Policy Gradient for Coherent Risk Measures

Aviv Tamar
UC Berkeley
avivt@berkeley.edu

Yinlam Chow
Stanford University
ychow@stanford.edu

Mohammad Ghavamzadeh
Adobe Research & INRIA
mohammad.ghavamzadeh@inria.fr

Shie Mannor
Technion
shie@ee.technion.ac.il

Abstract

Several authors have recently developed risk-sensitive policy gradient methods that augment the standard expected cost minimization problem with a measure of *variability* in cost. These studies have focused on *specific* risk-measures, such as the variance or conditional value at risk (CVaR). In this work, we extend the policy gradient method to *the whole class* of coherent risk measures, which is widely accepted in finance and operations research, among other fields. We consider both static and time-consistent dynamic risk measures. For static risk measures, our approach is in the spirit of *policy gradient* algorithms and combines a standard sampling approach with convex programming. For dynamic risk measures, our approach is *actor-critic* style and involves explicit approximation of value function. Most importantly, our contribution presents a *unified* approach to risk-sensitive reinforcement learning that generalizes and extends previous results.

1 Introduction

Risk-sensitive optimization considers 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 of finance and operations research. In reinforcement learning (RL) [33], risk-sensitive objectives have gained popularity as a means to regularize the variability of the total (discounted) cost/reward in a Markov decision process (MDP).

Many risk objectives have been investigated in the literature and applied to RL, such as the celebrated Markowitz mean-variance model [19], Value-at-Risk (VaR) and Conditional Value at Risk (CVaR) [22, 35, 26, 12, 10, 36]. The view taken in this paper is that the preference of one risk measure over another is *problem-dependent* and depends on factors such as the cost distribution, sensitivity to rare events, ease of estimation from data, and computational tractability of the optimization problem. However, the highly influential paper of Artzner et al. [2] identified a set of natural properties that are desirable for a risk measure to satisfy. Risk measures that satisfy these properties are termed *coherent* and have obtained widespread acceptance in financial applications, among others. We focus on such coherent measures of risk in this work.

For sequential decision problems, such as MDPs, 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 [5]). The recently proposed class of dynamic Markov coherent risk measures [30] satisfies both the coherence and time consistency properties.

In this work, we present policy gradient algorithms for RL with a coherent risk objective. Our approach applies to *the whole class* of coherent risk measures, thereby generalizing and unifying previous approaches that have focused on individual risk measures. We consider both *static* coherent

risk of the total discounted return from an MDP and time-consistent *dynamic* Markov coherent risk. Our main contribution is formulating the risk-sensitive policy-gradient under the coherent-risk framework. More specifically, we provide:

- A new formula for the gradient of static coherent risk that is convenient for approximation using sampling.
- An algorithm for the gradient of general static coherent risk that involves sampling with convex programming and a corresponding consistency result.
- A new policy gradient theorem for Markov coherent risk, relating the gradient to a suitable *value function* and a corresponding actor-critic algorithm.

Several previous results are special cases of the results presented here; our approach allows to re-derive them in greater generality and simplicity.

Related Work Risk-sensitive optimization in RL for specific risk functions has been studied recently by several authors. [8] studied exponential utility functions, [22], [35], [26] studied mean-variance models, [10], [36] studied CVaR in the static setting, and [25], [11] 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 not limited to a specific choice within that class, nor to particular system dynamics.

Reference [24] showed that an MDP with a dynamic coherent risk objective is essentially a robust MDP. The planning for large scale MDPs was considered in [37], using an approximation of the value function. For many problems, approximation in the policy space is more suitable (see, e.g., [18]). 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 of [37] in a part of our method.

Optimization of coherent risk measures was thoroughly investigated by Ruszczynski and Shapiro [31] (see also [32]) for the stochastic programming case in which the policy parameters do not affect the distribution of the stochastic system (i.e., the MDP trajectory), but only the reward function, and thus, this approach is not suitable for most RL problems. For the case of MDPs and dynamic risk, [30] proposed a dynamic programming approach. This approach does not scale-up to large MDPs, due to the "curse of dimensionality". For further motivation of risk-sensitive policy gradient methods, we refer the reader to [22, 35, 26, 10, 36].

