = x^0\}$ for some given initial state $x^0 \in \{1, \ldots, n\}$. Generalizations to non-atomic initial state distributions are straightforward, for which the details are omitted for the sake of brevity. A *stationary policy* $\mu(\cdot|x)$ for an MDP is a probability distribution over actions, conditioned on the current state. In policy gradient methods, such policies are parameterized by a κ-dimensional vector θ, so the space of policies can be written as $\{\mu(\cdot|x; \theta), x \in \mathcal{X}, \theta \in \Theta \subseteq R^\kappa\}$. Since in this setting a policy μ is uniquely defined by its parameter vector θ, policy-dependent functions can be written as a function of μ or θ, and we use μ and θ interchangeably in the paper.

Given a fixed $\gamma \in (0, 1)$, we denote by $d^\mu_\gamma(x|x^0) = (1-\gamma)\sum_{k=0}^\infty \gamma^k \mathbb{P}(x_k = x|x_0 = x^0; \mu)$ and $\pi^\mu_\gamma(x, a|x^0) = d^\mu_\gamma(x|x^0)\mu(a|x)$, the γ-discounted occupation measure of state x and state-action pair (x, a) under policy μ, respectively. This occupation measure is a γ-discounted probability distribution for visiting each state and action pair, and it plays an important role in sampling states and actions from the real system in policy gradient and actor-critic algorithms, and in guaranteeing their convergence. Because the state and action spaces are finite, Theorem 3.1 in Altman (1999) shows that the occupation measure $d^\mu_\gamma(x|x^0)$ is a well-defined probability distribution. On the other hand, when $\gamma = 1$ the occupation measure of state x and state-action pair (x, a) under policy μ are respectively defined by $d^\mu(x|x^0) = \sum_{t=0}^\infty \mathbb{P}(x_t = x|x^0; \mu)$ and $\pi^\mu(x, a|x^0) = d^\mu(x|x^0)\mu(a|x)$. In this case the occupation measures characterize the total sums of visiting probabilities (although they are not in general probability distributions themselves) of state x and state-action pair (x, a). To study the well-posedness of the occupation measure, we define the following notion of a transient MDP.

Definition 1 *Define $\mathcal{X}' = \mathcal{X} \setminus \{x_{Tar}\} = \{1, \ldots, n\}$ as a state space of transient states. An MDP is said to be transient if,*

1. $\sum_{k=0}^\infty \mathbb{P}(x_k = x|x^0, \mu) < \infty$ *for every $x \in \mathcal{X}'$ and every stationary policy μ,*
2. $P(x_{Tar}|x_{Tar}, a) = 1$ *for every admissible control action $a \in \mathcal{A}$.*

Furthermore let $T_{\mu,x}$ be the first-hitting time of the target state x_{Tar} from an arbitrary initial state $x \in \mathcal{X}$ in the Markov chain induced by transition probability $P(\cdot|x, a)$ and policy μ. Although transience implies the first-hitting time is square integrable and finite almost surely, we will make the stronger assumption (which implies transience) on the uniform boundedness of the first-hitting time.

1. Without loss of generality, we set the cost function $C(x, a)$ and constraint cost function $D(x, a)$ to zero when $x = x_{\text{Tar}}$.

Assumption 2 *The first-hitting time $T_{\mu,x}$ is bounded almost surely over all stationary policies μ and all initial states $x \in \mathcal{X}$. We will refer to this upper bound as T, i.e., $T_{\mu,x} \leq T$ almost surely.*

The above assumption can be justified by the fact that sample trajectories collected in most reinforcement learning algorithms (including policy gradient and actor-critic methods) consist of a finite stopping time (also known as a time-out). If nothing else, such a bound ensures that the computation time is not unbounded. Note that although a bounded stopping time would seem to conflict with the time-stationarity of the transition probabilities, this can be resolved by augmenting the state space with a time-counter state, analogous to the arguments given in Section 4.7 in Bertsekas (1995).

Finally, we define the constraint and cost functions. Let Z be a finite-mean ($\mathbb{E}[|Z|] < \infty$) random variable representing cost, with the cumulative distribution function $F_Z(z) = \mathbb{P}(Z \leq z)$ (e.g., one may think of Z as the total cost of an investment strategy μ). We define the *value-at-risk* at confidence level $\alpha \in (0, 1)$ as

$$\text{VaR}_\alpha(Z) = \min\big\{z \mid F_Z(z) \geq \alpha\big\}.$$

Here the minimum is attained because F_Z is non-decreasing and right-continuous in z. When F_Z is continuous and strictly increasing, $\text{VaR}_\alpha(Z)$ is the unique z satisfying $F_Z(z) = \alpha$. As mentioned, we refer to a constraint on the VaR as a chance constraint.

Although VaR is a popular risk measure, it is not a *coherent* risk measure (Artzner et al., 1999) and does not quantify the costs that might be suffered beyond its value in the α-tail of the distribution (Rockafellar and Uryasev, 2000), Rockafellar and Uryasev (2002). In many *financial applications* such as portfolio optimization where the probability of undesirable events could be small but the cost incurred could still be significant, besides describing risk as the probability of incurring costs, it will be more interesting to study the cost in the tail of the risk distribution. In this case, an alternative measure that addresses most of the VaR's shortcomings is the *conditional value-at-risk*, defined as (Rockafellar and Uryasev, 2000)

$$\text{CVaR}_\alpha(Z) := \min_{\nu \in \mathbb{R}} \Big\{\nu + \frac{1}{1-\alpha}\mathbb{E}\big[(Z-\nu)^+\big]\Big\}, \tag{1}$$

where $(x)^+ = \max(x, 0)$ represents the positive part of x. Although this definition is somewhat opaque, CVaR can be thought of as the average of the worst-case α-fraction of losses. Define $H_\alpha(Z, \nu) := \nu + \frac{1}{1-\alpha}\mathbb{E}\big[(Z-\nu)^+\big]$, so that $\text{CVaR}_\alpha(Z) = \min_{\nu \in \mathbb{R}} H_\alpha(Z, \nu)$.

We define the parameter $\gamma \in (0, 1]$ as the *discounting factor* for the cost and constraint cost functions. When $\gamma < 1$, we are aiming to solve the MDP problem with more focus on optimizing current costs over future costs. For a policy μ, we define the cost of a state x (state-action pair (x, a)) as the sum of (discounted) costs encountered by the decision-maker when it starts at state x (state-action pair (x, a)) and then follows policy μ, i.e.,

$$\mathcal{C}^\theta(x) = \sum_{k=0}^{T-1} \gamma^k C(x_k, a_k) \mid x_0 = x,\ \mu(\cdot|\cdot, \theta), \qquad \mathcal{D}^\theta(x) = \sum_{k=0}^{T-1} \gamma^k D(x_k, a_k) \mid x_0 = x,\ \mu(\cdot|\cdot, \theta),$$

and

$$\mathcal{C}^\theta(x, a) = \sum_{k=0}^{T-1} \gamma^k C(x_k, a_k) \mid x_0 = x,\ a_0 = a,\ \mu(\cdot|\cdot, \theta),$$

$$\mathcal{D}^\theta(x, a) = \sum_{k=0}^{T-1} \gamma^k D(x_k, a_k) \mid x_0 = x,\ a_0 = a,\ \mu(\cdot|\cdot, \theta).$$

The expected values of the random variables $\mathcal{C}^\theta(x)$ and $\mathcal{C}^\theta(x,a)$ are known as the value and action-value functions of policy μ, and are denoted by

$$V^\theta(x) = \mathbb{E}\big[\mathcal{C}^\theta(x)\big], \qquad\qquad Q^\theta(x,a) = \mathbb{E}\big[\mathcal{C}^\theta(x,a)\big].$$

2.2 Problem Statement

The goal for standard discounted MDPs is to find an optimal policy that solves

$$\theta^* = \operatorname*{argmin}_\theta V^\theta(x^0).$$

For *CVaR-constrained* optimization in MDPs, we consider the discounted cost optimization problem with $\gamma \in (0,1)$, i.e., for a given confidence level $\alpha \in (0,1)$ and cost tolerance $\beta \in \mathbb{R}$,

$$\min_\theta V^\theta(x^0) \qquad \text{subject to} \qquad \text{CVaR}_\alpha\big(\mathcal{D}^\theta(x^0)\big) \leq \beta. \tag{2}$$

Using the definition of $H_\alpha(Z,\nu)$, one can reformulate (2) as:

$$\min_{\theta,\nu} V^\theta(x^0) \qquad \text{subject to} \qquad H_\alpha\big(\mathcal{D}^\theta(x^0),\nu\big) \leq \beta. \tag{3}$$

It is shown in Rockafellar and Uryasev (2000) and Rockafellar and Uryasev (2002) that the optimal ν actually equals VaR_α, so we refer to this parameter as the VaR parameter. Here we choose to analyze the discounted-cost CVaR-constrained optimization problem, i.e., with $\gamma \in (0,1)$, as in many financial and marketing applications where CVaR constraints are used, it is more intuitive to put more emphasis on current costs rather than on future costs. The analysis can be easily generalized for the case where $\gamma = 1$.

For *chance-constrained* optimization in MDPs, we consider the stopping cost optimization problem with $\gamma = 1$, i.e., for a given confidence level $\beta \in (0,1)$ and cost tolerance $\alpha \in \mathbb{R}$,

$$\min_\theta V^\theta(x^0) \qquad \text{subject to} \qquad \mathbb{P}\big(\mathcal{D}^\theta(x^0) \geq \alpha\big) \leq \beta. \tag{4}$$

Here we choose $\gamma = 1$ because in many engineering applications, where chance constraints are used to ensure overall safety, there is no notion of discounting since future threats are often as important as the current one. Similarly, the analysis can be easily extended to the case where $\gamma \in (0,1)$.

There are a number of mild technical and notational assumptions which we will make throughout the paper, so we state them here:

Assumption 3 (Differentiability) *For any state-action pair (x,a), $\mu(a|x;\theta)$ is continuously differentiable in θ and $\nabla_\theta \mu(a|x;\theta)$ is a Lipschitz function in θ for every $a \in \mathcal{A}$ and $x \in \mathcal{X}$.*[2]

Assumption 4 (Strict Feasibility) *There exists a transient policy $\mu(\cdot|x;\theta)$ such that $H_\alpha\big(\mathcal{D}^\theta(x^0),\nu\big) < \beta$ in the CVaR-constrained optimization problem, and $P\big(\mathcal{D}^\theta(x^0) \geq \alpha\big) < \beta$ in the chance-constrained problem.*

2. In actor-critic algorithms, the assumption on continuous differentiability holds for the augmented state Markovian policies $\mu(a|x,s;\theta)$.

Note that Assumption 3 imposes smoothness on the optimal policy. Assumption 4 guarantees the existence of a locally optimal policy for the CVaR-constrained optimization problem via the Lagrangian analysis introduced in the next subsection.

In the remainder of the paper we first focus on studying stochastic approximation algorithms for the CVaR-constrained optimization problem (Sections 3 and 4) and then adapt the results to the chance-constrained optimization problem in Section 5. Our solution approach relies on a Lagrangian relaxation procedure, which is discussed next.

2.3 Lagrangian Approach and Reformulation

To solve (3), we employ a Lagrangian relaxation procedure (Bertsekas, 1999), which leads to the unconstrained problem:

$$\max_{\lambda \geq 0} \min_{\theta,\nu} \Big(L(\nu, \theta, \lambda) := V^{\theta}(x^0) + \lambda \Big(H_{\alpha}\big(\mathcal{D}^{\theta}(x^0), \nu\big) - \beta \Big) \Big), \tag{5}$$

where λ is the Lagrange multiplier. Notice that $L(\nu, \theta, \lambda)$ is a linear function in λ and $H_{\alpha}\big(\mathcal{D}^{\theta}(x^0), \nu\big)$ is a continuous function in ν. Corollary 4 in (Vilkov, 1986) implies the existence of a locally optimal policy θ^* for the CVaR-constrained optimization problem, which corresponds to the existence of the local saddle point $(\nu^*, \theta^*, \lambda^*)$ for the minimax optimization problem $\max_{\lambda \geq 0} \min_{\theta,\nu} L(\nu, \theta, \lambda)$, defined as follows.

Definition 5 *A local saddle point of $L(\nu, \theta, \lambda)$ is a point $(\nu^*, \theta^*, \lambda^*)$ such that for some $r > 0$, $\forall (\theta, \nu) \in \Theta \times \left[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma} \right] \cap \mathcal{B}_{(\theta^*,\nu^*)}(r)$ and $\forall \lambda \geq 0$, we have*

$$L(\nu, \theta, \lambda^*) \geq L(\nu^*, \theta^*, \lambda^*) \geq L(\nu^*, \theta^*, \lambda), \tag{6}$$

where $\mathcal{B}_{(\theta^*,\nu^*)}(r)$ is a hyper-dimensional ball centered at (θ^*, ν^*) with radius $r > 0$.

In (Ott, 2010; Bäuerle and Ott, 2011) it is shown that there exists a *deterministic history-dependent* optimal policy for CVaR-constrained optimization. The important point is that this policy does not depend on the complete history, but only on the current time step k, current state of the system x_k, and accumulated discounted constraint cost $\sum_{i=0}^{k} \gamma^i D(x_i, a_i)$.

In the following two sections, we present a policy gradient (PG) algorithm (Section 3) and several actor-critic (AC) algorithms (Section 4) to optimize (5) (and hence find a locally optimal solution to problem (3)). While the PG algorithm updates its parameters after observing several trajectories, the AC algorithms are incremental and update their parameters at each time-step.

3. A Trajectory-based Policy Gradient Algorithm

In this section, we present a policy gradient algorithm to solve the optimization problem (5). The idea of the algorithm is to descend in (θ, ν) and ascend in λ using the gradients of $L(\nu, \theta, \lambda)$ w.r.t. θ,

ν, and λ, i.e.,[3]

$$\nabla_\theta L(\nu,\theta,\lambda) = \nabla_\theta V^\theta(x^0) + \frac{\lambda}{(1-\alpha)}\nabla_\theta \mathbb{E}\Big[\big(\mathcal{D}^\theta(x^0)-\nu\big)^+\Big], \tag{7}$$

$$\partial_\nu L(\nu,\theta,\lambda) = \lambda\bigg(1+\frac{1}{(1-\alpha)}\partial_\nu \mathbb{E}\Big[\big(\mathcal{D}^\theta(x^0)-\nu\big)^+\Big]\bigg) \ni \lambda\bigg(1-\frac{1}{(1-\alpha)}\mathbb{P}\big(\mathcal{D}^\theta(x^0)\geq\nu\big)\bigg), \tag{8}$$

$$\nabla_\lambda L(\nu,\theta,\lambda) = \nu + \frac{1}{(1-\alpha)}\mathbb{E}\Big[\big(\mathcal{D}^\theta(x^0)-\nu\big)^+\Big] - \beta. \tag{9}$$

The unit of observation in this algorithm is a system trajectory generated by following the current policy. At each iteration, the algorithm generates N trajectories by following the current policy, uses them to estimate the gradients in (7)–(9), and then uses these estimates to update the parameters ν, θ, λ.

Let $\xi = \{x_0, a_0, c_0, x_1, a_1, c_1, \ldots, x_{T-1}, a_{T-1}, c_{T-1}, x_T\}$ be a trajectory generated by following the policy θ, where $x_T = x_{\text{Tar}}$ is the target state of the system. The cost, constraint cost, and probability of ξ are defined as $\mathcal{C}(\xi) = \sum_{k=0}^{T-1}\gamma^k C(x_k,a_k)$, $\mathcal{D}(\xi) = \sum_{k=0}^{T-1}\gamma^k D(x_k,a_k)$, and $\mathbb{P}_\theta(\xi) = P_0(x_0)\prod_{k=0}^{T-1}\mu(a_k|x_k;\theta)P(x_{k+1}|x_k,a_k)$, respectively. Based on the definition of $\mathbb{P}_\theta(\xi)$, one obtains $\nabla_\theta \log \mathbb{P}_\theta(\xi) = \sum_{k=0}^{T-1}\nabla_\theta \log \mu(a_k|x_k;\theta)$.

Algorithm 1 contains the pseudo-code of our proposed policy gradient algorithm. What appears inside the parentheses on the right-hand-side of the update equations are the estimates of the gradients of $L(\nu,\theta,\lambda)$ w.r.t. θ,ν,λ (estimates of (7)–(9)). Gradient estimates of the Lagrangian function can be found in Appendix A.1. In the algorithm, Γ_Θ is an operator that projects a vector $\theta \in \mathbb{R}^\kappa$ to the closest point in a compact and convex set $\Theta \subset \mathbb{R}^\kappa$, i.e., $\Gamma_\Theta(\theta) = \arg\min_{\hat{\theta}\in\Theta}\|\theta-\hat{\theta}\|_2^2$, Γ_N is a projection operator to $[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$, i.e., $\Gamma_N(\nu) = \arg\min_{\hat{\nu}\in[-\frac{D_{\max}}{1-\gamma},\frac{D_{\max}}{1-\gamma}]}\|\nu-\hat{\nu}\|_2^2$, and Γ_Λ is a projection operator to $[0,\lambda_{\max}]$, i.e., $\Gamma_\Lambda(\lambda) = \arg\min_{\hat{\lambda}\in[0,\lambda_{\max}]}\|\lambda-\hat{\lambda}\|_2^2$. These projection operators are necessary to ensure the convergence of the algorithm. Next we introduce the following assumptions for the step-sizes of the policy gradient method in Algorithm 1.

Assumption 6 (Step Sizes for Policy Gradient) *The step size schedules* $\{\zeta_1(k)\}$, $\{\zeta_2(k)\}$, *and* $\{\zeta_3(k)\}$ *satisfy*

$$\sum_k \zeta_1(k) = \sum_k \zeta_2(k) = \sum_k \zeta_3(k) = \infty, \tag{10}$$

$$\sum_k \zeta_1(k)^2, \quad \sum_k \zeta_2(k)^2, \quad \sum_k \zeta_3(k)^2 < \infty, \tag{11}$$

$$\zeta_1(k) = o\big(\zeta_2(k)\big), \quad \zeta_2(i) = o\big(\zeta_3(k)\big). \tag{12}$$

These step-size schedules satisfy the standard conditions for stochastic approximation algorithms, and ensure that the ν update is on the fastest time-scale $\big\{\zeta_3(k)\big\}$, the policy θ update is on the intermediate time-scale $\big\{\zeta_2(k)\big\}$, and the Lagrange multiplier λ update is on the slowest time-scale $\big\{\zeta_1(k)\big\}$. This results in a three time-scale stochastic approximation algorithm.

In the following theorem, we prove that our policy gradient algorithm converges to a locally optimal policy for the CVaR-constrained optimization problem.

3. The notation $\ni$ in (8) means that the right-most term is a member of the sub-gradient set $\partial_\nu L(\nu,\theta,\lambda)$.

Theorem 7 *Under Assumptions 2–6, the sequence of policy updates in Algorithm 1 converges almost surely to a locally optimal policy θ^* for the CVaR-constrained optimization problem.*

While we refer the reader to Appendix A.2 for the technical details of this proof, a high level overview of the proof technique is given as follows.

1. First we show that each update of the multi-time scale discrete stochastic approximation algorithm $(\nu_i, \theta_i, \lambda_i)$ converges almost surely, but at different speeds, to the stationary point $(\nu^*, \theta^*, \lambda^*)$ of the corresponding continuous time system.

2. Then by using Lyapunov analysis, we show that the continuous time system is locally asymptotically stable at the stationary point $(\nu^*, \theta^*, \lambda^*)$.

3. Since the Lyapunov function used in the above analysis is the Lagrangian function $L(\nu, \theta, \lambda)$, we finally conclude that the stationary point $(\nu^*, \theta^*, \lambda^*)$ is also a local saddle point, which implies θ^* is the locally optimal policy for the CVaR-constrained optimization problem.

This convergence proof procedure is rather standard for stochastic approximation algorithms, see (Bhatnagar et al., 2009; Bhatnagar and Lakshmanan, 2012; Prashanth and Ghavamzadeh, 2013) for more details, and represents the structural backbone for the convergence analysis of the other policy gradient and actor-critic methods provided in this paper.

Notice that the difference in convergence speeds between θ_i, ν_i, and λ_i is due to the step-size schedules. Here ν converges faster than θ and θ converges faster than λ. This multi-time scale convergence property allows us to simplify the convergence analysis by assuming that θ and λ are fixed in ν's convergence analysis, assuming that ν converges to $\nu^*(\theta)$ and λ is fixed in θ's convergence analysis, and finally assuming that ν and θ have already converged to $\nu^*(\lambda)$ and $\theta^*(\lambda)$ in λ's convergence analysis. To illustrate this idea, consider the following two-time scale stochastic approximation algorithm for updating $(x_i, y_i) \in \mathbf{X} \times \mathbf{Y}$:

$$x_{i+1} = x_i + \zeta_1(i)\big(f(x_i, y_i) + M_{i+1}\big), \tag{13}$$

$$y_{i+1} = y_i + \zeta_2(i)\big(g(x_i, y_i) + N_{i+1}\big), \tag{14}$$

where $f(x_i, y_i)$ and $g(x_i, y_i)$ are Lipschitz continuous functions, M_{i+1}, N_{i+1} are square integrable Martingale differences w.r.t. the σ-fields $\sigma(x_k, y_k, M_k, k \leq i)$ and $\sigma(x_k, y_k, N_k, k \leq i)$, and $\zeta_1(i)$ and $\zeta_2(i)$ are non-summable, square summable step sizes. If $\zeta_2(i)$ converges to zero faster than $\zeta_1(i)$, then (13) is a faster recursion than (14) after some iteration i_0 (i.e., for $i \geq i_0$), which means (13) has uniformly larger increments than (14). Since (14) can be written as

$$y_{i+1} = y_i + \zeta_1(i)\Big(\frac{\zeta_2(i)}{\zeta_1(i)}\big(g(x_i, y_i) + N_{i+1}\big)\Big),$$

and by the fact that $\zeta_2(i)$ converges to zero faster than $\zeta_1(i)$, (13) and (14) can be viewed as noisy Euler discretizations of the ODEs $\dot{x} = f(x, y)$ and $\dot{y} = 0$. Note that one can consider the ODE $\dot{x} = f(x, y_0)$ in place of $\dot{x} = f(x, y)$, where y_0 is constant, because $\dot{y} = 0$. One can then show (see e.g., Theorem 6.2 of Borkar 2008) the main two-timescale convergence result, i.e., under the above assumptions associated with (14), the sequence (x_i, y_i) converges to $\big(\mu(y^\star), y^\star\big)$ as $i \to \infty$, with probability one, where $\mu(y_0)$ is a globally asymptotically stable equilibrium of the ODE $\dot{x} = f(x, y_0)$, μ is a Lipschitz continuous function, and $y^\star$ is a globally asymptotically stable equilibrium of the ODE $\dot{y} = g\big(\mu(y), y\big)$.

