Sequential Decision Making with Coherent Risk

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, Shie Mannor

***Abstract*—We provide sampling-based algorithms for optimization under a coherent-risk objective. The class of coherent-risk measures is widely accepted in finance and operations research, among other fields, and encompasses popular risk-measures such as conditional value at risk and mean-semi-deviation. Our approach is suitable for problems in which tuneable parameters control the *distribution* of the cost, such as in reinforcement learning or approximate dynamic programming with a parameterized policy. Such problems cannot be solved using previous approaches. We consider both static risk measures and time-consistent dynamic risk measures. For static risk measures, our approach is in the spirit of *policy gradient* methods, while for the dynamic risk measures, we use *actor-critic* type algorithms.**

***Index Terms*—Coherent Risk, Dynamic Programming, Markov Decision Processes, Policy Gradient.**

I. Introduction

We consider stochastic optimization problems in which the objective involves a *risk measure* of the random cost, in contrast to the typical *expected* cost objective. Such problems are important when the decision-maker wishes to manage the *variability* of the cost, in addition to its expected outcome, and are standard in various applications in finance and operations research (OR).

There are various approaches to quantify the risk of a random cost, such as the celebrated Markowitz mean-variance model [1], or the more recent Value at Risk (VaR) and Conditional Value at Risk (CVaR) [2]. The preference of one risk measure over another depends on factors such as sensitivity to rare events, ease of estimation from data, and computational tractability of the optimization problem, and in general, there is no single choice that dominates the rest. However, the highly influential paper of Artzner et al. [3] identified a set of natural properties that are desirable for a risk measure. Risk measures that satisfy these properties are termed *coherent* and have obtained widespread acceptance in finance and OR applications, among others.

When the optimization problem is sequential, such as when we would like to solve a Markov decision process (MDP), another desirable property of a risk measure is *time consistency*. A time-consistent risk measure satisfies a “dynamic programming” style property: if a strategy is risk-optimal for an n-stage problem, then the component of the policy from the t-th time until the end (where $t < n$) is also risk-optimal (see principle of optimality in [4]). The recently proposed class of dynamic Markov coherent risk measures [5] satisfies both the coherence and time consistency properties.

In this work, we are interested in solving general problems of the form

$$\min_{\theta} \ \rho(C;\theta),$$

where C is a random cost, controlled by the tuneable parameter vector θ, and ρ is a *coherent* risk measure. We consider both time-consistent *dynamic* Markov coherent risk measures and standard *static* coherent risk-measures without explicit temporal dependence.

For the static case and when the cost is of the form $C = f_\theta(Z)$, where f_θ is a deterministic function of the random variable Z that is independent of θ, the optimization may be formulated as a stochastic program (owing to the special mathematical properties of coherent risk measures) and solved using standard sampling approaches [6]. Such a cost structure is appropriate for certain domains, such as portfolio optimization, in which the investment strategy generally does not affect the asset prices. However, in many important domains, such as queueing systems, resource allocation, and reinforcement learning, the tuneable parameters also control the *distribution* of the random outcomes. This is the case we consider in this paper for which the existing approaches do not apply.

In this work, we develop sampling-based algorithms for estimating the gradient $\nabla_\theta \rho(C;\theta)$, when ρ is either a static or dynamic coherent risk measure, and θ controls the distribution of Z. The optimization is then carried out using the standard stochastic gradient descent techniques. A particular application of our approach is to risk-sensitive MDPs, where the optimization is over a parametric set of polices (θ is the policy parameter). Our proposed algorithm for the static risk is in the spirit of *policy gradient* algorithms [7], while the one for the dynamic risk has *actor-critic* style [8]. Policy gradient and actor critic algorithms have been applied to various domains such as robotics, network routing, and finance [9], [10], [11], and are particularly suitable for problems with large or continuous state and action spaces [12]. Such problems often pose a challenge for standard dynamic programming algorithms due to the curse of dimensionality [13].

Our contributions can be listed as follows:

- A new formula for the gradient of static coherent risk that is convenient for sampling-based approximation.
- A sampling-based algorithm for the gradient of general static coherent risk and a consistency proof.
- A new policy-gradient theorem for Markov coherent risk that relates the gradient to a suitable *value function*.
- A corresponding actor-critic algorithm for the gradient of dynamic Markov coherent risk with value-function approximation. We prove the consistency of the gradient and analyze the sensitivity of the value-function to approximation errors.

Related Work: For the case of static-risk, our approach is similar in spirit to *policy gradient methods* ([7]; a.k.a. the

Manuscript received... revised... accepted... This work was supported by the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement 306638 (SUPREL). Yinlam Chow was partially supported by the Croucher Foundation Doctoral Scholarship. Aviv Tamar was partially supported by the Viterbi Scholarship, Technion. (A. Tamar and Y. Chow contributed equally equally to this work.)

A. Tamar is with Department of Electrical Engineering & Computer Sciences, UC Berkeley, email: avivt@berkeley.edu.

Y. Chow is with Institute for Computational and Mathematical Engineering, Stanford University, email: ychow@stanford.edu.

M. Ghavamzadeh is with Adobe Research, on leave of absence from INRIA, email: mohammad.ghavamzadeh@inria.fr.

S. Mannor is with Department of Electrical Engineering, Technion, email: shie@ee.technion.ac.il.

DOI:...

likelihood-ratio method in the simulation-based optimization literature; [14]), and may be seen as an extension of these methods to coherent risk objectives. Optimization of coherent risk measures was thoroughly investigated by [6], [15] for the case discussed above, in which θ does not control the distribution of Z. For the case of MDPs and dynamic risk, [5] proposed a dynamic programming approach. This approach does not scale-up to large MDPs, due to the "curse of dimensionality". The work of [16], [17], [18], [19], [20], [21] on robust MDPs (sometime also referred to as distributionally robust MDPs) is relevant since an MDP with a dynamic coherent risk objective is essentially a robust MDP. For robust MDPs in which the state-space is finite and moderately-sized, the approaches in [16], [17], [18], [19] may be used to calculate an optimal solution. Our method, on the other hand, can be applied to large or continuous state spaces, by employing a policy gradient approach to search for a policy in a parameterized policy space. We note that [20], [21] considered function approximation, but only in the value function. For many problems, approximation in the policy space is more suitable (see, e.g., [22]). Our sampling-based RL-style approach is suitable for approximations both in the policy and value function, and scales-up to large or continuous MDPs. We do, however, make use of a technique from [20] in a part of our approach.

Risk-sensitive optimization in RL for specific risk functions has been studied recently by several authors. [23] studied exponential utility functions, [24] and [25] studied mean-variance models, [26] and [27] studied CVaR in the static setting, and [28] and [29] studied dynamic coherent risk for systems with linear dynamics. Our paper presents a general method *for the whole class* of coherent risk measures (both static and dynamic) and is neither limited to a specific choice within that class nor to particular system dynamics. For the special case of CVaR, we obtain results similar to [26], [27], but under weaker assumptions and simpler derivations.

II. Preliminaries

Consider a probability space $(\Omega, \mathcal{F}, P_\theta)$, where Ω is the set of outcomes (sample space), $\mathcal{F}$ is a σ-algebra over Ω representing the set of events we are interested in, and $P_\theta \in \mathcal{B}$, where $\mathcal{B} := \left\{\xi : \int_{\omega\in\Omega} \xi(\omega) = 1, \xi \geq 0\right\}$ is the set of probability distributions, is a probability measure over $\mathcal{F}$ parameterized by some tuneable parameter $\theta \in \mathbb{R}^K$. In the following, we suppress the notation of θ in θ-dependent quantities.

To ease the technical exposition, in this paper we restrict our attention to finite probability spaces, i.e., Ω has a finite number of elements. Our results can be extended to the L_p-normed spaces without loss of generality, but the details are omitted for brevity.

We denote by $\mathcal{Z}$ the space of random variables $Z : \Omega \mapsto (-\infty, \infty)$ defined over the probability space $(\Omega, \mathcal{F}, P_\theta)$. In this paper, a random variable $Z \in \mathcal{Z}$ is interpreted as a cost, and thus, the smaller the realization of Z, the better. For $Z, W \in \mathcal{Z}$, we denote by $Z \leq W$ the point-wise partial order, i.e., $Z(\omega) \leq W(\omega)$, $\forall \omega \in \Omega$, and by $\mathbb{E}_\xi[Z] \doteq \sum_{\omega\in\Omega} P_\theta(\omega)\xi(\omega)Z(\omega)$ a ξ-weighted expectation of Z.

An MDP is a tuple $\mathcal{M} = (\mathcal{X}, \mathcal{A}, C, P, \gamma, x_0)$, where $\mathcal{X}$ and $\mathcal{A}$ are the state and action spaces; $C(x) \in [-C_{\max}, C_{\max}]$ is a bounded, deterministic, and state-dependent cost; $P(\cdot|x, a)$ is the transition probability distribution; γ is a discount factor; and x_0 is the initial state.[1] Actions are chosen according to a θ-parameterized stationary Markov[2] policy $\mu_\theta(\cdot|x)$. 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 θ or μ, and we use μ and θ interchangeably in the paper. We denote by $x_0, a_0, \ldots, x_T, a_T$ a trajectory of length T drawn by following the policy μ_θ in the MDP.

A. Coherent Risk Measures

A *risk measure* is a function $\rho : \mathcal{Z} \to \mathbb{R}$ that maps an uncertain outcome Z to the extended real line $\mathbb{R} \cup \{+\infty, -\infty\}$, e.g., the expectation $\mathbb{E}[Z]$ or the conditional value-at-risk (CVaR) $\min_{\nu\in\mathbb{R}} \left\{\nu + \frac{1}{\alpha}\mathbb{E}\big[(Z - \nu)^+\big]\right\}$. A risk measure is called *coherent*, if it satisfies the following conditions for all $Z, W \in \mathcal{Z}$ [3]:

A1 Convexity: $\forall \lambda \in [0, 1]$, $\rho\big(\lambda Z + (1 - \lambda)W\big) \leq \lambda\rho(Z) + (1 - \lambda)\rho(W)$;
A2 Monotonicity: if $Z \leq W$, then $\rho(Z) \leq \rho(W)$;
A3 Translation invariance: $\forall a \in \mathbb{R}$, $\rho(Z+a) = \rho(Z)+a$;
A4 Positive homogeneity: if $\lambda \geq 0$, then $\rho(\lambda Z) = \lambda\rho(Z)$.

Intuitively, these condition ensure the "rationality" of single-period risk assessments: A1 ensures that diversifying an investment will reduce its risk; A2 guarantees that an asset with a higher cost for every possible scenario is indeed riskier; A3, also known as 'cash invariance', means that the deterministic part of an investment portfolio does not contribute to its risk; the intuition behind A4 is that doubling a position in an asset doubles its risk. We refer the readers to [3] for a more detailed motivation of coherent risk.

The following representation theorem [6] shows an important property of coherent risk measures that is fundamental to our gradient-based approach.

Theorem II.1. *A risk measure $\rho : \mathcal{Z} \to \mathbb{R}$ is coherent if and only if there exists a convex bounded and closed set $\mathcal{U} \subset \mathcal{B}$ such that*[3]

$$\rho(Z) = \max_{\xi\,:\,\xi P_\theta \in \mathcal{U}(P_\theta)} \mathbb{E}_\xi[Z]. \tag{1}$$

The result essentially states that any coherent risk measure is an expectation w.r.t. a worst-case density function ξP_θ, i.e., a re-weighting of P_θ by ξ, chosen adversarially from a suitable set of test density functions $\mathcal{U}(P_\theta)$, referred to as *risk envelope*. Moreover, a coherent risk measure is *uniquely represented* by its risk envelope. In the sequel, we shall interchangeably refer to coherent risk measures either by their explicit functional representation, or by their corresponding risk-envelope.

In this paper, we assume that the risk envelope $\mathcal{U}(P_\theta)$ is given in a canonical convex programming formulation and satisfies the following conditions.

[1] Our results may be easily extended to random costs, state-action dependent costs, and random initial states.

[2] For Markov coherent risk, the class of optimal policies is stationary Markov [5], while this is not necessarily true for static risk. Our results can be extended to history-dependent policies or stationary Markov policies on a state space augmented with accumulated cost. The latter has shown to be sufficient for optimizing the CVaR risk [30].

[3] When we study risk in MDPs, the risk envelope $\mathcal{U}(P_\theta)$ in Eq. 1 also depends on the state x.

Assumption II.2 (The General Form of a Risk Envelope)**.** *For each given policy parameter $\theta \in \mathbb{R}^K$, the risk envelope $\mathcal{U}$ of a coherent risk measure can be written as*

$$\mathcal{U}(P_\theta) = \Big\{\xi P_\theta : \; g_e(\xi, P_\theta) = 0, \; \forall e \in \mathcal{E}, \; f_i(\xi, P_\theta) \leq 0, \\ \forall i \in \mathcal{I}, \sum_{\omega \in \Omega} \xi(\omega) P_\theta(\omega) = 1, \; \xi(\omega) \geq 0 \Big\}, \tag{2}$$

where each constraint $g_e(\xi, P_\theta)$ is an affine function in ξ, each constraint $f_i(\xi, P_\theta)$ is a convex function in ξ, and there exists a strictly feasible point $\overline{\xi}$. $\mathcal{E}$ and $\mathcal{I}$ here denote the finite sets of equality and inequality constraints, respectively[4]*. Furthermore, for any given $\xi \in \mathcal{B}$, $f_i(\xi, p)$ and $g_e(\xi, p)$ are twice differentiable in p, and there exists a $M > 0$ such that*

$$\max\left\{\max_{i \in \mathcal{I}} \left| \frac{df_i(\xi, p)}{dp(\omega)} \right|, \max_{e \in \mathcal{E}} \left| \frac{dg_e(\xi, p)}{dp(\omega)} \right| \right\} \leq M, \; \forall \omega \in \Omega.$$

Assumption II.2 implies that the risk envelope $\mathcal{U}(P_\theta)$ is known in an *explicit* form. From Theorem 6.6 of [6], in the case of a finite probability space, ρ is a coherent risk measure if and only if $\mathcal{U}(P_\theta)$ is a convex and compact set. This justifies the affine assumption of g_e and the convex assumption of f_i. Moreover, the additional assumption on the smoothness of the constraints holds for many popular coherent risk measures, such as CVaR, mean-semi-deviation, and spectral risk measures [31].

B. Dynamic Risk Measures

The risk measures defined above do not take into account the temporal structure of the random variable, such as when it is associated with the return of a trajectory in the case of MDPs. In this sense, such risk measures are called *static*. On the other hand, *dynamic* risk measures explicitly take into account the temporal nature of the stochastic outcome. A primary motivation for considering such measures is the issue of *time consistency*, usually defined as follows [5]: if a certain outcome is considered less risky in all states of the world at stage $t + 1$, then it should also be considered less risky at stage t. Example 2.1 in [32] shows the importance of time consistency in the evaluation of risk in a dynamic setting. It illustrates that for multi-period decision-making, optimizing a static measure can lead to "time-inconsistent" behavior. Similar paradoxical results could be obtained with other risk metrics; we refer the readers to [5] and [32] for further insights.

Markov Coherent Risk Measures.: Markov risk measures were first introduced in [5] and have constituted a useful class of dynamic time-consistent risk measures that are important to our study of risk in MDPs. For a T-length horizon and MDP $\mathcal{M}$, the Markov coherent risk measure $\rho_T(\mathcal{M})$ is

$$\rho_T(\mathcal{M}) = C(x_0) + \gamma\rho\bigg(C(x_1) + \ldots + \gamma\rho\Big(C(x_{T-1}) + \gamma\rho\big(C(x_T)\big)\Big)\bigg). \tag{3}$$

Here ρ is a static coherent risk measure that satisfies Assumption II.2 and $x_0, \ldots, x_T$ is a trajectory drawn from the MDP $\mathcal{M}$ under policy μ_θ. It is important to note that in (3), each static coherent risk measure ρ at state $x \in \mathcal{X}$ is induced by the transition probability $P_\theta(\cdot|x) = \sum_{a \in \mathcal{A}} P(\cdot|x, a)\mu_\theta(a|x)$. We also define $\rho_\infty(\mathcal{M}) \doteq \lim_{T\to\infty} \rho_T(\mathcal{M})$, which is well-defined since $\gamma < 1$ and the cost is bounded. We further assume that ρ in (3) is a *Markov risk* measure, i.e., the evaluation of each static coherent risk measure ρ is not allowed to depend on the whole past.

III. Problem Formulation

In this paper, we are interested in solving two risk-sensitive optimization problems. Given a random variable Z and a static coherent risk measure ρ as defined in Section II, the static risk problem (SRP) is given by

$$\min_\theta \quad \rho(Z). \tag{4}$$

For example, in an RL setting, Z may correspond to the cumulative discounted cost $Z = C(x_0) + \gamma C(x_1) + \cdots + \gamma^T C(x_T)$ of a trajectory $x_0, \ldots, x_T$ induced by an MDP with a policy parameterized by θ.

For an MDP $\mathcal{M}$ and a dynamic Markov coherent risk measure ρ_T as defined by Eq. 3, the dynamic risk problem (DRP) is given by

$$\min_\theta \quad \rho_\infty(\mathcal{M}). \tag{5}$$

Except for very limited cases, there is no reason to hope that neither the SRP in (4) nor the DRP in (5) should be tractable problems, since the dependence of the risk measure on θ may be complex and non-convex. In this work, we aim towards a more modest goal and search for a *locally* optimal θ. Thus, the main problem that we are trying to solve is how to calculate the gradients of the SRP's and DRP's objective functions

$$\nabla_\theta \rho(Z) \qquad \text{and} \qquad \nabla_\theta \rho_\infty(\mathcal{M}).$$

We are interested in non-trivial cases in which the gradients cannot be calculated analytically. In the static case, this would correspond to a non-trivial dependence of Z on θ. For dynamic risk, we also consider cases where the state space is too large for a tractable computation. Our approach to deal with such difficult cases is through sampling. We assume that in the static case, we may obtain i.i.d. samples of the random variable Z. For the dynamic case, we assume that for each state and action (x, a) of the MDP, we may obtain i.i.d. samples of the next state $x' \sim P(\cdot|x, a)$. We show that sampling may indeed be used in both cases to devise suitable gradient estimators.