2 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 tunable 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.

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, i.e., 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., $Z(\omega) \leq W(\omega)$ for all $\omega \in \Omega$. We denote 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 $\mu_\theta(\cdot|x)$. 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.

[1] Our results may easily be 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 [30], while this is not necessarily true for static risk. Our results can be extended to history-dependent policies or stationary Markov

2.1 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) $\min_{\nu \in \mathbb{R}} \left\{ \nu + \frac{1}{\alpha} \mathbb{E}\left[(Z-\nu)^+\right] \right\}$. A risk measure is called *coherent*, if it satisfies the following conditions for all $Z, W \in \mathcal{Z}$ [2]:

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 further refer the reader to [2] for a more detailed motivation of coherent risk.

The following representation theorem [32] shows an important property of coherent risk measures that is fundamental to our gradient-based approach.

Theorem 2.1. *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.

Assumption 2.2 (The General Form of 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*

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

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 $\bar{\xi}$. $\mathcal{E}$ and $\mathcal{I}$ here denote the sets of equality and inequality constraints, respectively. 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 2.2 implies that the risk envelope $\mathcal{U}(P_\theta)$ is known in an *explicit* form. From Theorem 6.6 of [32], in the case of a finite probability space, ρ is a coherent risk 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 the CVaR, the mean-semi-deviation, and spectral risk measures [1].

2.2 Dynamic Risk Measures

The risk measures defined above do not take into account any temporal structure that the random variable might have, 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*. *Dynamic* risk measures, on the other hand,

policies on a state space augmented with accumulated cost. The latter has shown to be sufficient for optimizing the CVaR risk [4].

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

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 [30]: 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 [16] 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 [30] and [16] for further insights.

Markov Coherent Risk Measures. Markov risk measures were introduced in [30] and constitute 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

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

where ρ is a static coherent risk measure that satisfies Assumption 2.2 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 ρ at state $x \in \mathcal{X}$ is induced by the transition probability $P_\theta(\cdot|x) = \sum_{a\in\mathcal{A}} P(x'|x,a)\mu_\theta(a|x)$. 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.

3 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 2, 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 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 in this paper 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 for dealing 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 estimators for the gradients.

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

4 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 [18].

Assumption 4.1. *The likelihood ratio $\nabla_\theta \log P(\omega)$ is well-defined and bounded for all $\omega \in \Omega$.*

Moreover, our approach implicitly assumes that given some $\omega \in \Omega$, $\nabla_\theta \log P(\omega)$ may be easily calculated. This is also a standard requirement for policy gradient algorithms [18] and is satisfied in various applications such as queueing systems, inventory management, and financial engineering (see, e.g., the survey by Fu [14]).

Using Theorem 2.1 and Assumption 2.2, 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

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

The convexity of (1) and its strict feasibility due to Assumption 2.2 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 4.2. *Let Assumptions 2.2 and 4.1 hold. For any saddle point* $(\xi_\theta^*, \lambda_\theta^{*,\mathcal{P}}, \lambda_\theta^{*,\mathcal{E}}, \lambda_\theta^{*,\mathcal{I}}) \in \mathcal{S}$ *of* (6)*, we have*

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

The proof of this theorem, given in the supplementary material, involves an application of the Envelope theorem [21] and a standard 'likelihood-ratio' trick. We now demonstrate the utility of Theorem 4.2 with several examples in which we show that it generalizes previously known results, and also enables deriving new useful gradient formulas.

4.1 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 [28], 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, $\mathbb{E}\left[Z \middle| Z > q_\alpha \right]$, 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 [32] $\mathcal{U} = \left\{ \xi P_\theta : \xi(\omega) \in [0, \alpha^{-1}], \quad \sum_{\omega\in\Omega} \xi(\omega) P_\theta(\omega) = 1 \right\}$. Furthermore, [32] show that the saddle points of (6) satisfy $\xi_\theta^*(\omega) = \alpha^{-1}$ when $Z(\omega) > \lambda_\theta^{*,\mathcal{P}}$, and $\xi_\theta^*(\omega) = 0$ when $Z(\omega) < \lambda_\theta^{*,\mathcal{P}}$, where $\lambda_\theta^{*,\mathcal{P}}$ is any $(1-\alpha)$-quantile of Z. Plugging this result into Theorem 4.2, we can easily show that

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

This formula was recently proved in [36] 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.