Algorithm 1 Trajectory-based Policy Gradient Algorithm for CVaR Optimization

Input: parameterized policy $\mu(\cdot|\cdot;\theta)$, confidence level α, and cost tolerance β
Initialization: policy $\theta = \theta_0$, VaR parameter $\nu = \nu_0$, and the Lagrangian parameter $\lambda = \lambda_0$
while TRUE **do**
  for $i = 0, 1, 2, \ldots$ **do**
    for $j = 1, 2, \ldots$ **do**
      Generate N trajectories $\{\xi_{j,i}\}_{j=1}^{N}$ by starting at $x_0 = x^0$ and following the current policy θ_i.
    end for

$$\nu \textbf{ Update:} \quad \nu_{i+1} = \Gamma_N \left[\nu_i - \zeta_3(i) \left(\lambda_i - \frac{\lambda_i}{(1-\alpha)N} \sum_{j=1}^{N} \mathbf{1}\{\mathcal{D}(\xi_{j,i}) \geq \nu_i\} \right) \right]$$

$$\theta \textbf{ Update:} \quad \theta_{i+1} = \Gamma_\Theta \left[\theta_i - \zeta_2(i) \left(\frac{1}{N} \sum_{j=1}^{N} \nabla_\theta \log \mathbb{P}_\theta(\xi_{j,i})|_{\theta=\theta_i} \mathcal{C}(\xi_{j,i}) \right. \right.$$
$$\left. \left. + \frac{\lambda_i}{(1-\alpha)N} \sum_{j=1}^{N} \nabla_\theta \log \mathbb{P}_\theta(\xi_{j,i})|_{\theta=\theta_i} \big(\mathcal{D}(\xi_{j,i}) - \nu_i\big) \mathbf{1}\{\mathcal{D}(\xi_{j,i}) \geq \nu_i\} \right) \right]$$

$$\lambda \textbf{ Update:} \quad \lambda_{i+1} = \Gamma_\Lambda \left[\lambda_i + \zeta_1(i) \left(\nu_i - \beta + \frac{1}{(1-\alpha)N} \sum_{j=1}^{N} \big(\mathcal{D}(\xi_{j,i}) - \nu_i\big) \mathbf{1}\{\mathcal{D}(\xi_{j,i}) \geq \nu_i\} \right) \right]$$

  end for
  if $\{\lambda_i\}$ converges to $\lambda_{\max}$, i.e., $|\lambda_{i^*} - \lambda_{\max}| \leq \epsilon$ for some tolerance parameter $\epsilon > 0$ **then**
    Set $\lambda_{\max} \leftarrow 2\lambda_{\max}$.
  else
    return parameters ν, θ, λ and **break**
  end if
end while

4. Actor-Critic Algorithms

As mentioned in Section 3, the unit of observation in our policy gradient algorithm (Algorithm 1) is a system trajectory. This may result in high variance for the gradient estimates, especially when the length of the trajectories is long. To address this issue, in this section, we propose two actor-critic algorithms that approximate some quantities in the gradient estimates by linear combinations of basis functions and update the parameters (linear coefficients) incrementally (after each state-action transition). We present two actor-critic algorithms for optimizing (5). These algorithms are based on the gradient estimates of Sections 4.1-4.3. While the first algorithm (SPSA-based) is fully incremental and updates all the parameters θ, ν, λ at each time-step, the second one updates θ at each time-step and updates ν and λ only at the end of each trajectory, thus is regarded as a semi-trajectory-based method. Algorithm 2 contains the pseudo-code of these algorithms. The projection operators Γ_Θ, Γ_N, and Γ_Λ are defined as in Section 3 and are necessary to ensure the convergence of the algorithms. Next, we introduce the following assumptions for the step-sizes of the actor-critic method in Algorithm 2.

Assumption 8 (Step Sizes) *The step size schedules* $\{\zeta_1(k)\}$, $\{\zeta_2(k)\}$, $\{\zeta_3(k)\}$, *and* $\{\zeta_4(k)\}$ *satisfy*

$$\sum_k \zeta_1(k) = \sum_k \zeta_2(k) = \sum_k \zeta_3(k) = \sum_k \zeta_4(k) = \infty, \tag{15}$$

$$\sum_k \zeta_1(k)^2, \quad \sum_k \zeta_2(k)^2, \quad \sum_k \zeta_3(k)^2, \quad \sum_k \zeta_4(k)^2 < \infty, \tag{16}$$

$$\zeta_1(k) = o\big(\zeta_2(k)\big), \quad \zeta_2(i) = o\big(\zeta_3(k)\big), \quad \zeta_3(k) = o\big(\zeta_4(k)\big). \tag{17}$$

Furthermore, the SPSA step size $\{\Delta_k\}$ *in the actor-critic algorithm satisfies* $\Delta_k \to 0$ *as* $k \to \infty$ *and* $\sum_k (\zeta_2(k)/\Delta_k)^2 < \infty$.

These step-size schedules satisfy the standard conditions for stochastic approximation algorithms, and ensure that the critic update is on the fastest time-scale $\big\{\zeta_4(k)\big\}$, the policy and VaR parameter updates are on the intermediate time-scale, with the ν-update $\big\{\zeta_3(k)\big\}$ being faster than the θ-update $\big\{\zeta_2(k)\big\}$, and finally the Lagrange multiplier update is on the slowest time-scale $\big\{\zeta_1(k)\big\}$. This results in four time-scale stochastic approximation algorithms.

4.1 Gradient w.r.t. the Policy Parameters θ

The gradient of the objective function w.r.t. the policy θ in (7) may be rewritten as

$$\nabla_\theta L(\nu, \theta, \lambda) = \nabla_\theta \left(\mathbb{E}\big[\mathcal{C}^\theta(x^0)\big] + \frac{\lambda}{(1-\alpha)} \mathbb{E}\Big[\big(\mathcal{D}^\theta(x^0) - \nu\big)^+\Big] \right). \tag{24}$$

Given the original MDP $\mathcal{M} = (\mathcal{X}, \mathcal{A}, C, D, P, P_0)$ and the parameter λ, we define the augmented MDP $\bar{\mathcal{M}} = (\bar{\mathcal{X}}, \bar{\mathcal{A}}, \bar{C}_\lambda, \bar{P}, \bar{P}_0)$ as $\bar{\mathcal{X}} = \mathcal{X} \times \mathcal{S}$, $\bar{\mathcal{A}} = \mathcal{A}$, $\bar{P}_0(x, s) = P_0(x)\mathbf{1}\{s_0 = s\}$, and

$$\bar{C}_\lambda(x, s, a) = \begin{cases} \lambda(-s)^+/(1-\alpha) & \text{if } x = x_{\text{Tar}}, \\ C(x, a) & \text{otherwise}, \end{cases}$$

$$\bar{P}(x', s'|x, s, a) = \begin{cases} P(x'|x, a)\mathbf{1}\{s' = \big(s - D(x, a)\big)/\gamma\} & \text{if } x \in \mathcal{X}', \\ \mathbf{1}\{x' = x_{\text{Tar}}, s' = 0\} & \text{if } x = x_{\text{Tar}}, \end{cases}$$

where $\mathcal{S}$ is the finite state space of the augmented state s, s_0 is the initial state of the augmented MDP, x_{Tar} is the target state of the original MDP $\mathcal{M}$ and s_{Tar} is the s part of the state when a policy θ reaches a target state x_{Tar}, which we assume occurs before an upper-bound T number of steps , i.e., $s_{\text{Tar}} = \frac{1}{\gamma^T}\Big(\nu - \sum_{k=0}^{T-1} \gamma^k D(x_k, a_k)\Big)$, such that the initial state is given by $s_0 = \nu$. We will now use n to indicate the size of the *augmented* state space $\bar{\mathcal{X}}$ instead of the size of the original state space $\mathcal{X}$. It can be later seen that the augmented state s in the MDP $\bar{\mathcal{M}}$ keeps track of the cumulative CVaR constraint cost, and allows one to reformulate the CVaR Lagrangian problem as an MDP (with respect to $\bar{\mathcal{M}}$).

We define a class of parameterized stochastic policies $\big\{\mu(\cdot|x, s; \theta), (x, s) \in \bar{\mathcal{X}}, \theta \in \Theta \subseteq R^{\kappa_1}\big\}$ for this augmented MDP. Recall that $\mathcal{C}^\theta(x)$ is the discounted cumulative cost and $\mathcal{D}^\theta(x)$ is the discounted cumulative constraint cost. Therefore, the total (discounted) cost of a trajectory can be written as

$$\sum_{k=0}^{T} \gamma^k \bar{C}_\lambda(x_k, s_k, a_k) \mid x_0 = x, s_0 = s, \mu \;\; = \;\; \mathcal{C}^\theta(x) + \frac{\lambda}{(1-\alpha)}\big(\mathcal{D}^\theta(x) - s\big)^+. \tag{25}$$

Algorithm 2 Actor-Critic Algorithms for CVaR Optimization

Input: Parameterized policy $\mu(\cdot|\cdot;\theta)$ and value function feature vector $\phi(\cdot)$ (both over the augmented MDP $\bar{\mathcal{M}}$), confidence level α, and cost tolerance β

Initialization: policy $\theta = \theta_0$; VaR parameter $\nu = \nu_0$; Lagrangian parameter $\lambda = \lambda_0$; value function weight vector $v = v_0$; initial condition $(x_0, s_0) = (x^0, \nu)$

while TRUE **do**

// (1) SPSA-based Algorithm:

for $k = 0, 1, 2, \ldots$ **do**

Draw action $a_k \sim \mu(\cdot|x_k, s_k; \theta_k)$; Observe cost $\bar{C}_{\lambda_k}(x_k, s_k, a_k)$;

Observe next state $(x_{k+1}, s_{k+1}) \sim \bar{P}(\cdot|x_k, s_k, a_k)$; *// note that* $s_{k+1} = (s_k - D(x_k, a_k))/\gamma$

// AC Algorithm:

TD Error: $$\delta_k(v_k) = \bar{C}_{\lambda_k}(x_k, s_k, a_k) + \gamma v_k^\top \phi(x_{k+1}, s_{k+1}) - v_k^\top \phi(x_k, s_k) \tag{18}$$

Critic Update: $$v_{k+1} = v_k + \zeta_4(k)\delta_k(v_k)\phi(x_k, s_k) \tag{19}$$

ν **Update:** $$\nu_{k+1} = \Gamma_N\left(\nu_k - \zeta_3(k)\Big(\lambda_k + \frac{v_k^\top\big[\phi\big(x^0, \nu_k + \Delta_k\big) - \phi(x^0, \nu_k - \Delta_k)\big]}{2\Delta_k}\Big)\right) \tag{20}$$

θ **Update:** $$\theta_{k+1} = \Gamma_\Theta\Big(\theta_k - \frac{\zeta_2(k)}{1-\gamma}\nabla_\theta \log \mu(a_k|x_k, s_k; \theta)\cdot \delta_k(v_k)\Big) \tag{21}$$

λ **Update:** $$\lambda_{k+1} = \Gamma_\Lambda\Big(\lambda_k + \zeta_1(k)\big(\nu_k - \beta + \frac{1}{(1-\alpha)(1-\gamma)}\mathbf{1}\{x_k = x_{\text{Tar}}\}(-s_k)^+\big)\Big) \tag{22}$$

if $x_k = x_{\text{Tar}}$ (reach a target state), **then** set $(x_{k+1}, s_{k+1}) = (x^0, \nu_{k+1})$

end for

// (2) Semi Trajectory-based Algorithm:

Initialize $t = 0$

for $k = 0, 1, 2, \ldots$ **do**

Draw action $a_k \sim \mu(\cdot|x_k, s_k; \theta_k)$, observe cost $\bar{C}_{\lambda_k}(x_k, s_k, a_k)$, and next state $(x_{k+1}, s_{k+1}) \sim \bar{P}(\cdot|x_k, s_k, a_k)$; Update $(\delta_k, v_k, \theta_k, \lambda_k)$ using Eqs. 18, 19, 21, and 22

if $x_k = x_{\text{Tar}}$ **then**

Update ν as

ν **Update:** $$\nu_{k+1} = \Gamma_N\left(\nu_k - \zeta_3(k)\Big(\lambda_k - \frac{\lambda_k}{1-\alpha}\mathbf{1}\big\{x_k = x_{\text{Tar}}, s_k \le 0\big\}\Big)\right) \tag{23}$$

Set $(x_{k+1}, s_{k+1}) = (x^0, \nu_{k+1})$ and $t = 0$

else

$t \leftarrow t + 1$

end if

end for

if $\{\lambda_i\}$ converges to $\lambda_{\max}$, i.e., $|\lambda_{i^*} - \lambda_{\max}| \le \epsilon$ for some tolerance parameter $\epsilon > 0$ **then**

Set $\lambda_{\max} \leftarrow 2\lambda_{\max}$.

else

return parameters $v, w, \nu, \theta, \lambda$, and **break**

end if

end while

From (25), it is clear that the quantity in the parenthesis of (24) is the value function of the policy θ at state (x^0, ν) in the augmented MDP $\bar{\mathcal{M}}$, i.e., $V^\theta(x^0, \nu)$. Thus, it is easy to show that[4]

$$\nabla_\theta L(\nu, \theta, \lambda) = \nabla_\theta V^\theta(x^0, \nu) = \frac{1}{1-\gamma} \sum_{x,s,a} \pi_\gamma^\theta(x, s, a|x^0, \nu) \, \nabla \log \mu(a|x, s; \theta) \, Q^\theta(x, s, a),^5 \tag{26}$$

where π_γ^θ is the discounted occupation measure (defined in Section 2) and Q^θ is the action-value function of policy θ in the augmented MDP $\bar{\mathcal{M}}$. We can show that $\frac{1}{1-\gamma} \nabla \log \mu(a_k|x_k, s_k; \theta) \cdot \delta_k$ is an unbiased estimate of $\nabla_\theta L(\nu, \theta, \lambda)$, where

$$\delta_k = \bar{C}_\lambda(x_k, s_k, a_k) + \gamma \widehat{V}(x_{k+1}, s_{k+1}) - \widehat{V}(x_k, s_k)$$

is the temporal-difference (TD) error in the MDP $\bar{\mathcal{M}}$ from (18), and $\widehat{V}$ is an unbiased estimator of V^θ (see e.g., Bhatnagar et al. 2009). In our actor-critic algorithms, the critic uses linear approximation for the value function $V^\theta(x, s) \approx v^\top \phi(x, s) = \widetilde{V}^{\theta,v}(x, s)$, where the feature vector $\phi(\cdot)$ belongs to a low-dimensional space $\mathbb{R}^{\kappa_1}$ with dimension κ_1. The linear approximation $\widetilde{V}^{\theta,v}$ belongs to a low-dimensional subspace $S_V = \{\Phi v | v \in \mathbb{R}^{\kappa_1}\}$, where Φ is the $n \times \kappa_1$ matrix whose rows are the transposed feature vectors $\phi^\top(\cdot)$. To ensure that the set of feature vectors forms a well-posed linear approximation to the value function, we impose the following assumption to the basis functions.

Assumption 9 (Independent Basis Functions) *The basis functions $\left\{\phi^{(i)}\right\}_{i=1}^{\kappa_1}$ are linearly independent. In particular, $\kappa_1 \leq n$ and Φ is full column rank. Moreover, for every $v \in \mathbb{R}^{\kappa_1}$, $\Phi v \neq e$, where e is the n-dimensional vector with all entries equal to one.*

The following theorem shows that the critic update v_k converges almost surely to v^*, the minimizer of the Bellman residual. Details of the proof can be found in Appendix B.2.

Theorem 10 *Define $v^* \in \arg\min_v \|B_\theta[\Phi v] - \Phi v\|_{d_\gamma^\theta}^2$ as the minimizer to the Bellman residual, where the Bellman operator is given by*

$$B_\theta[V](x, s) = \sum_a \mu(a|x, s; \theta) \left\{ \bar{C}_\lambda(x, s, a) + \sum_{x', s'} \gamma \bar{P}(x', s'|x, s, a) V(x', s') \right\}$$

and $\tilde{V}^*(x, s) = (v^*)^\top \phi(x, s)$ is the projected Bellman fixed point of $V^\theta(x, s)$, i.e., $\tilde{V}^*(x, s) = \Pi B_\theta[\tilde{V}^*](x, s)$. Suppose the γ-occupation measure π_γ^θ is used to generate samples of (x_k, s_k, a_k) for any $k \in \{0, 1, \ldots, \}$. Then under Assumptions 8–9, the v-update in the actor-critic algorithm converges to v^* almost surely.

4. Note that the second equality in Equation (26) is the result of the policy gradient theorem (Sutton et al., 2000; Peters et al., 2005).
5. Notice that the state and action spaces of the original MDP are finite, and there is only a finite number of outcomes in the transition of s (due to the assumption of a bounded first hitting time). Therefore the augmented state s belongs to a finite state space as well.

4.2 Gradient w.r.t. the Lagrangian Parameter λ

We may rewrite the gradient of the objective function w.r.t. the Lagrangian parameters λ in (9) as

$$\nabla_\lambda L(\nu,\theta,\lambda) = \nu - \beta + \nabla_\lambda \left(\mathbb{E}\big[\mathcal{C}^\theta(x^0)\big] + \frac{\lambda}{(1-\alpha)} \mathbb{E}\Big[\big(\mathcal{D}^\theta(x^0) - \nu\big)^+\Big] \right) \overset{\text{(a)}}{=} \nu - \beta + \nabla_\lambda V^\theta(x^0,\nu). \tag{27}$$

Similar to Section 4.1, equality (a) comes from the fact that the quantity in parenthesis in (27) is $V^\theta(x^0,\nu)$, the value function of the policy θ at state (x^0,ν) in the augmented MDP $\bar{\mathcal{M}}$. Note that the dependence of $V^\theta(x^0,\nu)$ on λ comes from the definition of the cost function $\bar{C}_\lambda$ in $\bar{\mathcal{M}}$. We now derive an expression for $\nabla_\lambda V^\theta(x^0,\nu)$, which in turn will give us an expression for $\nabla_\lambda L(\nu,\theta,\lambda)$.

Lemma 11 *The gradient of $V^\theta(x^0,\nu)$ w.r.t. the Lagrangian parameter λ may be written as*

$$\nabla_\lambda V^\theta(x^0,\nu) = \frac{1}{1-\gamma} \sum_{x,s,a} \pi_\gamma^\theta(x,s,a|x^0,\nu) \frac{1}{(1-\alpha)} \mathbf{1}\{x = x_{Tar}\}(-s)^+. \tag{28}$$

Proof. See Appendix B.1. ■

From Lemma 11 and (27), it is easy to see that $\nu - \beta + \frac{1}{(1-\gamma)(1-\alpha)} \mathbf{1}\{x = x_{\text{Tar}}\}(-s)^+$ is an unbiased estimate of $\nabla_\lambda L(\nu,\theta,\lambda)$. An issue with this estimator is that its value is fixed to $\nu_k - \beta$ all along a system trajectory, and only changes at the end to $\nu_k - \beta + \frac{1}{(1-\gamma)(1-\alpha)}(-s_{\text{Tar}})^+$. This may affect the incremental nature of our actor-critic algorithm. To address this issue, Chow and Ghavamzadeh (2014) previously proposed a different approach to estimate the gradients w.r.t. θ and λ which involves another value function approximation to the constraint. However this approach is less desirable in many practical applications as it increases the approximation error and impedes the speed of convergence.

Another important issue is that the above estimator is unbiased only if the samples are generated from the distribution $\pi_\gamma^\theta(\cdot|x^0,\nu)$. If we just follow the policy θ, then we may use $\nu_k - \beta + \frac{\gamma^k}{(1-\alpha)} \mathbf{1}\{x_k = x_{\text{Tar}}\}(-s_k)^+$ as an estimate for $\nabla_\lambda L(\nu,\theta,\lambda)$. Note that this is an issue for all discounted actor-critic algorithms: their (likelihood ratio based) estimate for the gradient is unbiased only if the samples are generated from π_γ^θ, and not when we simply follow the policy. This might also be the reason why, to the best of our knowledge, no rigorous convergence analysis can be found in the literature for (likelihood ratio based) discounted actor-critic algorithms under the sampling distribution.[6]

4.3 Sub-Gradient w.r.t. the VaR Parameter ν

We may rewrite the sub-gradient of our objective function w.r.t. the VaR parameter ν in (8) as

$$\partial_\nu L(\nu,\theta,\lambda) \ni \lambda \left(1 - \frac{1}{(1-\alpha)} \mathbb{P}\Big(\sum_{k=0}^{\infty} \gamma^k D(x_k,a_k) \geq \nu \mid x_0 = x^0; \theta \Big) \right). \tag{29}$$

From the definition of the augmented MDP $\bar{\mathcal{M}}$, the probability in (29) may be written as $\mathbb{P}(s_{\text{Tar}} \leq 0 \mid x_0 = x^0, s_0 = \nu; \theta)$, where s_{Tar} is the s part of the state in $\bar{\mathcal{M}}$ when we reach a target state, i.e., $x = x_{\text{Tar}}$ (see Section 4.1). Thus, we may rewrite (29) as

$$\partial_\nu L(\nu,\theta,\lambda) \ni \lambda \Big(1 - \frac{1}{(1-\alpha)} \mathbb{P}\big(s_{\text{Tar}} \leq 0 \mid x_0 = x^0, s_0 = \nu; \theta \big) \Big). \tag{30}$$

6. Note that the discounted actor-critic algorithm with convergence proof in (Bhatnagar, 2010) is based on SPSA.

From (30), it is easy to see that $\lambda - \lambda \mathbf{1}\{s_{\text{Tar}} \leq 0\}/(1-\alpha)$ is an unbiased estimate of the sub-gradient of $L(\nu, \theta, \lambda)$ w.r.t. ν. An issue with this (unbiased) estimator is that it can only be applied at the end of a system trajectory (i.e., when we reach the target state x_{Tar}), and thus, using it prevents us from having a fully incremental algorithm. In fact, this is the estimator that we use in our *semi-trajectory-based* actor-critic algorithm.