Finally, to solve the SRP and DRP problems, a gradient estimate may be plugged into a standard stochastic gradient descent (SGD) algorithm to learn a locally optimal solution to (4) and (5). From the structure of the dynamic risk in (3), one may think that a gradient estimator for $\rho(Z)$ may help us to estimate the gradient $\nabla_\theta \rho_\infty(\mathcal{M})$. We follow this idea and begin with estimating the gradient in the case of static risk.

IV. Gradient Formula for Static Risk

In this section, we consider a static coherent risk measure $\rho(Z)$ and propose sampling-based estimators for $\nabla_\theta \rho(Z)$. We make the following assumption on the policy parametrization, which is standard in the policy gradient literature [22].

Assumption IV.1. *The likelihood ratio $\nabla_\theta \log P(\omega)$ is well-defined (i.e., $\log P(\omega)$ is a differentiable function in θ in cases when $P(\omega) \neq 0$) and bounded for all $\omega \in \Omega$.*

[4] While generalizing $\mathcal{E}$ and $\mathcal{I}$ to countably infinite sets is an interesting research direction from the theoretical standpoint, for practical purposes we assume the risk envelope only contains a finite number of inequality and equality constraints. Notice that all coherent risk metrics we are aware of in the literature (see, e.g., [6]) are already captured by the above risk envelope.

Moreover, our approach implicitly assumes that given some $\omega \in \Omega$, $\nabla_\theta \log P(\omega)$ may be easily calculated, which is a standard requirement for policy-gradient algorithms. In many applications, such as in [9], [10], [11], the probability $P(\omega)$ can be decomposed into a product of an *unknown* stochastic variable that is independent of the control parameter θ, encompassing the stochastic dynamics in the problem, and a *known* random variable that represents decisions and depends on θ. In such cases, $\nabla_\theta \log P(\omega)$ will only require the gradient of the *known* decision component in $P(\omega)$. Our experiments in Section VI provide an illustration of this idea, and a comprehensive treatment can be found in the survey by Fu [33].

Using Theorem II.1 and Assumption II.2, for each θ, we have that $\rho(Z)$ is the solution to the convex optimization problem (1) (for that value of θ). The Lagrangian function of (1), denoted by $L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$, may be written as

$$\begin{aligned} &L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \\ &= \sum_{\omega\in\Omega} \xi(\omega) P_\theta(\omega) Z(\omega) - \lambda^{\mathcal{P}} \left(\sum_{\omega\in\Omega} \xi(\omega) P_\theta(\omega) - 1 \right) \\ &- \sum_{e\in\mathcal{E}} \lambda^{\mathcal{E}}(e) g_e(\xi, P_\theta) - \sum_{i\in\mathcal{I}} \lambda^{\mathcal{I}}(i) f_i(\xi, P_\theta). \end{aligned} \tag{6}$$

The convexity of (1) and its strict feasibility due to Assumption II.2 implies that $L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$ has a non-empty set of saddle points $\mathcal{S}$. The next theorem presents a formula for the gradient $\nabla_\theta \rho(Z)$. As we shall subsequently show, this formula is particularly convenient for devising sampling-based estimators for $\nabla_\theta \rho(Z)$.

Theorem IV.2. *Let Assumptions II.2 and IV.1 hold. For any saddle point $(\xi^*_\theta, \lambda^{*,\mathcal{P}}_\theta, \lambda^{*,\mathcal{E}}_\theta, \lambda^{*,\mathcal{I}}_\theta) \in \mathcal{S}$ of* (6)*, we have that*

$$\begin{aligned} \nabla_\theta \rho(Z) = &\mathbb{E}_{\xi^*_\theta} \left[\nabla_\theta \log P(\omega)(Z - \lambda^{*,\mathcal{P}}_\theta) \right] \\ &- \sum_{e\in\mathcal{E}} \lambda^{*,\mathcal{E}}_\theta(e) \nabla_\theta g_e(\xi^*_\theta; P_\theta) \\ &- \sum_{i\in\mathcal{I}} \lambda^{*,\mathcal{I}}_\theta(i) \nabla_\theta f_i(\xi^*_\theta; P_\theta). \end{aligned}$$

The proof of this theorem, given in Appendix B, involves an application of the Envelope theorem [34] and a standard "likelihood-ratio" trick. We now demonstrate the utility of Theorem IV.2 with several examples. The details of deriving these results are in Appendix A.

A. Example 1: CVaR

The CVaR at level $\alpha \in [0, 1]$ of a random variable Z, denoted by $\rho_{\text{CVaR}}(Z; \alpha)$, is a very popular coherent risk measure [2], defined as

$$\rho_{\text{CVaR}}(Z; \alpha) \doteq \inf_{t\in\mathbb{R}} \left\{ t + \alpha^{-1} \mathbb{E}\left[(Z - t)_+\right] \right\}.$$

When Z is continuous, $\rho_{\text{CVaR}}(Z; \alpha)$ is well-known to be the mean of the α-tail distribution of Z, $\mathbb{E}\left[Z \mid Z > q_\alpha\right]$, where q_α is a $(1-\alpha)$-quantile of Z. Thus, selecting a small α makes CVaR particularly sensitive to rare, but very high costs.

The risk envelope for CVaR is known to be [6]

$$\mathcal{U} = \left\{ \xi P_\theta : \xi(\omega) \in [0, \alpha^{-1}], \quad \sum_{\omega\in\Omega} \xi(\omega) P_\theta(\omega) = 1 \right\}.$$

Furthermore, it is shown in [6] that the saddle points of (6) satisfy $\xi^*_\theta(\omega) = \alpha^{-1}$ when $Z(\omega) > \lambda^{*,\mathcal{P}}_\theta$, and $\xi^*_\theta(\omega) = 0$ when $Z(\omega) < \lambda^{*,\mathcal{P}}_\theta$, where $\lambda^{*,\mathcal{P}}_\theta$ is any $(1-\alpha)$-quantile of Z. Plugging this result into Theorem IV.2, we can show that

$$\nabla_\theta \rho_{\text{CVaR}}(Z; \alpha) = \mathbb{E}\left[\nabla_\theta \log P(\omega)(Z - q_\alpha) \mid Z(\omega) > q_\alpha\right]. \tag{7}$$

This formula was recently proved in [27] for the case of continuous distributions by an explicit calculation of the conditional expectation and under several additional smoothness assumptions. Here we show that it holds regardless of these assumptions and in the discrete case as well. Our proof is also considerably simpler.

B. Example 2: Mean-Semideviation

The semi-deviation of a random variable Z is defined as $\mathbb{SD}[Z] \doteq \left(\mathbb{E}\left[(Z - \mathbb{E}[Z])_+^2\right]\right)^{1/2}$. The semi-deviation captures the variation of the cost only *above its mean*, and is an appealing alternative to the standard deviation, which does not distinguish between the variability of upside and downside deviations. For some $\alpha \in [0, 1]$, the *mean-semideviation* risk measure is defined as $\rho_{\text{MSD}}(Z; \alpha) \doteq \mathbb{E}[Z] + \alpha \mathbb{SD}[Z]$, and is a coherent risk measure [6]. We have the following result:

Proposition IV.3. *Under Assumption IV.1, with $\nabla_\theta \mathbb{E}[Z] = \mathbb{E}[\nabla_\theta \log P(\omega) Z]$, we have*

$$\nabla_\theta \rho_{MSD}(Z; \alpha) = \nabla_\theta \mathbb{E}[Z] + \frac{\alpha \mathbb{E}\left[(Z - \mathbb{E}[Z])_+ (\nabla_\theta \log P(\omega)(Z - \mathbb{E}[Z]) - \nabla_\theta \mathbb{E}[Z])\right]}{\mathbb{SD}(Z)}.$$

The proof of Proposition IV.3 is given in Appendix A. Proposition IV.3 can be used to devise a sampling-based estimator for $\nabla_\theta \rho_{\text{MSD}}(Z; \alpha)$ by replacing all the expectations with sample averages. The resulting algorithm, which we term `GMSD` (Gradient of Mean Semi-Deviation), is described next. Let $z_1, \dots, z_N \sim P_\theta$ denote an i.i.d. sequence of samples. We propose the following estimates:

$$\begin{aligned} \widehat{\mathbb{E}[Z]} &= \frac{1}{N} \sum_{i=1}^N z_i, \\ \widehat{\mathbb{SD}(Z)} &= \left(\frac{1}{N} \sum_{i=1}^N (z_i - \widehat{\mathbb{E}[Z]})_+^2 \right)^{1/2}, \\ \widehat{\nabla_\theta \mathbb{E}[Z]} &= \frac{1}{N} \sum_{i=1}^N \nabla_\theta \log P(z_i) z_i, \\ \widehat{\nabla_\theta \rho(Z)} &= \widehat{\nabla_\theta \mathbb{E}[Z]} + \frac{\alpha}{\widehat{\mathbb{SD}(Z)}} \frac{1}{N} \sum_{i=1}^N (z_i - \widehat{\mathbb{E}[Z]})_+ \times \\ &\quad \times \left(\nabla_\theta \log P(z_i)(z_i - \widehat{\mathbb{E}[Z]}) - \widehat{\nabla_\theta \mathbb{E}[Z]} \right). \end{aligned}$$

In Section VI, we provide a numerical illustration of optimization with a mean-semideviation objective, using the `GMSD` algorithm.

C. General Gradient Estimation Algorithm

In the two previous examples, we obtained a gradient formula by *analytically* calculating the Lagrangian saddle point (6) and plugging it into the formula of Theorem IV.2. We now consider a general coherent risk $\rho(Z)$ for which, in contrast to the CVaR and mean-semideviation cases, the

Lagrangian saddle-point is not known analytically. *We only assume that we know the structure of the risk-envelope* as given by (2). We show that in this case, $\nabla_\theta \rho(Z)$ may be estimated using a *sample average approximation* (SAA; [6]) of the formula in Theorem IV.2.

Assume that we are given N i.i.d. samples $\omega_i \sim P_\theta$, $i = 1, \dots, N$, and let $P_{\theta;N}(\omega) \doteq \frac{1}{N}\sum_{i=1}^N \mathbb{I}\{\omega_i = \omega\}$ denote the corresponding empirical distribution. Also, let the *sample risk envelope* $\mathcal{U}(P_{\theta;N})$ be defined according to Eq. 2 with P_θ replaced by $P_{\theta;N}$. Consider the following SAA version of the optimization in Eq. 1:

$$\rho_N(Z) = \max_{\xi : \xi P_{\theta;N} \in \mathcal{U}(P_{\theta;N})} \sum_{i \in 1,\dots,N} P_{\theta;N}(\omega_i)\xi(\omega_i)Z(\omega_i). \quad (8)$$

Note that (8) defines a convex optimization problem with $\mathcal{O}(N)$ variables and constraints. In the following, we assume that a solution to (8) may be computed efficiently using standard convex programming tools such as interior point methods [35]. Let $\xi^*_{\theta;N}$ denote a solution to (8) and $\lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}$ denote the corresponding KKT multipliers, which can be obtained from the convex programming algorithm [35]. We propose the following estimator for the gradient in Theorem IV.2:

$$\nabla_{\theta;N}\rho(Z) = \sum_{i=1}^N P_{\theta;N}(\omega_i)\xi^*_{\theta;N}(\omega_i)\nabla_\theta \log P(\omega_i)(Z(\omega_i) - \lambda^{*,\mathcal{P}}_{\theta;N})$$
$$-\sum_{e\in\mathcal{E}}\lambda^{*,\mathcal{E}}_{\theta;N}(e)\nabla_\theta g_e(\xi^*_{\theta;N}; P_{\theta;N}) - \sum_{i\in\mathcal{I}}\lambda^{*,\mathcal{I}}_{\theta;N}(i)\nabla_\theta f_i(\xi^*_{\theta;N}; P_{\theta;N}).$$

Thus, our gradient estimation algorithm is a two-step procedure involving *both sampling and convex programming*. In the following, we show that under some conditions on the set $\mathcal{U}(P_\theta)$, $\nabla_{\theta;N}\rho(Z)$ is a consistent estimator of $\nabla_\theta\rho(Z)$. The proof is in Appendix C.

Proposition IV.4. *Let Assumptions II.2 and IV.1 hold. Suppose there exists a compact set $C = C_\xi \times C_\lambda$ such that: (I) The set of Lagrangian saddle points $\mathcal{S} \subset C$ is non-empty and bounded. (II) The functions $f_e(\xi, P_\theta)$ for all $e \in \mathcal{E}$ and $f_i(\xi, P_\theta)$ for all $i \in \mathcal{I}$ are finite-valued and continuous (in ξ) on C_ξ. (III) For N large enough, the set $\mathcal{S}_N$ is non-empty and $\mathcal{S}_N \subset C$ w.p. 1. Further assume that: (IV) If $\xi_N P_{\theta;N} \in \mathcal{U}(P_{\theta;N})$ and ξ_N converges w.p. 1 to a point ξ, then $\xi P_\theta \in \mathcal{U}(P_\theta)$. We then have that* $\lim_{N\to\infty}\rho_N(Z) = \rho(Z)$ *and* $\lim_{N\to\infty}\nabla_{\theta;N}\rho(Z) = \nabla_\theta\rho(Z)$ *w.p. 1.*

The set of assumptions for Proposition IV.4 is large, but rather mild. Note that (I) is implied by the Slater condition of Assumption II.2. For satisfying (III), we need that the risk be well-defined for every empirical distribution, which is a natural requirement. Since $P_{\theta;N}$ always converges to P_θ uniformly on Ω, (IV) essentially requires smoothness of the constraints. We remark that in particular, constraints (I) to (IV) are satisfied for the popular CVaR, mean-semideviation, and spectral risk.

It is interesting to compare the performance of the SAA estimator (9) with the analytical-solution based estimator, as in Sections IV-A and IV-B. In Section VI-B, we report an empirical comparison between the two approaches for the case of CVaR risk, which shows that the two approaches performed very similarly. This is well-expected, since in general, both SAA and standard likelihood-ratio based estimators obey the law-of-large-numbers with variance bound of order $1/\sqrt{N}$ [6].

To summarize this section, we have seen that by exploiting the special structure of coherent risk measures in Theorem II.1 and by the envelope-theorem style result of Theorem IV.2, we are able to derive sampling-based likelihood-ratio style algorithms for estimating the policy gradient $\nabla_\theta\rho(Z)$ of coherent static risk measures. The gradient estimation algorithms developed here for static risk measures will be used as a subroutine in our subsequent treatment of dynamic risk measures.

V. Gradient Formula for Dynamic Risk

In this section, we first derive a new formula for the gradient of a general Markov-coherent dynamic risk measure $\nabla_\theta\rho_\infty(\mathcal{M})$ that involves the *value function* of the risk objective $\rho_\infty(\mathcal{M})$ (e.g., the value function proposed by [5]). This formula extends the well-known "policy gradient theorem" [36], [8] developed for the expected return to Markov-coherent dynamic risk measures. Using this formula, we suggest the following actor-critic style algorithm for estimating $\nabla_\theta\rho_\infty(\mathcal{M})$:

- **Critic:** For a given policy θ, calculate the *risk-sensitive value function* of $\rho_\infty(\mathcal{M})$ (see Section V-B), and
- **Actor:** Using the critic's value function, estimate $\nabla_\theta\rho_\infty(\mathcal{M})$ by sampling (see Section V-C).

The value function proposed by [5] assigns to each state a particular value that encodes the long-term risk starting from that state. When the state space $\mathcal{X}$ is large, calculating the value function by dynamic programming (as suggested by [5]) becomes intractable due to the "curse of dimensionality". For the risk-neutral case, a standard solution to this problem is to approximate the value function by a set of state-dependent features, and use sampling to calculate the parameters of this approximation [37]. In particular, *temporal difference* (TD) learning methods [38] are popular for this purpose, which have been recently extended to robust MDPs by [20]. We use their (robust) TD algorithm and show how our critic use it to approximates the *risk-sensitive* value function. We then discuss how the error introduced by this approximation affects the gradient estimate of the actor.

A. Risk-Sensitive Bellman Equation

Our value-function estimation method is driven by a Bellman-style equation for Markov coherent risks. Let $B(\mathcal{X})$ denote the space of real-valued bounded functions on $\mathcal{X}$, we now define the risk sensitive Bellman operator $T_\theta[V] : B(\mathcal{X}) \mapsto B(\mathcal{X})$ as

$$T_\theta[V](x) = \max_{\xi\in\mathcal{U}(x,P(\cdot|x,\cdot)\mu_\theta(\cdot|x))} \mathbb{E}_{\xi P(\cdot|x_t,\cdot)\mu_\theta(\cdot|x_t)}[C(x,\widehat{a}) + \gamma V(\widehat{x})], \quad (9)$$

where $\widehat{a} \in \mathcal{A}$ and $\widehat{x} \in \mathcal{X}$ are random variables such that $(\widehat{a}, \widehat{x}) \sim \mu_\theta(a|x)P(x'|x,a)$. According to Theorem 1 in [5], the operator T_θ has a unique fixed-point V_θ, i.e., $T_\theta[V_\theta](x) = V_\theta(x), \ \forall x \in \mathcal{X}$, that is equal to the risk objective function induced by θ, i.e., $V_\theta(x_0) = \rho_\infty(\mathcal{M})$. However, when the state space $\mathcal{X}$ is large, exact enumeration of the Bellman equation is intractable due to "curse of dimensionality". Next, we provide an iterative approach to approximate the risk sensitive value function.