4.2 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 [32]. We have the following result:

Proposition 4.3. *Under Assumption 4.1, with* $\nabla_\theta \mathbb{E}[Z] = \mathbb{E}[\nabla_\theta \log P(\omega) Z]$*, we have*

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

This proposition 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 algorithm along with the proof of the proposition are in the supplementary material. In Section 6 we provide a numerical illustration of optimization with a mean-semideviation objective.

4.3 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 4.2. 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; [32]) of the formula in Theorem 4.2.

Assume that we are given N i.i.d. samples $\omega_i \sim P_\theta$, $i = 1, \ldots, N$, 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:

$$\rho_N(Z) = \max_{\xi : \xi P_{\theta;N} \in \mathcal{U}(P_{\theta;N})} \sum_{i \in 1,\ldots,N} P_{\theta;N}(\omega_i)\xi(\omega_i)Z(\omega_i). \tag{7}$$

Note that (7) defines a convex optimization problem with $\mathcal{O}(N)$ variables and constraints. In the following, we assume that a solution to (7) may be computed efficiently using standard convex programming tools such as interior point methods [9]. Let $\xi^*_{\theta;N}$ denote a solution to (7) 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 [9]. We propose the following estimator for the gradient-based on Theorem 4.2:

$$\begin{aligned}\nabla_{\theta;N}\rho(Z) = &\sum_{i=1}^{N} P_{\theta;N}(\omega_i)\xi^*_{\theta;N}(\omega_i)\nabla_\theta \log P(\omega_i)(Z(\omega_i) - \lambda^{*,\mathcal{P}}_{\theta;N}) \\ &- \sum_{e\in\mathcal{E}} \lambda^{*,\mathcal{E}}_{\theta;N}(e)\nabla_\theta g_e(\xi^*_{\theta;N}; P_{\theta;N}) - \sum_{i\in\mathcal{I}} \lambda^{*,\mathcal{I}}_{\theta;N}(i)\nabla_\theta f_i(\xi^*_{\theta;N}; P_{\theta;N}).\end{aligned} \tag{8}$$

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 has been reported in the supplementary material.

Proposition 4.4. *Let Assumptions 2.2 and 4.1 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 4.4 is large, but rather mild. Note that (I) is implied by the Slater condition of Assumption 2.2. 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 (8) with the analytical-solution based estimator, as in Sections 4.1 and 4.2. In the supplementary material, we report an empirical comparison between the two approaches for the case of CVaR risk, which showed that the two approaches performed very similarly. This is well-expected, since in general, both SAA and standard likelihood-ratio based estimators obey a law-of-large-numbers variance bound of order $1/\sqrt{N}$ [32].

To summarize this section, we have seen that by exploiting the special structure of coherent risk measures in Theorem 2.1 and by the envelope-theorem style result of Theorem 4.2, 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 sub-routine in our subsequent treatment of dynamic risk measures.

5 Gradient Formula for Dynamic Risk

In this section, we derive a new formula for the gradient of the Markov coherent dynamic risk measure, $\nabla_\theta\rho_\infty(\mathcal{M})$. Our approach is based on combining the static gradient formula of Theorem 4.2, with a dynamic-programming decomposition of $\rho_\infty(\mathcal{M})$.

The risk-sensitive *value-function* for an MDP $\mathcal{M}$ under the policy θ is defined as $V_\theta(x) = \rho_\infty(\mathcal{M}|x_0 = x)$, where with a slight abuse of notation, $\rho_\infty(\mathcal{M}|x_0 = x)$ denotes the Markov-coherent dynamic risk in (3) when the initial state x_0 is x. It is shown in [30] that due to the structure of the Markov dynamic risk $\rho_\infty(\mathcal{M})$, the value function is the unique solution to the *risk-sensitive Bellman equation*

$$V_\theta(x) = C(x) + \gamma \max_{\xi P_\theta(\cdot|x) \in \mathcal{U}(x, P_\theta(\cdot|x))} \mathbb{E}_\xi[V_\theta(x')], \tag{9}$$

where the expectation is taken over the next state transition. Note that by definition, we have $\rho_\infty(\mathcal{M}) = V_\theta(x_0)$, and thus, $\nabla_\theta \rho_\infty(\mathcal{M}) = \nabla_\theta V_\theta(x_0)$.