One approach to estimate this sub-gradient incrementally is to use the *simultaneous perturbation stochastic approximation* (SPSA) method (Bhatnagar et al., 2013). The idea of SPSA is to estimate the sub-gradient $g(\nu) \in \partial_\nu L(\nu, \theta, \lambda)$ using two values of g at $\nu^- = \nu - \Delta$ and $\nu^+ = \nu + \Delta$, where $\Delta > 0$ is a positive perturbation (see Bhatnagar et al. 2013; Prashanth and Ghavamzadeh 2013 for the detailed description of Δ).[7] In order to see how SPSA can help us to estimate our sub-gradient incrementally, note that

$$\partial_\nu L(\nu, \theta, \lambda) = \lambda + \partial_\nu \left(\mathbb{E}\big[D^\theta(x^0)\big] + \frac{\lambda}{(1-\alpha)} \mathbb{E}\Big[\big(D^\theta(x^0) - \nu\big)^+\Big] \right) \overset{\text{(a)}}{=} \lambda + \partial_\nu V^\theta(x^0, \nu). \quad (31)$$

Similar to Sections 4.1–4.3, equality (a) comes from the fact that the quantity in parenthesis in (31) is $V^\theta(x^0, \nu)$, the value function of the policy θ at state (x^0, ν) in the augmented MDP $\bar{\mathcal{M}}$. Since the critic uses a linear approximation for the value function, i.e., $V^\theta(x, s) \approx v^\top \phi(x, s)$, in our actor-critic algorithms (see Section 4.1 and Algorithm 2), the SPSA estimate of the sub-gradient would be of the form $g(\nu) \approx \lambda + v^\top \big[\phi(x^0, \nu^+) - \phi(x^0, \nu^-)\big]/2\Delta$.

4.4 Convergence of Actor-Critic Methods

In this section, we will prove that the actor-critic algorithms converge to a locally optimal policy for the CVaR-constrained optimization problem. Define

$$\epsilon_\theta(v_k) = \|B_\theta[\Phi v_k] - \Phi v_k\|_\infty$$

as the residual of the value function approximation at step k, induced by policy $\mu(\cdot|\cdot, \cdot; \theta)$. By the triangle inequality and fixed point theorem $B_\theta[V^*] = V^*$, it can be easily seen that $\|V^* - \Phi v_k\|_\infty \leq \epsilon_\theta(v_k) + \|B_\theta[\Phi v_k] - B_\theta[V^*]\|_\infty \leq \epsilon_\theta(v_k) + \gamma \|\Phi v_k - V^*\|_\infty$. The last inequality follows from the contraction property of the Bellman operator. Thus, one concludes that $\|V^* - \Phi v_k\|_\infty \leq \epsilon_\theta(v_k)/(1-\gamma)$. Now, we state the main theorem for the convergence of actor-critic methods.

Theorem 12 *Suppose $\epsilon_{\theta_k}(v_k) \to 0$ and the γ-occupation measure π_γ^θ is used to generate samples of (x_k, s_k, a_k) for any $k \in \{0, 1, \ldots\}$. For the SPSA-based algorithms, suppose the feature vector satisfies the technical Assumption 21 (provided in Appendix B.2.2) and suppose the SPSA step-size satisfies the condition $\epsilon_{\theta_k}(v_k) = o(\Delta_k)$, i.e., $\epsilon_{\theta_k}(v_k)/\Delta_k \to 0$. Then under Assumptions 2–4 and 8–9, the sequence of policy updates in Algorithm 2 converges almost surely to a locally optimal policy for the CVaR-constrained optimization problem.*

Details of the proof can be found in Appendix B.2.

7. SPSA-based gradient estimate was first proposed in (Spall, 1992) and has been widely used in various settings, especially those involving a high-dimensional parameter. The SPSA estimate described above is two-sided. It can also be implemented single-sided, where we use the values of the function at ν and ν^+. We refer the readers to (Bhatnagar et al., 2013) for more details on SPSA and to (Prashanth and Ghavamzadeh, 2013) for its application to learning in mean-variance risk-sensitive MDPs.

5. Extension to Chance-Constrained Optimization of MDPs

In many applications, in particular in engineering (see, for example, (Ono et al., 2015)), *chance constraints* are imposed to ensure mission success with high probability. Accordingly, in this section we extend the analysis of CVaR-constrained MDPs to chance-constrained MDPs (i.e., (4)). As for CVaR-constrained MDPs, we employ a Lagrangian relaxation procedure (Bertsekas, 1999) to convert a chance-constrained optimization problem into the following unconstrained problem:

$$\max_{\lambda} \min_{\theta,\alpha} \Big(L(\theta,\lambda) := \mathcal{C}^{\theta}(x^0) + \lambda \Big(\mathbb{P}\big(\mathcal{D}^{\theta}(x^0) \geq \alpha\big) - \beta \Big) \Big), \tag{32}$$

where λ is the Lagrange multiplier. Recall Assumption 4 which assumed strict feasibility, i.e., there exists a transient policy $\mu(\cdot|x;\theta)$ such that $\mathbb{P}\big(\mathcal{D}^{\theta}(x^0) \geq \alpha\big) < \beta$. This is needed to guarantee the existence of a local saddle point.

5.1 Policy Gradient Method

In this section we propose a policy gradient method for chance-constrained MDPs (similar to Algorithm 1). Since we do not need to estimate the ν-parameter in chance-constrained optimization, the corresponding policy gradient algorithm can be simplified and at each inner loop of Algorithm 1 we only perform the following updates at the end of each trajectory:

$$\theta \textbf{ Update:} \quad \theta_{i+1} = \Gamma_{\Theta}\bigg[\theta_i - \frac{\zeta_2(i)}{N}\Big(\sum_{j=1}^{N} \nabla_{\theta} \log \mathbb{P}(\xi_{j,i}) \mathcal{C}(\xi_{j,i}) + \lambda_i \nabla_{\theta} \log \mathbb{P}(\xi_{j,i}) \mathbf{1}\big\{\mathcal{D}(\xi_{j,i}) \geq \alpha\big\}\Big)\bigg]$$

$$\lambda \textbf{ Update:} \quad \lambda_{i+1} = \Gamma_{\Lambda}\bigg[\lambda_i + \zeta_1(i)\Big(-\beta + \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}\big\{\mathcal{D}(\xi_{j,i}) \geq \alpha\big\}\Big)\bigg]$$

Considering the multi-time-scale step-size rules in Assumption 6, the θ update is on the fast time-scale $\big\{\zeta_2(i)\big\}$ and the Lagrange multiplier λ update is on the slow time-scale $\big\{\zeta_1(i)\big\}$. This results in a two time-scale stochastic approximation algorithm. In the following theorem, we prove that our policy gradient algorithm converges to a locally optimal policy for the chance-constrained problem.

Theorem 13 *Under Assumptions 2–6, the sequence of policy updates in Algorithm 1 converges to a locally optimal policy θ^* for the chance-constrained optimization problem almost surely.*

Proof. [Sketch] By taking the gradient of $L(\theta,\lambda)$ w.r.t. θ, we have

$$\nabla_{\theta} L(\theta,\lambda) = \nabla_{\theta} \mathcal{C}^{\theta}(x^0) + \lambda \nabla_{\theta} \mathbb{P}\big(\mathcal{D}^{\theta}(x^0) \geq \alpha\big) = \sum_{\xi} \nabla_{\theta} \mathbb{P}_{\theta}(\xi) \mathcal{C}(\xi) + \lambda \sum_{\xi} \nabla_{\theta} \mathbb{P}_{\theta}(\xi) \mathbf{1}\big\{\mathcal{D}(\xi) \geq \alpha\big\}.$$

On the other hand, the gradient of $L(\theta,\lambda)$ w.r.t. λ is given by

$$\nabla_{\lambda} L(\theta,\lambda) = \mathbb{P}\big(\mathcal{D}^{\theta}(x^0) \geq \alpha\big) - \beta.$$

One can easily verify that the θ and λ updates are therefore unbiased estimates of $\nabla_{\theta} L(\theta,\lambda)$ and $\nabla_{\lambda} L(\theta,\lambda)$, respectively. Then the rest of the proof follows analogously from the convergence proof of Algorithm 1 in steps 2 and 3 of Theorem 7. ■

5.2 Actor-Critic Method

In this section, we present an actor-critic algorithm for the chance-constrained optimization. Given the original MDP $\mathcal{M} = (\mathcal{X}, \mathcal{A}, C, D, P, P_0)$ and parameter λ, we define the augmented MDP $\bar{\mathcal{M}} = (\bar{\mathcal{X}}, \bar{\mathcal{A}}, \bar{C}_\lambda, \bar{P}, \bar{P}_0)$ as in the CVaR counterpart, except that $\bar{P}_0(x,s) = P_0(x)\mathbf{1}\{s = \alpha\}$ and

$$\bar{C}_\lambda(x,s,a) = \begin{cases} \lambda \mathbf{1}\{s \le 0\} & \text{if } x = x_{\text{Tar}}, \\ C(x,a) & \text{otherwise.} \end{cases}$$

Thus, the total cost of a trajectory can be written as

$$\sum_{k=0}^{T} \bar{C}_\lambda(x_k, s_k, a_k) \mid x_0 = x, s_0 = \beta,\ \mu = \mathcal{C}^\theta(x) + \lambda \mathbb{P}(\mathcal{D}^\theta(x) \ge \beta). \tag{33}$$

Unlike the actor-critic algorithms for CVaR-constrained optimization, here the value function approximation parameter v, policy θ, and Lagrange multiplier estimate λ are updated episodically, i.e., after each episode ends by time T when $(x_k, s_k) = (x_{\text{Tar}}, s_{\text{Tar}})$[8], as follows:

$$\textbf{Critic Update:} \quad v_{k+1} = v_k + \zeta_3(k) \sum_{h=0}^{T} \phi(x_h, s_h)\delta_h(v_k) \tag{34}$$

$$\textbf{Actor Updates:} \quad \theta_{k+1} = \Gamma_\Theta\Big(\theta_k - \zeta_2(k) \sum_{h=0}^{T} \nabla_\theta \log \mu(a_h|x_h, s_h; \theta)|_{\theta=\theta_k} \cdot \delta_h(v_k)\Big) \tag{35}$$

$$\lambda_{k+1} = \Gamma_\Lambda\Big(\lambda_k + \zeta_1(k)\big(-\beta + \mathbf{1}\{s_{\text{Tar}} \le 0\}\big)\Big) \tag{36}$$

From analogous analysis as for the CVaR actor-critic method, the following theorem shows that the critic update v_k converges almost surely to v^*.

Theorem 14 *Let $v^* \in \arg\min_v \|B_\theta[\Phi v] - \Phi v\|^2_{d^\theta}$ be a minimizer of the Bellman residual, where the undiscounted Bellman operator at every $(x,s) \in \bar{\mathcal{X}}'$ is given by*

$$B_\theta[V](x,s) = \sum_{a \in \mathcal{A}} \mu(a|x,s;\theta)\Big\{\bar{C}_\lambda(x,s,a) + \sum_{(x',s') \in \bar{\mathcal{X}}'} \bar{P}(x',s'|x,s,a)V(x',s')\Big\}$$

and $\tilde{V}^*(x,s) = \phi^\top(x,s)v^*$ is the projected Bellman fixed point of $V^\theta(x,s)$, i.e., $\tilde{V}^*(x,s) = \Pi B_\theta[\tilde{V}^*](x,s)$ for $(x,s) \in \bar{\mathcal{X}}'$. Then under Assumptions 8–9, the v-update in the actor-critic algorithm converges to v^* almost surely.

Proof. [Sketch] The proof of this theorem follows the same steps as those in the proof of Theorem 10, except replacing the γ-occupation measure d^θ_γ with the occupation measure d^θ (the total visiting probability). Similar analysis can also be found in the proof of Theorem 10 in Tamar and Mannor (2013). Under Assumption 2, the occupation measure of any transient states $x \in \mathcal{X}'$ (starting at an arbitrary initial transient state $x_0 \in \mathcal{X}'$) can be written as $d^\mu(x|x^0) = \sum_{t=0}^{T_{\mu,x}} \mathbb{P}(x_t = x|x^0;\mu)$ when $\gamma = 1$. This further implies the total visiting probabilities are bounded as follows: $d^\mu(x|x^0) \le T_{\mu,x}$ and $\pi^\mu(x,a|x^0) \le T_{\mu,x}$ for any $x, x_0 \in \mathcal{X}'$. Therefore, when the sequence of

8. Note that s_{Tar} is the state of s_t when x_t hits the (recurrent) target state x_{Tar}.

states $\{(x_h, s_h)\}_{h=0}^T$ is sampled by the h-step transition distribution $\mathbb{P}(x_h, s_h \mid x^0, s^0, \theta)$, $\forall h \leq T$, the unbiased estimators of

$$A := \sum_{(y,s') \in \bar{\mathcal{X}}', a' \in \mathcal{A}} \pi^\theta(y, s', a'|x, s)\phi(y, s')\Big(\phi^\top(y, s') - \sum_{(z,s'') \in \bar{\mathcal{X}}'} \bar{P}(z, s''|y, s', a)\phi^\top(z, s'')\Big)$$

and

$$b := \sum_{(y,s') \in \bar{\mathcal{X}}', a' \in \mathcal{A}} \pi^\theta(y, s', a'|x, s)\phi(y, s')\bar{C}_\lambda(y, s', a')$$

are given by $\sum_{h=0}^T \phi(x_h, s_h)(\phi^\top(x_h, s_h) - \phi^\top(x_{h+1}, s_{h+1}))$ and $\sum_{h=0}^T \phi(x_h, s_h)\bar{C}_\lambda(x_h, s_h, a_h)$, respectively. Note that in this theorem, we directly use the results from Theorem 7.1 in (Bertsekas, 1995) to show that every eigenvalue of matrix A has positive real part, instead of using the technical result in Lemma 20. ■

Recall that $\epsilon_\theta(v_k) = \|B_\theta[\Phi v_k] - \Phi v_k\|_\infty$ is the residual of the value function approximation at step k induced by policy $\mu(\cdot|\cdot, \cdot; \theta)$. By the triangle inequality and fixed-point theorem of stochastic stopping problems, i.e., $B_\theta[V^*] = V^*$ from Theorem 3.1 in (Bertsekas, 1995), it can be easily seen that $\|V^* - \Phi v_k\|_\infty \leq \epsilon_\theta(v_k) + \|B_\theta[\Phi v_k] - B_\theta[V^*]\|_\infty \leq \epsilon_\theta(v_k) + \kappa\|\Phi v_k - V^*\|_\infty$ for some $\kappa \in (0, 1)$. Similar to the actor-critic algorithm for CVaR-constrained optimization, the last inequality also follows from the contraction mapping property of B_θ from Theorem 3.2 in (Bertsekas, 1995). Now, we state the main theorem for the convergence of the actor-critic method.

Theorem 15 *Under Assumptions 2–9, if $\epsilon_{\theta_k}(v_k) \to 0$, then the sequence of policy updates converges almost surely to a locally optimal policy θ^* for the chance-constrained optimization problem.*

Proof. [Sketch] From Theorem 14, the critic update converges to the minimizer of the Bellman residual. Since the critic update converges on the fastest scale, as in the proof of Theorem 12, one can replace v_k by $v^*(\theta_k)$ in the convergence proof of the actor update. Furthermore, by sampling the sequence of states $\{(x_h, s_h)\}_{h=0}^T$ with the h-step transition distribution $\mathbb{P}(x_h, s_h \mid x^0, s^0, \theta)$, $\forall h \leq T$, the unbiased estimator of the gradient of the linear approximation to the Lagrangian function is given by

$$\nabla_\theta \tilde{L}^v(\theta, \lambda) := \sum_{(x,s) \in \bar{\mathcal{X}}', a \in \mathcal{A}} \pi^\theta(x, s, a|x_0 = x^0, s_0 = \nu)\nabla_\theta \log \mu(a|x, s; \theta)\tilde{A}^{\theta,v}(x, s, a),$$

where $\tilde{Q}^{\theta,v}(x, s, a) - v^\top\phi(x, s)$ is given by $\sum_{h=0}^T \nabla_\theta \log \mu(a_h|x_h, s_h; \theta)|_{\theta=\theta_k} \cdot \delta_h(v^*)$ and the unbiased estimator of $\nabla_\lambda L(\theta, \lambda) = -\beta + \mathbb{P}(s_{\text{Tar}} \leq 0)$ is given by $-\beta + \mathbf{1}\{s_{\text{Tar}} \leq 0\}$. Analogous to equation (75) in the proof of Theorem 24, by convexity of quadratic functions, we have for any value function approximation v,

$$\sum_{(y,s') \in \bar{\mathcal{X}}', a' \in \mathcal{A}} \pi^\theta(y, s', a'|x, s)(A_\theta(y, s', a') - \tilde{A}^v_\theta(y, s', a')) \leq 2T\frac{\epsilon_\theta(v)}{1 - \kappa},$$

which further implies that $\nabla_\theta L(\theta, \lambda) - \nabla_\theta \tilde{L}^v(\theta, \lambda) \to 0$ when $\epsilon_\theta(v) \to 0$ at $v = v^*(\theta_k)$. The rest of the proof follows identical arguments as in steps 3 to 5 of the proof of Theorem 12. ■

6. Examples

In this section we illustrate the effectiveness of our risk-constrained policy gradient and actor-critic algorithms by testing them on an American option stopping problem and on a long-term personalized advertisement-recommendation (ad-recommendation) problem.

6.1 The Optimal Stopping Problem

We consider an optimal stopping problem in which the state at each time step $k \leq T$ consists of the cost c_k and time k, i.e., $x = (c_k, k)$, where T is the stopping time. The agent (buyer) should decide either to accept the present cost ($u_k = 1$) or wait ($u_k = 0$). If he/she accepts or when $k = T$, the system reaches a terminal state and the cost $\max(K, c_k)$ is received (K is the maximum cost threshold), otherwise, she receives a holding cost p_h and the new state is $(c_{k+1}, k+1)$, where c_{k+1} is $f_u c_k$ w.p. p and $f_d c_k$ w.p. $1-p$ ($f_u > 1$ and $f_d < 1$ are constants). Moreover, there is a discount factor $\gamma \in (0,1)$ to account for the increase in the buyer's affordability. Note that if we change cost to reward and minimization to maximization, this is exactly the American option pricing problem, a standard testbed to evaluate risk-sensitive algorithms (e.g., see Tamar et al. 2012). Since the state space size n is exponential in T, finding an exact solution via dynamic programming (DP) quickly becomes infeasible, and thus the problem requires approximation and sampling techniques.

The optimal stopping problem can be reformulated as follows

$$\min_{\theta} \mathbb{E}\left[\mathcal{C}^{\theta}(x^0)\right] \qquad \text{subject to} \qquad \text{CVaR}_{\alpha}\big(\mathcal{C}^{\theta}(x^0)\big) \leq \beta \quad \text{or} \quad \mathbb{P}\left(\mathcal{C}^{\theta}(x^0) \geq \alpha\right) \leq \beta, \tag{37}$$

where the discounted cost and constraint cost functions are identical ($\mathcal{C}^{\theta}(x) = \mathcal{D}^{\theta}(x)$) and are both given by $\mathcal{C}^{\theta}(x) = \sum_{k=0}^{T} \gamma^k \left(\mathbf{1}\{u_k = 1\}\max(K, c_k) + \mathbf{1}\{u_k = 0\}p_h\right) \mid x_0 = x,\ \mu$. We set the parameters of the MDP as follows: $x_0 = [1; 0]$, $p_h = 0.1$, $T = 20$, $K = 5$, $\gamma = 0.95$, $f_u = 2$, $f_d = 0.5$, and $p = 0.65$. The confidence interval and constraint threshold are given by $\alpha = 0.95$ and $\beta = 3$. The number of sample trajectories N is set to $500,000$ and the parameter bounds are $\lambda_{\max} = 5,000$ and $\Theta = [-20, 20]^{\kappa_1}$, where the dimension of the basis functions is $\kappa_1 = 1024$. We implement radial basis functions (RBFs) as feature functions and search over the class of Boltzmann policies $\left\{\theta : \theta = \{\theta_{x,a}\}_{x \in \mathcal{X}, a \in \mathcal{A}},\ \mu_\theta(a|x) = \frac{\exp(\theta_{x,a}^\top x)}{\sum_{a \in \mathcal{A}} \exp(\theta_{x,a}^\top x)}\right\}$.

We consider the following trajectory-based algorithms:

1. **PG:** This is a policy gradient algorithm that minimizes the expected discounted cost function without considering any risk criteria.

2. **PG-CVaR/PG-CC:** These are the CVaR/chance-constrained simulated trajectory-based policy gradient algorithms given in Section 3.

The experiments for each algorithm comprise the following two phases:

1. **Tuning phase:** We run the algorithm and update the policy until (ν, θ, λ) converges.

2. **Converged run:** Having obtained a converged policy θ^* in the tuning phase, in the converged run phase, we perform a Monte Carlo simulation of $10,000$ trajectories and report the results as averages over these trials.

We also consider the following incremental algorithms:

1. **AC:** This is an actor-critic algorithm that minimizes the expected discounted cost function without considering any risk criteria. This is similar to Algorithm 1 in (Bhatnagar, 2010).

2. **AC-CVaR/AC-VaR:** These are the CVaR/chance-constrained semi-trajectory actor-critic algorithms given in Section 4.

3. **AC-CVaR-SPSA:** This is the CVaR-constrained SPSA actor-critic algorithm given in Section 4.

Similar to the trajectory-based algorithms, we use RBF features for $[x; s]$ and consider the family of augmented state Boltzmann policies. Similarly, the experiments comprise two phases: 1) the tuning phase, where the set of parameters $(v, \nu, \theta, \lambda)$ is obtained after the algorithm converges, and 2) the converged run, where the policy is simulated with $10,000$ trajectories.

We compare the performance of PG-CVaR and PG-CC (given in Algorithm 1), and AC-CVaR-SPSA, AC-CVaR, and AC-VaR (given in Algorithm 2), with PG and AC, their risk-neutral counterparts. Figures 1 and 2 show the distribution of the discounted cumulative cost $\mathcal{C}^\theta(x^0)$ for the policy θ learned by each of these algorithms. The results indicate that the risk-constrained algorithms yield a higher expected cost, but less worst-case variability, compared to the risk-neutral methods. More precisely, the cost distributions of the risk-constrained algorithms have lower right-tail (worst-case) distribution than their risk-neutral counterparts. Table 1 summarizes the performance of these algorithms. The numbers reiterate what we concluded from Figures 1 and 2.