B. Value Function Approximation

Consider the linear approximation of the risk-sensitive value function $V_\theta(x) \approx v^\top\phi(x)$, where $\phi(\cdot) \in \mathbb{R}^{\kappa_2}$ is the κ_2-dimensional state-dependent feature vector. Thus, the approximate value function belongs to the low dimensional subspace $\mathcal{V} = \{\Phi v | v \in \mathbb{R}^{\kappa_2}\}$, where $\Phi : \mathcal{X} \to \mathbb{R}^{\kappa_2}$ is a function mapping such that $\Phi(x) = \phi(x)$. The goal of our critric is to find a good approximation of V_θ from simulated trajectories of the MDP. In order to have a well-defined approximation scheme, we first impose the following standard assumption [37].

Assumption V.1. *The mapping Φ has full column rank.*

For a function $y : \mathcal{X} \to \mathbb{R}$, we define its weighted (by d) ℓ_2-norm as $\|y\|_d = \sqrt{\sum_{x'} d(x'|x)y(x')^2}$, where d is a distribution over $\mathcal{X}$. Using this, we define $\Pi : \mathcal{X} \to \mathcal{V}$, the orthogonal projection from $\mathbb{R}$ to $\mathcal{V}$, w.r.t. a norm weighted by the stationary distribution of the policy, $d_\theta(x'|x)$.

Note that the TD methods approximate the value function V_θ with the fixed-point of the projected Bellman operator ΠT_θ, i.e., $\widetilde{V}_\theta(x) = v_\theta^{*\top}\phi(x)$, such that

$$\forall x \in \mathcal{X}, \qquad \widetilde{V}_\theta(x) = \Pi T_\theta[\widetilde{V}_\theta](x).^5 \tag{10}$$

From Eq. 1 that has been derived from Theorem II.1 for dynamic risks, it is easy to see that the risk-sensitive Bellman equation (9) is a robust Bellman equation [17] with uncertainty set $\mathcal{U}(x, P(\cdot|x,\cdot)\mu_\theta(\cdot|x))$. Thus, we may use the TD approximation of the robust Bellman equation proposed by [20] to find an approximation of V_θ. We will need the following assumption analogous to Assumption 2 in [20].

Assumption V.2. *There exists $\kappa \in (0,1)$ such that $\xi(a,x') \le \kappa/\gamma$, for all $\xi(\cdot)P(\cdot|x,\cdot)\mu_\theta(\cdot|x) \in \mathcal{U}(x, P(\cdot|x,\cdot)\mu_\theta(\cdot|x))$ and all $x, x' \in \mathcal{X}$, $a \in \mathcal{A}$.*

Given Assumption V.2, Proposition 3 in [20] guarantees that the projected risk-sensitive Bellman operator ΠT_θ is a contraction w.r.t. the d_θ-norm. Therefore, Eq. 10 has a unique fixed-point solution $\widetilde{V}_\theta(x) = v_\theta^{*\top}\phi(x)$. This means that $v_\theta^* \in \mathbb{R}^{\kappa_2}$ satisfies $v_\theta^* \in \arg\min_v \|T_\theta[\Phi v] - \Phi v\|_{d_\theta}^2$. By the projection theorem on Hilbert spaces, the orthogonality condition for v_θ^* becomes

$$\sum_{x\in\mathcal{X}} d_\theta(x|x_0)\phi(x)\phi(x)^\top v_\theta^* = \sum_{x\in\mathcal{X}} d_\theta(x|x_0)\phi(x) \max_{\xi\in\mathcal{U}(x,P(\cdot|x,\cdot)\mu_\theta(\cdot|x))} \mathbb{E}_{\xi P(\cdot|x,\cdot)\mu_\theta(\cdot|x)}[C(x,\widehat{a}) + \gamma\Phi(\widehat{x})v_\theta^*].$$

As a result, given a long enough trajectory x_0, a_0, x_1, a_1, …, x_{N-1}, a_{N-1} generated by policy θ, we may estimate the fixed-point solution v_θ^* using the projected risk sensitive value iteration (PRSVI) algorithm with the update rule

$$v_{k+1} = \left(\frac{1}{N}\sum_{t=0}^{N-1}\phi(x_t)\phi(x_t)^\top\right)^{-1}\left[\frac{1}{N}\sum_{t=0}^{N-1}\phi(x_t) \max_{\xi\in\mathcal{U}(x_t,P(\cdot|x_t,\cdot)\mu_\theta(\cdot|x_t))} \mathbb{E}_{\xi P(\cdot|x_t,\cdot)\mu_\theta(\cdot|x_t)}[C(x,\widehat{a})+\gamma\Phi(\widehat{x})v_k]\right]. \tag{11}$$

Note that using the law of large numbers, as both N and k tend to infinity, v_k converges w.p. 1 to v_θ^*, the unique solution of the fixed point equation $\Pi T_\theta[\Phi v] = \Phi v$.

In order to implement the iterative algorithm (11), one must repeatedly solve the inner optimization problem

$$\max_{\xi:\xi\circ P\circ\mu_\theta\in\mathcal{U}(x,P(\cdot|x,\cdot)\mu_\theta(\cdot|x))} \mathbb{E}_{\xi P(\cdot|x,\cdot)\mu_\theta(\cdot|x)}[C(x,\widehat{a})+\gamma\Phi(\widehat{x})v].$$

When the state space $\mathcal{X}$ is large, solving this optimization problem is often computationally expensive or even intractable. Similar to Section 3.4 of [20], we propose the following SAA approach to solve this problem. For the trajectory, x_0, a_0, x_1, a_1, …, x_{N-1}, a_{N-1}, we define the empirical transition probability $P_N(x'|x,a) \doteq \frac{\sum_{t=0}^{N-1}\mathbf{1}\{x_t=x,\, a_t=a,\, x_{t+1}=x'\}}{\sum_{t=0}^{N-1}\mathbf{1}\{x_t=x,\, a_t=a\}}$.[6] Consider the following ℓ_2-regularized empirical robust optimization problem[7]

$$\rho_N\big(C(x,\widehat{a}) + \gamma\Phi(\widehat{x})v\big) = \max_{\xi\in\mathcal{U}\left(x,P_{\theta;N}(\cdot|x,\cdot)\mu_\theta(\cdot|x)\right)} \sum_{a\in\mathcal{A},x'\in\mathcal{X}} P_{\theta;N}(x'|x,a)\mu_\theta(a|x)\xi(a,x') \Big[C(x,a) + \gamma\phi^\top(x')v + \frac{1}{2N}\xi(a,x')\Big]. \tag{12}$$

As in [39], the ℓ_2-regularization term in this optimization problem guarantees convergence of the optimizers ξ^* and the corresponding KKT multipliers, when $N \to \infty$. Convergence of these parameters is crucial for the policy gradient analysis in the next sections. We denote by $\xi^*_{\theta,x;N}$, the solution of the above empirical optimization problem, and by $\lambda^{*,\mathcal{P}}_{\theta,x;N}$, $\lambda^{*,\mathcal{E}}_{\theta,x;N}$, and $\lambda^{*,\mathcal{I}}_{\theta,x;N}$, the corresponding KKT multipliers. We obtain the empirical PRSVI algorithm by replacing

$$\max_{\xi\in\mathcal{U}(x_t,P(\cdot|x_t,\cdot)\mu_\theta(\cdot|x_t))} \mathbb{E}_{\xi P(\cdot|x_t,\cdot)\mu_\theta(\cdot|x_t)}[C(x_t,\widehat{a}) + \gamma\Phi(\widehat{x})v_\theta^*]$$

in Eq. 11 with its SAA approximation $\rho_N(C(x_t,\widehat{a})+\gamma\Phi(\widehat{x})v)$ from Eq. 12. Similarly, as both N and k tend to infinity, v_k converges w.p. 1 to v_θ^*. Details can be found in Appendix D.

C. Gradient Estimation

In Section V-B, we showed that we may effectively approximate the value function of a fixed policy θ using the (empirical) PRSVI algorithm in Eq. 11. In this section, we first derive a formula for the gradient of the Markov-coherent dynamic risk measure $\rho_\infty(\mathcal{M})$, and then propose a SAA algorithm for estimating this gradient, in which we use the SAA approximation of value function from Section V-B. As described in Section V-A, $\rho_\infty(\mathcal{M}) = V_\theta(x_0)$, and thus, we shall first derive a formula for $\nabla_\theta V_\theta(x_0)$.

Let $(\xi^*_{\theta,x}, \lambda^{*,\mathcal{P}}_{\theta,x}, \lambda^{*,\mathcal{E}}_{\theta,x}, \lambda^{*,\mathcal{I}}_{\theta,x})$ be the saddle point of (6) corresponding to the state $x \in \mathcal{X}$. In many common coherent risk measures such as CVaR and mean semi-deviation, there are closed-form formulas for $\xi^*_{\theta,x}$ and KKT multipliers $(\lambda^{*,\mathcal{P}}_{\theta,x}, \lambda^{*,\mathcal{E}}_{\theta,x}, \lambda^{*,\mathcal{I}}_{\theta,x})$. We will briefly discuss the case when the saddle point does not have an explicit solution later in this section. Before analyzing the gradient estimation, we have the following standard assumption in analogous to Assumption IV.1 of the static case.

Assumption V.3. *The likelihood ratio $\nabla_\theta \log \mu_\theta(a|x)$ is well-defined and bounded for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$.*

As in Theorem IV.2 for static case, we may use the envelope theorem and the risk-sensitive Bellman equation,

$$V_\theta(x) = \max_{\xi\in\mathcal{U}(x,P(\cdot|x,\cdot)\mu_\theta(\cdot|x))} \mathbb{E}_{\xi P(\cdot|x,\cdot)\mu_\theta(\cdot|x)}[C(x,\widehat{a})+\gamma V_\theta(\widehat{x})],$$

[5] Here we denote by ΠT_θ the projected Bellman operator. In particular, for the parameterized value function $\widetilde{v}^\top\phi(x) = \widetilde{V}_\theta(x)$, the expression $\widetilde{V}_\theta(x) = \Pi T_\theta[\widetilde{V}_\theta](x)$ is a short hand of the following optimization problem: $\widetilde{v} \in \mathrm{argmin}_v \sum_x d_\theta(x|x_0)(v^\top\phi(x) - T_\theta[v^\top\phi](x))^2$.

[6] In the case when the sizes of state and action spaces are huge or when these spaces are continuous, the empirical transition probability can be found by kernel density estimation.

[7] In the SAA approach, we only sum over the elements for which $P_{\theta;N}(x'|x,a) > 0$, thus, the sum has at most N elements.

to derive a formula for $\nabla_\theta V_\theta(x)$. We report this result in Theorem V.4, which is analogous to the risk-neutral policy gradient theorem [36], [8], [40]. The proof is in Appendix E.

Theorem V.4. *Under Assumptions II.2 and V.3, the gradient of the value function can be expressed as*

$$\nabla V_\theta(x) = \mathbb{E}_{\xi^*_\theta}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta\log\mu_\theta(a_t|x_t)h_\theta(x_t,a_t)\,|\,x_0{=}x\right],$$

where $\mathbb{E}_{\xi^*_\theta}[\cdot]$ denotes the expectation w.r.t. trajectories generated by a Markov decision process with action probability $\mu_\theta(\cdot|x)$, transition probability $P(\cdot|x,\cdot)\xi^*_{\theta,x}(\cdot,\cdot)$, and the stage-wise cost function $h_\theta(x,a)$ is defined as

$$h_\theta(x,a) = C(x,a) + \sum_{x'\in\mathcal{X}} P(x'|x,a)\xi^*_{\theta,x}(a,x')\Big[\gamma V_\theta(x') - \lambda^{*,\mathcal{P}}_{\theta,x} - \sum_{i\in\mathcal{I}}\lambda^{*,\mathcal{I}}_{\theta,x}(i)\frac{df_i(\xi^*_{\theta,x},p)}{dp(x')} - \sum_{e\in\mathcal{E}}\lambda^{*,\mathcal{E}}_{\theta,x}(e)\frac{dg_e(\xi^*_{\theta,x},p)}{dp(x')}\Big]. \quad (13)$$

Notice that the dynamic risk presented in this paper is a composition of static risks. In order to derive the stage-wise cost function $h_\theta(x,a)$ in (13), one simply applies the result from Theorem IV.2 as follows: The sample space corresponds to the state space $\mathcal{X}$ of the MDP, the probability distribution is given by the transition kernel $P(\cdot|x,a)$ (conditioned on current state x and action a), and the random variable Z is the sum of stage-wise cost and value function ($C(x,a)+\gamma V_\theta(x')$).

Theorem V.4 indicates that the policy gradient of the Markov-coherent dynamic risk measure $\rho_\infty(\mathcal{M})$, i.e., $\nabla_\theta\rho_\infty(\mathcal{M}) = \nabla_\theta V_\theta$, is equivalent to the risk-neutral value function of policy θ in a MDP with the stage-wise cost function $\nabla_\theta\log\mu_\theta(a|x)h_\theta(x,a)$ (which is well-defined and bounded), action probability $\mu_\theta(\cdot|x)$, and transition probability $P(\cdot|x,\cdot)\xi^*_{\theta,x}(\cdot,\cdot)$. Thus, when the saddle points are known and the state space $\mathcal{X}$ is not too large, we can compute $\nabla_\theta V_\theta$ using a policy evaluation algorithm. However, when the state space is large, exact calculation of ∇V_θ by policy evaluation becomes impossible, and our goal would be to derive a sampling method to estimate ∇V_θ. Unfortunately, since the risk envelop depends on the policy parameter θ, unlike the risk-neutral case, the risk sensitive (or robust) Bellman equation $T_\theta[V_\theta](x)$ in (9) is nonlinear in the stationary Markov policy μ_θ. Thus, h_θ cannot be considered using the action-value function (Q-function) of the robust MDP. Therefore, even if the exact formulation of the value function V_θ is known, it is computationally intractable to enumerate the summation over x' to compute $h_\theta(x,a)$. On top of that, in most applications with a large or continuous state space, the value function V_θ cannot be accurately calculated, due to the 'curse of dimensionality' [13]; this further complicates the gradient estimation process. To estimate the policy gradient when the value function is unknown, we approximate it by the projected risk sensitive value function Φv^*_θ. To address the sampling issues, we propose the following *two-phase sampling procedure* for estimating ∇V_θ.

(1) Generate N trajectories $\{x^{(j)}_0, a^{(j)}_0, x^{(j)}_1, a^{(j)}_1, \ldots\}_{j=1}^N$ from the Markov decision process with action probability $\mu_\theta(\cdot|x)$ and transition probability $P^\xi(\cdot|x,\cdot) := \xi^*_{\theta,x}(\cdot,\cdot)P(\cdot|x,\cdot)$.

(2) For each state-action pair $(x^{(j)}_t, a^{(j)}_t) = (x,a)$, generate N samples $\{y^{(k)}\}_{k=1}^N$ using the transition probability $P(\cdot|x,a)$ and calculate the following empirical average estimate of $h_\theta(x,a)$:

$$h_{\theta,N}(x,a) := C(x,a) + \frac{1}{N}\sum_{k=1}^{N}\xi^*_{\theta,x}(a,y^{(k)})\Bigg[\gamma {v^*_\theta}^\top\phi(y^{(k)}) - \lambda^{*,\mathcal{P}}_{\theta,x} - \sum_{i\in\mathcal{I}}\lambda^{*,\mathcal{I}}_{\theta,x}(i)\frac{df_i(\xi^*_{\theta,x},p)}{dp(y^{(k)})} - \sum_{e\in\mathcal{E}}\lambda^{*,\mathcal{E}}_{\theta,x}(e)\frac{dg_e(\xi^*_{\theta,x},p)}{dp(y^{(k)})}\Bigg].$$

(3) Calculate an estimate of ∇V_θ using the following average over all the samples:

$$\frac{1}{N}\sum_{j=1}^{N}\sum_{t=0}^{\infty}\gamma^t\nabla_\theta\log\mu_\theta(a^{(j)}_t|x^{(j)}_t)h_{\theta,N}(x^{(j)}_t, a^{(j)}_t).$$

Indeed, by the definition of empirical transition probability $P_N(x'|x,a)$, $h_{\theta,N}(x,a)$ can be re-written as in the same structure of $h_\theta(x,a)$, except by replacing the transition probability $P(x'|x,a)$ with $P_N(x'|x,a)$.

To compare the above algorithm with the policy gradient algorithm in static risk optimization, notice that with the gradient of the static risk objective function, one can immediately apply the policy gradient algorithm to compute the optimal policy using the Monte-Carlo estimate of the objective function. On the other hand, under the assumption of a Markov decision process and the framework of dynamic risk optimization, the algorithm in steps 1-3 resembles an actor-critic method where the function approximation of the risk sensitive value function, instead of the Monte-Carlo estimate, is used to calculate the policy gradient.

Furthermore, in the case that the saddle points $(\xi^*_{\theta,x}, \lambda^{*,\mathcal{P}}_{\theta,x}, \lambda^{*,\mathcal{E}}_{\theta,x}, \lambda^{*,\mathcal{I}}_{\theta,x})$ do not have a closed-form solution, we may follow the SAA procedure of Section V-B and replace them and the transition probabilities $P(x'|x,a)$ with their sample estimates $(\xi^*_{\theta,x;N}, \lambda^{*,\mathcal{P}}_{\theta,x;N}, \lambda^{*,\mathcal{E}}_{\theta,x;N}, \lambda^{*,\mathcal{I}}_{\theta,x;N})$ and $P_N(x'|x,a)$, respectively.