We now develop a formula for $\nabla_\theta V_\theta(x)$; this formula extends the well-known "policy gradient theorem" [34, 17], developed for the expected return, to Markov-coherent dynamic risk measures. We make a standard assumption, analogous to Assumption 4.1 of the static case.

Assumption 5.1. *The likelihood ratio $\nabla_\theta \log \mu_\theta(a|x)$ is well-defined and bounded for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$.*

For each state $x \in \mathcal{X}$, let $(\xi^*_{\theta,x}, \lambda^{*,\mathcal{P}}_{\theta,x}, \lambda^{*,\mathcal{E}}_{\theta,x}, \lambda^{*,\mathcal{I}}_{\theta,x})$ denote a saddle point of (6), corresponding to the state x, with $P_\theta(\cdot|x)$ replacing P_θ in (6) and V_θ replacing Z. The next theorem presents a formula for $\nabla_\theta V_\theta(x)$; the proof is in the supplementary material.

Theorem 5.2. *Under Assumptions 2.2 and 5.1, we have*

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

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

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

Theorem 5.2 may be used to develop an *actor-critic* style [34, 17] sampling-based algorithm for solving the DRP problem (5), composed of two interleaved procedures:

Critic: For a given policy θ, calculate the risk-sensitive value function V_θ, and
Actor: Using the critic's V_θ and Theorem 5.2, estimate $\nabla_\theta \rho_\infty(\mathcal{M})$ and update θ.

Space limitation restricts us from specifying the full details of our actor-critic algorithm and its analysis. In the following, we highlight only the key ideas and results. For the full details, we refer the reader to the full paper version, provided in the supplementary material.

For the critic, the main challenge is calculating the value function when the state space $\mathcal{X}$ is large and dynamic programming cannot be applied due to the 'curse of dimensionality'. To overcome this, we exploit the fact that V_θ is equivalent to the value function in a robust MDP [24] and modify a recent algorithm in [37] to estimate it using function approximation.

For the actor, the main challenge is that in order to estimate the gradient using Thm. 5.2, we need to sample from an MDP with ξ^*_θ-weighted transitions. Also, $h_\theta(x,a)$ involves an expectation for each s and a. Therefore, we propose a *two-phase sampling procedure* to estimate ∇V_θ in which we first use the critic's estimate of V_θ to derive ξ^*_θ, and sample a trajectory from an MDP with ξ^*_θ-weighted transitions. For each state in the trajectory, we then sample several next states to estimate $h_\theta(x,a)$.

The convergence analysis of the actor-critic algorithm and the gradient error incurred from function approximation of V_θ are reported in the supplementary material. We remark that our actor-critic algorithm requires a simulator for sampling multiple state-transitions from each state. Extending our approach to work with a single trajectory roll-out is an interesting direction for future research.

6 Numerical Illustration

In this section, we illustrate our approach with a numerical example. 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.

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 $A2 \sim$

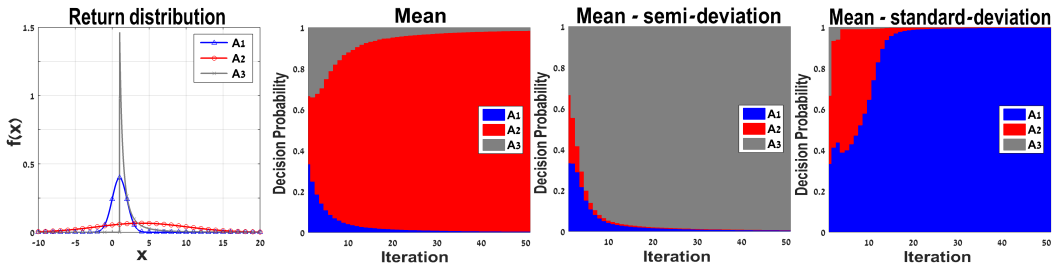

Figure 1: Numerical illustration - selection between 3 assets. A: Probability density of asset return. B,C,D: Bar plots of the probability of selecting each asset vs. training iterations, for policies π_1, π_2, and π_3, respectively. At each iteration, 10,000 samples were used for gradient estimation.