Notice that while the risk averse policy satisfies the CVaR constraint, it is not tight (i.e., the constraint is not matched). In fact this is a problem of local optimality, and other experiments in the literature (for example see the numerical results in Prashanth and Ghavamzadeh (2013) and in Bhatnagar and Lakshmanan (2012)) have the same problem of producing solutions which obey the constraints but not tightly. However, since both the expectation and CVaR risk metrics are sub-additive and convex, one can always construct a policy that is a linear combination of the risk neutral optimal policy and the risk averse policy, such that it matches the constraint threshold and has a lower cost compared to the risk averse policy.

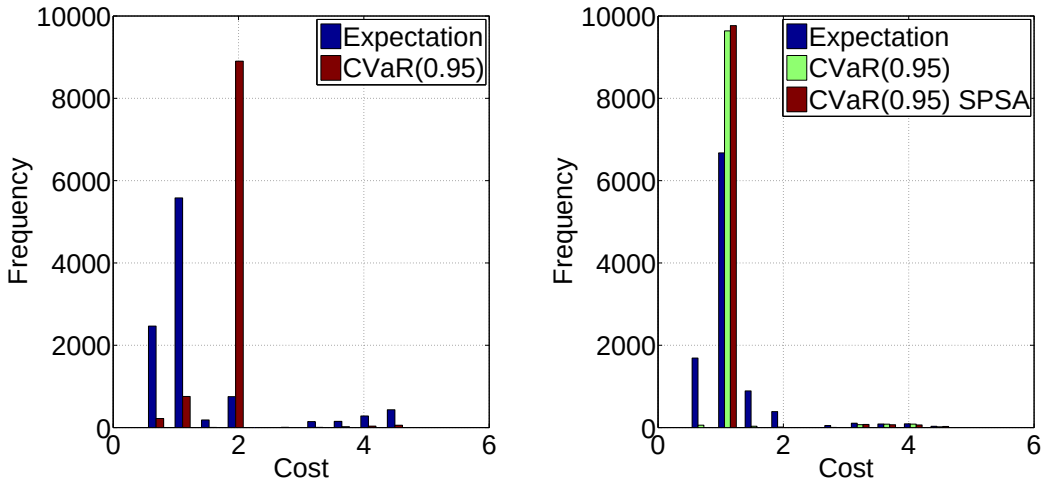

Figure 1: Cost distributions for the policies learned by the CVaR-constrained and risk-neutral policy gradient and actor-critic algorithms. The left figure corresponds to the PG methods and the right figure corresponds to the AC algorithms.

6.2 A Personalized Ad-Recommendation System

Many companies such as banks and retailers use user-specific targeting of advertisements to attract more customers and increase their revenue. When a user requests a webpage that contains a box for

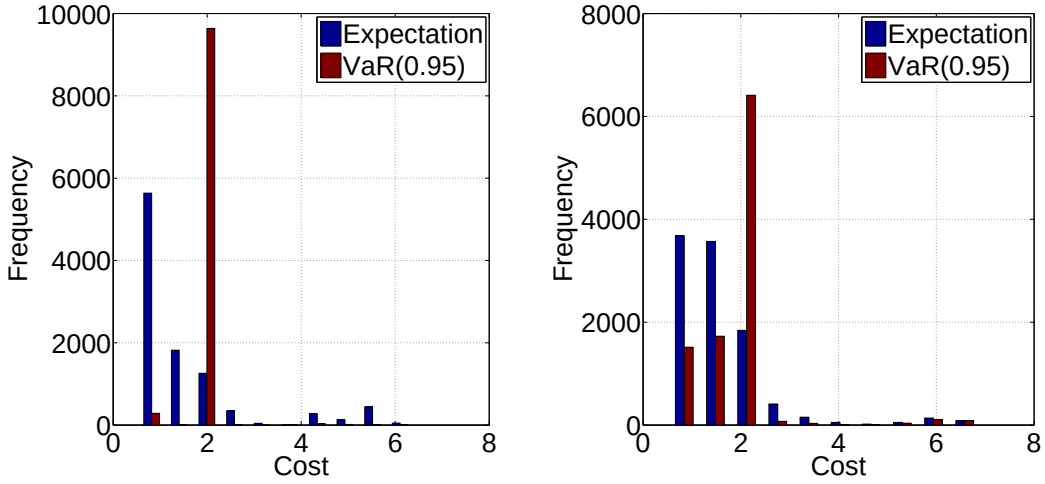

Figure 2: Cost distributions for the policies learned by the chance-constrained and risk-neutral policy gradient and actor-critic algorithms. The left figure corresponds to the PG methods and the right figure corresponds to the AC algorithms.

	$\mathbb{E}\big(\mathcal{C}^\theta(x^0)\big)$	$\sigma\big(\mathcal{C}^\theta(x^0)\big)$	$\text{CVaR}\big(\mathcal{C}^\theta(x^0)\big)$	$\text{VaR}\big(\mathcal{C}^\theta(x^0)\big)$
PG	1.177	1.065	4.464	4.005
PG-CVaR	1.997	0.060	2.000	2.000
PG-CC	1.994	0.121	2.058	2.000
AC	1.113	0.607	3.331	3.220
AC-CVaR-SPSA	1.326	0.322	2.145	1.283
AC-CVaR	1.343	0.346	2.208	1.290
AC-VaR	1.817	0.753	4.006	2.300

Table 1: Performance comparison of the policies learned by the risk-constrained and risk-neutral algorithms. In this table $\sigma\big(\mathcal{C}^\theta(x^0)\big)$ stands for the standard deviation of the total cost.

an advertisement, the system should decide which advertisement (among those in the current campaign) to show to this particular user based on a vector containing all her features, often collected by a cookie. Our goal here is to generate a strategy that for each user of the website selects an ad that when it is presented to her has the highest probability to be clicked on. These days, almost all the industrial personalized ad recommendation systems use supervised learning or contextual bandits algorithms. These methods are based on the i.i.d. assumption of the visits (to the website) and do not discriminate between a visit and a visitor, i.e., each visit is considered as a new visitor that has been sampled i.i.d. from the population of the visitors. As a result, these algorithms are myopic and do not try to optimize for the long-term performance. Despite their success, these methods seem to be insufficient as users establish longer-term relationship with the websites they visit, i.e., the ad recommendation systems should deal with more and more returning visitors. The increase in returning visitors violates (more) the main assumption underlying the supervised learning and bandit algorithms, i.e., there is no difference between a visit and a visitor, and thus, shows the need for a new class of solutions.

The reinforcement learning (RL) algorithms that have been designed to optimize the long-term performance of the system (expected sum of rewards/costs) seem to be suitable candidates for ad recommendation systems (Shani et al., 2002). The nature of these algorithms allows them to take into account all the available knowledge about the user at the current visit, and then selects an offer to maximize the total number of times she will click over multiple visits, also known as the user's life-time value (LTV). Unlike myopic approaches, RL algorithms differentiate between a visit and a visitor, and consider all the visits of a user (in chronological order) as a system trajectory generated by her. In this approach, while the visitors are i.i.d. samples from the population of the users, their visits are not. This long-term approach to the ad recommendation problem allows us to make decisions that are not usually possible with myopic techniques, such as to propose an offer to a user that might be a loss to the company in the short term, but has the effect that makes the user engaged with the website/company and brings her back to spend more money in the future.

For our second case study, we use an Adobe personalized ad-recommendation (Theocharous and Hallak, 2013) simulator that has been trained based on real data captured with permission from the website of a Fortune 50 company that receives hundreds of visitors per day. The simulator produces a vector of 31 real-valued features that provide a compressed representation of all of the available information about a user. The advertisements are clustered into four high-level classes that the agent must select between. After the agent selects an advertisement, the user either clicks (reward of $+1$) or does not click (reward of 0) and the feature vector describing the user is updated. In this case, we test our algorithm by maximizing the customers' life-time value in 15 time steps subject to a bounded tail risk.

Instead of using the cost-minimization framework from the main paper, by defining the return random variable (under a fixed policy θ) $\mathcal{R}^\theta(x^0)$ as the (discounted) total number of clicks along a user's trajectory, here we formulate the personalized ad-recommendation problem as a return maximization problem where the tail risk corresponds to the worst case return distribution:

$$\max_\theta \mathbb{E}\left[\mathcal{R}^\theta(x^0)\right] \qquad \text{subject to} \qquad \text{CVaR}_{1-\alpha}\big(-\mathcal{R}^\theta(x^0)\big) \leq \beta. \tag{38}$$

We set the parameters of the MDP as $T = 15$ and $\gamma = 0.98$, the confidence interval and constraint threshold as $\alpha = 0.05$ and $\beta = 0.12$, the number of sample trajectories N to $1,000,000$, and the parameter bounds as $\lambda_{\max} = 5,000$ and $\Theta = [-60, 60]^{\kappa_1}$, where the dimension of the basis functions is $\kappa_1 = 4096$. Similar to the optimal stopping problem, we implement both the trajectory based algorithm (PG, PG-CVaR) and the actor-critic algorithms (AC, AC-CVaR) for risk-neutral and risk sensitive optimal control. Here we used the 3rd order Fourier basis with cross-products in (Konidaris et al., 2011) as features and search over the family of Boltzmann policies. We compared the performance of PG-CVaR and AC-CVaR, our risk-constrained policy gradient (Algorithm 1) and actor-critic (Algorithms 2) algorithms, with their risk-neutral counterparts (PG and AC). Figure 3 shows the distribution of the discounted cumulative return $\mathcal{R}^\theta(x^0)$ for the policy θ learned by each of these algorithms. The results indicate that the risk-constrained algorithms yield a lower expected reward, but have higher left tail (worst-case) reward distributions. Table 2 summarizes the findings of this experiment.

7. Conclusions and Future Work

We proposed novel policy gradient and actor-critic algorithms for CVaR-constrained and chance-constrained optimization in MDPs, and proved their convergence. Using an optimal stopping problem and a personalized ad-recommendation problem, we showed that our algorithms resulted in

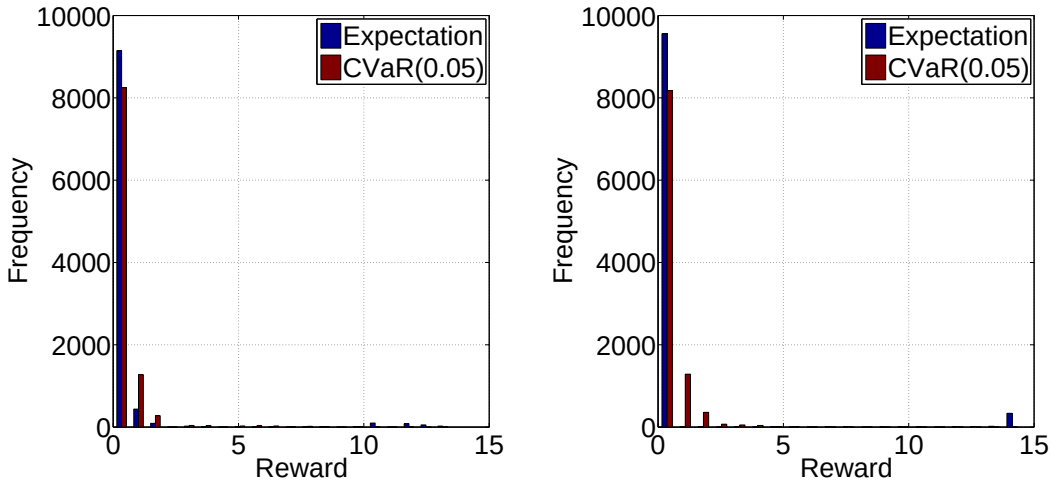

Figure 3: Reward distributions for the policies learned by the CVaR-constrained and risk-neutral policy gradient and actor-critic algorithms. The left figure corresponds to the PG methods and the right figure corresponds to the AC algorithms.

	$\mathbb{E}\big(\mathcal{R}^\theta(x^0)\big)$	$\sigma\big(\mathcal{R}^\theta(x^0)\big)$	$\text{CVaR}\big(\mathcal{R}^\theta(x^0)\big)$	$\text{VaR}\big(\mathcal{R}^\theta(x^0)\big)$
PG	0.396	1.898	0.037	1.000
PG-CVaR	0.287	0.914	0.126	1.795
AC	0.581	2.778	0	0
AC-CVaR	0.253	0.634	0.137	1.890

Table 2: Performance comparison of the policies learned by the CVaR-constrained and risk-neutral algorithms. In this table $\sigma\big(\mathcal{R}^\theta(x^0)\big)$ stands for the standard deviation of the total reward.

policies whose cost distributions have lower right-tail compared to their risk-neutral counterparts. This is important for a risk-averse decision-maker, especially if the right-tail contains catastrophic costs. Future work includes: 1) Providing convergence proofs for our AC algorithms when the samples are generated by following the policy and not from its discounted occupation measure, 2) Using importance sampling methods (Bardou et al., 2009; Tamar et al., 2015) to improve gradient estimates in the right-tail of the cost distribution (worst-case events that are observed with low probability), and 3) Applying the algorithms presented in this paper to a variety of applications ranging from operations research to robotics and finance.

Acknowledgments

We would like to thank Csaba Szepesvari for his comments that helped us with the derivation of the algorithms, Georgios Theocharous for sharing his ad-recommendation simulator with us, and Philip Thomas for helping us with the experiments with the simulator. Y-L. Chow is partially supported by The Croucher Foundation doctoral scholarship. L. Janson was partially supported by NIH training grant T32GM096982. M. Pavone was partially supported by the Office of Naval Research, Science of Autonomy Program, under Contract N00014-15-1-2673.

References

E. Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

E. Altman, K. Avrachenkov, and R. Núñez-Queija. Perturbation analysis for denumerable Markov chains with application to queueing models. *Advances in Applied Probability*, pages 839–853, 2004.

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

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

N. Bäuerle and A. Mundt. Dynamic mean-risk optimization in a binomial model. *Mathematical Methods of Operations Research*, 70(2):219–239, 2009.

N. Bäuerle and J. Ott. Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74(3):361–379, 2011.

J. Baxter and P. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

M. Benaim, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions, Part II: Applications. *Mathematics of Operations Research*, 31(4):673–695, 2006.

D. Bertsekas. *Dynamic programming and optimal control*. Athena Scientific, 1995.

D. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

D. Bertsekas. Min common/max crossing duality: A geometric view of conjugacy in convex optimization. *Lab. for Information and Decision Systems, MIT, Tech. Rep. Report LIDS-P-2796*, 2009.

D. Bertsekas and J. Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

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

S. Bhatnagar and K. Lakshmanan. An online actor-critic algorithm with function approximation for constrained Markov decision processes. *Journal of Optimization Theory and Applications*, 153(3):688–708, 2012.

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

S. Bhatnagar, H. Prasad, and L. Prashanth. *Stochastic recursive algorithms for optimization*, volume 434. Springer, 2013.

K. Boda and J. Filar. Time consistent dynamic risk measures. *Mathematical Methods of Operations Research*, 63(1):169–186, 2006.

K. Boda, J. Filar, Y. Lin, and L. Spanjers. Stochastic target hitting time and the problem of early retirement. *Automatic Control, IEEE Transactions on*, 49(3):409–419, 2004.

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

V. Borkar. Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 27:294–311, 2002.

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

V. Borkar. *Stochastic approximation: a dynamical systems viewpoint*. Cambridge University Press, 2008.

V. Borkar and R. Jain. Risk-constrained Markov decision processes. *IEEE Transaction on Automatic Control*, 2014.

Y. Chow and M. Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In *Advances in Neural Information Processing Systems*, pages 3509–3517, 2014.

Y. Chow and M. Pavone. Stochastic Optimal Control with Dynamic, Time-Consistent Risk Constraints. In *American Control Conference*, pages 390–395, Washington, DC, June 2013. doi: 10.1109/ACC.2013.6579868. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6579868`.

E. Collins. Using Markov decision processes to optimize a nonlinear functional of the final distribution, with manufacturing applications. In *Stochastic Modelling in Innovative Manufacturing*, pages 30–45. Springer, 1997.

B. Derfer, N. Goodyear, K. Hung, C. Matthews, G. Paoni, K. Rollins, R. Rose, M. Seaman, and J. Wiles. Online marketing platform, August 17 2007. US Patent App. 11/893,765.

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

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

R. Howard and J. Matheson. Risk sensitive Markov decision processes. *Management Science*, 18(7):356–369, 1972.

H. Khalil and J. Grizzle. *Nonlinear systems*, volume 3. Prentice hall Upper Saddle River, 2002.

V. Konda and J. Tsitsiklis. Actor-Critic algorithms. In *Proceedings of Advances in Neural Information Processing Systems 12*, pages 1008–1014, 2000.

G. Konidaris, S. Osentoski, and P. Thomas. Value function approximation in reinforcement learning using the Fourier basis. In *AAAI*, 2011.

H. Kushner and G. Yin. *Stochastic approximation algorithms and applications*. Springer, 1997.

P. Marbach. *Simulated-Based Methods for Markov Decision Processes*. PhD thesis, Massachusetts Institute of Technology, 1998.

P. Milgrom and I. Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

T. Morimura, M. Sugiyama, M. Kashima, H. Hachiya, and T. Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning*, pages 799–806, 2010.

M. Ono, M. Pavone, Y. Kuwata, and J. Balaram. Chance-constrained dynamic programming with application to risk-aware robotic space exploration. *Autonomous Robots*, 39(4):555–571, 2015.

J. Ott. *A Markov decision model for a surveillance application and risk-sensitive Markov decision processes*. PhD thesis, Karlsruhe Institute of Technology, 2010.

J. Peters, S. Vijayakumar, and S. Schaal. Natural actor-critic. In *Proceedings of the Sixteenth European Conference on Machine Learning*, pages 280–291, 2005.

M. Petrik and D. Subramanian. An approximate solution method for large risk-averse Markov decision processes. In *Proceedings of the 28th International Conference on Uncertainty in Artificial Intelligence*, 2012.

L. Prashanth and M. Ghavamzadeh. Actor-critic algorithms for risk-sensitive MDPs. In *Proceedings of Advances in Neural Information Processing Systems 26*, pages 252–260, 2013.

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

R. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, 26(7):1443 – 1471, 2002.

G. Shani, R. Brafman, and D. Heckerman. An MDP-based recommender system. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 453–460. Morgan Kaufmann Publishers Inc., 2002.

A. Shapiro, W. Tekaya, J. da Costa, and M. Soares. Risk neutral and risk averse stochastic dual dynamic programming method. *European journal of operational research*, 224(2):375–391, 2013.

T. Shardlow and A. Stuart. A perturbation theory for ergodic Markov chains and application to numerical approximations. *SIAM journal on numerical analysis*, 37(4):1120–1137, 2000.

M. Sobel. The variance of discounted Markov decision processes. *Applied Probability*, pages 794–802, 1982.

J. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

R. Sutton and A. Barto. *Introduction to reinforcement learning*. MIT Press, 1998.

R. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of Advances in Neural Information Processing Systems 12*, pages 1057–1063, 2000.

Y. Le Tallec. *Robust, risk-sensitive, and data-driven control of Markov decision processes*. PhD thesis, Massachusetts Institute of Technology, 2007.

A. Tamar and S. Mannor. Variance adjusted actor critic algorithms. *arXiv preprint arXiv:1310.3697*, 2013.

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

A. Tamar, Y. Glassner, and S. Mannor. Policy gradients beyond expectations: Conditional value-at-risk. In *AAAI*, 2015.

G. Theocharous and A. Hallak. Lifetime value marketing using reinforcement learning. *RLDM 2013*, page 19, 2013.

V. Vilkov. Some properties of the Lagrange function in mathematical programming problems. *Cybernetics and Systems Analysis*, 22(1):75–81, 1986.

D. White. Mean, variance, and probabilistic criteria in finite Markov decision processes: A review. *Journal of Optimization Theory and Applications*, 56(1):1–29, 1988.

R. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

C. Wu and Y. Lin. Minimizing risk models in Markov decision processes with policies depending on target values. *Journal of Mathematical Analysis and Applications*, 231(1):47–67, 1999.