Finally, we show the convergence of the above two-phase sampling procedure. Let $d_\theta(x|x_0)$ and $\pi_\theta(x,a|x_0)$ be the state and state-action occupancy measure induced by the transition probability $P^\xi(\cdot|x,\cdot)$ and action probability $\mu_\theta(\cdot|x)$, respectively. Similarly, let $d_{\theta;N}(x|x_0)$ and $\pi_{\theta;N}(x,a|x_0)$ be the state and state-action occupancy measure induced by the action probability $\mu_\theta(\cdot|x)$ and estimated transition probability function $P^\xi_{\theta;N}(\cdot|x,\cdot) := \xi^*_{\theta,x;N}(\cdot,\cdot)P_{\theta;N}(\cdot|x,\cdot)$. From the two-phase sampling procedure for policy gradient estimation and by the strong law of large numbers, when $N\to\infty$, with probability 1, we have that

$$\frac{1}{N}\sum_{j=1}^{N}\sum_{t=0}^{\infty}\gamma^t\mathbf{1}\{x^{(j)}_t = x, a^{(j)}_t = a\} = \pi_{\theta;N}(x,a|x_0).$$

Based on the strongly convex property of the ℓ_2-regularized objective function in the inner robust optimization problem $\rho_N(\Phi v)$, we can show that both the state-action occupancy measure $\pi_{\theta;N}(x,a|x_0)$ and the stage-wise cost $h_{\theta;N}(x,a)$ converge to the their true values within a value function approximation error bound $\Delta = \|\Phi v^*_\theta - V_\theta\|_\infty$. We refer the readers to Appendix F for these technical results. These results together with Theorem V.4 imply the consistency of the policy gradient estimation.

Theorem V.5. *The following expression holds almost surely:*

$$\left|\lim_{N\to\infty}\frac{1}{N}\sum_{j=1}^{N}\sum_{t=0}^{\infty}\gamma^t\,\nabla\log\mu_\theta(a_t^{(j)}|x_t^{(j)})\;h_{\theta,N}(x_t^{(j)},a_t^{(j)}) - \nabla V_\theta(x_0)\right| = O(\Delta),\ \forall x_0\in\mathcal{X}.$$

Thm. V.5 guarantees that as the value function approximation error decreases and the number of samples increases, the sampled gradient converges to the true gradient.

VI. Numerical Illustration

In this section, we illustrate our approach with several numerical examples. The purpose of this illustration is to emphasize the importance of *flexibility* in designing risk criteria for selecting an *appropriate* risk-measure – such that suits both the user's risk preference *and* the problem-specific properties.

A. Example 1: Single-Step Horizon Asset Allocation

We consider a trading agent that can invest in one of three assets (see Figure 1 for their distributions). The returns of the first two assets, $A1$ and $A2$, are normally distributed: $A1 \sim \mathcal{N}(1,1)$ and $A2 \sim \mathcal{N}(4,6)$. The return of the third asset $A3$ has a Pareto distribution: $f(z)=\frac{\alpha}{z^{\alpha+1}}\ \forall z>1$, with $\alpha=1.5$. The mean of the return from $A3$ is 3 and its variance is infinite; such heavy-tailed distributions are widely used in financial modeling [41]. The agent selects an action randomly, with probability $P(A_i)\propto\exp(\theta_i)$, where $\theta\in\mathbb{R}^3$ is the policy parameter. We trained three different policies π_1, π_2, and π_3. Policy π_1 is risk-neutral, i.e., $\max_\theta \mathbb{E}[Z]$, and it was trained using standard policy gradient [22]. Policy π_2 is risk-averse and had a mean-semideviation objective $\max_\theta \mathbb{E}[Z]-\mathbb{SD}[Z]$, and was trained using the algorithm in Section IV. Policy π_3 is also risk-averse, with a mean-standard-deviation objective, as proposed in [24], [25], $\max_\theta \mathbb{E}[Z]-\sqrt{\mathrm{Var}[Z]}$, and was trained using the algorithm of [24]. For each of these policies, Figure 1 shows the probability of selecting each asset vs. training iterations. Although $A2$ has the highest mean return, the risk-averse policy π_2 chooses $A3$, since it has a lower downside, as expected. However, because of the heavy upper-tail of $A3$, policy π_3 opted to choose $A1$ instead. This is counter-intuitive as a rational investor should not avert high returns. In fact, in this case $A3$ stochastically dominates $A1$ [42].

We clarify that in these experiments, the probability distribution of the returns was not given to the algorithm, which only requires samples from this distribution. This is since the probability of the return can be written as $P(Z)=P(A_i)P(Z|A_i)$, and $P(Z|A_i)$ does not depend on θ, therefore the term $\nabla_\theta\log P(Z)$ in the algorithms satisfies $\nabla_\theta\log P(Z)=\nabla_\theta\log P(A_i)$. Thus, algorithmically, any black-box simulator of returns could have been used instead of the distributions reported here.

B. Example 2: Empirical Comparison of Analytical-Solution-Based and SAA-Based Policy Gradient

In this example, we compare the CVaR policy gradient as obtained by the analytical result in Eq. 7 with the general sampling-based algorithm of Eq. 9.

For the analytical-solution-based policy gradient, we use the `GCVaR` algorithm [27], which is the sampling-based version of Eq. 7. For the sampling-based algorithm, we solve the linear program in Eq. 8 from Section IV-A with the risk envelope for CVaR. The resultant numerical values for $\xi^*_{\theta;N}$ and $\lambda^{*,\mathcal{P}}_{\theta;N}$ were plugged into Eq. 9 for the gradient estimate (using the CVaR risk envelope, all other terms in Eq. 9 vanish).

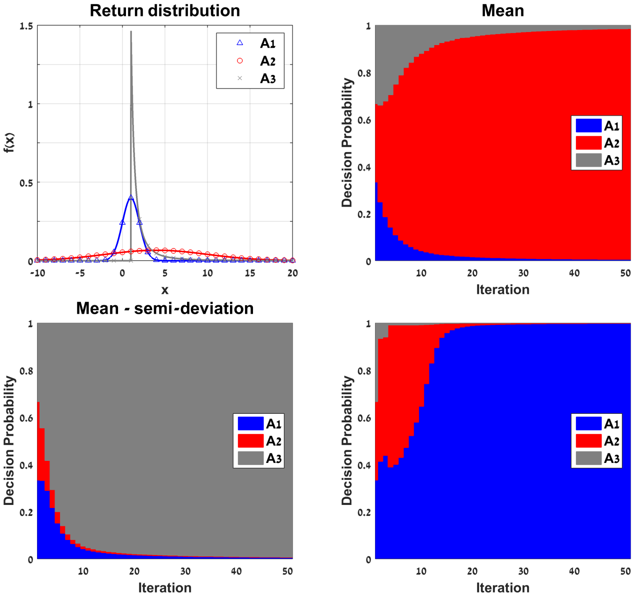


Fig. 1. Numerical illustration - selection between 3 assets. A: Probability density of asset return. B,C,D: Bar plots of the probability of selecting each asset vs. training iterations, for policies π_1, π_2, and π_3, respectively. At each iteration, 10,000 samples were used for gradient estimation.

We present empirical results for the asset selection domain of Example 1. We chose a CVaR level of $\alpha=0.05$ (corresponding to the average of the worst 5% outcomes), and trained policies with either the analytical-solution-based policy gradient (labeled CVaR), and the general sampling-based algorithm (labeled CVaRS). In Figure 2, we plot the learning curves (the θ values vs. training episodes) of both policies, for different values of the sampling budget N.

As one can observe, both policies exhibit similar learning performance and the differences diminish as N grows. This verifies our theoretical findings on consistency of policy gradient with sample average approximation (i.e., Proposition IV.4).

C. Example 3: American Option Trading

We empirically evaluate our algorithms on the American put option domain: a standard testbed for risk-sensitive RL [43], [20], [26]. In our setting, the state is continuous and represents the price of some stock. It evolves according to a geometric Brownian motion (GBM), i.e., $x_{t+1}/x_t \sim \ln\mathcal{N}\left(\mu_t-\sigma_t^2/2,\sigma_t^2\right)$, where $\ln\mathcal{N}$ is the log-normal distribution and μ_t and σ_t are parameters. The action at each time t is binary. An execution action generates reward $\max\{0,K-x_t\}$, where K is fixed and known as the *strike price*, and terminates the episode; A hold action generates zero reward, and the price transitions to x_{t+1} as described above. Unless an execution occurred, the episode ends after T steps, with reward $\max\{0,K-x_T\}$. For the expected return, the optimal policy is a time dependent threshold policy that holds if $x_t>\theta_t$ [44], where θ_t is the threshold, and executes, otherwise. Accordingly, we search in the space of soft-threshold policies of the form $\mu_\theta(hold|x_t)=\frac{1}{1+\exp(-\beta(x_t-\theta_t))}$, for some softness parameter $\beta>0$.

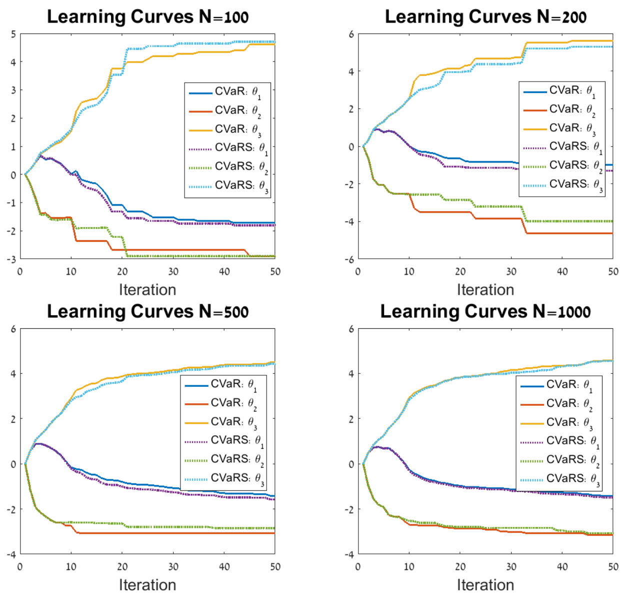


Fig. 2. Learning curves (θ vs. training episodes) of the analytical-solution-based policy gradient (labeled CVaR), and the general sampling-based algorithm (labeled CVaRS), for different values of the sampling budget N.

We consider a case where the option is 'deep in the money', i.e., $x_0 < K$. For such a case, the decision-maker may execute immediately and earn reward $K - x_0$, but may also wait for a better price with the risk of never getting it on time.

We trained policies to optimize the static CVaR($\alpha = 0.3$), CVaR($\alpha = 0.6$), $0.9 \cdot \text{Expectation} + 0.1 \cdot \text{CVaR}(\alpha = 0.3)$, and Expectation $- \ 0.1 \cdot$ Semideviation, using stochastic gradient descent. The gradient for the expectation was calculated using standard (episodic) policy gradient [22]; the gradient for CVaR was calculated according to the derivation in Section IV-A, using the `GCVaR` algorithm [27]; and the gradient for mean-semideviation was calculated using the `GMSD` algorithm of Section IV-B. In Figure 3, we plot the histograms of the payoff of the different policies. The risk-averse nature of the policies trained with a risk-sensitive objective may be observed. In our experiments we set $K = 1$ and $x_0 = 0.5$. The CVaR($\alpha = 0.3$) was very conservative, chose to execute immediately, and received a reward of 0.5. The CVaR($\alpha = 0.6$) was less conservative, but still had lower variability than the standard expectation objective policy.

We also trained policies to optimize the dynamic CVaR($\alpha = 0.6$) risk, dynamic $0.95 \cdot \text{Expectation} + 0.05 \cdot \text{CVaR}(\alpha = 0.3)$ risk, and dynamic $0.98 \cdot \text{Expectation} + 0.02 \cdot \text{CVaR}(\alpha = 0.3)$ risk using stochastic gradient descent, with the gradient calculated according to the algorithm of Section V. We used RBF features for estimating the value function. In Figure 4, we plot the histograms of the payoff of different policies. The dynamic CVaR($\alpha = 0.6$) was very conservative and chose to execute immediately. This also occurred with higher values of α, such as $\alpha = 0.8$. One way to reduce conservatism in dynamic risk, while still maintaining control of reward variability, is to use the combination of expectation and CVaR. As Figure 4 demonstrates, this is indeed a practical approach.

D. Example 4: Optimal Execution in Portfolio Optimization

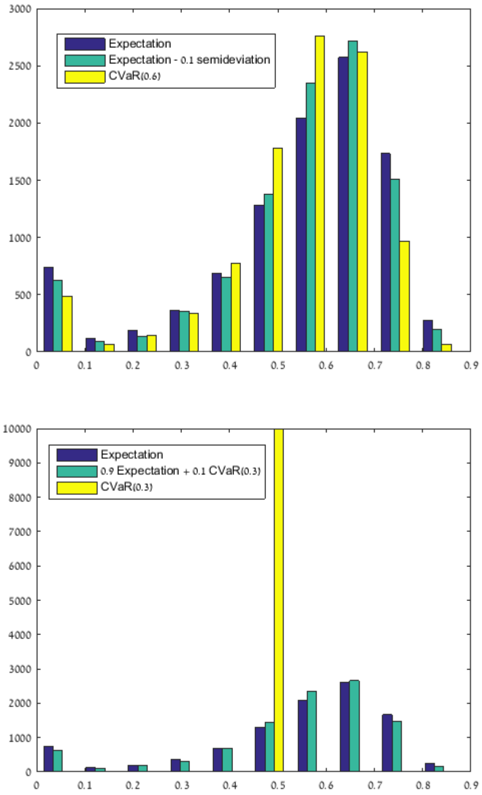


Fig. 3. Reward histogram for various *static* risk-sensitive policies.

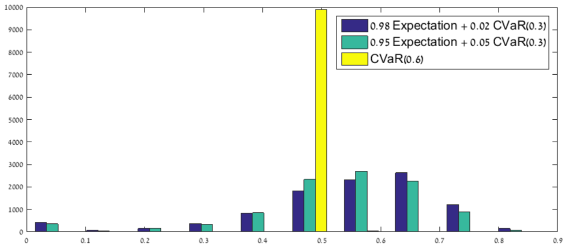


Fig. 4. Reward histogram for various *dynamic* risk-sensitive policies.

In this example, we consider the optimal trade execution in quantitative finance, where the objective is to design a strategy that sells (respectively, buys) X shares of a given stock within a fixed time period (or horizon) T, in a manner that maximizes the revenue received (respectively, minimizes the capital spent). Similar to [45], [46], we describe stock evolution using the classical Almgren-Chriss [47] model with linear price impact. This model is popular amongst sell-side institutions as a basis for arrival price benchmark algorithms. By formulating our problem as a finite-horizon MDP, we hereby apply the risk sensitive reinforcement learning algorithms to find *near-optimal* trade-execution strategies that maximize expected reward while controlling 1) tail risk (modeled by CVaR) or 2) downside variance (modeled by semi-deviation).

Based on the descriptions of [47], the trading model and price dynamics are characterized by the following finite-horizon MDP $(\mathcal{X}, \mathcal{A}, R, P, x_0)$,[8] where $\mathcal{X} = \{0, \ldots, X\} \times \{0, \ldots, T\}$ is the combined state space of the unliquidated financial securities and passage of time, $\mathcal{A} = \{0, \ldots, X\}$

[8] In contrary to the cost minimization MDP used in the main paper, in this example we adopt the reward maximization framework.

is the action space of executable units at each time step, $R : \mathcal{X} \times \mathcal{A} \times \Omega \to \mathbb{R}$ is the immediate reward incurred from trading, P is the evolution of the trading trajectory, and $x_0 = (X, 0)$ is the combined initial state of security block-size and starting time. For simplicity, we assume deterministic portfolio execution, i.e., given state x_t and action a_t, the next state is given by

$$x_{t+1}(1) = x_t(1) - a_t,\ x_{t+1}(2) = x_{t+1}(2) + 1.$$

Furthermore the immediate reward incurred by executing a trade is given by

$$R(x_t, a_t, \omega_t) = (\sigma\omega_t - g(a_t))x_t - a_t h(a_t),$$

where ω_t is the uncertain price fluctuation modeled by an independent white noise process with mean 0 and variance 1, σ is the random price volatility (variance), $g : \mathcal{A} \to \mathbb{R}$ is the linear permanent impact function given by $g(a_t) = La_t$ with linear impact factor $L > 0$, and $h : \mathcal{A} \to \mathbb{R}$ is the temporary impact function given by $g(a_t) = \epsilon \text{sgn}(a_t) + \eta a_t$ with absolute impact factor $\epsilon > 0$ and quadratic impact factor $\eta > 0$. Therefore, the objective function is to find an optimal trading strategy that maximizes $\rho(\mathcal{R}) := \rho\left(\sum_{t=0}^{T} \gamma^t R(x_t, a_t, \omega_t)\right)$, where the discounting factor is $\gamma = 1$ and ρ is the coherent risk of interest. Followed from the benchmark example in [47], we set $\sigma = 0.95$, $L = 0.25$, $\epsilon = 0.725$, and $\eta = 0.25$. In the following experiments, we set the horizon of the trading model to $T = 15$ and the size of financial security to $X = 10$. Also for the RL algorithms of this experiment, we use RBF features for value function approximation and the class of Boltzmann policies for policy parameterization.