$\mathcal{N}(4, 6)$. 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 [27]. 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 [18]. 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 4. Policy π_3 is also risk-averse, with a mean-standard-deviation objective, as proposed in [35, 26], $\max_\theta \mathbb{E}[Z] - \sqrt{\text{Var}[Z]}$, and was trained using the algorithm of [35]. 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 $A1$ [15].

7 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 the CVaR [3].

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 example shows, 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 2.1), naturally relates the risk to model uncertainty. In [24], a similar connection was made between model-uncertainty in MDPs and dynamic Markov coherent risk. 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.

Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Unions Seventh Framework Program (FP7/2007-2013) / ERC Grant Agreement n. 306638. Yinlam Chow is partially supported by Croucher Foundation Doctoral Scholarship.

References

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

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

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

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

[5] D. Bertsekas. *Dynamic programming and optimal control*. Athena Scientific, 4th edition, 2012.

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

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

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

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

[10] Y. Chow and M. Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In *NIPS 27*, 2014.

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

[12] E. Delage and S. Mannor. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, 58(1):203213, 2010.

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

[14] M. Fu. Gradient estimation. In *Simulation*, volume 13 of *Handbooks in Operations Research and Management Science*, pages 575 – 616. Elsevier, 2006.

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

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

[17] V. Konda and J. Tsitsiklis. Actor-critic algorithms. In *NIPS*, 2000.

[18] P. Marbach and J. Tsitsiklis. Simulation-based optimization of Markov reward processes. *IEEE Transactions on Automatic Control*, 46(2):191–209, 1998.

[19] H. Markowitz. *Portfolio selection: Efficient diversification of investment*. John Wiley and Sons, 1959.

[20] F. Meng and H. Xu. A regularized sample average approximation method for stochastic mathematical programs with nonsmooth equality constraints. *SIAM Journal on Optimization*, 17(3):891–919, 2006.

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

[22] J. Moody and M. Saffell. Learning to trade via direct reinforcement. *Neural Networks, IEEE Transactions on*, 12(4):875–889, 2001.

[23] A. Nilim and L. El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[24] T. Osogami. Robustness and risk-sensitivity in Markov decision processes. In *NIPS*, 2012.

[25] M. Petrik and D. Subramanian. An approximate solution method for large risk-averse Markov decision processes. In *UAI*, 2012.

[26] L. Prashanth and M. Ghavamzadeh. Actor-critic algorithms for risk-sensitive MDPs. In *NIPS 26*, 2013.

[27] S. Rachev and S. Mittnik. *Stable Paretian models in finance*. John Willey & Sons, New York, 2000.

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

[29] R. Rockafellar, R. Wets, and M. Wets. *Variational analysis*, volume 317. Springer, 1998.

[30] A. Ruszczyński. Risk-averse dynamic programming for Markov decision processes. *Mathematical Programming*, 125(2):235–261, 2010.

[31] A. Ruszczyński and A. Shapiro. Optimization of convex risk functions. *Math. OR*, 31(3):433–452, 2006.

[32] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming*, chapter 6, pages 253–332. SIAM, 2009.

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

[34] R. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS 13*, 2000.

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

[36] A. Tamar, Y. Glassner, and S. Mannor. Optimizing the CVaR via sampling. In *AAAI*, 2015.

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

A Proof of Theorem 4.2

First note from Assumption 2.2 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 [9], 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 2.2, 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 Envelop theorem for saddle-point problems (see Theorem 4 of [21]), we have

$$\nabla_\theta \max_{\xi\geq 0} \min_{\lambda^{\mathcal{P}},\lambda^{\mathcal{I}}\geq 0,\lambda^{\mathcal{E}}} L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) = \nabla_\theta L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})|_{(\xi^*_\theta,\lambda^{*,\mathcal{P}}_\theta,\lambda^{*,\mathcal{E}}_\theta,\lambda^{*,\mathcal{I}}_\theta)} . \tag{10}$$

The result follows by writing the gradient in (10) explicitly, and using the likelihood-ratio trick:

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

where the last equality is justified by Assumption 4.1.