Appendix A. Convergence of Policy Gradient Methods

A.1 Computing the Gradients

i) $\nabla_\theta L(\nu,\theta,\lambda)$: Gradient of $L(\nu,\theta,\lambda)$ w.r.t. θ By expanding the expectations in the definition of the objective function $L(\nu,\theta,\lambda)$ in (5), we obtain

$$L(\nu,\theta,\lambda) = \sum_{\xi} \mathbb{P}_\theta(\xi)\mathcal{C}(\xi) + \lambda\nu + \frac{\lambda}{1-\alpha}\sum_{\xi}\mathbb{P}_\theta(\xi)\big(\mathcal{D}(\xi)-\nu\big)^+ - \lambda\beta.$$

By taking the gradient with respect to θ, we have

$$\nabla_\theta L(\nu,\theta,\lambda) = \sum_{\xi} \nabla_\theta\mathbb{P}_\theta(\xi)\mathcal{C}(\xi) + \frac{\lambda}{1-\alpha}\sum_{\xi}\nabla_\theta\mathbb{P}_\theta(\xi)\big(\mathcal{D}(\xi)-\nu\big)^+.$$

This gradient can be rewritten as

$$\nabla_\theta L(\nu,\theta,\lambda) = \sum_{\xi:\mathbb{P}_\theta(\xi)\neq 0} \mathbb{P}_\theta(\xi)\cdot\nabla_\theta\log\mathbb{P}_\theta(\xi)\left(\mathcal{C}(\xi) + \frac{\lambda}{1-\alpha}\big(\mathcal{D}(\xi)-\nu\big)\mathbf{1}\big\{\mathcal{D}(\xi)\geq\nu\big\}\right), \quad (39)$$

where in the case of $\mathbb{P}_\theta(\xi)\neq 0$, the term $\nabla_\theta\log\mathbb{P}_\theta(\xi)$ is given by:

$$\begin{aligned}\nabla_\theta\log\mathbb{P}_\theta(\xi) =& \nabla_\theta\left\{\sum_{k=0}^{T-1}\log P(x_{k+1}|x_k,a_k) + \log\mu(a_k|x_k;\theta) + \log\mathbf{1}\{x_0=x^0\}\right\}\\ =& \sum_{k=0}^{T-1}\nabla_\theta\log\mu(a_k|x_k;\theta)\\ =& \sum_{k=0}^{T-1}\frac{1}{\mu(a_k|x_k;\theta)}\nabla_\theta\mu(a_k|x_k;\theta).\end{aligned}$$

ii) $\partial_\nu L(\nu,\theta,\lambda)$: Sub-differential of $L(\nu,\theta,\lambda)$ w.r.t. ν From the definition of $L(\nu,\theta,\lambda)$, we can easily see that $L(\nu,\theta,\lambda)$ is a convex function in ν for any fixed $\theta\in\Theta$. Note that for every fixed ν and any ν', we have

$$\big(\mathcal{D}(\xi)-\nu'\big)^+ - \big(\mathcal{D}(\xi)-\nu\big)^+ \geq g\cdot(\nu'-\nu),$$

where g is any element in the set of sub-derivatives:

$$g\in\partial_\nu\big(\mathcal{D}(\xi)-\nu\big)^+ := \begin{cases}-1 & \text{if } \nu<\mathcal{D}(\xi),\\ -q:q\in[0,1] & \text{if } \nu=\mathcal{D}(\xi),\\ 0 & \text{otherwise.}\end{cases}$$

Since $L(\nu,\theta,\lambda)$ is finite-valued for any $\nu\in\mathbb{R}$, by the additive rule of sub-derivatives, we have

$$\partial_\nu L(\nu,\theta,\lambda) = \left\{-\frac{\lambda}{1-\alpha}\sum_{\xi}\mathbb{P}_\theta(\xi)\mathbf{1}\big\{\mathcal{D}(\xi)>\nu\big\} - \frac{\lambda q}{1-\alpha}\sum_{\xi}\mathbb{P}_\theta(\xi)\mathbf{1}\big\{\mathcal{D}(\xi)=\nu\big\} + \lambda \mid q\in[0,1]\right\}. \quad (40)$$

In particular for $q = 1$, we may write the sub-gradient of $L(\nu, \theta, \lambda)$ w.r.t. ν as

$$\partial_\nu L(\nu, \theta, \lambda)|_{q=0} = \lambda - \frac{\lambda}{1-\alpha} \sum_{\xi} \mathbb{P}_\theta(\xi) \cdot \mathbf{1}\{\mathcal{D}(\xi) \geq \nu\}$$

or

$$\lambda - \frac{\lambda}{1-\alpha} \sum_{\xi} \mathbb{P}_\theta(\xi) \cdot \mathbf{1}\{\mathcal{D}(\xi) \geq \nu\} \in \partial_\nu L(\nu, \theta, \lambda).$$

iii) $\nabla_\lambda L(\nu, \theta, \lambda)$: Gradient of $L(\nu, \theta, \lambda)$ w.r.t. λ Since $L(\nu, \theta, \lambda)$ is a linear function in λ, one can express the gradient of $L(\nu, \theta, \lambda)$ w.r.t. λ as follows:

$$\nabla_\lambda L(\nu, \theta, \lambda) = \nu - \beta + \frac{1}{1-\alpha} \sum_{\xi} \mathbb{P}_\theta(\xi) \cdot \big(\mathcal{D}(\xi) - \nu\big)\mathbf{1}\{\mathcal{D}(\xi) \geq \nu\}. \tag{41}$$

A.2 Proof of Convergence of the Policy Gradient Algorithm

In this section, we prove the convergence of the policy gradient algorithm (Algorithm 1).

Since ν converges on the faster timescale than θ and λ, the ν-update can be rewritten by assuming (θ, λ) as invariant quantities, i.e.,

$$\nu_{i+1} = \Gamma_N \bigg[\nu_i - \zeta_3(i) \bigg(\lambda - \frac{\lambda}{(1-\alpha)N} \sum_{j=1}^{N} \mathbf{1}\{\mathcal{D}(\xi_{j,i}) \geq \nu_i\} \bigg) \bigg]. \tag{42}$$

Consider the continuous time dynamics of ν defined using differential inclusion

$$\dot{\nu} \in \Upsilon_\nu \left[-g(\nu) \right], \qquad \forall g(\nu) \in \partial_\nu L(\nu, \theta, \lambda), \tag{43}$$

where

$$\Upsilon_\nu[K(\nu)] := \lim_{0 < \eta \to 0} \frac{\Gamma_N(\nu + \eta K(\nu)) - \Gamma_N(\nu)}{\eta}.$$

Here $\Upsilon_\nu[K(\nu)]$ is the left directional derivative of the function $\Gamma_N(\nu)$ in the direction of $K(\nu)$. By using the left directional derivative $\Upsilon_\nu \left[-g(\nu) \right]$ in the sub-gradient descent algorithm for ν, the gradient will point in the descent direction along the boundary of ν whenever the ν-update hits its boundary.

Furthermore, since ν converges on a faster timescale than θ, and λ is on the slowest time-scale, the θ-update can be rewritten using the converged $\nu^*(\theta)$, assuming λ as an invariant quantity, i.e.,

$$\begin{aligned} \theta_{i+1} = \Gamma_\Theta \bigg[\theta_i - \zeta_2(i) \bigg(& \frac{1}{N} \sum_{j=1}^{N} \nabla_\theta \log \mathbb{P}_\theta(\xi_{j,i})|_{\theta=\theta_i} \mathcal{C}(\xi_{j,i}) \\ & + \frac{\lambda}{(1-\alpha)N} \sum_{j=1}^{N} \nabla_\theta \log \mathbb{P}_\theta(\xi_{j,i})|_{\theta=\theta_i} \big(\mathcal{D}(\xi_{j,i}) - \nu\big) \mathbf{1}\{\mathcal{D}(\xi_{j,i}) \geq \nu^*(\theta_i)\} \bigg) \bigg]. \end{aligned}$$

Consider the continuous time dynamics of $\theta \in \Theta$:

$$\dot{\theta} = \Upsilon_\theta \left[-\nabla_\theta L(\nu, \theta, \lambda) \right]|_{\nu = \nu^*(\theta)}, \tag{44}$$

where

$$\Upsilon_\theta[K(\theta)] := \lim_{0<\eta\to 0} \frac{\Gamma_\Theta(\theta+\eta K(\theta)) - \Gamma_\Theta(\theta)}{\eta}.$$

Similar to the analysis of ν, $\Upsilon_\theta[K(\theta)]$ is the left directional derivative of the function $\Gamma_\Theta(\theta)$ in the direction of $K(\theta)$. By using the left directional derivative $\Upsilon_\theta\left[-\nabla_\theta L(\nu,\theta,\lambda)\right]$ in the gradient descent algorithm for θ, the gradient will point in the descent direction along the boundary of Θ whenever the θ-update hits its boundary.

Finally, since the λ-update converges in the slowest time-scale, the λ-update can be rewritten using the converged $\theta^*(\lambda)$ and $\nu^*(\lambda)$, i.e.,

$$\lambda_{i+1} = \Gamma_\Lambda\left(\lambda_i + \zeta_1(i)\left(\nu^*(\lambda_i) + \frac{1}{1-\alpha}\frac{1}{N}\sum_{j=1}^{N}\left(\mathcal{D}(\xi_{j,i}) - \nu^*(\lambda_i)\right)^+ - \beta\right)\right). \tag{45}$$

Consider the continuous time system

$$\dot{\lambda}(t) = \Upsilon_\lambda\left[\nabla_\lambda L(\nu,\theta,\lambda)\Big|_{\theta=\theta^*(\lambda),\nu=\nu^*(\lambda)}\right], \qquad \lambda(t)\geq 0, \tag{46}$$

where

$$\Upsilon_\lambda[K(\lambda)] := \lim_{0<\eta\to 0} \frac{\Gamma_\Lambda\big(\lambda+\eta K(\lambda)\big) - \Gamma_\Lambda(\lambda)}{\eta}.$$

Again, similar to the analysis of (ν,θ), $\Upsilon_\lambda[K(\lambda)]$ is the left directional derivative of the function $\Gamma_\Lambda(\lambda)$ in the direction of $K(\lambda)$. By using the left directional derivative $\Upsilon_\lambda\left[\nabla_\lambda L(\nu,\theta,\lambda)\right]$ in the gradient ascent algorithm for λ, the gradient will point in the ascent direction along the boundary of $[0,\lambda_{\max}]$ whenever the λ-update hits its boundary.

Define

$$L^*(\lambda) = L(\nu^*(\lambda),\theta^*(\lambda),\lambda),$$

for $\lambda\geq 0$ where $(\theta^*(\lambda),\nu^*(\lambda))\in\Theta\times[-\frac{D_{\max}}{1-\gamma},\frac{D_{\max}}{1-\gamma}]$ is a local minimum of $L(\nu,\theta,\lambda)$ for fixed $\lambda\geq 0$, i.e., $L(\nu,\theta,\lambda)\geq L(\nu^*(\lambda),\theta^*(\lambda),\lambda)$ for any $(\theta,\nu)\in\Theta\times[-\frac{D_{\max}}{1-\gamma},\frac{D_{\max}}{1-\gamma}]\cap\mathcal{B}_{(\theta^*(\lambda),\nu^*(\lambda))}(r)$ for some $r>0$.

Next, we want to show that the ODE (46) is actually a gradient ascent of the Lagrangian function using the envelope theorem from mathematical economics (Milgrom and Segal, 2002). The envelope theorem describes sufficient conditions for the derivative of L^* with respect to λ to equal the partial derivative of the objective function L with respect to λ, holding (θ,ν) at its local optimum $(\theta,\nu)=(\theta^*(\lambda),\nu^*(\lambda))$. We will show that $\nabla_\lambda L^*(\lambda)$ coincides with $\nabla_\lambda L(\nu,\theta,\lambda)|_{\theta=\theta^*(\lambda),\nu=\nu^*(\lambda)}$ as follows.

Theorem 16 *The value function L^* is absolutely continuous. Furthermore,*

$$L^*(\lambda) = L^*(0) + \int_0^\lambda \nabla_{\lambda'} L(\nu,\theta,\lambda')\Big|_{\theta=\theta^*(s),\nu=\nu^*(s),\lambda'=s} ds,\ \ \lambda\geq 0. \tag{47}$$

Proof. The proof follows from analogous arguments to Lemma 4.3 in (Borkar, 2005). From the definition of L^*, observe that for any $\lambda', \lambda'' \geq 0$ with $\lambda' < \lambda''$,

$$\begin{aligned}
|L^*(\lambda'') - L^*(\lambda')| &\leq \sup_{\theta\in\Theta,\nu\in[-\frac{D_{\max}}{1-\gamma},\frac{D_{\max}}{1-\gamma}]} |L(\nu,\theta,\lambda'') - L(\nu,\theta,\lambda')| \\
&= \sup_{\theta\in\Theta,\nu\in[-\frac{D_{\max}}{1-\gamma},\frac{D_{\max}}{1-\gamma}]} \left| \int_{\lambda'}^{\lambda''} \nabla_\lambda L(\nu,\theta,s) ds \right| \\
&\leq \int_{\lambda'}^{\lambda''} \sup_{\theta\in\Theta,\nu\in[\frac{-D_{\max}}{1-\gamma},\frac{D_{\max}}{1-\gamma}]} |\nabla_\lambda L(\nu,\theta,s)|\, ds \leq \frac{3D_{\max}}{(1-\alpha)(1-\gamma)}(\lambda'' - \lambda').
\end{aligned}$$

This implies that L^* is absolutely continuous. Therefore, L^* is continuous everywhere and differentiable almost everywhere.

By the Milgrom–Segal envelope theorem in mathematical economics (Theorem 1 of (Milgrom and Segal, 2002)), one concludes that the derivative of $L^*(\lambda)$ coincides with the derivative of $L(\nu,\theta,\lambda)$ at the point of differentiability λ and $\theta = \theta^*(\lambda)$, $\nu = \nu^*(\lambda)$. Also since L^* is absolutely continuous, the limit of $(L^*(\lambda) - L^*(\lambda'))/(\lambda - \lambda')$ at $\lambda \uparrow \lambda'$ (or $\lambda \downarrow \lambda'$) coincides with the lower/upper directional derivatives if λ' is a point of non-differentiability. Thus, there is only a countable number of non-differentiable points in L^* and the set of non-differentiable points of L^* has measure zero. Therefore, expression (47) holds and one concludes that $\nabla_\lambda L^*(\lambda)$ coincides with $\nabla_\lambda L(\nu,\theta,\lambda)|_{\theta=\theta^*(\lambda),\nu=\nu^*(\lambda)}$. ■

Before getting into the main result, we have the following technical proposition whose proof directly follows from the definition of $\log \mathbb{P}_\theta(\xi)$ and Assumption 3 that $\nabla_\theta \mu(a_k|x_k;\theta)$ is Lipschitz in θ.

Proposition 17 *$\nabla_\theta L(\nu,\theta,\lambda)$ is Lipschitz in θ.*

Remark 18 *The fact that $\nabla_\theta L(\nu,\theta,\lambda)$ is Lipschitz in θ implies that $\|\nabla_\theta L(\nu,\theta,\lambda)\|^2 \leq 2(\|\nabla_\theta L(\nu,\theta_0,\lambda)\| + \|\theta_0\|)^2 + 2\|\theta\|^2$ which further implies that*

$$\|\nabla_\theta L(\nu,\theta,\lambda)\|^2 \leq K_1(1 + \|\theta\|^2).$$

for $K_1 = 2\max(1, (\|\nabla_\theta L(\nu,\theta_0,\lambda)\| + \|\theta_0\|)^2) > 0$. Similarly, the fact that $\nabla_\theta \log \mathbb{P}_\theta(\xi)$ is Lipschitz implies that

$$\|\nabla_\theta \log \mathbb{P}_\theta(\xi)\|^2 \leq K_2(\xi)(1 + \|\theta\|^2)$$

for a positive random variable $K_2(\xi)$. Furthermore, since $T < \infty$ w.p. 1, $\mu(a_k|x_k;\theta) \in (0,1]$ and $\nabla_\theta \mu(a_k|x_k;\theta)$ is Lipschitz for any $k < T$, $K_2(\xi) < \infty$ w.p. 1.

Remark 19 *For any given $\theta \in \Theta$, $\lambda \geq 0$, and $g(\nu) \in \partial_\nu L(\nu,\theta,\lambda)$, we have*

$$|g(\nu)| \leq 3\lambda(1 + |\nu|)/(1-\alpha). \tag{48}$$

To see this, recall that the set of $g(\nu) \in \partial_\nu L(\nu,\theta,\lambda)$ can be parameterized by $q \in [0,1]$ as

$$g(\nu;q) = -\frac{\lambda}{(1-\alpha)} \sum_\xi \mathbb{P}_\theta(\xi) \mathbf{1}\{\mathcal{D}(\xi) > \nu\} - \frac{\lambda q}{1-\alpha} \sum_\xi \mathbb{P}_\theta(\xi) \mathbf{1}\{\mathcal{D}(\xi) = \nu\} + \lambda.$$

It is obvious that $|\mathbf{1}\{\mathcal{D}(\xi)=\nu\}|, |\mathbf{1}\{\mathcal{D}(\xi)>\nu\}| \leq 1+|\nu|$. *Thus,* $\left|\sum_\xi \mathbb{P}_\theta(\xi)\mathbf{1}\{\mathcal{D}(\xi)>\nu\}\right| \leq \sup_\xi |\mathbf{1}\{\mathcal{D}(\xi)>\nu\}| \leq 1+|\nu|$, *and* $\left|\sum_\xi \mathbb{P}_\theta(\xi)\mathbf{1}\{\mathcal{D}(\xi)=\nu\}\right| \leq 1+|\nu|$. *Recalling that* $0<(1-q),(1-\alpha)<1$, *these arguments imply the claim of* (48).

We are now in a position to prove the convergence analysis of Theorem 7.

Proof. [Proof of Theorem 7] We split the proof into the following four steps:

Step 1 (Convergence of ν-update) Since ν converges on a faster time scale than θ and λ, one can take both θ and λ as fixed quantities in the ν-update, i.e.,

$$\nu_{i+1} = \Gamma_N\left(\nu_i + \zeta_3(i)\left(\frac{\lambda}{(1-\alpha)N}\sum_{j=1}^{N}\mathbf{1}\big\{\mathcal{D}(\xi_{j,i})\geq\nu_i\big\} - \lambda + \delta\nu_{i+1}\right)\right), \tag{49}$$

and the Martingale difference term with respect to ν is given by

$$\delta\nu_{i+1} = \frac{\lambda}{1-\alpha}\left(-\frac{1}{N}\sum_{j=1}^{N}\mathbf{1}\big\{\mathcal{D}(\xi_{j,i})\geq\nu_i\big\} + \sum_{\xi}\mathbb{P}_\theta(\xi)\mathbf{1}\{\mathcal{D}(\xi)\geq\nu_i\}\right). \tag{50}$$

First, one can show that $\delta\nu_{i+1}$ is square integrable, i.e.,

$$\mathbb{E}[\|\delta\nu_{i+1}\|^2 \mid F_{\nu,i}] \leq 4\left(\frac{\lambda_{\max}}{1-\alpha}\right)^2$$

where $\mathcal{F}_{\nu,i} = \sigma\big(\nu_m,\, \delta\nu_m,\, m\leq i\big)$ is the filtration of ν_i generated by different independent trajectories.

Second, since the history trajectories are generated based on the sampling probability mass function $\mathbb{P}_\theta(\xi)$, expression (40) implies that $\mathbb{E}\left[\delta\nu_{i+1} \mid \mathcal{F}_{\nu,i}\right] = 0$. Therefore, the ν-update is a stochastic approximation of the ODE (43) with a Martingale difference error term, i.e.,

$$\frac{\lambda}{1-\alpha}\sum_{\xi}\mathbb{P}_\theta(\xi)\mathbf{1}\{\mathcal{D}(\xi)\geq\nu_i\} - \lambda \in -\partial_\nu L(\nu,\theta,\lambda)|_{\nu=\nu_i}.$$

Then one can invoke Corollary 4 in Chapter 5 of Borkar (2008) (stochastic approximation theory for non-differentiable systems) to show that the sequence $\{\nu_i\},\ \nu_i \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$ converges almost surely to a fixed point $\nu^* \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$ of the differential inclusion (43), where

$$\nu^* \in N_c := \left\{\nu \in \left[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}\right] : \Upsilon_\nu[-g(\nu)] = 0,\ g(\nu)\in\partial_\nu L(\nu,\theta,\lambda)\right\}.$$

To justify the assumptions of this corollary, 1) from Remark 19, the Lipschitz property is satisfied, i.e., $\sup_{g(\nu)\in\partial_\nu L(\nu,\theta,\lambda)}|g(\nu)| \leq 3\lambda(1+|\nu|)/(1-\alpha)$, 2) $[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$ and $\partial_\nu L(\nu,\theta,\lambda)$ are convex compact sets by definition, which implies $\{(\nu, g(\nu)) \mid g(\nu)\in\partial_\nu L(\nu,\theta,\lambda)\}$ is a closed set, and further implies $\partial_\nu L(\nu,\theta,\lambda)$ is an upper semi-continuous set valued mapping, 3) the step-size rule follows from Assumption 6, 4) the Martingale difference assumption follows from (50), and 5) $\nu_i \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}], \forall i$ implies that $\sup_i \|\nu_i\| < \infty$ almost surely.

Consider the ODE for $\nu \in \mathbb{R}$ in (43), we define the set-valued derivative of L as follows:

$$D_t L(\nu,\theta,\lambda) = \big\{ g(\nu)\Upsilon_\nu\big[-g(\nu)\big] \mid \forall g(\nu) \in \partial_\nu L(\nu,\theta,\lambda)\big\}.$$

One can conclude that

$$\max_{g(\nu)} D_t L(\nu,\theta,\lambda) = \max\big\{ g(\nu)\Upsilon_\nu\big[-g(\nu)\big] \mid g(\nu) \in \partial_\nu L(\nu,\theta,\lambda)\big\}.$$

We now show that $\max_{g(\nu)} D_t L(\nu,\theta,\lambda) \le 0$ and this quantity is non-zero if $\Upsilon_\nu\big[-g(\nu)\big] \neq 0$ for every $g(\nu) \in \partial_\nu L(\nu,\theta,\lambda)$ by considering three cases. To distinguish the latter two cases, we need to define,

$$\mathcal{G}(\nu) := \left\{ g(\nu) \in \partial L_\nu(\nu,\theta,\lambda) \middle| \forall \eta_0 > 0,\ \exists \eta \in (0,\eta_0] \text{ such that } \theta - \eta g(\nu) \notin \left[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}\right] \right\}.$$

Case 1: $\nu \in (-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma})$.
For every $g(\nu) \in \partial_\nu L(\nu,\theta,\lambda)$, there exists a sufficiently small $\eta_0 > 0$ such that $\nu - \eta_0 g(\nu) \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$ and

$$\Gamma_N\big(\theta - \eta_0 g(\nu)\big) - \theta = -\eta_0 g(\nu).$$

Therefore, the definition of $\Upsilon_\theta[-g(\nu)]$ implies

$$\max_{g(\nu)} D_t L(\nu,\theta,\lambda) = \max\big\{ -g^2(\nu) \mid g(\nu) \in \partial_\nu L(\nu,\theta,\lambda)\big\} \le 0. \tag{51}$$

The maximum is attained because $\partial_\nu L(\nu,\theta,\lambda)$ is a convex compact set and $g(\nu)\Upsilon_\nu\big[-g(\nu)\big]$ is a continuous function. At the same time, we have $\max_{g(\nu)} D_t L(\nu,\theta,\lambda) < 0$ whenever $0 \notin \partial_\nu L(\nu,\theta,\lambda)$.
Case 2: $\nu \in \{-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}\}$ *and* $\mathcal{G}(\nu)$ *is empty.*
The condition $\nu - \eta g(\nu) \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$ implies that

$$\Upsilon_\nu\big[-g(\nu)\big] = -g(\nu).$$

Then we obtain

$$\max_{g(\nu)} D_t L(\nu,\theta,\lambda) = \max\big\{ -g^2(\nu) \mid g(\nu) \in \partial_\nu L(\nu,\theta,\lambda)\big\} \le 0. \tag{52}$$

Furthermore, we have $\max_{g(\nu)} D_t L(\nu,\theta,\lambda) < 0$ whenever $0 \notin \partial_\nu L(\nu,\theta,\lambda)$.
Case 3: $\nu \in \{-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}\}$ *and* $\mathcal{G}(\nu)$ *is nonempty*.
First, consider any $g(\nu) \in \mathcal{G}(\nu)$. For any $\eta > 0$, define $\nu_\eta := \nu - \eta g(\nu)$. The above condition implies that when $0 < \eta \to 0$, $\Gamma_N\big[\nu_\eta\big]$ is the projection of ν_η to the tangent space of $[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$. For any element $\hat{\nu} \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$, since the set $\{\nu \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}] : \|\nu - \nu_\eta\|_2 \le \|\hat{\nu} - \nu_\eta\|_2\}$ is compact, the projection of ν_η on $[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$ exists. Furthermore, since $f(\nu) := \frac{1}{2}(\nu - \nu_\eta)^2$ is a strongly convex function and $\nabla f(\nu) = \nu - \nu_\eta$, by the first order optimality condition, one obtains

$$\nabla f(\nu_\eta^*)(\nu - \nu_\eta^*) = (\nu_\eta^* - \nu_\eta)(\nu - \nu_\eta^*) \ge 0, \quad \forall \nu \in \left[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}\right]$$

where ν_η^* is the unique projection of ν_η (the projection is unique because $f(\nu)$ is strongly convex and $[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$ is a convex compact set). Since the projection (minimizer) is unique, the above equality holds if and only if $\nu = \nu_\eta^*$.