Similar to the optimal stopping example, we trained policies to optimize static CVaR($\alpha = 0.25$), $0.5 \cdot$ Expectation $+ 0.5 \cdot$ CVaR($\alpha = 0.25$), and $0.5 \cdot$ Expectation $- 0.5 \cdot$ Semideviation. In Table I, we provide the statistics obtained from different policies. Again the risk-averse nature of the policies was observed. The CVaR($\alpha = 0.25$) was conservative (reward was 23% lower than risk neutral policies), but the worst case expectation corresponding to this policy was well-controlled (CVaR was 15% lower than the risk-neutral policy). On the other hand, the risk-sensitive policies from $0.5 \cdot$ Expectation $+ 0.5 \cdot$ CVaR($\alpha = 0.25$) and $0.5 \cdot$Expectation$-0.5\cdot$Semideviation balanced the trade-offs between reward/CVaR and reward/semi-deviation, respectively. Note that there is indeed a tradeoff – the CVaR-sensitive policy had better (higher) CVaR and worse (higher) semi-deviation compared to the semi-deviation sensitive policy, and vice-versa. Thus, compared to the theoretical mean-variance optimization in [47], our risk-sensitive policy-gradient approach has an advantage when the investor's desired risk-profile is different than a mean-variance-based risk criterion.

	$\mathbb{E}(\mathcal{R})$	$\sigma(\mathcal{R})$	CVaR($\mathcal{R}$)	$\mathbb{SD}[\mathcal{R}]$
PG	-21.14	26.18	-54.29	18.45
PG-CVaR	-26.41	13.46	-42.63	9.52
PG-Mean-CVaR	-23.84	14.79	-43.95	10.42
PG-Mean-SD	-27.52	12.49	-44.17	8.81

TABLE I

PERFORMANCE COMPARISON OF THE POLICIES LEARNED BY THE RISK-SENSITIVE ALGORITHMS. HERE OPTIMIZATION WITH EXPECTATION RISK YIELDS A LOWEST MEAN COST, WHILE OPTIMIZATION WITH CVAR AND MEAN-CVAR TRADES-OFF MEAN COST AND WORST-CASE COST, AND OPTIMIZATION WITH MEAN-SEMI-DEVIATION YIELDS THE LOWEST VARIABILITY.

VII. DISCUSSION AND CONCLUSION

We presented algorithms for estimating the gradient of both static and dynamic coherent risk measures using two new policy gradient style formulas that combine sampling with convex programming. Thereby, our approach extends risk-sensitive RL to the whole class of coherent risk measures and generalizes several recent studies that focused on specific risk measures.

On the technical side, an important future direction is to improve the convergence rate of gradient estimates using importance sampling methods. This is especially important for risk criteria that are sensitive to rare events, such as CVaR [48].

From a more conceptual point of view, the coherent-risk framework explored in this work provides the decision-maker with *flexibility* in designing risk preference. As our numerical examples show, such flexibility is important for selecting appropriate *problem-specific* risk measures for managing the cost variability. However, we believe that our approach has much more potential than that.

In almost every real-world application, uncertainty emanates from stochastic dynamics, but also, and perhaps more importantly, from modeling errors (model uncertainty). A prudent policy should protect against *both* types of uncertainties. The representation duality of coherent-risk (Theorem II.1), naturally relates the risk to model uncertainty. For model-uncertainty in MDPs, a similar connection was made with dynamic Markov coherent risk [49] and static CVaR risk [50]. Therefore, we believe that by carefully shaping the risk-criterion, the decision-maker may be able to take uncertainty into account in a *broad* sense. Designing a principled procedure for such *risk-shaping* is not trivial and is beyond the scope of this paper. However, we believe that there is much potential to risk-shaping as it may be the key for handling model misspecification in dynamic decision-making.

REFERENCES

[1] H. Markowitz, *Portfolio Selection: Efficient Diversification of Investment*. John Wiley and Sons, 1959.

[2] R. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *Journal of risk*, vol. 2, pp. 21–42, 2000.

[3] P. Artzner, F. Delbaen, J. Eber, and D. Heath, "Coherent measures of risk," *Mathematical finance*, vol. 9, no. 3, pp. 203–228, 1999.

[4] D. Bertsekas, *Dynamic Programming and Optimal Control, Vol I*, 3rd ed. Athena Scientific, 2005.

[5] A. Ruszczyński, "Risk-averse dynamic programming for Markov decision processes," *Mathematical Programming*, vol. 125, no. 2, pp. 235–261, 2010.

[6] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming*. SIAM, 2009, ch. 6, pp. 253–332.

[7] J. Baxter and P. Bartlett, "Infinite-horizon policy-gradient estimation," *JAIR*, vol. 15, pp. 319–350, 2001.

[8] V. Konda and J. Tsitsiklis, "Actor-critic algorithms," in *Advances in Neural Information Processing Systems*, 2000, pp. 1008–1014.

[9] J. Peters and S. Schaal, "Policy gradient methods for robotics," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2006, pp. 2219–2225.

[10] N. Tao, J. Baxter, and L. Weaver, "A multi-agent, policy-gradient approach to network routing," in *International Conference on Machine Learning*, 2001.

[11] J. Moody and M. Saffell, "Learning to trade via direct reinforcement," *IEEE Transactions on Neural Networks*, vol. 12, no. 4, pp. 875–889, 2001.

[12] M. Deisenroth, G. Neumann, and J. Peters, "A survey on policy search for robotics." *Foundations and Trends in Robotics*, vol. 2, no. 1-2, pp. 1–142, 2013.

[13] D. Bertsekas, *Dynamic Programming and Optimal Control, Vol II*, 4th ed. Athena Scientific, 2012.

[14] P. Glynn, "Likelihood ratio gradient estimation for stochastic systems," *Communications of the ACM*, vol. 33, no. 10, pp. 75–84, 1990.

[15] A. Ruszczyński and A. Shapiro, "Optimization of convex risk functions," *Math. OR*, vol. 31, no. 3, pp. 433–452, 2006.

[16] G. N. Iyengar, "Robust dynamic programming," *Mathematics of Operations Research*, vol. 30, no. 2, pp. 257–280, 2005.

[17] A. Nilim and L. El Ghaoui, "Robust control of Markov decision processes with uncertain transition matrices," *Operations Research*, vol. 53, no. 5, pp. 780–798, 2005.

[18] D. L. Kaufman and A. J. Schaefer, "Robust modified policy iteration," *INFORMS Journal on Computing*, vol. 25, no. 3, pp. 396–410, 2013.

[19] W. Wiesemann, D. Kuhn, and B. Rustem, "Robust Markov decision processes," *Mathematics of Operations Research*, vol. 38, no. 1, pp. 153–183, 2013.

[20] A. Tamar, S. Mannor, and H. Xu, "Scaling up robust MDPs using function approximation," in *International Conference on Machine Learning*, 2014.

[21] J. Perolat, B. Scherrer, B. Piot, and O. Pietquin, "Approximate dynamic programming for two-player zero-sum markov games," in *International Conference on Machine Learning*, 2015.

[22] P. Marbach and J. Tsitsiklis, "Simulation-based optimization of Markov reward processes," *IEEE Transactions on Automatic Control*, vol. 46, no. 2, pp. 191–209, 1998.

[23] V. Borkar, "A sensitivity formula for risk-sensitive cost and the actor–critic algorithm," *Systems & Control Letters*, vol. 44, no. 5, pp. 339–346, 2001.

[24] A. Tamar, D. Di Castro, and S. Mannor, "Policy gradients with variance related risk criteria," in *International Conference on Machine Learning*, 2012.

[25] L. Prashanth and M. Ghavamzadeh, "Actor-critic algorithms for risk-sensitive MDPs," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 252–260.

[26] Y. Chow and M. Ghavamzadeh, "Algorithms for CVaR optimization in MDPs," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 3509–3517.

[27] A. Tamar, Y. Glassner, and S. Mannor, "Optimizing the CVaR via sampling," in *AAAI*, 2015.

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

[29] Y. Chow and M. Pavone, "A unifying framework for time-consistent, risk-averse model predictive control: theory and algorithms," in *American Control Conference*, 2014.

[30] N. Bäuerle and J. Ott, "Markov decision processes with average-value-at-risk criteria," *Mathematical Methods of Operations Research*, vol. 74, no. 3, pp. 361–379, 2011.

[31] C. Acerbi, "Spectral measures of risk: a coherent representation of subjective risk aversion," *Journal of Banking & Finance*, vol. 26, no. 7, pp. 1505–1518, 2002.

[32] D. Iancu, M. Petrik, and D. Subramanian, "Tight approximations of dynamic risk measures," *arXiv:1106.6102*, 2011.

[33] M. Fu, "Gradient estimation," in *Simulation*, ser. Handbooks in Operations Research and Management Science. Elsevier, 2006, vol. 13, pp. 575 – 616.

[34] P. Milgrom and I. Segal, "Envelope theorems for arbitrary choice sets," *Econometrica*, vol. 70, no. 2, pp. 583–601, 2002.

[35] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2009.

[36] R. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in Neural Information Processing Systems 13*, 2000.

[37] D. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[38] R. Sutton and A. Barto, *Reinforcement learning: An introduction*. Cambridge Univ Press, 1998.

[39] F. Meng and H. Xu, "A regularized sample average approximation method for stochastic mathematical programs with nonsmooth equality constraints," *SIAM Journal on Optimization*, vol. 17, no. 3, pp. 891–919, 2006.

[40] S. Bhatnagar, R. Sutton, M. Ghavamzadeh, and M. Lee, "Natural actor-critic algorithms," *Automatica*, vol. 45, no. 11, pp. 2471–2482, 2009.

[41] S. Mittnik, S. Rachev, and M. Paolella, "Stable Paretian modeling in finance: Some empirical and theoretical aspects," *A Practical Guide to Heavy Tails*, pp. 79–110, 1998.

[42] J. Hadar and W. R. Russell, "Rules for ordering uncertain prospects," *The American Economic Review*, pp. 25–34, 1969.

[43] Y. Li, C. Szepesvari, and D. Schuurmans, "Learning exercise policies for american options," in *Proc. of the Twelfth International Conference on Artificial Intelligence and Statistics, JMLR: W&CP*, vol. 5, 2009, pp. 352–359.

[44] J. Hull, *Options, Futures, and Other Derivatives (6th edition)*. Prentice Hall, 2006.

[45] D. Hendricks and D. Wilcox, "A reinforcement learning extension to the Almgren-Chriss framework for optimal trade execution," in *Computational Intelligence for Financial Engineering & Economics (CIFEr), 2104 IEEE Conference on*. IEEE, 2014, pp. 457–464.

[46] Y. Nevmyvaka, Y. Feng, and M. Kearns, "Reinforcement learning for optimized trade execution," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 673–680.

[47] R. Almgren and N. Chriss, "Optimal execution of portfolio transactions," *Journal of Risk*, vol. 3, pp. 5–40, 2001.

[48] O. Bardou, N. Frikha, and G. Pagès, "Computing VaR and CVaR using stochastic approximation and adaptive unconstrained importance sampling," *Monte Carlo Methods and Applications*, vol. 15, no. 3, pp. 173–210, 2009.

[49] T. Osogami, "Robustness and risk-sensitivity in Markov decision processes," in *Advances in Neural Information Processing Systems*, 2012, pp. 233–241.

[50] Y. Chow, A. Tamar, S. Mannor, and M. Pavone, "Risk-sensitive and robust decision-making: a CVaR optimization approach," in *Advances in Neural Information Processing Systems 28*, 2015.

[51] R. Rockafellar, R. Wets, and M. Wets, *Variational analysis*. Springer, 1998, vol. 317.

[52] A. Fiacco, *Introduction to sensitivity and stability analysis in nonlinear programming*. Elsevier, 1983.

Appendix

A. Gradient Results for Static Mean-Semideviation

In this section we consider the mean-semideviation risk measure, defined as follows:

$$\rho_{\text{MSD}}(Z) = \mathbb{E}\left[Z\right] + \alpha\left(\mathbb{E}\left[(Z-\mathbb{E}\left[Z\right])_+^2\right]\right)^{1/2}, \qquad (14)$$

Following the derivation in [6], note that $\left(\mathbb{E}\left[|Z|^2\right]\right)^{1/2} = \|Z\|_2$, where $\|\cdot\|_2$ denotes the L_2 norm of the space $\mathcal{L}_2(\Omega,\mathcal{F},P_\theta)$. The norm may also be written as:

$$\|Z\|_2 = \sup_{\|\xi\|_2\le 1}\langle \xi, Z\rangle,$$

and hence

$$\begin{aligned}\left(\mathbb{E}\left[(Z-\mathbb{E}\left[Z\right])_+^2\right]\right)^{1/2} &= \sup_{\|\xi\|_2\le 1}\langle \xi, (Z-\mathbb{E}\left[Z\right])_+\rangle \\ &= \sup_{\|\xi\|_2\le 1,\xi\ge 0}\langle \xi, Z-\mathbb{E}\left[Z\right]\rangle \\ &= \sup_{\|\xi\|_2\le 1,\xi\ge 0}\langle \xi-\mathbb{E}\left[\xi\right], Z\rangle.\end{aligned}$$

It follows that Eq. (1) holds with

$$\mathcal{U} = \left\{\xi' \in \mathcal{Z}^* : \quad \xi' = 1+\alpha\xi-\alpha\mathbb{E}\left[\xi\right], \quad \|\xi\|_q \le 1, \quad \xi\ge 0\right\}.$$

For this case it will be more convenient to write Eq. (1) in the following form

$$\rho_{\text{MSD}}(Z) = \sup_{\|\xi\|_q\le 1,\xi\ge 0}\langle 1+\alpha\xi-\alpha\mathbb{E}\left[\xi\right], Z\rangle. \qquad (15)$$

Let $\bar{\xi}$ denote an optimal solution for (15). In [6] it is shown that $\bar{\xi}$ is a contact point of $(Z-\mathbb{E}\left[Z\right])_+$, that is

$$\bar{\xi} \in \arg\max\left\{\langle \xi, (Z-\mathbb{E}\left[Z\right])_+\rangle : \|\xi\|_2 \le 1\right\},$$

and we have that

$$\bar{\xi} = \frac{(Z-\mathbb{E}\left[Z\right])_+}{\|(Z-\mathbb{E}\left[Z\right])_+\|_2} = \frac{(Z-\mathbb{E}\left[Z\right])_+}{\mathbb{SD}(Z)}. \qquad (16)$$

Note that $\bar{\xi}$ is not necessarily a probability distribution, but for $c\in[0,1]$, it can be shown [6] that $1+\alpha\bar{\xi}-\alpha\mathbb{E}\left[\bar{\xi}\right]$ always is.

In the following we show that $\bar{\xi}$ may be used to write the gradient $\nabla_\theta\rho_{\text{MSD}}(Z)$ as an expectation, which will lead to a sampling algorithm for the gradient.

Proposition A.1. *Under Assumption IV.1, we have that*

$$\nabla_\theta\rho_{MSD}(Z) = \nabla_\theta\mathbb{E}\left[Z\right] + \frac{\alpha}{\mathbb{SD}(Z)}\mathbb{E}\left[(Z-\mathbb{E}\left[Z\right])_+(\nabla_\theta\log P(\omega)(Z-\mathbb{E}\left[Z\right]) - \nabla_\theta\mathbb{E}\left[Z\right])\right],$$

and, according to the standard likelihood-ratio method,

$$\nabla_\theta \mathbb{E}\left[Z\right] = \mathbb{E}\left[\nabla_\theta \log P(\omega) Z\right].$$

Proof. Note that in Eq. (15) the constraints do not depend on θ. Therefore, using the envelope theorem we obtain that

$$\begin{aligned}\nabla_\theta \rho(Z) &= \nabla_\theta \langle 1 + \alpha\bar{\xi} - \alpha\mathbb{E}\left[\bar{\xi}\right], Z\rangle \\ &= \nabla_\theta\langle 1, Z\rangle + \alpha\nabla_\theta\langle\bar{\xi}, Z\rangle - \alpha\nabla_\theta\langle\mathbb{E}\left[\bar{\xi}\right], Z\rangle.\end{aligned} \tag{17}$$

We now write each of the terms in Eq. (17) as an expectation. We start with the following standard likelihood-ratio result:

$$\nabla_\theta\langle 1, Z\rangle = \nabla_\theta\mathbb{E}\left[Z\right] = \mathbb{E}\left[\nabla_\theta \log P(\omega) Z\right].$$

Also, we have that

$$\langle\mathbb{E}\left[\bar{\xi}\right], Z\rangle = \mathbb{E}\left[\bar{\xi}\right]\mathbb{E}\left[Z\right],$$

therefore, by the derivative of a product rule:

$$\nabla_\theta\langle\mathbb{E}\left[\bar{\xi}\right], Z\rangle = \nabla_\theta\mathbb{E}\left[\bar{\xi}\right]\mathbb{E}\left[Z\right] + \mathbb{E}\left[\bar{\xi}\right]\nabla_\theta\mathbb{E}\left[Z\right].$$

By the likelihood-ratio trick and Eq. (16) we have that

$$\nabla_\theta\mathbb{E}\left[\bar{\xi}\right] = \frac{1}{\mathbb{SD}(Z)}\mathbb{E}\left[\nabla_\theta \log P(\omega)(Z - \mathbb{E}\left[Z\right])_+\right].$$

Also, by the likelihood-ratio trick

$$\nabla_\theta\mathbb{E}\left[\bar{\xi}Z\right] = \mathbb{E}\left[\nabla_\theta \log P(\omega)\bar{\xi}Z\right].$$

Plugging these terms back in Eq. (17), we have that

$$\begin{aligned}&\nabla_\theta\rho(Z) \\ =&\nabla_\theta\mathbb{E}\left[Z\right] + \alpha\nabla_\theta\mathbb{E}\left[\bar{\xi}Z\right] - \alpha\nabla_\theta\mathbb{E}\left[\bar{\xi}\right]\mathbb{E}\left[Z\right] - \alpha\mathbb{E}\left[\bar{\xi}\right]\nabla_\theta\mathbb{E}\left[Z\right] \\ =&\nabla_\theta\mathbb{E}\left[Z\right] + \alpha\mathbb{E}\left[\bar{\xi}\left(\nabla_\theta \log P(\omega)Z - \nabla_\theta\mathbb{E}\left[Z\right]\right)\right] - \alpha\nabla_\theta\mathbb{E}\left[\bar{\xi}\right]\mathbb{E}\left[Z\right] \\ =&\nabla_\theta\mathbb{E}\left[Z\right] + \frac{\alpha}{\mathbb{SD}(Z)}\mathbb{E}\left[(Z - \mathbb{E}\left[Z\right])_+(\nabla_\theta \log P(\omega)Z - \nabla_\theta\mathbb{E}\left[Z\right])\right] \\ &-\alpha\nabla_\theta\mathbb{E}\left[\bar{\xi}\right]\mathbb{E}\left[Z\right] \\ =&\nabla_\theta\mathbb{E}\left[Z\right] + \frac{\alpha}{\mathbb{SD}(Z)}\mathbb{E}\Big[\big(Z - \mathbb{E}\left[Z\right]\big)_+(\nabla_\theta \log P(\omega)(Z - \mathbb{E}\left[Z\right]) \\ &\qquad - \nabla_\theta\mathbb{E}\left[Z\right])\Big].\end{aligned}$$

□

B. Proof of Theorem IV.2

First note from Assumption II.2 that

(i) Slater's condition holds in the primal optimization problem (1),

(ii) $L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$ is convex in ξ and concave in $(\lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$.