B Gradient Results for Static Mean-Semideviation

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

$$\rho_{\text{MSD}}(Z) = \mathbb{E}[Z] + c\left(\mathbb{E}\left[(Z - \mathbb{E}[Z])_+^2\right]\right)^{1/2}, \tag{11}$$

Following the derivation in [32], 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:

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

and hence

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

It follows that Eq. (1) holds with

$$\mathcal{U} = \{\xi' \in \mathcal{Z}^* : \quad \xi' = 1 + c\xi - c\mathbb{E}[\xi], \quad \|\xi\|_q \leq 1, \quad \xi \geq 0\}.$$

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

$$\rho_{\text{MSD}}(Z) = \sup_{\|\xi\|_q\leq 1,\xi\geq 0} \langle 1 + c\xi - c\mathbb{E}[\xi], Z\rangle. \tag{12}$$

Let $\bar{\xi}$ denote an optimal solution for (12). In [32] it is shown that $\bar{\xi}$ is a contact point of $(Z - \mathbb{E}[Z])_+$, that is

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

and we have that

$$\bar{\xi} = \frac{(Z - \mathbb{E}[Z])_+}{\|(Z - \mathbb{E}[Z])_+\|_2} = \frac{(Z - \mathbb{E}[Z])_+}{\mathbb{SD}(Z)}. \tag{13}$$

Note that $\bar{\xi}$ is not necessarily a probability distribution, but for $c \in [0, 1]$, it can be shown [32] that $1 + c\bar{\xi} - c\mathbb{E}\left[\bar{\xi}\right]$ 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 B.1. *Under Assumption 4.1, we have that*

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

and, according to the standard likelihood-ratio method,

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

Proof. Note that in Eq. (12) the constraints do not depend on θ. Therefore, using the envelope theorem we obtain that

$$\begin{aligned} \nabla_\theta \rho(Z) &= \nabla_\theta \langle 1 + c\bar{\xi} - c\mathbb{E}\left[\bar{\xi}\right], Z\rangle \\ &= \nabla_\theta \langle 1, Z\rangle + c\nabla_\theta \langle \bar{\xi}, Z\rangle - c\nabla_\theta \langle \mathbb{E}\left[\bar{\xi}\right], Z\rangle. \end{aligned} \tag{14}$$

We now write each of the terms in Eq. (14) as an expectation. We start with the following standard likelihood-ratio result:

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

Also, we have that

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

therefore, by the derivative of a product rule:

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

By the likelihood-ratio trick and Eq. (13) we have that

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

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. (14), we have that

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

□

Proposition 4.3 naturally leads to a sampling-based gradient estimation algortihm, which we term `GMSD` (Gradient of Mean Semi-Deviation). The algorithm is described in Algorithm 1.

C Consistency Proof

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

$$\begin{aligned} L_{\theta;N}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) = &\sum_{\omega \in \Omega} \xi(\omega) P_{\theta;N}(\omega) Z(\omega) - \lambda^{\mathcal{P}} \left(\sum_{\omega \in \Omega} \xi(\omega) P_{\theta;N}(\omega) - 1\right) \\ &- \sum_{e \in \mathcal{E}} \lambda^{\mathcal{E}}(e) f_e(\xi, P_{\theta;N}) - \sum_{i \in \mathcal{I}} \lambda^{\mathcal{I}}(i) f_i(\xi, P_{\theta;N}). \end{aligned} \tag{15}$$

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 (15) 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:

Algorithm 1 GMSD

1: **Given:**

- Risk level c
- An i.i.d. sequence $z_1, \ldots, z_N \sim P_\theta$.