Therefore, for any $\nu \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$ and $\eta > 0$,

$$g(\nu)\Upsilon_\nu\big[-g(\nu)\big] = g(\nu)\left(\lim_{0<\eta\to 0}\frac{\nu_\eta^* - \nu}{\eta}\right)$$
$$= \left(\lim_{0<\eta\to 0}\frac{\nu - \nu_\eta}{\eta}\right)\left(\lim_{0<\eta\to 0}\frac{\nu_\eta^* - \nu}{\eta}\right) = \lim_{0<\eta\to 0}\frac{-\|\nu_\eta^* - \nu\|^2}{\eta^2} + \lim_{0<\eta\to 0}\big(\nu_\eta^* - \nu_\eta\big)\left(\frac{\nu_\eta^* - \nu}{\eta^2}\right) \leq 0.$$

Second, for any $g(\nu) \in \partial_\nu L(\nu,\theta,\lambda) \cap \mathcal{G}(\nu)^c$, one obtains $\nu - \eta g(\nu) \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$, for any $\eta \in (0, \eta_0]$ and some $\eta_0 > 0$. In this case, the arguments follow from case 2 and the following expression holds: $\Upsilon_\nu\big[-g(\nu)\big] = -g(\nu)$.

Combining these arguments, one concludes that

$$\max_{g(\nu)} D_t L(\nu,\theta,\lambda)$$
$$\leq \max\Big\{\max\big\{g(\nu)\ \Upsilon_\nu\big[-g(\nu)\big] \mid g(\nu) \in \mathcal{G}(\nu)\big\}, \max\big\{-g^2(\nu) \mid g(\nu) \in \partial_\nu L(\nu,\theta,\lambda) \cap \mathcal{G}(\nu)^c\big\}\Big\} \leq 0. \tag{53}$$

This quantity is non-zero whenever $0 \notin \{g(\nu)\ \Upsilon_\nu\big[-g(\nu)\big] \mid \forall g(\nu) \in \partial_\nu L(\nu,\theta,\lambda)\}$ (this is because, for any $g(\nu) \in \partial_\nu L(\nu,\theta,\lambda) \cap \mathcal{G}(\nu)^c$, one obtains $g(\nu)\ \Upsilon_\nu\big[-g(\nu)\big] = -g(\nu)^2$). Thus, by similar arguments one may conclude that $\max_{g(\nu)} D_t L(\nu,\theta,\lambda) \leq 0$ and it is non-zero if $\Upsilon_\nu\big[-g(\nu)\big] \neq 0$ for every $g(\nu) \in \partial_\nu L(\nu,\theta,\lambda)$.

Now for any given θ and λ, define the following Lyapunov function

$$\mathcal{L}_{\theta,\lambda}(\nu) = L(\nu,\theta,\lambda) - L(\nu^*,\theta,\lambda)$$

where ν^* is a minimum point (for any given (θ,λ), L is a convex function in ν). Then $\mathcal{L}_{\theta,\lambda}(\nu)$ is a positive definite function, i.e., $\mathcal{L}_{\theta,\lambda}(\nu) \geq 0$. On the other hand, by the definition of a minimum point, one easily obtains $0 \in \{g(\nu^*)\ \Upsilon_\nu\big[-g(\nu^*)\big]|_{\nu=\nu^*} \mid \forall g(\nu^*) \in \partial_\nu L(\nu,\theta,\lambda)|_{\nu=\nu^*}\}$ which means that ν^* is also a stationary point, i.e., $\nu^* \in N_c$.

Note that $\max_{g(\nu)} D_t\mathcal{L}_{\theta,\lambda}(\nu) = \max_{g(\nu)} D_t L(\nu,\theta,\lambda) \leq 0$ and this quantity is non-zero if $\Upsilon_\nu\big[-g(\nu)\big] \neq 0$ for every $g(\nu) \in \partial_\nu L(\nu,\theta,\lambda)$. Therefore, by the Lyapunov theory for asymptotically stable differential inclusions (see Theorem 3.10 and Corollary 3.11 in Benaim et al. (2006), where the Lyapunov function $\mathcal{L}_{\theta,\lambda}(\nu)$ satisfies Hypothesis 3.1 and the property in (53) is equivalent to Hypothesis 3.9 in the reference), the above arguments imply that with any initial condition $\nu(0)$, the state trajectory $\nu(t)$ of (43) converges to ν^*, i.e., $L(\nu^*,\theta,\lambda) \leq L(\nu(t),\theta,\lambda) \leq L(\nu(0),\theta,\lambda)$ for any $t \geq 0$.

As stated earlier, the sequence $\{\nu_i\},\ \nu_i \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$ constitutes a stochastic approximation to the differential inclusion (43), and thus converges almost surely its solution (Borkar, 2008), which further converges almost surely to $\nu^* \in N_c$. Also, it can be easily seen that N_c is a closed subset of the compact set $[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$, and therefore a compact set itself.

Step 2 (Convergence of θ-update) Since θ converges on a faster time scale than λ and ν converges faster than θ, one can take λ as a fixed quantity and ν as a converged quantity $\nu^*(\theta)$ in the θ-update. The θ-update can be rewritten as a stochastic approximation, i.e.,

$$\theta_{i+1} = \Gamma_\Theta\left(\theta_i + \zeta_2(i)\Big(-\nabla_\theta L(\nu,\theta,\lambda)|_{\theta=\theta_i,\nu=\nu^*(\theta_i)} + \delta\theta_{i+1}\Big)\right), \tag{54}$$

where

$$\begin{aligned}\delta\theta_{i+1} =&\nabla_\theta L(\nu,\theta,\lambda)|_{\theta=\theta_i,\nu=\nu^*(\theta_i)} - \frac{1}{N}\sum_{j=1}^{N}\nabla_\theta \log \mathbb{P}_\theta(\xi_{j,i})\ |_{\theta=\theta_i}\ \mathcal{C}(\xi_{j,i})\\ &- \frac{\lambda}{(1-\alpha)N}\sum_{j=1}^{N}\nabla_\theta \log \mathbb{P}_\theta(\xi_{j,i})|_{\theta=\theta_i}\big(\mathcal{D}(\xi_{j,i})-\nu^*(\theta_i)\big)\mathbf{1}\big\{\mathcal{D}(\xi_{j,i})\geq \nu^*(\theta_i)\big\}.\end{aligned} \tag{55}$$

First, one can show that $\delta\theta_{i+1}$ is square integrable, i.e., $\mathbb{E}[\|\delta\theta_{i+1}\|^2 \mid F_{\theta,i}] \leq K_i(1+\|\theta_i\|^2)$ for some $K_i > 0$, where $\mathcal{F}_{\theta,i} = \sigma\big(\theta_m,\, \delta\theta_m,\, m \leq i\big)$ is the filtration of θ_i generated by different independent trajectories. To see this, notice that

$$\begin{aligned}&\|\delta\theta_{i+1}\|^2\\ \leq& 2\left(\nabla_\theta L(\nu,\theta,\lambda)|_{\theta=\theta_i,\nu=\nu^*(\theta_i)}\right)^2 + \frac{2}{N^2}\left(\frac{C_{\max}}{1-\gamma}+\frac{2\lambda D_{\max}}{(1-\alpha)(1-\gamma)}\right)^2\left(\sum_{j=1}^{N}\nabla_\theta \log \mathbb{P}_\theta(\xi_{j,i})\ |_{\theta=\theta_i}\right)^2\\ \leq& 2K_{1,i}(1+\|\theta_i\|^2) + \frac{2^N}{N^2}\left(\frac{C_{\max}}{1-\gamma}+\frac{2\lambda_{\max} D_{\max}}{(1-\alpha)(1-\gamma)}\right)^2\left(\sum_{j=1}^{N}\|\nabla_\theta \log \mathbb{P}_\theta(\xi_{j,i})\ |_{\theta=\theta_i}\|^2\right)\\ \leq& 2K_{1,i}(1+\|\theta_i\|^2) + \frac{2^N}{N^2}\left(\frac{C_{\max}}{1-\gamma}+\frac{2\lambda_{\max} D_{\max}}{(1-\alpha)(1-\gamma)}\right)^2\left(\sum_{j=1}^{N}K_2(\xi_{j,i})(1+\|\theta_i\|^2)\right)\\ \leq& 2\left(K_{1,i}+\frac{2^{N-1}}{N}\left(\frac{C_{\max}}{1-\gamma}+\frac{2\lambda_{\max} D_{\max}}{(1-\alpha)(1-\gamma)}\right)^2\max_{1\leq j\leq N}K_2(\xi_{j,i})\right)(1+\|\theta_i\|^2).\end{aligned}$$

The Lipschitz upper bounds are due to the results in Remark 18. Since $K_2(\xi_{j,i}) < \infty$ w.p. 1, there exists $K_{2,i} < \infty$ such that $\max_{1\leq j\leq N} K_2(\xi_{j,i}) \leq K_{2,i}$. By combining these results, one concludes that $\mathbb{E}[\|\delta\theta_{i+1}\|^2 \mid F_{\theta,i}] \leq K_i(1+\|\theta_i\|^2)$ where

$$K_i = 2\left(K_{1,i}+\frac{2^{N-1}K_{2,i}}{N}\left(\frac{C_{\max}}{1-\gamma}+\frac{2\lambda_{\max} D_{\max}}{(1-\alpha)(1-\gamma)}\right)^2\right) < \infty.$$

Second, since the history trajectories are generated based on the sampling probability mass function $\mathbb{P}_{\theta_i}(\xi)$, expression (39) implies that $\mathbb{E}\left[\delta\theta_{i+1} \mid \mathcal{F}_{\theta,i}\right] = 0$. Therefore, the θ-update is a stochastic approximation of the ODE (44) with a Martingale difference error term. In addition, from the convergence analysis of the ν-update, $\nu^*(\theta)$ is an asymptotically stable equilibrium point for the sequence $\{\nu_i\}$. From (40), $\partial_\nu L(\nu,\theta,\lambda)$ is a Lipschitz set-valued mapping in θ (since $\mathbb{P}_\theta(\xi)$ is Lipschitz in θ), and thus it can be easily seen that $\nu^*(\theta)$ is a Lipschitz continuous mapping of θ.

Now consider the continuous time dynamics for $\theta \in \Theta$, given in (44). We may write

$$\left.\frac{dL(\nu,\theta,\lambda)}{dt}\right|_{\nu=\nu^*(\theta)} = \left(\nabla_\theta L(\nu,\theta,\lambda)|_{\nu=\nu^*(\theta)}\right)^\top \Upsilon_\theta\big[-\nabla_\theta L(\nu,\theta,\lambda)|_{\nu=\nu^*(\theta)}\big]. \tag{56}$$

By considering the following cases, we now show that $dL(\nu,\theta,\lambda)/dt|_{\nu=\nu^*(\theta)} \leq 0$ and this quantity is non-zero whenever $\left\|\Upsilon_\theta\left[-\nabla_\theta L(\nu,\theta,\lambda)|_{\nu=\nu^*(\theta)}\right]\right\| \neq 0$.

Case 1: When $\theta \in \Theta^\circ = \Theta \setminus \partial\Theta$.
Since Θ° is the interior of the set Θ and Θ is a convex compact set, there exists a sufficiently small $\eta_0 > 0$ such that $\theta - \eta_0 \nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)} \in \Theta$ and

$$\Gamma_\Theta\big(\theta - \eta_0 \nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}\big) - \theta = -\eta_0 \nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}.$$

Therefore, the definition of $\Upsilon_\theta\big[-\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}\big]$ implies

$$\left.\frac{dL(\nu, \theta, \lambda)}{dt}\right|_{\nu=\nu^*(\theta)} = -\left\|\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}\right\|^2 \leq 0. \tag{57}$$

At the same time, we have $dL(\nu, \theta, \lambda)/dt|_{\nu=\nu^*(\theta)} < 0$ whenever $\|\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}\| \neq 0$.

Case 2: When $\theta \in \partial\Theta$ and $\theta - \eta \nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)} \in \Theta$ for any $\eta \in (0, \eta_0]$ and some $\eta_0 > 0$.
The condition $\theta - \eta \nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)} \in \Theta$ implies that

$$\Upsilon_\theta\big[-\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}\big] = -\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}.$$

Then we obtain

$$\left.\frac{dL(\nu, \theta, \lambda)}{dt}\right|_{\nu=\nu^*(\theta)} = -\left\|\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}\right\|^2 \leq 0. \tag{58}$$

Furthermore, $dL(\nu, \theta, \lambda)/dt|_{\nu=\nu^*(\theta)} < 0$ when $\|\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}\| \neq 0$.

Case 3: When $\theta \in \partial\Theta$ and $\theta - \eta \nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)} \notin \Theta$ for some $\eta \in (0, \eta_0]$ and any $\eta_0 > 0$.
For any $\eta > 0$, define $\theta_\eta := \theta - \eta \nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}$. The above condition implies that when $0 < \eta \to 0$, $\Gamma_\Theta\big[\theta_\eta\big]$ is the projection of θ_η to the tangent space of Θ. For any element $\hat{\theta} \in \Theta$, since the set $\{\theta \in \Theta : \|\theta - \theta_\eta\|_2 \leq \|\hat{\theta} - \theta_\eta\|_2\}$ is compact, the projection of θ_η on Θ exists. Furthermore, since $f(\theta) := \frac{1}{2}\|\theta - \theta_\eta\|_2^2$ is a strongly convex function and $\nabla f(\theta) = \theta - \theta_\eta$, by the first order optimality condition, one obtains

$$\nabla f(\theta_\eta^*)^\top (\theta - \theta_\eta^*) = (\theta_\eta^* - \theta_\eta)^\top (\theta - \theta_\eta^*) \geq 0, \quad \forall \theta \in \Theta,$$

where θ_η^* is the unique projection of θ_η (the projection is unique because $f(\theta)$ is strongly convex and Θ is a convex compact set). Since the projection (minimizer) is unique, the above equality holds if and only if $\theta = \theta_\eta^*$.

Therefore, for any $\theta \in \Theta$ and $\eta > 0$,

$$\big(\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}\big)^\top \Upsilon_\theta\big[-\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}\big] = \big(\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}\big)^\top \left(\lim_{0<\eta\to 0} \frac{\theta_\eta^* - \theta}{\eta}\right)$$

$$= \left(\lim_{0<\eta\to 0} \frac{\theta - \theta_\eta}{\eta}\right)^\top \left(\lim_{0<\eta\to 0} \frac{\theta_\eta^* - \theta}{\eta}\right) = \lim_{0<\eta\to 0} \frac{-\|\theta_\eta^* - \theta\|^2}{\eta^2} + \lim_{0<\eta\to 0} \big(\theta_\eta^* - \theta_\eta\big)^\top \left(\frac{\theta_\eta^* - \theta}{\eta^2}\right) \leq 0.$$

By combining these arguments, one concludes that $dL(\nu, \theta, \lambda)/dt|_{\nu=\nu^*(\theta)} \leq 0$ and this quantity is non-zero whenever $\big\|\Upsilon_\theta\left[-\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}\right]\big\| \neq 0$.

Now, for any given λ, define the Lyapunov function

$$\mathcal{L}_\lambda(\theta) = L(\nu^*(\theta), \theta, \lambda) - L(\nu^*(\theta^*), \theta^*, \lambda),$$

where θ^* is a local minimum point. Then there exists a ball centered at θ^* with radius r such that for any $\theta \in \mathcal{B}_{\theta^*}(r)$, $\mathcal{L}_\lambda(\theta)$ is a locally positive definite function, i.e., $\mathcal{L}_\lambda(\theta) \geq 0$. On the other hand, by the definition of a local minimum point, one obtains $\Upsilon_\theta[-\nabla_\theta L(\theta^*, \nu, \lambda)|_{\nu=\nu^*(\theta^*)}]|_{\theta=\theta^*} = 0$ which means that θ^* is a stationary point, i.e., $\theta^* \in \Theta_c$.

Note that $d\mathcal{L}_\lambda(\theta(t))/dt = dL(\theta(t), \nu^*(\theta(t)), \lambda)/dt \leq 0$ and the time-derivative is non-zero whenever $\left\|\Upsilon_\theta\left[-\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}\right]\right\| \neq 0$. Therefore, by the Lyapunov theory for asymptotically stable systems (Khalil and Grizzle, 2002), the above arguments imply that with any initial condition $\theta(0) \in \mathcal{B}_{\theta^*}(r)$, the state trajectory $\theta(t)$ of (44) converges to θ^*, i.e., $L(\theta^*, \nu^*(\theta^*), \lambda) \leq L(\theta(t), \nu^*(\theta(t)), \lambda) \leq L(\theta(0), \nu^*(\theta(0)), \lambda)$ for any $t \geq 0$.

Based on the above properties and noting that 1) from Proposition 17, $\nabla_\theta L(\nu, \theta, \lambda)$ is a Lipschitz function in θ, 2) the step-size rule follows from Assumption 6, 3) expression (60) implies that $\delta\theta_{i+1}$ is a square integrable Martingale difference, and 4) $\theta_i \in \Theta$, $\forall i$ implies that $\sup_i \|\theta_i\| < \infty$ almost surely, one can invoke Theorem 2 in Chapter 6 of (Borkar, 2008) (multi-time scale stochastic approximation theory) to show that the sequence $\{\theta_i\},\ \theta_i \in \Theta$ converges almost surely to the solution of the ODE (44), which further converges almost surely to $\theta^* \in \Theta$.

Step 3 (Local Minimum) Now, we want to show that the sequence $\{\theta_i, \nu_i\}$ converges to a local minimum of $L(\nu, \theta, \lambda)$ for any fixed λ. Recall that $\{\theta_i, \nu_i\}$ converges to $(\theta^*, \nu^*) := (\theta^*, \nu^*(\theta^*))$. Previous arguments on the (ν, θ)-convergence imply that with any initial condition $(\theta(0), \nu(0))$, the state trajectories $\theta(t)$ and $\nu(t)$ of (43) and (44) converge to the set of stationary points (θ^*, ν^*) in the positive invariant set $\Theta_c \times N_c$ and $L(\theta^*, \nu^*, \lambda) \leq L(\theta(t), \nu^*(\theta(t)), \lambda) \leq L(\theta(0), \nu^*(\theta(0)), \lambda) \leq L(\theta(0), \nu(t), \lambda) \leq L(\theta(0), \nu(0), \lambda)$ for any $t \geq 0$.

By contradiction, suppose (θ^*, ν^*) is not a local minimum. Then there exists $(\bar{\theta}, \bar{\nu}) \in \Theta \times [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}] \cap \mathcal{B}_{(\theta^*, \nu^*)}(r)$ such that

$$L(\bar{\theta}, \bar{\nu}, \lambda) = \min_{(\theta,\nu)\in\Theta\times[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]\cap\mathcal{B}_{(\theta^*,\nu^*)}(r)} L(\nu, \theta, \lambda).$$

The minimum is attained by the Weierstrass extreme value theorem. By putting $\theta(0) = \bar{\theta}$, the above arguments imply that

$$L(\bar{\theta}, \bar{\nu}, \lambda) = \min_{(\theta,\nu)\in\Theta\times[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]\cap\mathcal{B}_{(\theta^*,\nu^*)}(r)} L(\nu, \theta, \lambda) < L(\theta^*, \nu^*, \lambda) \leq L(\bar{\theta}, \bar{\nu}, \lambda)$$

which is a contradiction. Therefore, the stationary point (θ^*, ν^*) is a local minimum of $L(\nu, \theta, \lambda)$ as well.

Step 4 (Convergence of λ-update) Since the λ-update converges in the slowest time scale, it can be rewritten using the converged $\theta^*(\lambda) = \theta^*(\nu^*(\lambda), \lambda)$ and $\nu^*(\lambda)$, i.e.,

$$\lambda_{i+1} = \Gamma_\Lambda\left(\lambda_i + \zeta_1(i)\left(\nabla_\lambda L(\nu, \theta, \lambda)\Big|_{\theta=\theta^*(\lambda_i),\nu=\nu^*(\lambda_i),\lambda=\lambda_i} + \delta\lambda_{i+1}\right)\right) \tag{59}$$

where

$$\delta\lambda_{i+1} = -\nabla_\lambda L(\nu, \theta, \lambda)\Big|_{\theta=\theta^*(\lambda),\nu=\nu^*(\lambda),\lambda=\lambda_i} + \left(\nu^*(\lambda_i) + \frac{1}{1-\alpha}\frac{1}{N}\sum_{j=1}^{N}\left(\mathcal{D}(\xi_{j,i}) - \nu^*(\lambda_i)\right)^+ - \beta\right). \tag{60}$$

From (41), we see that $\nabla_\lambda L(\nu, \theta, \lambda)$ is a constant function of λ. Similar to the θ-update, one can easily show that $\delta\lambda_{i+1}$ is square integrable, i.e.,

$$\mathbb{E}[\|\delta\lambda_{i+1}\|^2 \mid \mathcal{F}_{\lambda,i}] \leq 2\left(\beta + \frac{3D_{\max}}{(1-\gamma)(1-\alpha)}\right)^2,$$

where $\mathcal{F}_{\lambda,i} = \sigma\big(\lambda_m,\, \delta\lambda_m,\, m \leq i\big)$ is the filtration of λ generated by different independent trajectories. Furthermore, expression (41) implies that $\mathbb{E}\left[\delta\lambda_{i+1} \mid \mathcal{F}_{\lambda,i}\right] = 0$. Therefore, the λ-update is a stochastic approximation of the ODE (46) with a Martingale difference error term. In addition, from the convergence analysis of the (θ, ν)-update, $(\theta^*(\lambda), \nu^*(\lambda))$ is an asymptotically stable equilibrium point for the sequence $\{\theta_i, \nu_i\}$. From (39), $\nabla_\theta L(\nu, \theta, \lambda)$ is a linear mapping in λ, and $(\theta^*(\lambda), \nu^*(\lambda))$ is a Lipschitz continuous mapping of λ.