Thus by the duality result in convex optimization [35], the above conditions imply strong duality and we have $\rho(Z) = \max_{\xi\geq 0}\min_{\lambda^{\mathcal{P}},\lambda^{\mathcal{I}}\geq 0,\lambda^{\mathcal{E}}} L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) = \min_{\lambda^{\mathcal{P}},\lambda^{\mathcal{I}}\geq 0,\lambda^{\mathcal{E}}}\max_{\xi\geq 0} L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$. From Assumption II.2, one can also see that the family of functions $\{L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})\}_{(\xi,\lambda^{\mathcal{P}},\lambda^{\mathcal{E}},\lambda^{\mathcal{I}})\in\mathbb{R}^{|\Omega|}\times\mathbb{R}\times\mathbb{R}^{|\mathcal{E}|}\times\mathbb{R}^{|\mathcal{I}|}}$ is equi-differentiable in θ, $L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$ is Lipschitz, as a result, an absolutely continuous function in θ, and thus, $\nabla_\theta L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$ is continuous and bounded at each $(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$. Then for every selection of saddle point $(\xi^*_\theta, \lambda^{*,\mathcal{P}}_\theta, \lambda^{*,\mathcal{E}}_\theta, \lambda^{*,\mathcal{I}}_\theta) \in \mathcal{S}$ of (6), using the envelope theorem for saddle-point problems (see Theorem 4 of [34]), we have

$$\begin{aligned}&\nabla_\theta \max_{\xi\geq 0}\min_{\lambda^{\mathcal{P}},\lambda^{\mathcal{I}}\geq 0,\lambda^{\mathcal{E}}} L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \\ =&\nabla_\theta L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})|_{(\xi^*_\theta,\lambda^{*,\mathcal{P}}_\theta,\lambda^{*,\mathcal{E}}_\theta,\lambda^{*,\mathcal{I}}_\theta)}.\end{aligned} \tag{18}$$

The result follows by writing the gradient in (18) explicitly, and using the likelihood-ratio trick:

$$\begin{aligned}&\sum_{\omega\in\Omega}\xi(\omega)\nabla_\theta P_\theta(\omega)Z(\omega) - \lambda^{\mathcal{P}}\sum_{\omega\in\Omega}\xi(\omega)\nabla_\theta P_\theta(\omega) \\ =&\sum_{\omega\in\Omega}\xi(\omega)P(\omega)\nabla_\theta \log P(\omega)\left(Z(\omega) - \lambda^{\mathcal{P}}\right),\end{aligned}$$

where the last equality is justified by Assumption IV.1.

C. Proof of Proposition IV.4

Let $(\Omega_{SAA}, \mathcal{F}_{SAA}, P_{SAA})$ denote the probability space of the SAA functions (i.e., the randomness due to sampling).

Let $L_{\theta;N}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$ denote the Lagrangian of the SAA problem

$$\begin{aligned}&L_{\theta;N}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \\ =&\sum_{\omega\in\Omega}\xi(\omega)P_{\theta;N}(\omega)Z(\omega) - \lambda^{\mathcal{P}}\left(\sum_{\omega\in\Omega}\xi(\omega)P_{\theta;N}(\omega) - 1\right) \\ &- \sum_{e\in\mathcal{E}}\lambda^{\mathcal{E}}(e)f_e(\xi, P_{\theta;N}) - \sum_{i\in\mathcal{I}}\lambda^{\mathcal{I}}(i)f_i(\xi, P_{\theta;N}).\end{aligned} \tag{19}$$

Recall that $\mathcal{S} \subset \mathbb{R}^{|\Omega|}\times\mathbb{R}\times\mathbb{R}^{|\mathcal{E}|}\times\mathbb{R}^{|\mathcal{I}|}_+$ denotes the set of saddle points of the true Lagrangian (6). Let $\mathcal{S}_N \subset \mathbb{R}^{|\Omega|}\times\mathbb{R}\times\mathbb{R}^{|\mathcal{E}|}\times\mathbb{R}^{|\mathcal{I}|}_+$ denote the set of SAA Lagrangian (19) saddle points.

Suppose that there exists a compact set $C \equiv C_\xi \times C_\lambda$, where $C_\xi \subset \mathbb{R}^{|\Omega|}$ and $C_\lambda \subset \mathbb{R}\times\mathbb{R}^{|\mathcal{E}|}\times\mathbb{R}^{|\mathcal{I}|}_+$ such that:

(i) The set of Lagrangian saddle points $\mathcal{S} \subset C$ is non-empty and bounded.

(ii) The functions $f_e(\xi, P_\theta)$ for all $e \in \mathcal{E}$ and $f_i(\xi, P_\theta)$ for all $i \in \mathcal{I}$ are finite valued and continuous (in ξ) on C_ξ.

(iii) For N large enough the set $\mathcal{S}_N$ is non-empty and $\mathcal{S}_N \subset C$ w.p. 1.

Recall from Assumption II.2 that for each fixed $\xi \in \mathcal{B}$, both $f_i(\xi, p)$ and $g_e(\xi, p)$ are continuous in p. Furthermore, by the strong law of large numbers (S.L.L.N.) for Markov chains, for each policy parameter, we have $P_{\theta,N} \to P_\theta$ w.p. 1. From the definition of the Lagrangian function and continuity of constraint functions, one can easily see that for each $(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \in \mathbb{R}^{|\Omega|}\times\mathbb{R}\times\mathbb{R}^{|\mathcal{E}|}\times\mathbb{R}^{|\mathcal{I}|}_+$, $L_{\theta;N}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \to L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$ w.p. 1. Denote with $\mathbb{D}\left\{A, B\right\}$ the deviation of set A from set B, i.e., $\mathbb{D}\left\{A, B\right\} = \sup_{x\in A}\inf_{y\in B}\|x - y\|$. Further assume that:

(iv) If $\xi_N \in \mathcal{U}(P_{\theta;N})$ and ξ_N converges w.p. 1 to a point ξ, then $\xi \in \mathcal{U}(P_\theta)$.

According to the discussion in Page 161 of [6], the Slater condition of Assumption II.2 guarantees the following condition:

(v) For some point $\xi \in \mathcal{P}$ there exists a sequence $\xi_N \in \mathcal{U}(P_{\theta;N})$ such that $\xi_N \to \xi$ w.p. 1,

and from Theorem 6.6 in [6], we know that both sets $\mathcal{U}(P_{\theta;N})$ and $\mathcal{U}(P_\theta)$ are convex and compact. Furthermore, note that we have

(vi) The objective function on (1) is linear, finite valued and continuous in ξ on C_ξ (these conditions obviously hold for almost all $\omega \in \Omega$ in the integrand function $\xi(\omega)Z(\omega)$).

(vii) S.L.L.N. holds point-wise for any ξ.

From (i,iv,v,vi,vii), and under the same lines of proof as in Theorem 5.5 of [6], we have that

$$\rho_N(Z) \to \rho(Z) \text{ w.p. 1 as } N \to \infty, \tag{20}$$

$$\mathbb{D}\{\mathcal{P}_N, \mathcal{P}\} \to 0 \text{ w.p. 1 as } N \to \infty, \tag{21}$$

In part 1 and part 2 of the following proof, we show, by following similar derivations as in Theorem 5.2, Theorem 5.3 and Theorem 5.4 of [6], that $L_{\theta;N}(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) \to L_\theta(\xi^*_\theta, \lambda^{*,\mathcal{P}}_\theta, \lambda^{*,\mathcal{E}}_\theta, \lambda^{*,\mathcal{I}}_\theta)$ w.p. 1 and $\mathbb{D}\{\mathcal{S}_N, \mathcal{S}\} \to 0$ w.p. 1 as $N \to \infty$. Based on the definition of the deviation of sets, the limit point of any element in $\mathcal{S}_N$ is also an element in $\mathcal{S}$.

Assumptions (i) and (iii) imply that we can restrict our attention to the set C.

Part 1: We first show that $L_{\theta;N}(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N})$ converges to $L_\theta(\xi^*_\theta, \lambda^{*,\mathcal{P}}_\theta, \lambda^{*,\mathcal{E}}_\theta, \lambda^{*,\mathcal{I}}_\theta)$ w.p. 1 as $N \to \infty$.

For each fixed $(\lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \in C_\lambda$, the function $L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$ is convex and continuous in ξ. Together with the point-wise S.L.L.N. property, Theorem 7.49 of [6] implies that $L_{\theta;N}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) - L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \xrightarrow{e} 0$, where $\xrightarrow{e}$ denotes epi-convergence. Furthermore, since the objective and constraint functions are convex in ξ and are finite valued on C_ξ, the set $\mathrm{dom} L_\theta(\cdot, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$ has non-empty interior. It follows from Theorem 7.27 of [6] that epi-convergence of $L_{\theta,N}$ to L_θ implies uniform convergence on C_ξ, i.e., $\sup_{\xi \in C_\xi} \left| L_{\theta;N}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) - L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \right| \le \epsilon$. On the other hand, for each fixed $\xi \in C_\xi$, the function $L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$ is linear and thus continuous in $(\lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$ and $\mathrm{dom} L_\theta(\xi, \cdot, \cdot, \cdot) = \mathbb{R} \times \mathbb{R}^{|\mathcal{E}|} \times \mathbb{R}^{|\mathcal{I}|}$ has non-empty interior. It follows from analogous arguments that $\sup_{(\lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \in C_\lambda} \left| L_{\theta;N}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) - L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \right| \le \epsilon$. Combining these results implies that for any $\epsilon > 0$ and a.e. $\omega_{SAA} \in \Omega_{SAA}$ there is a $N^*(\epsilon, \omega_{SAA})$ such that

$$\sup_{(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \in C} \left| L_{\theta;N}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) - L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \right| \le \epsilon. \tag{22}$$

Now, assume by contradiction that for some $N > N^*(\epsilon, \omega_{SAA})$ we have $L_{\theta;N}(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) - L_\theta(\xi^*_\theta, \lambda^{*,\mathcal{P}}_\theta, \lambda^{*,\mathcal{E}}_\theta, \lambda^{*,\mathcal{I}}_\theta) > \epsilon$. Then by definition of the saddle points

$$\begin{aligned} L_{\theta;N}(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_\theta, \lambda^{*,\mathcal{E}}_\theta, \lambda^{*,\mathcal{I}}_\theta) \ge & L_{\theta;N}(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) \\ > & L_\theta(\xi^*_\theta, \lambda^{*,\mathcal{P}}_\theta, \lambda^{*,\mathcal{E}}_\theta, \lambda^{*,\mathcal{I}}_\theta) + \epsilon \\ \ge & L_\theta(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_\theta, \lambda^{*,\mathcal{E}}_\theta, \lambda^{*,\mathcal{I}}_\theta) + \epsilon, \end{aligned}$$

contradicting (22).

Similarly, assuming by contradiction that $L_\theta(\xi^*_\theta, \lambda^{*,\mathcal{P}}_\theta, \lambda^{*,\mathcal{E}}_\theta, \lambda^{*,\mathcal{I}}_\theta) - L_{\theta;N}(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) > \epsilon$ gives

$$\begin{aligned} L_\theta(\xi^*_\theta, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) \ge & L_\theta(\xi^*_\theta, \lambda^{*,\mathcal{P}}_\theta, \lambda^{*,\mathcal{E}}_\theta, \lambda^{*,\mathcal{I}}_\theta) \\ > & L_{\theta;N}(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) + \epsilon \\ \ge & L_{\theta;N}(\xi^*_\theta, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) + \epsilon, \end{aligned}$$

also contradicting (22).

It follows that

$$\left| L_{\theta;N}(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) - L_\theta(\xi^*_\theta, \lambda^{*,\mathcal{P}}_\theta, \lambda^{*,\mathcal{E}}_\theta, \lambda^{*,\mathcal{I}}_\theta) \right| \le \epsilon$$

for all $N > N^*(\epsilon, \omega_{SAA})$, and therefore

$$\lim_{N \to \infty} L_{\theta;N}(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) = L_\theta(\xi^*_\theta, \lambda^{*,\mathcal{P}}_\theta, \lambda^{*,\mathcal{E}}_\theta, \lambda^{*,\mathcal{I}}_\theta), \tag{23}$$

w.p. 1.

Part 2: Let us now show that $\mathbb{D}\{\mathcal{S}_N, \mathcal{S}\} \to 0$. We argue by a contradiction. Suppose that $\mathbb{D}\{\mathcal{S}_N, \mathcal{S}\} \nrightarrow 0$. Since C is compact, we can assume that there exists a sequence $(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) \in \mathcal{S}_N$ that converges to a point $(\bar{\xi}^*, \bar{\lambda}^{*,\mathcal{P}}, \bar{\lambda}^{*,\mathcal{E}}, \bar{\lambda}^{*,\mathcal{I}}) \in C$ and $(\bar{\xi}^*, \bar{\lambda}^{*,\mathcal{P}}, \bar{\lambda}^{*,\mathcal{E}}, \bar{\lambda}^{*,\mathcal{I}}) \notin \mathcal{S}$. However, from (21) we must have that $\bar{\xi}^* \in \mathcal{P}$. Therefore, we must have that

$$L_\theta(\bar{\xi}^*, \bar{\lambda}^{*,\mathcal{P}}, \bar{\lambda}^{*,\mathcal{E}}, \bar{\lambda}^{*,\mathcal{I}}) > L_\theta(\bar{\xi}^*, \lambda^{*,\mathcal{P}}_\theta, \lambda^{*,\mathcal{E}}_\theta, \lambda^{*,\mathcal{I}}_\theta),$$

by definition of the saddle point set.

Now,

$$\begin{aligned} & L_{\theta;N}(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) - L_\theta(\bar{\xi}^*, \bar{\lambda}^{*,\mathcal{P}}, \bar{\lambda}^{*,\mathcal{E}}, \bar{\lambda}^{*,\mathcal{I}}) \\ = & \left[L_{\theta;N}(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) - L_\theta(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) \right] \\ & + \left[L_\theta(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) - L_\theta(\bar{\xi}^*, \bar{\lambda}^{*,\mathcal{P}}, \bar{\lambda}^{*,\mathcal{E}}, \bar{\lambda}^{*,\mathcal{I}}) \right]. \end{aligned} \tag{24}$$

The first term in the r.h.s. of (24) tends to zero, using the argument from (22), and the second by continuity of L_θ guaranteed by (ii). We thus obtain that $L_{\theta;N}(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N})$ tends to $L_\theta(\bar{\xi}^*, \bar{\lambda}^{*,\mathcal{P}}, \bar{\lambda}^{*,\mathcal{E}}, \bar{\lambda}^{*,\mathcal{I}}) > L_\theta(\xi^*_\theta, \lambda^{*,\mathcal{P}}_\theta, \lambda^{*,\mathcal{E}}_\theta, \lambda^{*,\mathcal{I}}_\theta)$, which is a contradiction to (23).

Part 3: We now show the consistency of $\nabla_{\theta;N}\rho(Z)$. Consider Eq. (9). Since $\nabla_\theta \log P(\cdot)$ is bounded by Assumption IV.1, and $\nabla_\theta f_i(\cdot; P_\theta)$ and $\nabla_\theta g_e(\cdot; P_\theta)$ are bounded by Assumption II.2, and using our previous result $\mathbb{D}\{\mathcal{S}_N, \mathcal{S}\} \to 0$, we have that for a.e. $\omega_{SAA} \in \Omega_{SAA}$

$$\begin{aligned} \lim_{N \to \infty} \nabla_{\theta;N}\rho(Z) = & \sum_{\omega \in \Omega} P_\theta(\omega)\xi^*_\theta(\omega)\nabla_\theta \log P_\theta(\omega)(Z(\omega) - \lambda^{*,\mathcal{P}}_\theta) \\ & - \sum_{e \in \mathcal{E}} \lambda^{*,\mathcal{E}}_\theta(e)\nabla_\theta g_e(\xi^*_\theta; P_\theta) \\ & - \sum_{i \in \mathcal{I}} \lambda^{*,\mathcal{I}}_\theta(i)\nabla_\theta f_i(\xi^*_\theta; P_\theta) \\ = & \nabla_\theta \rho(Z). \end{aligned}$$

where the first equality is obtained from the envelope theorem (see Theorem IV.2) with $(\xi^*_\theta, \lambda^{*,\mathcal{P}}_\theta, \lambda^{*,\mathcal{E}}_\theta, \lambda^{*,\mathcal{I}}_\theta) \in \mathcal{S}_N \cap \mathcal{S}$ the limit point of the converging sequence $\{(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N})\}_{N \in \mathbb{N}}$.