2: Set

$$\widehat{\mathbb{E}[Z]} = \frac{1}{N}\sum_{i=1}^{N} z_i.$$

3: Set

$$\widehat{\mathbb{SD}(Z)} = \left(\frac{1}{N}\sum_{i=1}^{N}(z_i - \widehat{\mathbb{E}[Z]})_+^2\right)^{1/2}.$$

4: Set

$$\widehat{\nabla_\theta \mathbb{E}[Z]} = \frac{1}{N}\sum_{i=1}^{N}\nabla_\theta \log P(z_i) z_i.$$

5: **Return:**

$$\nabla_\theta \hat{\rho}(Z) = \widehat{\nabla_\theta \mathbb{E}[Z]} + \frac{c}{\widehat{\mathbb{SD}(Z)}}\frac{1}{N}\sum_{i=1}^{N}(z_i - \widehat{\mathbb{E}[Z]})_+ \left(\nabla_\theta \log P(z_i)(z_i - \widehat{\mathbb{E}[Z]}) - \widehat{\nabla_\theta \mathbb{E}[Z]}\right).$$

(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 2.2 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 S.L.L.N. of 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 $\mathbb{D}\{A, B\}$ the deviation of set A from set B, i.e., $\mathbb{D}\{A, B\} = \sup_{x\in A}\inf_{y\in B}\|x - y\|$. 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 [32], the Slater condition of Assumption 2.2 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 [32], 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 [32], we have that

$$\rho_N(Z) \to \rho(Z) \text{ w.p. 1 as } N \to \infty, \tag{16}$$

$$\mathbb{D}\{\mathcal{P}_N, \mathcal{P}\} \to 0 \text{ w.p. 1 as } N \to \infty, \tag{17}$$

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 [32], 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 [32] implies that $L_{\theta;N}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) - L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \overset{e}{\to} 0$, where $\overset{e}{\to}$ 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 [32] that epi-convergence of $L_{\theta,N}$ to L_θ implies uniform convergence on C_ξ, i.e., $\sup_{\xi \in C_\xi} \left| L_{\theta;N}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) - L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \right| \leq \epsilon$. 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 $\sup_{(\lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \in C_\lambda} \left| L_{\theta;N}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) - L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \right| \leq \epsilon$. 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

$$\sup_{(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \in C} \left| L_{\theta;N}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) - L_\theta(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) \right| \leq \epsilon. \tag{18}$$

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

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

contradicting (18).

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

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

also contradicting (18).

It follows that $\left| L_{\theta;N}(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) - L_\theta(\xi^*_\theta, \lambda^{*,\mathcal{P}}_\theta, \lambda^{*,\mathcal{E}}_\theta, \lambda^{*,\mathcal{I}}_\theta) \right| \leq \epsilon$ for all $N > N^*(\epsilon, \omega_{SAA})$, and therefore

$$\lim_{N \to \infty} L_{\theta;N}(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) = L_\theta(\xi^*_\theta, \lambda^{*,\mathcal{P}}_\theta, \lambda^{*,\mathcal{E}}_\theta, \lambda^{*,\mathcal{I}}_\theta), \tag{19}$$

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 $\mathbb{D}\{\mathcal{S}_N, \mathcal{S}\} \not\to 0$. 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 (17) 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,

$$\begin{aligned} &L_{\theta;N}(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) - L_\theta(\bar{\xi}^*, \bar{\lambda}^{*,\mathcal{P}}, \bar{\lambda}^{*,\mathcal{E}}, \bar{\lambda}^{*,\mathcal{I}}) \\ &= \left[L_{\theta;N}(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) - L_\theta(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) \right] + \\ &\quad + \left[L_\theta(\xi^*_{\theta;N}, \lambda^{*,\mathcal{P}}_{\theta;N}, \lambda^{*,\mathcal{E}}_{\theta;N}, \lambda^{*,\mathcal{I}}_{\theta;N}) - L_\theta(\bar{\xi}^*, \bar{\lambda}^{*,\mathcal{P}}, \bar{\lambda}^{*,\mathcal{E}}, \bar{\lambda}^{*,\mathcal{I}}) \right]. \end{aligned} \tag{20}$$

The first term in the r.h.s. of (20) tends to zero, using the argument from (18), 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 (19).

Part 3 We now show the consistency of $\nabla_{\theta;N}\rho(Z)$.