Consider the ODE for $\lambda \in [0, \lambda_{\max}]$ in (46). Analogous to the arguments for the θ-update, we can write

$$\frac{d(-L(\nu,\theta,\lambda))}{dt}\bigg|_{\theta=\theta^*(\lambda),\nu=\nu^*(\lambda)} = -\nabla_\lambda L(\nu,\theta,\lambda)\bigg|_{\theta=\theta^*(\lambda),\nu=\nu^*(\lambda)} \Upsilon_\lambda\left[\nabla_\lambda L(\nu,\theta,\lambda)\bigg|_{\theta=\theta^*(\lambda),\nu=\nu^*(\lambda)}\right],$$

and show that $-dL(\nu,\theta,\lambda)/dt|_{\theta=\theta^*(\lambda),\nu=\nu^*(\lambda)} \leq 0$. This quantity is non-zero whenever

$$\left\|\Upsilon_\lambda\left[dL(\nu,\theta,\lambda)/d\lambda|_{\theta=\theta^*(\lambda),\nu=\nu^*(\lambda)}\right]\right\| \neq 0.$$

Consider the Lyapunov function

$$\mathcal{L}(\lambda) = -L(\theta^*(\lambda), \nu^*(\lambda), \lambda) + L(\theta^*(\lambda^*), \nu^*(\lambda^*), \lambda^*)$$

where λ^* is a local maximum point. Then there exists a ball centered at λ^* with radius r such that for any $\lambda \in \mathcal{B}_{\lambda^*}(r)$, $\mathcal{L}(\lambda)$ is a locally positive definite function, i.e., $\mathcal{L}(\lambda) \geq 0$. On the other hand, by the definition of a local maximum point, one obtains

$$\Upsilon_\lambda\left[dL(\nu,\theta,\lambda)/d\lambda|_{\theta=\theta^*(\lambda),\nu=\nu^*(\lambda),\lambda=\lambda^*}\right]|_{\lambda=\lambda^*} = 0$$

which means that λ^* is also a stationary point, i.e., $\lambda^* \in \Lambda_c$. Since

$$\frac{d\mathcal{L}(\lambda(t))}{dt} = -\frac{dL(\theta^*(\lambda(t)), \nu^*(\lambda(t)), \lambda(t))}{dt} \leq 0$$

and the time-derivative is non-zero whenever $\left\|\Upsilon_\lambda[\nabla_\lambda L(\nu,\theta,\lambda)\,|_{\nu=\nu^*(\lambda),\theta=\theta^*(\lambda)}]\right\| \neq 0$, the Lyapunov theory for asymptotically stable systems implies that $\lambda(t)$ converges to λ^*.

Given the above results and noting that the step size rule is selected according to Assumption 6, one can apply the multi-time scale stochastic approximation theory (Theorem 2 in Chapter 6 of Borkar (2008)) to show that the sequence $\{\lambda_i\}$ converges almost surely to the solution of the ODE (46), which further converges almost surely to $\lambda^* \in [0, \lambda_{\max}]$. Since $[0, \lambda_{\max}]$ is a compact set, following the same lines of arguments and recalling the envelope theorem (Theorem 16) for local optima, one further concludes that λ^* is a local maximum of $L(\theta^*(\lambda), \nu^*(\lambda), \lambda) = L^*(\lambda)$.

Step 5 (Local Optima) By letting $\theta^* = \theta^*\big(\nu^*(\lambda^*), \lambda^*\big)$ and $\nu^* = \nu^*(\lambda^*)$, we will show that θ^* is a locally optimal policy for the CVaR-constrained optimization problem, which constitutes a (local) saddle point $(\theta^*, \nu^*, \lambda^*)$ of the Lagrangian function $L(\nu, \theta, \lambda)$ if $\lambda^* \in [0, \lambda_{\max})$.

Suppose the sequence $\{\lambda_i\}$ generated from (59) converges to a stationary point $\lambda^* \in [0, \lambda_{\max})$. Since step 3 implies that (θ^*, ν^*) is a local minimum of $L(\nu, \theta, \lambda^*)$ over the feasible set $(\theta, \nu) \in \Theta \times [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$, there exists a $r > 0$ such that

$$L(\theta^*, \nu^*, \lambda^*) \leq L(\nu, \theta, \lambda^*), \quad \forall (\theta, \nu) \in \Theta \times \left[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}\right] \cap \mathcal{B}_{(\theta^*, \nu^*)}(r).$$

In order to complete the proof, we must show

$$\nu^* + \frac{1}{1-\alpha}\mathbb{E}\left[\left(\mathcal{D}^{\theta^*}(x^0) - \nu^*\right)^+\right] \leq \beta, \tag{61}$$

and

$$\lambda^*\left(\nu^* + \frac{1}{1-\alpha}\mathbb{E}\left[\left(\mathcal{D}^{\theta^*}(x^0) - \nu^*\right)^+\right] - \beta\right) = 0. \tag{62}$$

These two equations imply

$$\begin{aligned}
L(\theta^*, \nu^*, \lambda^*) =& V^{\theta^*}(x^0) + \lambda^*\left(\nu^* + \frac{1}{1-\alpha}\mathbb{E}\left[\left(\mathcal{D}^{\theta^*}(x^0) - \nu^*\right)^+\right] - \beta\right) \\
=& V^{\theta^*}(x^0) \\
\geq& V^{\theta^*}(x^0) + \lambda\left(\nu^* + \frac{1}{1-\alpha}\mathbb{E}\left[\left(\mathcal{D}^{\theta^*}(x^0) - \nu^*\right)^+\right] - \beta\right) = L(\theta^*, \nu^*, \lambda),
\end{aligned}$$

which further implies that $(\theta^*, \nu^*, \lambda^*)$ is a saddle point of $L(\nu, \theta, \lambda)$. We now show that (61) and (62) hold.

Recall that

$$\Upsilon_\lambda\left[\nabla_\lambda L(\nu, \theta, \lambda)|_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda), \lambda=\lambda^*}\right]|_{\lambda=\lambda^*} = 0.$$

We show (61) by contradiction. Suppose

$$\nu^* + \frac{1}{1-\alpha}\mathbb{E}\left[\left(\mathcal{D}^{\theta^*}(x^0) - \nu^*\right)^+\right] > \beta.$$

This implies that for $\lambda^* \in [0, \lambda_{\max})$, we have

$$\Gamma_\Lambda\left(\lambda^* - \eta\left(\beta - \left(\nu^* + \frac{1}{1-\alpha}\mathbb{E}\left[\left(\mathcal{D}^{\theta^*}(x^0) - \nu^*\right)^+\right]\right)\right)\right) = \lambda^* - \eta\left(\beta - \left(\nu^* + \frac{1}{1-\alpha}\mathbb{E}\left[\left(\mathcal{D}^{\theta^*}(x^0) - \nu^*\right)^+\right]\right)\right)$$

for any $\eta \in (0, \eta_{\max}]$, for some sufficiently small $\eta_{\max} > 0$. Therefore,

$$\Upsilon_\lambda\left[\nabla_\lambda L(\nu, \theta, \lambda)\Big|_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda), \lambda=\lambda^*}\right]\Bigg|_{\lambda=\lambda^*} = \nu^* + \frac{1}{1-\alpha}\mathbb{E}\left[\left(\mathcal{D}^{\theta^*}(x^0) - \nu^*\right)^+\right] - \beta > 0.$$

This is in contradiction with the fact that $\Upsilon_\lambda\left[\nabla_\lambda L(\nu, \theta, \lambda)|_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda), \lambda=\lambda^*}\right]|_{\lambda=\lambda^*} = 0$. Therefore, (61) holds.

To show that (62) holds, we only need to show that $\lambda^* = 0$ if

$$\nu^* + \frac{1}{1-\alpha}\mathbb{E}\left[\left(\mathcal{D}^{\theta^*}(x^0) - \nu^*\right)^+\right] < \beta.$$

Suppose $\lambda^* \in (0, \lambda_{\max})$, then there exists a sufficiently small $\eta_0 > 0$ such that

$$\begin{aligned}&\frac{1}{\eta_0}\left(\Gamma_\Lambda\left(\lambda^* - \eta_0\left(\beta - \left(\nu^* + \frac{1}{1-\alpha}\mathbb{E}\left[\left(\mathcal{D}^{\theta^*}(x^0) - \nu^*\right)^+\right]\right)\right)\right) - \Gamma_\Lambda(\lambda^*)\right)\\ =&\nu^* + \frac{1}{1-\alpha}\mathbb{E}\left[\left(\mathcal{D}^{\theta^*}(x^0) - \nu^*\right)^+\right] - \beta < 0.\end{aligned}$$

This again contradicts the assumption $\Upsilon_\lambda\left[\nabla_\lambda L(\nu, \theta, \lambda)|_{\theta=\theta^*(\lambda),\nu=\nu^*(\lambda),\lambda=\lambda^*}\right]|_{\lambda=\lambda^*} = 0$. Therefore (62) holds.

When $\lambda^* = \lambda_{\max}$ and $\nu^* + \frac{1}{1-\alpha}\mathbb{E}\left[\left(\mathcal{D}^{\theta^*}(x^0) - \nu^*\right)^+\right] > \beta$,

$$\Gamma_\Lambda\left(\lambda^* - \eta\left(\beta - \left(\nu^* + \frac{1}{1-\alpha}\mathbb{E}\left[\left(\mathcal{D}^{\theta^*}(x^0) - \nu^*\right)^+\right]\right)\right)\right) = \lambda_{\max}$$

for any $\eta > 0$ and

$$\Upsilon_\lambda\left[\nabla_\lambda L(\nu, \theta, \lambda)|_{\theta=\theta^*(\lambda),\nu=\nu^*(\lambda),\lambda=\lambda^*}\right]|_{\lambda=\lambda^*} = 0.$$

In this case one cannot guarantee feasibility using the above analysis, and $(\theta^*, \nu^*, \lambda^*)$ is not a local saddle point. Such a λ^* is referred to as a spurious fixed point (Kushner and Yin, 1997). Notice that λ^* is bounded (otherwise we can conclude that the problem is infeasible), so that by incrementally increasing $\lambda_{\max}$ in Algorithm 1, we can always prevent ourselves from obtaining a spurious fixed point solution.

Combining the above arguments, we finally conclude that θ^* is a locally optimal policy for the CVaR-constrained optimization problem. ■

Appendix B. Convergence of Actor-Critic Algorithms

Recall from Assumption 6 that the SPSA step size $\{\Delta_k\}$ satisfies $\Delta_k \to 0$ as $k \to \infty$ and $\sum_k (\zeta_2(k)/\Delta_k)^2 < \infty$.

B.1 Gradient with Respect to λ (Proof of Lemma 11)

By taking the gradient of $V^\theta(x^0,\nu)$ w.r.t. λ (recall that both V and Q depend on λ through the cost function $\bar{C}$ of the augmented MDP $\bar{\mathcal{M}}$), we obtain

$$\begin{aligned}
\nabla_\lambda V^\theta(x^0,\nu) &= \sum_{a\in\bar{\mathcal{A}}} \mu(a|x^0,\nu;\theta)\nabla_\lambda Q^\theta(x^0,\nu,a) \\
&= \sum_{a\in\bar{\mathcal{A}}} \mu(a|x^0,\nu;\theta)\nabla_\lambda\Big[\bar{C}(x^0,\nu,a) + \sum_{(x',s')\in\bar{\mathcal{X}}} \gamma\bar{P}(x',s'|x^0,\nu,a)V^\theta(x',s')\Big] \\
&= \underbrace{\sum_{a} \mu(a|x^0,\nu;\theta)\nabla_\lambda\bar{C}(x^0,\nu,a)}_{h(x^0,\nu)} + \gamma \sum_{a,x',s'} \mu(a|x^0,\nu;\theta)\bar{P}(x',s'|x^0,\nu,a)\nabla_\lambda V^\theta(x',s') \\
&= h(x^0,\nu) + \gamma \sum_{a,x',s'} \mu(a|x^0,\nu;\theta)\bar{P}(x',s'|x^0,\nu,a)\nabla_\lambda V^\theta(x',s') \qquad (63)\\
&= h(x^0,\nu) + \gamma \sum_{a,x',s'} \mu(a|x^0,\nu;\theta)\bar{P}(x',s'|x^0,\nu,a)\Big[h(x',s') \\
&\qquad + \gamma \sum_{a',x'',s''} \mu(a'|x',s';\theta)\bar{P}(x'',s''|x',s',a')\nabla_\lambda V^\theta(x'',s'')\Big].
\end{aligned}$$

By unrolling the last equation using the definition of $\nabla_\lambda V^\theta(x,s)$ from (63), we obtain

$$\begin{aligned}
\nabla_\lambda V^\theta(x^0,\nu) &= \sum_{k=0}^{\infty}\gamma^k \sum_{x,s} \mathbb{P}(x_k=x, s_k=s \mid x_0=x^0, s_0=\nu;\theta)h(x,s) \\
&= \frac{1}{1-\gamma}\sum_{x,s} d_\gamma^\theta(x,s|x^0,\nu)h(x,s) = \frac{1}{1-\gamma}\sum_{x,s,a} d_\gamma^\theta(x,s|x^0,\nu)\mu(a|x,s)\nabla_\lambda\bar{C}(x,s,a) \\
&= \frac{1}{1-\gamma}\sum_{x,s,a} \pi_\gamma^\theta(x,s,a|x^0,\nu)\nabla_\lambda\bar{C}(x,s,a) \\
&= \frac{1}{1-\gamma}\sum_{x,s,a} \pi_\gamma^\theta(x,s,a|x^0,\nu)\frac{1}{1-\alpha}\mathbf{1}\{x=x_{\text{Tar}}\}(-s)^+.
\end{aligned}$$

This completes the proof.

B.2 Proof of Convergence of the Actor-Critic Algorithms

B.2.1 Proof of Theorem 10: Critic Update (v-update)

By the step size conditions, one notices that $\{v_k\}$ converges on a faster time scale than $\{\nu_k\}$, $\{\theta_k\}$, and $\{\lambda_k\}$. Thus, one can take (ν,θ,λ) in the v-update as fixed quantities. The critic update can be re-written as follows:

$$v_{k+1} = v_k + \zeta_4(k)\phi(x_k,s_k)\delta_k(v_k), \qquad (64)$$

where the scalar

$$\delta_k(v_k) = -v_k^\top\phi(x_k,s_k) + \gamma v_k^\top\phi(x_{k+1},s_{k+1}) + \bar{C}_\lambda(x_k,s_k,a_k)$$

is the temporal difference (TD) from (18). Define

$$A := \sum_{y,a',s'} \pi_\gamma^\theta(y,s',a'|x,s)\phi(y,s')\left(\phi^\top(y,s') - \gamma \sum_{z,s''} \bar{P}(z,s''|y,s',a)\phi^\top\left(z,s''\right)\right), \tag{65}$$

and

$$b := \sum_{y,a',s'} \pi_\gamma^\theta(y,s',a'|x,s)\phi(y,s')\bar{C}_\lambda(y,s',a'). \tag{66}$$

It is easy to see that the critic update v_k in (64) can be re-written as the following stochastic approximation scheme:

$$v_{k+1} = v_k + \zeta_4(k)(b - Av_k + \delta A_{k+1}), \tag{67}$$

where the noise term δA_{k+1} is a square integrable Martingale difference, i.e., $\mathbb{E}[\delta A_{k+1} \mid \mathcal{F}_k] = 0$ if the γ-occupation measure π_γ^θ is used to generate samples of (x_k, s_k, a_k)—with $\mathcal{F}_k$ being the filtration generated by different independent trajectories. By writing

$$\delta A_{k+1} = -(b - Av_k) + \phi(x_k, s_k)\delta_k(v_k)$$

and noting $\mathbb{E}_{\pi_\gamma^\theta}[\phi(x_k, s_k)\delta_k(v_k) \mid \mathcal{F}_k] = -Av_k + b$, one can easily verify that the stochastic approximation scheme in (67) is equivalent to the critic iterates in (64) and δA_{k+1} is a Martingale difference, i.e., $\mathbb{E}_{\pi_\gamma^\theta}[\delta A_{k+1} \mid \mathcal{F}_k] = 0$. Let

$$h\left(v\right) := -Av + b.$$

Before getting into the convergence analysis, we present a technical lemma whose proof can be found in (Bertsekas and Tsitsiklis, 1996, Lemma 6.10).

Lemma 20 *Every eigenvalue of the matrix A has positive real part.*

We now turn to the analysis of the critic iteration. Note that the following properties hold for the critic update scheme in (64): 1) $h\left(v\right)$ is Lipschitz, 2) the step size satisfies the properties in Assumption 8, 3) the noise term δA_{k+1} is a square integrable Martingale difference, 4) the function $h_c\left(v\right) := h\left(cv\right)/c$, $c \geq 1$ converges uniformly to a continuous function $h_\infty\left(v\right)$ for any v in a compact set, i.e., $h_c\left(v\right) \to h_\infty\left(v\right)$ as $c \to \infty$, and 5) the ordinary differential equation (ODE) $\dot{v} = h_\infty\left(v\right)$ has the origin as its unique globally asymptotically stable equilibrium. The fourth property can be easily verified from the fact that the magnitude of b is finite and $h_\infty\left(v\right) = -Av$. The fifth property follows directly from the facts that $h_\infty\left(v\right) = -Av$ and all eigenvalues of A have positive real parts.

By Theorem 3.1 in (Borkar, 2008), these five properties imply:

$$\text{The critic iterates } \{v_k\} \text{ are bounded almost surely, i.e., } \sup_k \|v_k\| < \infty \text{ almost surely.}$$

The convergence of the critic iterates in (64) can be related to the asymptotic behavior of the ODE

$$\dot{v} = h\left(v\right) = b - Av. \tag{68}$$

Specifically, Theorem 2 in Chapter 2 of (Borkar, 2008) and the above conditions imply $v_k \to v^*$ with probability 1, where the limit v^* depends on (ν, θ, λ) and is the unique solution satisfying $h\left(v^*\right) = 0$, i.e., $Av^* = b$. Therefore, the critic iterates converge to the unique fixed point v^* almost surely, as $k \to \infty$.

B.2.2 PROOF OF THEOREM 12

Step 1 (Convergence of v-update) The proof of convergence for the critic parameter follows directly from Theorem 10.

Step 2 (Convergence of SPSA based ν-update) In this section, we analyze the ν-update for the incremental actor-critic method. This update is based on the SPSA perturbation method. The idea of this method is to estimate the sub-gradient $g(\nu) \in \partial_\nu L(\nu, \theta, \lambda)$ using two simulated value functions corresponding to $\nu^- = \nu - \Delta$ and $\nu^+ = \nu + \Delta$. Here $\Delta \geq 0$ is a positive random perturbation that vanishes asymptotically. The SPSA-based estimate for a sub-gradient $g(\nu) \in \partial_\nu L(\nu, \theta, \lambda)$ is given by

$$g(\nu) \approx \lambda + \frac{1}{2\Delta}\left(\phi^\top\left(x^0, \nu + \Delta\right) - \phi^\top\left(x^0, \nu - \Delta\right)\right) v.$$

We turn to the convergence analysis of the sub-gradient estimation and ν-update. Since v converges faster than ν, and ν converges faster than θ and λ, the ν-update in (20) can be rewritten using the converged critic parameter $v^*(\nu)$, i.e.,

$$\nu_{k+1} = \Gamma_N\left(\nu_k - \zeta_3(k)\left(\lambda + \frac{1}{2\Delta_k}\left(\phi^\top\left(x^0, \nu_k + \Delta_k\right) - \phi^\top\left(x^0, \nu_k - \Delta_k\right)\right) v^*(\nu_k)\right)\right), \quad (69)$$

where (θ, λ) in this expression are viewed as constant quantities.

First, we consider the following assumption on the feature functions in order to prove that the SPSA approximation is asymptotically unbiased.

Assumption 21 *For any $v \in \mathbb{R}^{\kappa_1}$, the feature functions satisfy the following conditions*

$$|\phi_V^\top\left(x^0, \nu + \Delta\right) v - \phi_V^\top\left(x^0, \nu - \Delta\right) v| \leq K_1(v)(1 + \Delta).$$

Furthermore, the Lipschitz constants are uniformly bounded, i.e., $\sup_{v \in \mathbb{R}^{\kappa_1}} K_1^2(v) < \infty$.

This assumption is mild as the expected utility objective function implies that $L(\nu, \theta, \lambda)$ is Lipschitz in ν, and $\phi_V^\top\left(x^0, \nu\right) v$ is just a linear function approximation of $V^\theta(x^0, \nu)$.