D. Convergence Analysis of Empirical PRSVI

Lemma A.2 (Technical Lemma). *Let $P(\cdot|\cdot,\cdot)$ and $\widetilde{P}(\cdot|\cdot,\cdot)$ be two arbitrary transition probability matrices induced by policy $\mu(\cdot|\cdot)$. At state $x \in \mathcal{X}$, for any $\xi : \xi \circ P \circ \mu \in \mathcal{U}(x, P(\cdot|x,\cdot)\mu(\cdot|x))$, there exists a $M_\xi > 0$ such that for some $\tilde{\xi} : \tilde{\xi} \circ \widetilde{P} \circ \mu \in \mathcal{U}(x, \widetilde{P}(\cdot|x,\cdot)\mu(\cdot|x))$,*

$$\begin{aligned} & \sum_{x' \in \mathcal{X}, a \in \mathcal{A}} |\xi(a, x') - \tilde{\xi}(a, x')| \mu(a|x) \\ \le & M_\xi \sum_{x' \in \mathcal{X}, a \in \mathcal{A}} \left| P(x'|x, a) - \widetilde{P}(x'|x, a) \right| \mu(a|x). \end{aligned}$$

Proof. From Theorem II.1, we know that $\mathcal{U}(x, P(\cdot|x,\cdot)\mu(\cdot|x))$ is a closed, bounded, convex set of probability distribution functions. Since any conditional probability mass function P is in the interior of $\mathrm{dom}(\mathcal{U})$ and the graph of $\mathcal{U}(x, P(\cdot|x,\cdot)\mu(\cdot|x))$ is closed, by Theorem 2.7 in [51], $\mathcal{U}(x, P(\cdot|x,\cdot)\mu(\cdot|x))$ is a Lipschitz set-valued mapping with respect to the Hausdorff distance. Thus, for any $\xi : \xi \circ P \circ \mu \in \mathcal{U}(x, P(\cdot|x,\cdot)\mu(\cdot|x))$, the following expression holds for some $M_\xi > 0$:

$$\inf_{\hat{\xi}\,:\,\hat{\xi}\circ P\circ\mu\in\mathcal{U}(x,P(\cdot|x,\cdot)\mu(\cdot|x))} \sum_{x'\in\mathcal{X},a\in\mathcal{A}} |\xi(a,x') - \hat{\xi}(a,x')|\mu(a|x)$$
$$\leq M_\xi \sum_{x'\in\mathcal{X},a\in\mathcal{A}} \left|P(x'|x,a) - \widetilde{P}(x'|x,a)\right| \mu(a|x).$$

Next, we want to show that the infimum of the left side is attained. Since the objective function is convex, and $\mathcal{U}(x, \widetilde{P}(\cdot|x)\mu(\cdot|x))$ is a convex compact set, there exists $\hat{\xi}^* : \hat{\xi}^* \circ \widetilde{P} \circ \mu \in \mathcal{U}(x, \widetilde{P}(\cdot|x)\mu(\cdot|x))$ such that infimum is attained. □

Lemma A.3 (Strong Law of Large Number). *Consider the sampling based PRSVI algorithm with update sequence $\{\widehat{v}_k\}$. Then as both N and k tend to ∞, $\widehat{v}_k$ converges with probability 1 to v^*_θ, the unique solution of projected risk sensitive fixed point equation $\Pi T_\mu[\Phi v] = \Phi v$.*

Proof. By the strong law of large number of Markov process, the empirical visiting distribution and transition probability asymptotically converges to their statistical limits with probability 1, i.e., for any $x, x' \in \mathcal{X}$ and $a \in \mathcal{A}$,

$$\frac{\sum_{t=0}^{N-1} \mathbf{1}\{x_t = x\}}{N} \to d_\theta(x|x_0), \text{ and } \widehat{P}(x'|x,a) \to P(x'|x,a).$$

Therefore with probability 1,

$$\frac{1}{N}\sum_{t=0}^{N-1} \phi(x_t)\phi(x_t)^\top \to \sum_x d_\theta(x|x_0)\cdot\phi(x)\phi^\top(x).$$

Now we show that following expression holds with probability 1:

$$\max_{\xi\,:\,\xi\circ P_{\theta;N}\circ\mu_\theta\in\mathcal{U}(x_t,P_{\theta;N}\circ\mu_\theta)} \sum_{x'\in\mathcal{X},a\in\mathcal{A}} \xi(a,x')P_{\theta;N}(x'|x_t,a)\cdot$$
$$\left\{C(x_t,a) + \gamma v^\top \phi\left(x'\right)\right\} + \frac{1}{2N}(\xi(a,x')P_{\theta;N}(x'|x_t,a))^2$$
$$\to \max_{\xi\,:\,\xi\circ P\circ\mu_\theta\in\mathcal{U}(x_t,P\circ\mu_\theta)} \sum_{x',a} \xi(a,x')P(x'|x_t,a)\left\{C(x_t,a)+\gamma v^\top\phi\left(x'\right)\right\}. \tag{25}$$

Notice that for

$$\{\xi^*_{\theta,x_t;N}(a,x')\}_{a\in\mathcal{A},x'\in\mathcal{X}} \in \arg \max_{\xi\,:\,\xi\circ P_{\theta;N}\circ\mu_\theta\in\mathcal{U}(x_t,P_{\theta;N}\circ\mu_\theta)}$$
$$\sum_{x'\in\mathcal{X},a\in\mathcal{A}} \xi(a,x')P_{\theta;N}(x'|x_t,a)\mu_\theta(a|x_t)\left\{C(x_t,a)+\gamma v^\top\phi\left(x'\right)\right\},$$

Lemma A.2 implies

$$\max_{\xi\,:\,\xi\circ P_{\theta;N}\circ\mu_\theta\in\mathcal{U}(x_t,P_{\theta;N}\circ\mu_\theta)} \sum_{x',a} \xi(a,x')P_{\theta;N}(x'|x_t,a)\mu_\theta(a|x_t)\cdot$$
$$\left\{C(x_t,a) + \gamma v^\top \phi\left(x'\right)\right\} + \frac{1}{2N}(\xi(a,x')P_{\theta;N}(x'|x_t,a))^2$$
$$- \max_{\xi\,:\,\xi\circ P\circ\mu_\theta\in\mathcal{U}(x_t,P\circ\mu_\theta)} \sum_{x'\in\mathcal{X},a\in\mathcal{A}} \xi(a,x')P(x'|x_t,a)\mu_\theta(a|x_t)\cdot$$
$$\left\{C(x_t,a) + \gamma v^\top \phi\left(x'\right)\right\}$$
$$\leq \{C_{\max} + \gamma\|\Phi v\|_\infty\} \left(M_{\xi^*_{\theta,x_t;N}} + \max_{x\in\mathcal{X},a\in\mathcal{A}} |\xi^*_{\theta,x_t;N}(a,x)|\right)$$
$$\sum_{x'\in\mathcal{X},a\in\mathcal{A}} |P_\theta(x'|x_t,a) - P_{\theta;N}(x'|x_t,a)|\,\mu(a|x_t) + \frac{1}{2N}.$$

The quantity $\max_{x\in\mathcal{X},a\in\mathcal{A}} |\xi^*_{\theta,x_t;N}(a,x)|$ is bounded because $\mathcal{U}(x_t, P_{\theta;N}(\cdot|x_t,\cdot)\mu(\cdot|x_t))$ is a closed and bounded convex set from the definition of coherent risk measures. By repeating the above analysis by interchanging P_θ and $P_{\theta;N}$ and combining previous arguments, one obtains

$$\Bigg| \max_{\xi\,:\,\xi\circ P_{\theta;N}\circ\mu_\theta\in\mathcal{U}(x_t,P_{\theta;N}\circ\mu_\theta)} \sum_{x',a} \xi(a,x')P_{\theta;N}(x'|x_t,a)\mu_\theta(a|x_t)$$
$$\left\{C(x_t,a) + \gamma v^\top \phi\left(x'\right)\right\} + \frac{1}{2N}(\xi(a,x')P_{\theta;N}(x'|x_t,a))^2$$
$$- \max_{\xi\,:\,\xi\circ P\circ\mu_\theta\in\mathcal{U}(x_t,P\circ\mu_\theta)} \sum_{x',a} \xi(a,x')P(x'|x_t,a)\mu_\theta(a|x_t)\cdot$$
$$\left\{C(x_t,a) + \gamma v^\top \phi\left(x'\right)\right\}\Bigg| \leq \frac{1}{2N} + \{C_{\max} + \gamma\|\Phi v\|_\infty\}\cdot$$
$$\max\left\{\left(M_{\xi^*} + \max_{x,a}|\xi^*(a,x)|\right), \left(M_{\xi^*_{\theta,x_t;N}} + \max_{x,a}|\xi^*_{\theta,x_t;N}(a,x)|\right)\right\}\cdot$$
$$\sum_{x',a} \mu_\theta(a|x_t)\left|P(x'|x_t,a) - P_{\theta;N}(x'|x_t,a)\right|.$$

Therefore, the claim in expression (25) holds when $N \to \infty$ and $\sum_{x'\in\mathcal{X},a\in\mathcal{A}} \mu_\theta(a|x_t)\left|P_\theta(x'|x_t,a) - P_{\theta;N}(x'|x_t,a)\right| \to 0$. On the other hand, the strong law of large numbers also implies that with probability 1, for $x_0 = x$,

$$\frac{1}{N}\sum_{t=0}^{N-1} \phi(x_t)\rho(\Phi v_t) \to d_\theta(x|x_0)\phi(x) \max_{\xi\,:\,\xi\circ P\circ\mu_\theta\in\mathcal{U}(x,P(\cdot|x,\cdot)\mu_\theta(\cdot|x))} \cdot$$
$$\sum_{x'\in\mathcal{X},a\in\mathcal{A}} \xi(a,x')P(x'|x,a)\mu_\theta(a|x)\left\{C(x,a) + \gamma {v^*_\theta}^\top \phi\left(x'\right)\right\}.$$

As $N \to \infty$, the above arguments imply that $v_k - \widehat{v}_k \to 0$. On the other hand, Proposition 1 in [20] implies that the projected risk sensitive Bellman operator $\Pi T_\theta[V]$ is a contraction, it follows that from the analysis in Section 6.3 in [13] that the sequence $\{\Phi\widehat{v}_k\}$ generated by projected value iteration converges to the unique fixed point Φv^*_θ. This in turns implies that the sequence $\{\Phi v_k\}$ converges to Φv^*_θ. □

E. Proof of Theorem V.4

Similar to the proof of Theorem IV.2, recall the saddle point definition of $(\xi^*_{\theta,x}, \lambda^{*,\mathcal{P}}_{\theta,x}, \lambda^{*,\mathcal{E}}_{\theta,x}, \lambda^{*,\mathcal{I}}_{\theta,x}) \in \mathcal{S}$ and strong duality result, i.e.,

$$\max_{\xi\in\mathcal{U}(x,P(\cdot|x,\cdot)\mu_\theta(\cdot|x))} \mathbb{E}_{\xi P(\cdot|x,\cdot)\mu_\theta(\cdot|x)}[C_\theta(x,\cdot) + \gamma V_\theta]$$
$$= \max_{\xi\geq 0} \min_{\lambda^{\mathcal{P}},\lambda^{\mathcal{I}}\geq 0,\lambda^{\mathcal{E}}} L_{\theta,x}(\xi,\lambda^{\mathcal{P}},\lambda^{\mathcal{E}},\lambda^{\mathcal{I}})$$
$$= \min_{\lambda^{\mathcal{P}},\lambda^{\mathcal{I}}\geq 0,\lambda^{\mathcal{E}}} \max_{\xi\geq 0} L_{\theta,x}(\xi,\lambda^{\mathcal{P}},\lambda^{\mathcal{E}},\lambda^{\mathcal{I}}).$$

the gradient formula in (18) can be written as

$$\nabla_\theta V_\theta(x) = \nabla_\theta \left[\max_{\xi \in \mathcal{U}(x, P \circ \mu_\theta)} \mathbb{E}_{\xi P(\cdot|x,\cdot)\mu_\theta(\cdot|x)} [C_\theta(x,\cdot) + \gamma V_\theta] \right] = \gamma \sum_{x',a} \mu_\theta(a|x) \{ \xi^*_{\theta,x}(a,x') P(x'|x,a) \nabla_\theta V_\theta(x') + \nabla_\theta \log \mu_\theta(a|x) h_\theta(x,a) \},$$

where the stage-wise cost function $h_\theta(x,a)$ is defined in (13). By defining $\widehat{h}_\theta(x,a) = \nabla_\theta \log \mu_\theta(a|x) h_\theta(x,a)$ and unfolding the recursion, the above expression implies

$$\nabla_\theta V_\theta(x_0) = \sum_{a_0 \in \mathcal{A}} \widehat{h}_\theta(x_0,a_0) + \gamma \sum_{x_1 \in \mathcal{X}} P_\theta(x_1|x_0,a_0) \xi^*_\theta(a_0,x_1) \left[\sum_{a_1 \in \mathcal{A}} \widehat{h}_\theta(x_1,a_1) + \gamma \sum_{x_2 \in \mathcal{X}} P_\theta(x_2|x_1,a_1) \xi^*_\theta(a_1,x_2) \nabla_\theta V_\theta(x_2) \right].$$

Now since $\nabla_\theta V_\theta$ is continuously differentiable with bounded derivatives, when $t \to \infty$, one obtains $\gamma^t \nabla_\theta V_\theta(x) \to 0$ for any $x \in \mathcal{X}$. Therefore, by Bounded Convergence Theorem, $\lim_{t\to\infty} \rho(\gamma^t V_\theta(x_t)) = 0$, when $x_0 = x$ the above expression implies the result of this theorem.

F. Technical Results in Section V-C

Since by convention $\xi^*_{\theta,x;N}(a,x') = 0$ whenever $P_{\theta;N}(x'|x,a) = 0$. In this section, we simplify the analysis by letting $P_{\theta;N}(x'|x,a) > 0$ for any $x' \in \mathcal{X}$ without loss of generality. Consider the following empirical robust optimization problem:

$$\max_{\xi \in \mathcal{U}(x, P_{\theta;N}(\cdot|x,\cdot)\mu_\theta(\cdot|x))} \sum_{x' \in \mathcal{X}, a \in \mathcal{A}} \mu_\theta(a|x) P_{\theta;N}(x'|x,a) \xi(a,x') \{ C(x,a) + \gamma V_\theta(x') \}, \tag{26}$$

where the solution of the above empirical problem is $\bar{\xi}^*_{\theta,x;N}$ and the corresponding KKT multipliers are $(\bar{\lambda}^{*,\mathcal{P}}_{\theta,x;N}, \bar{\lambda}^{*,\mathcal{E}}_{\theta,x;N}, \bar{\lambda}^{*,\mathcal{I}}_{\theta,x;N})$. Comparing to the optimization problem for $\rho_N(C(x,\cdot) + \gamma \Phi v)$, i.e.,

$$\begin{aligned} &\rho_N(C(x,\cdot) + \gamma \Phi v) \\ &= \max_{\xi \in \mathcal{U}(x, P_{\theta;N}(\cdot|x,\cdot)\mu_\theta(\cdot|x))} \sum_{x' \in \mathcal{X}, a \in \mathcal{A}} \frac{1}{2N} (\xi(a,x') P_{\theta;N}(x'|x,a))^2 \\ &\quad + \mu_\theta(a|x) P_{\theta;N}(x'|x,a) \xi(a,x') \left\{ C(x,a) + \gamma \phi^\top(x') v \right\}, \end{aligned} \tag{27}$$

where the solution of the above empirical problem is $\xi^*_{\theta,x;N}$ and the corresponding KKT multipliers are $(\lambda^{*,\mathcal{P}}_{\theta,x;N}, \lambda^{*,\mathcal{E}}_{\theta,x;N}, \lambda^{*,\mathcal{I}}_{\theta,x;N})$, the optimization problem in (26) can be viewed as having a skewed objective function of the problem in (27), within the deviation of magnitude $\Delta + 1/2N$ where $\Delta = \|\Phi v^*_\theta - V_\theta\|_\infty$. Before getting into the main analysis, we have the following observations.

(i) Without loss of generality, we can also assume $(\xi^*_{\theta,x;N}, (\lambda^{*,\mathcal{P}}_{\theta,x;N}, \lambda^{*,\mathcal{E}}_{\theta,x;N}, \lambda^{*,\mathcal{I}}_{\theta,x;N}))$ follows the strict complementary slackness condition[9].

(ii) Recall from Assumption II.2 that the functions $f_i(\xi,p)$ and $g_e(\xi,p)$ are twice differentiable in ξ at $p = P_{\theta,N}(\cdot|x)$ for any $x \in \mathcal{X}$.

(iii) The Slater's condition in Assumption II.2 implies the linear independence constraint qualification (LICQ).

(iv) Since optimization problem (27) has a convex objective function and convex/affine constraints in $\xi \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{A}|}$, equipped with the Slater's condition we have that the first order KKT condition holds at $\xi^*_{\theta,x;N}$ with the corresponding KKT multipliers are $(\lambda^{*,\mathcal{P}}_{\theta,x;N}, \lambda^{*,\mathcal{E}}_{\theta,x;N}, \lambda^{*,\mathcal{I}}_{\theta,x;N})$. Furthermore, define the Lagrangian function

$$\begin{aligned} &\widehat{L}_{\theta;N}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \\ &\doteq \sum_{x' \in \mathcal{X}, a \in \mathcal{A}} \mu_\theta(a|x) P_{\theta;N}(x'|x,a) \xi(a,x') \cdot \\ &\quad \{ C(x,a) + \gamma \phi^\top(x') v \} + \frac{(P_{\theta;N}(x'|x,a)\xi(a,x'))^2}{2N} \\ &\quad - \lambda^{\mathcal{P}} \left(\sum_{x' \in \mathcal{X}, a \in \mathcal{A}} \mu_\theta(a|x) \xi(a,x') P_{\theta;N}(x'|x,a) - 1 \right) \\ &\quad - \sum_{e \in \mathcal{E}} \lambda^{\mathcal{E}}(e) f_e(\xi, P_{\theta;N}(\cdot|x,\cdot) \circ \mu_\theta(\cdot|x)) \\ &\quad - \sum_{i \in \mathcal{I}} \lambda^{\mathcal{I}}(i) f_i(\xi, P_{\theta;N}(\cdot|x,\cdot) \circ \mu_\theta(\cdot|x)). \end{aligned}$$

One can easily conclude that

$$\begin{aligned} \nabla^2 \widehat{L}_{\theta;N}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) &= -\frac{P_{\theta;N}(\cdot|x,\cdot)^\top P_{\theta;N}(\cdot|x,\cdot)}{N} \\ &\quad - \sum_{i \in \mathcal{I}} \lambda^{\mathcal{I}}(i) \nabla^2_\xi f_i(\xi, P_{\theta;N}(\cdot|x,\cdot) \circ \mu_\theta(\cdot|x)) \end{aligned}$$

such that for any vector $\nu \neq 0$,

$$\nu^\top \nabla^2 \widehat{L}_{\theta;N}(\xi^*_{\theta,x;N}, \lambda^{*,\mathcal{P}}_{\theta,x;N}, \lambda^{*,\mathcal{E}}_{\theta,x;N}, \lambda^{*,\mathcal{I}}_{\theta,x;N}) \nu < 0,$$

which further implies that the second order sufficient condition (SOSC) holds at $(\xi^*_{\theta,x;N}, \lambda^{*,\mathcal{P}}_{\theta,x;N}, \lambda^{*,\mathcal{E}}_{\theta,x;N}, \lambda^{*,\mathcal{I}}_{\theta,x;N})$.

Based on all the above analysis, we have the following sensitivity result from Corollary 3.2.4 in [52], derived based on Implicit Function Theorem.

Proposition A.4 (Basic Sensitivity Theorem)**.** *Under the Assumption II.2, for any $x \in \mathcal{X}$ there exists a bounded non-singular matrix $K_{\theta,x}$ and a bounded vector $L_{\theta,x}$, such that the difference between the optimizers and KKT multipliers of optimization problem* (26) *and* (27) *are bounded as follows:*

$$\begin{bmatrix} \bar{\xi}^*_{\theta,x;N} \\ \bar{\lambda}^{*,\mathcal{I}}_{\theta,x;N} \\ \bar{\lambda}^{*,\mathcal{P}}_{\theta,x;N} \\ \bar{\lambda}^{*,\mathcal{E}}_{\theta,x;N} \end{bmatrix} = \begin{bmatrix} \xi^*_{\theta,x;N} \\ \lambda^{*,\mathcal{I}}_{\theta,x;N} \\ \lambda^{*,\mathcal{P}}_{\theta,x;N} \\ \lambda^{*,\mathcal{E}}_{\theta,x;N} \end{bmatrix} + \Phi^{-1}_{\theta,x} \Psi_{\theta,x} \left(\Delta + \frac{1}{2N} \right) + o\left(\Delta + \frac{1}{2N} \right).$$

On the other hand, we know from Proposition IV.4 that $\bar{\xi}^*_{\theta,x;N} \to \xi^*_{\theta,x}$ and $(\bar{\lambda}^{*,\mathcal{P}}_{\theta,x;N}, \bar{\lambda}^{*,\mathcal{E}}_{\theta,x;N}, \bar{\lambda}^{*,\mathcal{I}}_{\theta,x;N}) \to (\lambda^{*,\mathcal{P}}_{\theta,x}, \lambda^{*,\mathcal{E}}_{\theta,x}, \lambda^{*,\mathcal{I}}_{\theta,x})$ with probability 1 as $N \to \infty$. Also recall from the law of large numbers that the sampled approximation error $\max_{x \in \mathcal{X}, a \in \mathcal{A}} \|P(\cdot|x,a) - P_N(\cdot|x,a)\|_1 \to 0$ almost surely as $N \to \infty$. Then we have the following error bound in the stage-wise cost approximation $\widehat{h}_{\theta;N}(x,a)$ and $\gamma-$visiting distribution $\pi_N(x,a)$.

Lemma A.5. *There exists a constant $M_h > 0$ such that* $\max_{x \in \mathcal{X}, a \in \mathcal{A}} |h_\theta(x,a) - \lim_{N\to\infty} \widehat{h}_{\theta;N}(x,a)| \le M_h \Delta$.

[9] The existence of strict complementary slackness solution follows from the KKT theorem and one can easily construct a strictly complementary pair using i.e. the Balinski-Tucker tableau with the linearized objective function and constraints, in finite time.

Proof. First we can easily see that for any state $x \in \mathcal{X}$ and action $a \in \mathcal{A}$,

$$
\begin{aligned}
&|\widehat{h}_{\theta;N}(x,a) - h_\theta(x,a)| \leq \\
&M \sum_{i\in\mathcal{I}} \left|\lambda^{*,\mathcal{I}}_{\theta,x;N}(i) - \lambda^{*,\mathcal{I}}_{\theta,x}(i)\right| \\
&+ M \sum_{e\in\mathcal{E}} \left|\lambda^{*,\mathcal{E}}_{\theta,x;N}(e) - \lambda^{*,\mathcal{E}}_{\theta,x}(e)\right| + \left|\lambda^{*,\mathcal{P}}_{\theta,x;N} - \lambda^{*,\mathcal{P}}_{\theta,x}\right| \\
&+ \gamma\|V_\theta\|_\infty \|\xi^*_{\theta,x;N} - \xi^*_{\theta,x}\|_1 + \gamma\|V_\theta - \Phi v^*_\theta\|_\infty \\
&+ \gamma\|V_\theta\|_\infty \max\{\|\xi^*_{\theta,x;N}\|_\infty, \|\xi^*_{\theta,x}\|_\infty\}\|P(\cdot|x,a) - P_N(\cdot|x,a)\|_1.
\end{aligned}
$$

Note that at $N \to \infty$, $\|P(\cdot|x,a) - P_N(\cdot|x,a)\|_1 \to 0$ with probability 1. Both $\|\xi^*_{\theta;N}\|_\infty$ and $\|\xi^*_{\theta,x}\|_\infty$ are finite valued because $\mathcal{U}(P_\theta)$ and $\mathcal{U}(P_{\theta;N})$ are convex compact sets of real vectors. Therefore, by noting that $\|V_\theta\|_\infty \leq C_{\max}/(1-\gamma)$ and applying Proposition IV.4 and A.4, the proof of this Lemma is completed by letting $N \to \infty$ and defining

$$
M_h(x) = \max\left\{1, M, \frac{\gamma C_{\max}}{1-\gamma}\right\}.
$$

$$
\begin{aligned}
&\left\| \begin{bmatrix} \xi^*_{\theta,x;N} - \bar{\xi}^*_{\theta,x;N} \\ \lambda^{*,\mathcal{I}}_{\theta,x;N} - \bar{\lambda}^{*,\mathcal{I}}_{\theta,x;N} \\ \lambda^{*,\mathcal{P}}_{\theta,x;N} - \bar{\lambda}^{*,\mathcal{P}}_{\theta,x;N} \\ \lambda^{*,\mathcal{E}}_{\theta,x;N} - \bar{\lambda}^{*,\mathcal{E}}_{\theta,x;N} \end{bmatrix} + \begin{bmatrix} \bar{\xi}^*_{\theta,x;N} - \xi^*_{\theta,x} \\ \bar{\lambda}^{*,\mathcal{I}}_{\theta,x;N} - \lambda^{*,\mathcal{I}}_{\theta,x} \\ \bar{\lambda}^{*,\mathcal{P}}_{\theta,x;N} - \lambda^{*,\mathcal{P}}_{\theta,x} \\ \bar{\lambda}^{*,\mathcal{E}}_{\theta,x;N} - \lambda^{*,\mathcal{E}}_{\theta,x} \end{bmatrix} \right\|_1 + \gamma\Delta \\
&\leq \left(\max\{1, M, \frac{\gamma C_{\max}}{1-\gamma}\}\|\Phi^{-1}_{\theta,x}\Psi_{\theta,x}\|_1 + \gamma\right)\Delta.
\end{aligned}
$$

□

Lemma A.6. *There exists a constant $M_\pi > 0$ such that $\|\pi - \lim_{N\to\infty}\pi_N\|_1 \leq M_\pi \Delta$.*

Proof. First, recall that the $\gamma-$visiting distribution satisfies the following identity:

$$
\begin{aligned}
&\gamma \sum_{x'\in\mathcal{X}} d_\theta(x'|x) \sum_{a\in\mathcal{A}} P(x'|x,a)\xi(a,x')\mu_\theta(a|x) \\
=&d_\theta(x) - (1-\gamma)\mathbf{1}\{x_0 = x\},
\end{aligned} \tag{28}
$$

By defining $P^\xi_\theta(x'|x) = \sum_{a\in\mathcal{A}} P(x'|x,a)\xi(a,x')\mu_\theta(a|x)$, from here one easily notice this expression can be rewritten as follows:

$$
\left(I - \gamma P^\xi_\theta\right)^\top d_\theta(\cdot|x) = \mathbf{1}\{x_0 = x\},\ \forall x \in \mathcal{X}.
$$

On the other hand, by repeating the analysis with $P_{\theta;N}(\cdot|x,\cdot) \circ \mu_\theta(\cdot|x)$ and defining

$$
P^\xi_{\theta;N}(x'|x) = \sum_{a\in\mathcal{A}} P_{\theta;N}(x'|x,a)\xi(a,x')\mu_\theta(a|x),
$$

we can also write $\left(I - \gamma P^\xi_{\theta;N}\right)^\top d_{\theta;N} = \{\mathbf{1}\{x_0 = z\}\}_{z\in\mathcal{X}}$. Combining the above expressions implies for any $x \in \mathcal{X}$,

$$
d_\theta - d_{\theta;N} - \gamma\left(\left(P^\xi_\theta\right)^\top d_\theta - (P^\xi_{\theta;N})^\top d_{\theta;N}\right) = 0,
$$

which further implies

$$
\begin{aligned}
&\left(I - \gamma P^\xi_\theta\right)^\top (d_\theta - d_{\theta;N}) = \gamma\left(P^\xi_\theta - P^\xi_{\theta;N}\right)^\top d_{\theta;N} \\
\Longleftrightarrow\ &(d_\theta - d_{\theta;N}) = \left(I - \gamma P^\xi_\theta\right)^{-\top} \gamma\left(P^\xi_\theta - P^\xi_{\theta;N}\right)^\top d_{\theta;N}.
\end{aligned}
$$

Notice that with transition probability matrix $P^\xi_\theta(\cdot|x)$, we have $(I - \gamma P^\xi_\theta)^{-1} = \sum_{t=0}^\infty \left(\gamma P^\xi_\theta\right)^k < \infty$. The series is summable because by Perron-Frobenius theorem, the maximum eigenvalue of P^ξ_θ is less than or equal to 1 and $I - \gamma P^\xi_\theta$ is invertible. On the other hand, for every given $x_0 \in \mathcal{X}$, and any $z' \in \mathcal{X}$,

$$
\begin{aligned}
&\left\{\left(P^\xi_\theta - P^\xi_{\theta;N}\right)^\top d_{\theta;N}\right\}(z') \\
=&\sum_{x\in\mathcal{X}}\sum_{k=0}^\infty \gamma^k(1-\gamma)\mathbb{P}_{P^\xi_{\theta;N}}(x_k = x|x_0)\left(P^\xi_\theta(z'|x) - P^\xi_{\theta;N}(z'|x)\right), \\
=&\mathbb{E}_{P^\xi_{\theta;N}}\left(\sum_{k=0}^\infty \gamma^k(1-\gamma)\left(P^\xi_\theta(z'|x_k) - P^\xi_{\theta;N}(z'|x_k)\right)|x_0\right), \\
\leq&\mathbb{E}_{P^\xi_{\theta;N}}\left(\sum_{k=0}^\infty \gamma^k(1-\gamma)\left|P^\xi_\theta(z'|x_k) - P^\xi_{\theta;N}(z'|x_k)\right||x_0\right) \doteq \mathcal{Q}(z').
\end{aligned}
$$

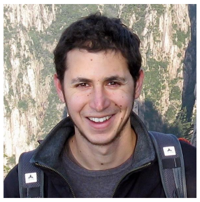
Note that every element in matrix $(I - \gamma P^\xi_\theta)^{-1} = \sum_{t=0}^\infty \left(\gamma P^\xi_\theta\right)^k$ is non-negative. This implies for any $z \in \mathcal{X}$,

$$
\begin{aligned}
&|\{d_\theta - d_{\theta;N}\}(z)| = \left|\left\{\left(I - \gamma P^\xi_\theta\right)^{-\top}\gamma\left(P^\xi_\theta - P^\xi_{\theta;N}\right)^\top d_{\theta;N}\right\}(z)\right|, \\
\leq&\left|\left\{\left(I - \gamma P^\xi_\theta\right)^{-\top}\gamma\mathcal{Q}\right\}(z)\right| = \left\{\left(I - \gamma P^\xi_\theta\right)^{-\top}\gamma\mathcal{Q}\right\}(z).
\end{aligned}
$$

The last equality is due to the fact that every element in vector $\mathcal{Q}$ is non-negative. Combining the above results with Proposition IV.4 and A.4, and noting that $(I - \gamma P^\xi_\theta)^{-1}e = \sum_{t=0}^\infty \left(\gamma P^\xi_\theta\right)^k e = \frac{1}{1-\gamma}e$, we further have that

$$
\begin{aligned}
&\|\pi_\theta - \pi_{\theta;N}\|_1 \leq \|d_\theta - d_{\theta;N}\|_1 \\
\leq&e^\top\left(I - \gamma P^\xi_\theta\right)^{-\top}\gamma\mathcal{Q} \\
=&\frac{\gamma}{1-\gamma}e^\top\mathcal{Q} \\
\leq&\frac{\gamma}{1-\gamma}\max_{x\in\mathcal{X}}\left\|P^\xi_\theta(\cdot|x) - P^\xi_{\theta;N}(\cdot|x)\right\|_1 \\
\leq&\frac{\gamma}{1-\gamma}\max_{x\in\mathcal{X},a\in\mathcal{A}}\Big(\|\xi^*_{\theta,x}(\cdot,\cdot) - \xi^*_{\theta,x;N}(\cdot,\cdot)\|_1\|P(\cdot|x,\cdot)\mu_\theta(\cdot|x)\|_\infty \\
&\quad + \max\{\|\xi^*_{\theta,x;N}\|_\infty, \|\xi^*_{\theta,x}\|_\infty\}\|P(\cdot|x,a) - P_N(\cdot|x,a)\|_1\Big).
\end{aligned}
$$

As in previous arguments, when $N \to \infty$, one obtains $\|P(\cdot|x,a) - P_N(\cdot|x,a)\|_1 \to 0$ with probability 1 and $\|\xi^*_{\theta,x}(\cdot,\cdot) - \xi^*_{\theta,x;N}(\cdot,\cdot)\|_1 \to 0$. We thus set the constant M_π as $\gamma\|\Phi^{-1}_{\theta,x}\Psi_{\theta,x}\|_1/(1-\gamma)$. □



Aviv Tamar received the M.Sc. and Ph.D. degrees in electrical engineering from the Technion - Israel Institute of Technology, Haifa, Israel, in 2011 and 2015, respectively. Since 2015, he is a Post-Doc scholar at the EECS department of the University of California, Berkeley. His research interests include reinforcement-learning, planning, machine learning, and risk-sensitive decision-making.

Yinlam Chow (S'09) received the B.Eng degree in mechanical engineering from the University of Hong Kong in 2009, the M.Sc. degree in aerospace engineering from Purdue University in 2011 and is currently a Ph.D. candidate in ICME, Stanford University. His research interests include control theory, machine learning, sequential decision making and reinforcement-learning.

Mohammad Ghavamzadeh received a Ph.D. degree in Computer Science from the University of Massachusetts Amherst in 2005. From 2005 to 2008, he was a postdoctoral fellow at the Department of Computing Science at the University of Alberta. He has been a permanent researcher (Charg de Recherche) at INRIA in France since November 2008. He was promoted to "Charg de Recherche premiere class" (CR1) in 2010, was the recipient of the "INRIA award for scientific excellence" in 2011, and successfully defended his Habilitation Diriger des Recherches (HDR) thesis in 2014. He is currently (from October 2013) on a leave of absence from INRIA working as a senior analytics researcher at Adobe Research in California, on projects in the area of digital marketing. His research is in the areas of machine learning, artificial intelligence, control, and learning theory; particularly to investigate the principles of scalable decision-making and to devise, analyze, and implement algorithms for sequential decision-making under uncertainty and reinforcement learning.

Shie Mannor (S'00-M'03-SM-09') received the B.Sc. degree in electrical engineering, the B.A. degree in mathematics, and the Ph.D. degree in electrical engineering from the Technion-Israel Institute of Technology, Haifa, Israel, in 1996, 1996, and 2002, respectively. From 2002 to 2004, he was a Fulbright scholar and a postdoctoral associate at M.I.T. He was with the Department of Electrical and Computer Engineering at McGill University from 2004 to 2010 where he was the Canada Research chair in Machine Learning. He has been with the Faculty of Electrical Engineering at the Technion since 2008 where he is currently a professor. His research interests include machine learning and pattern recognition, planning and control, multi-agent systems, and communications.