Next, we establish the bias and convergence of the stochastic sub-gradient estimate. Let

$$\overline{g}(\nu_k) \in \arg\max\left\{g : g \in \partial_\nu L(\nu, \theta, \lambda)|_{\nu=\nu_k}\right\},$$

and

$$\begin{aligned}
\Lambda_{1,k+1} =& \left(\frac{\left(\phi^\top\left(x^0, \nu_k + \Delta_k\right) - \phi^\top\left(x^0, \nu_k - \Delta_k\right)\right) v^*(\nu_k)}{2\Delta_k} - E_M(k)\right), \\
\Lambda_{2,k} =& \lambda_k + E_M^L(k) - \overline{g}(\nu_k), \\
\Lambda_{3,k} =& E_M(k) - E_M^L(k),
\end{aligned}$$

where

$$\begin{aligned}
E_M(k) :=& \mathbb{E}\left[\frac{1}{2\Delta_k}\left(\phi^\top\left(x^0, \nu_k + \Delta_k\right) - \phi^\top\left(x^0, \nu_k - \Delta_k\right)\right) v^*(\nu_k) \mid \Delta_k\right], \\
E_M^L(k) :=& \mathbb{E}\left[\frac{1}{2\Delta_k}\left(V^\theta\left(x^0, \nu_k + \Delta_k\right) - V^\theta\left(x^0, \nu_k - \Delta_k\right)\right) \mid \Delta_k\right].
\end{aligned}$$

Note that (69) is equivalent to

$$\nu_{k+1} = \Gamma_N \left(\nu_k - \zeta_3(k) \left(\overline{g}(\nu_k) + \Lambda_{1,k+1} + \Lambda_{2,k} + \Lambda_{3,k} \right) \right). \tag{70}$$

First, it is clear that $\Lambda_{1,k+1}$ is a Martingale difference as $\mathbb{E}[\Lambda_{1,k+1} \mid \mathcal{F}_k] = 0$, which implies that

$$M_{k+1} = \sum_{j=0}^{k} \zeta_3(j) \Lambda_{1,j+1}$$

is a Martingale w.r.t. the filtration $\mathcal{F}_k$. By the Martingale convergence theorem, we can show that if $\sup_{k \geq 0} \mathbb{E}[M_k^2] < \infty$, when $k \to \infty$, M_k converges almost surely and $\zeta_3(k)\Lambda_{1,k+1} \to 0$ almost surely. To show that $\sup_{k \geq 0} \mathbb{E}[M_k^2] < \infty$, for any $t \geq 0$ one observes that

$$\begin{aligned}
\mathbb{E}[M_{k+1}^2] &= \sum_{j=0}^{k} \left(\zeta_3(j)\right)^2 \mathbb{E}[\mathbb{E}[\Lambda_{1,j+1}^2 \mid \Delta_j]] \\
&\leq 2 \sum_{j=0}^{k} \mathbb{E}\Bigg[\left(\frac{\zeta_3(j)}{2\Delta_j} \right)^2 \Bigg\{ \mathbb{E}\left[\left(\left(\phi^\top \left(x^0, \nu_j + \Delta_j \right) - \phi^\top \left(x^0, \nu_j - \Delta_j \right) \right) v^*(\nu_j) \right)^2 \mid \Delta_j \right] \\
&\qquad + \mathbb{E}\left[\left(\phi^\top \left(x^0, \nu_j + \Delta_j \right) - \phi^\top \left(x^0, \nu_j - \Delta_j \right) \right) v^*(\nu_j) \mid \Delta_j \right]^2 \Bigg\} \Bigg].
\end{aligned}$$

Now based on Assumption 21, the above expression implies

$$\mathbb{E}[M_{k+1}^2] \leq 2 \sum_{j=0}^{k} \mathbb{E}\left[\left(\frac{\zeta_3(j)}{2\Delta_j} \right)^2 2K_1^2 (1 + \Delta_j)^2 \right].$$

Combining the above results with the step size conditions, there exists $K = 4K_1^2 > 0$ such that

$$\sup_{k \geq 0} \mathbb{E}[M_{k+1}^2] \leq K \sum_{j=0}^{\infty} \mathbb{E}\left[\left(\frac{\zeta_3(j)}{2\Delta_j} \right)^2 \right] + \left(\zeta_3(j)\right)^2 < \infty.$$

Second, by the Min Common/Max Crossing theorem in (Bertsekas, 2009), one can show that $\partial_\nu L(\nu, \theta, \lambda)|_{\nu = \nu_k}$ is a non-empty, convex, and compact set. Therefore, by duality of directional directives and sub-differentials, i.e.,

$$\max \left\{ g : g \in \partial_\nu L(\nu, \theta, \lambda)|_{\nu = \nu_k} \right\} = \lim_{\xi \downarrow 0} \frac{L(\nu_k + \xi, \theta, \lambda) - L(\nu_k - \xi, \theta, \lambda)}{2\xi},$$

one concludes that for $\lambda_k = \lambda$ (we can treat λ_k as a constant because it converges on a slower time scale than ν_k),

$$\lambda + E_M^L(k) = \overline{g}(\nu_k) + O(\Delta_k),$$

almost surely. This further implies that

$$\Lambda_{2,k} = O(\Delta_k), \quad \text{i.e.,} \quad \Lambda_{2,k} \to 0 \quad \text{as} \quad k \to \infty,$$

almost surely.

Third, since $d_\gamma^\theta(x^0, \nu | x^0, \nu) = 1$, from the definition of $\epsilon_\theta(v^*(\nu_k))$,

$$|\Lambda_{3,k}| \leq 2\epsilon_\theta(v^*(\nu_k))/\Delta_k.$$

As t goes to infinity, $\epsilon_\theta(v^*(\nu_k))/\Delta_k \to 0$ by assumption and $\Lambda_{3,k} \to 0$.

Finally, since $\zeta_3(k)\Lambda_{1,k+1} \to 0$, $\Lambda_{2,k} \to 0$, and $\Lambda_{3,k} \to 0$ almost surely, the ν-update in (70) is a noisy sub-gradient descent update with vanishing disturbance bias. Thus, the ν-update in (20) can be viewed as an Euler discretization of an element of the following differential inclusion,

$$\dot{\nu} \in \Upsilon_\nu \left[-g(\nu)\right], \qquad \forall g(\nu) \in \partial_\nu L(\nu, \theta, \lambda), \tag{71}$$

and the ν-convergence analysis is analogous to Step 1 of the proof of Theorem 7.

Step 2′ (Convergence of semi-trajectory ν-update) Since ν converges on a faster timescale than θ and λ, the ν-update in (23) can be rewritten using a fixed pair (θ, λ), i.e.,

$$\nu_{k+1} = \Gamma_N \left(\nu_i - \zeta_3(k) \left(\lambda - \frac{\lambda}{1-\alpha} \left(\mathbb{P}\left(s_{\text{Tar}} \leq 0 \mid x_0 = x^0, s_0 = \nu_k, \mu \right) + \delta\nu_{M,k+1} \right) \right) \right), \tag{72}$$

where

$$\delta\nu_{M,k+1} = -\mathbb{P}\left(s_{\text{Tar}} \leq 0 \mid x_0 = x^0, s_0 = \nu_i, \mu \right) + \mathbf{1}\left\{ x_k = x_{\text{Tar}}, s_k \leq 0 \right\} \tag{73}$$

is a square integrable stochastic term, specifically,

$$\mathbb{E}[(\delta\nu_{M,k+1})^2 \mid \mathcal{F}_{\nu,k}] \leq 2,$$

where $\mathcal{F}_{\nu,k} = \sigma(\nu_m,\, \delta\nu_m,\, m \leq k)$ is the filtration generated by ν. Since $\mathbb{E}\left[\delta\nu_{M,k+1} \mid \mathcal{F}_{\nu,k}\right] = 0$, $\delta\nu_{M,k+1}$ is a Martingale difference and the ν-update in (72) is a stochastic approximation of an element of the differential inclusion

$$\frac{\lambda}{1-\alpha}\mathbb{P}\left(s_{\text{Tar}} \leq 0 \mid x_0 = x^0, s_0 = \nu_k, \mu \right) - \lambda \in -\partial_\nu L(\nu, \theta, \lambda)|_{\nu=\nu_k}.$$

Thus, the ν-update in (23) can be viewed as an Euler discretization of the differential inclusion in (71), and the ν-convergence analysis is analogous to Step 1 of the proof of Theorem 7.

Step 3 (Convergence of θ-update) We first analyze the actor update (θ-update). Since θ converges on a faster time scale than λ, one can take λ in the θ-update as a fixed quantity. Furthermore, since v and ν converge on a faster scale than θ, one can also replace v and ν with their limits $v^*(\theta)$ and $\nu^*(\theta)$ in the convergence analysis. In the following analysis, we assume that the initial state $x^0 \in \mathcal{X}$ is given. Then the θ-update in (21) can be rewritten as follows:

$$\theta_{k+1} = \Gamma_\Theta \left(\theta_k - \zeta_2(k) \left(\nabla_\theta \log \mu(a_k | x_k, s_k; \theta)|_{\theta=\theta_k} \frac{\delta_k(v^*(\theta_k))}{1-\gamma} \right) \right). \tag{74}$$

Consider the case in which the value function for a fixed policy μ is approximated by a learned function approximator, $\phi^\top(x, s)v^*$. If the approximation is sufficiently good, we might hope to use it in place of $V^\theta(x, s)$ and still point roughly in the direction of the true gradient. Recall the temporal difference error (random variable) for a given pair $(x_k, s_k) \in \mathcal{X} \times \mathbb{R}$:

$$\delta_k(v) = -v^\top \phi(x_k, s_k) + \gamma v^\top \phi(x_{k+1}, s_{k+1}) + \bar{C}_\lambda(x_k, s_k, a_k).$$

Define the v-dependent approximated advantage function

$$\tilde{A}^{\theta,v}(x,s,a) := \tilde{Q}^{\theta,v}(x,s,a) - v^\top \phi(x,s),$$

where

$$\tilde{Q}^{\theta,v}(x,s,a) = \gamma \sum_{x',s'} \bar{P}(x',s'|x,s,a) v^\top \phi(x',s') + \bar{C}_\lambda(x,s,a).$$

The following lemma, whose proof follows from the proof of Lemma 3 in (Bhatnagar et al., 2009), shows that $\delta_k(v)$ is an unbiased estimator of $\tilde{A}^{\theta,v}$.

Lemma 22 *For any given policy μ and $v \in \mathbb{R}^{\kappa_1}$, we have*

$$\tilde{A}^{\theta,v}(x,s,a) = \mathbb{E}[\delta_k(v) \mid x_k = x, s_k = s, a_k = a].$$

Define

$$\nabla_\theta \tilde{L}_v(\nu,\theta,\lambda) := \frac{1}{1-\gamma} \sum_{x,a,s} \pi_\gamma^\theta(x,s,a|x_0 = x^0, s_0 = \nu) \nabla_\theta \log \mu(a|x,s;\theta) \tilde{A}^{\theta,v}(x,s,a)$$

as the linear function approximation of $\nabla_\theta \tilde{L}(\nu,\theta,\lambda)$. Similar to Proposition 17, we present the following technical lemma on the Lipschitz property of $\nabla_\theta \tilde{L}_v(\nu,\theta,\lambda)$.

Proposition 23 *$\nabla_\theta \tilde{L}_v(\nu,\theta,\lambda)$ is a Lipschitz function in θ.*

Proof. Consider the feature vector v. Recall that the feature vector satisfies the linear equation $Av = b$, where A and b are given by (65) and (66), respectively. From Lemma 1 in (Bhatnagar and Lakshmanan, 2012), by exploiting the inverse of A using Cramer's rule, one may show that v is continuously differentiable in θ. Now consider the γ-occupation measure π_γ^θ. By applying Theorem 2 in (Altman et al., 2004) (or Theorem 3.1 in Shardlow and Stuart 2000), it can be seen that the occupation measure π_γ^θ of the process (x_k, s_k) is continuously differentiable in θ. Recall from Assumption 3 in Section 2.2 that $\nabla_\theta \mu(a_k|x_k,s_k;\theta)$ is a Lipschitz function in θ for any $a \in \mathcal{A}$ and $k \in \{0,\ldots,T-1\}$, and $\mu(a_k|x_k,s_k;\theta)$ is differentiable in θ. By combining these arguments and noting that the sum of products of Lipschitz functions is Lipschitz, one concludes that $\nabla_\theta \tilde{L}_v(\nu,\theta,\lambda)$ is Lipschitz in θ. ■

We turn to the convergence proof of θ.

Theorem 24 *The sequence of θ-updates in* (21) *converges almost surely to an equilibrium point $\widehat{\theta}^*$ that satisfies $\Upsilon_\theta \left[-\nabla_\theta \tilde{L}_{v^*(\theta)}(\nu^*(\theta),\theta,\lambda)\right] = 0$, for a given $\lambda \in [0,\lambda_{\max}]$. Furthermore, if the function approximation error $\epsilon_\theta(v_k)$ vanishes as the feature vector v_k converges to v^*, then the sequence of θ-updates converges to θ^* almost surely, where θ^* is a local minimum point of $L(\nu^*(\theta),\theta,\lambda)$ for a given $\lambda \in [0,\lambda_{\max}]$.*

Proof. We will mainly focus on proving the convergence of $\theta_k \to \theta^*$ (second part of the theorem). Since we just showed in Proposition 23 that $\nabla_\theta \tilde{L}_{v^*(\theta)}(\nu^*(\theta),\theta,\lambda)$ is Lipschitz in θ, the convergence proof of $\theta_k \to \widehat{\theta}^*$ (first part of the theorem) follows from identical arguments.

Note that the θ-update in (74) can be rewritten as:

$$\theta_{k+1} = \Gamma_\Theta \left(\theta_k + \zeta_2(k) \left(-\nabla_\theta L(\nu,\theta,\lambda)|_{\nu=\nu^*(\theta),\theta=\theta_k} + \delta\theta_{k+1} + \delta\theta_\epsilon\right)\right),$$

where

$$\delta\theta_{k+1} = \sum_{x',a',s'} \pi_\gamma^{\theta_k}(x',s',a'|x_0 = x^0, s_0 = \nu^*(\theta_k))\nabla_\theta \log \mu(a'|x',s';\theta)|_{\theta=\theta_k} \frac{\tilde{A}^{\theta_k,v^*(\theta_k)}(x',s',a')}{1-\gamma}$$
$$- \nabla_\theta \log \mu(a_k|x_k,s_k;\theta)|_{\theta=\theta_k} \frac{\delta_k(v^*(\theta_k))}{1-\gamma}.$$

and

$$\delta\theta_\epsilon = \sum_{x',a',s'} \pi_\gamma^{\theta_k}(x',s',a'|x_0 = x^0, s_0 = \nu^*(\theta_k))\cdot$$
$$\frac{\nabla_\theta \log \mu(a'|x',s';\theta)|_{\theta=\theta_k}}{1-\gamma}(A^{\theta_k}(x',s',a') - \tilde{A}^{\theta_k,v^*(\theta_k)}(x',s',a'))$$

First, one can show that $\delta\theta_{k+1}$ is square integrable, specifically,

$$\begin{aligned}
&\mathbb{E}[\|\delta\theta_{k+1}\|^2 \mid \mathcal{F}_{\theta,k}] \\
&\leq \frac{2}{1-\gamma}\|\nabla_\theta \log \mu(u|x,s;\theta)|_{\theta=\theta_k} \mathbf{1}\{\mu(u|x,s;\theta_k) > 0\}\|_\infty^2 \left(\|\tilde{A}^{\theta_k,v^*(\theta_k)}(x,s,a)\|_\infty^2 + |\delta_k(v^*(\theta_k))|^2\right) \\
&\leq \frac{2}{1-\gamma} \cdot \frac{\|\nabla_\theta \mu(u|x,s;\theta)|_{\theta=\theta_k}\|_\infty^2}{\min\{\mu(u|x,s;\theta_k) \mid \mu(u|x,s;\theta_k) > 0\}^2} \left(\|\tilde{A}^{\theta_k,v^*(\theta_k)}(x,s,a)\|_\infty^2 + |\delta_k(v^*(\theta_k))|^2\right) \\
&\leq 64\frac{K^2\|\theta_k\|^2}{1-\gamma}\left(\max_{x,s,a}|\bar{C}_\lambda(x,s,a)|^2 + 2\max_{x,s}\|\phi(x,s)\|^2 \sup_k \|v_k\|^2\right) \\
&\leq 64\frac{K^2\|\theta_k\|^2}{1-\gamma}\left(\left|\max\left\{C_{\max}, \frac{2\lambda D_{\max}}{\gamma^T(1-\alpha)(1-\gamma)}\right\}\right|^2 + 2\max_{x,s}\|\phi(x,s)\|^2 \sup_k \|v_k\|^2\right),
\end{aligned}$$

for some Lipschitz constant K, where the indicator function in the second line can be explained by the fact that $\pi_\gamma^{\theta_k}(x,s,u) = 0$ whenever $\mu(u \mid x,s;\theta_k) = 0$ and because the expectation is taken with respect to $\pi_\gamma^{\theta_k}$. The third inequality uses Assumption 3 and the fact that μ takes on finitely-many values (and thus its nonzero values are bounded away from zero). Finally, $\sup_k \|v_k\| < \infty$ follows from the Lyapunov analysis in the critic update.

Second, note that

$$\delta\theta_\epsilon \leq \frac{(1+\gamma)\|\psi_{\theta_k}\|_\infty}{(1-\gamma)^2}\epsilon_{\theta_k}(v^*(\theta_k)), \tag{75}$$

where $\psi_\theta(x,s,a) = \nabla_\theta \log \mu(a|x,s;\theta)$ is the "compatible feature." The last inequality is due to the fact that since π_γ^θ is a probability measure, convexity of quadratic functions implies

$$
\begin{aligned}
&\sum_{x',a',s'} \pi_\gamma^\theta(x',s',a'|x_0=x^0,s_0=\nu^*(\theta))(A^\theta(x',s',a') - \tilde{A}^{\theta,v}(x',s',a')) \\
\leq &\sum_{x',a',s'} \pi_\gamma^\theta(x',s',a'|x_0=x^0,s_0=\nu^*(\theta))(Q^\theta(x',s',a') - \tilde{Q}^{\theta,v}(x',s',a')) \\
&+ \sum_{x',s'} d_\gamma^\theta(x',s'|x_0=x^0,s_0=\nu^*(\theta))(V^\theta(x',s') - \widetilde{V}^{\theta,v}(x',s')) \\
= &\gamma \sum_{x',a',s'} \pi_\gamma^\theta(x',s',a'|x_0=x^0,s_0=\nu^*(\theta)) \sum_{x'',s''} \bar{P}(x'',s''|x',s',a')(V^\theta(x'',s'') - \phi^\top(x'',s'')v) \\
&+ \sqrt{\sum_{x',s'} d_\gamma^\theta(x',s'|x_0=x^0,s_0=\nu^*(\theta))(V^\theta(x',s') - \widetilde{V}^{\theta,v}(x',s'))^2} \\
\leq &\gamma \sqrt{\sum_{x',a',s'} \pi_\gamma^\theta(x',s',a'|x_0=x^0,s_0=\nu^*(\theta)) \sum_{x'',s''} \bar{P}(x'',s''|x',s',a')(V^\theta(x'',s'') - \phi^\top(x'',s'')v)^2} \\
&+ \frac{\epsilon_\theta(v)}{1-\gamma} \\
\leq &\sqrt{\sum_{x'',s''} \left(d_\gamma^\theta(x'',s''|x^0,\nu^*(\theta)) - (1-\gamma)1\{x^0=x'',\nu=s''\}\right)(V^\theta(x'',s'') - \phi^\top(x'',s'')v)^2} + \frac{\epsilon_\theta(v)}{1-\gamma} \\
\leq &\left(\frac{1+\gamma}{1-\gamma}\right)\epsilon_\theta(v).
\end{aligned}
$$

Then by Lemma 22, if the γ-occupation measure π_γ^θ is used to generate samples (x_k, s_k, a_k), one obtains $\mathbb{E}\left[\delta\theta_{k+1} \mid \mathcal{F}_{\theta,k}\right] = 0$, where $\mathcal{F}_{\theta,k} = \sigma(\theta_m,\, \delta\theta_m,\, m \leq k)$ is the filtration generated by different independent trajectories. On the other hand, $|\delta\theta_\epsilon| \to 0$ as $\epsilon_{\theta_k}(v^*(\theta_k)) \to 0$. Therefore, the θ-update in (74) is a stochastic approximation of the continuous system $\theta(t)$, described by the ODE

$$\dot{\theta} = \Upsilon_\theta\left[-\nabla_\theta L(\nu,\theta,\lambda)|_{\nu=\nu^*(\theta)}\right],$$

with an error term that is a sum of a vanishing bias and a Martingale difference. Thus, the convergence analysis of θ follows analogously from Step 2 in the proof of Theorem 7, i.e., the sequence of θ-updates in (21) converges to θ^* almost surely, where θ^* is the equilibrium point of the continuous system θ satisfying

$$\Upsilon_\theta\left[-\nabla_\theta L(\nu,\theta,\lambda)|_{\nu=\nu^*(\theta)}\right] = 0. \tag{76}$$

■

Step 4 (Local minimum) The proof that (θ^*, ν^*) is a local minimum follows directly from the arguments in Step 3 in the proof of Theorem 7.

Step 5 (λ-update and convergence to saddle point) Note that the λ-update converges on the slowest time scale, thus, (20) may be rewritten using the converged $v^*(\lambda)$, $\theta^*(\lambda)$, and $\nu^*(\lambda)$ as

$$\lambda_{k+1} = \Gamma_\Lambda\left(\lambda_k + \zeta_1(k)\left(\nabla_\lambda L(\nu,\theta,\lambda)\Big|_{\theta=\theta^*(\lambda),\nu=\nu^*(\lambda),\lambda=\lambda_k} + \delta\lambda_{k+1}\right)\right), \tag{77}$$

where

$$\delta\lambda_{k+1} = -\nabla_\lambda L(\nu,\theta,\lambda)\Big|_{\theta=\theta^*(\lambda),\nu=\nu^*(\lambda),\lambda=\lambda_k} + \left(\nu^*(\lambda_k) + \frac{(-s_k)^+}{(1-\alpha)(1-\gamma)}\mathbf{1}\{x_k = x_{\text{Tar}}\} - \beta\right). \tag{78}$$

From (41), $\nabla_\lambda L(\nu,\theta,\lambda)$ does not depend on λ. Similar to the θ-update, one can easily show that $\delta\lambda_{k+1}$ is square integrable, specifically,

$$\mathbb{E}[\|\delta\lambda_{k+1}\|^2 \mid \mathcal{F}_{\lambda,k}] \le 8\left(\beta^2 + \left(\frac{D_{\max}}{1-\gamma}\right)^2 + \left(\frac{2D_{\max}}{(1-\gamma)^2(1-\alpha)}\right)^2\right),$$

where $\mathcal{F}_{\lambda,k} = \sigma\big(\lambda_m,\, \delta\lambda_m,\, m \le k\big)$ is the filtration of λ generated by different independent trajectories. Similar to the θ-update, using the γ-occupation measure π_γ^θ, one obtains $\mathbb{E}\left[\delta\lambda_{k+1} \mid \mathcal{F}_{\lambda,k}\right] = 0$. As above, the λ-update is a stochastic approximation for the continuous system $\lambda(t)$ described by the ODE

$$\dot\lambda = \Upsilon_\lambda\left[\nabla_\lambda L(\nu,\theta,\lambda)\Big|_{\theta=\theta^*(\lambda),\nu=\nu^*(\lambda)}\right],$$

with an error term that is a Martingale difference. Then the λ-convergence and the analysis of local optima follow from analogous arguments in Steps 4 and 5 in the proof of Theorem 7.