Consider Eq. (8). Since $\nabla_\theta \log P(\cdot)$ is bounded by Assumption 4.1, and $\nabla_\theta f_i(\cdot; P_\theta)$ and $\nabla_\theta g_e(\cdot; P_\theta)$ are bounded by Assumption 2.2, and using our previous result $\mathbb{D}\{\mathcal{S}_N, \mathcal{S}\} \to 0$, we have that for a.e. $\omega_{SAA} \in \Omega_{SAA}$

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

where the first equality is obtained from the Envelop theorem (see Theorem 4.2) with $(\xi^*_\theta, \lambda^{*,\mathcal{P}}_\theta, \lambda^{*,\mathcal{E}}_\theta, \lambda^{*,\mathcal{I}}_\theta) \in \mathcal{S}_N \cap \mathcal{S}$ is 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 Empirical Comparison of Analytic-Solution Based and SAA Based Policy Gradient

We compare the CVaR policy gradient as obtained by the analytical result in Section 4.1:

$$\nabla_\theta \rho_{\text{CVaR}}(Z;\alpha) = \mathbb{E}\left[\nabla_\theta \log P(\omega)(Z - q_\alpha) \middle| Z(\omega) > q_\alpha\right], \tag{21}$$

with the general sampling based algorithm of Eq. (8) in Section 4.3.

For the analytical-solution based policy gradient, we use the GCVaR algorithm of [30], which is the sampling-based version of Eq. (21). For the general sampling based algorithm, we used Matlab's 'linprog' to solve the linear program in Eq. (7), using the risk envelope for CVaR, as defined in Section 4.1. The resulting numerical values for ${\xi^*}_{\theta;N}$ and $\lambda^{*,\mathcal{P}}_{\theta;N}$ were plugged into Eq. (8) for the gradient estimate (the other terms in Eq. (8) cancel out by definition of the CVaR risk envelope).

We present empirical results for the asset selection domain of Section 6. 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 N - the sampling budget.

As may be observed, both policies exhibit similar learning performance, and the differences diminish as N grows large.

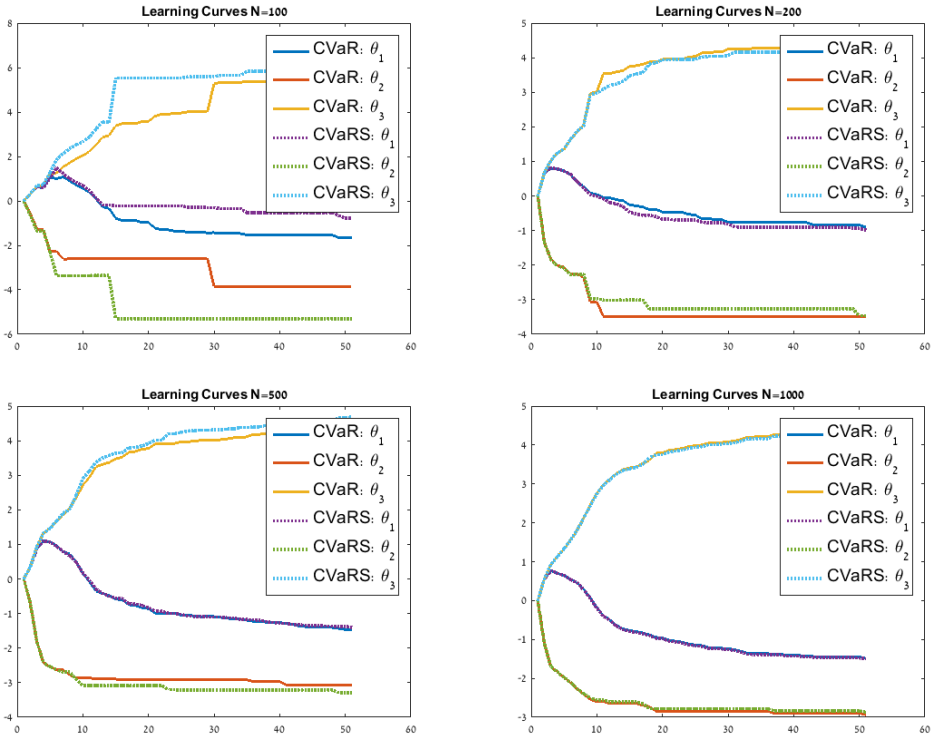

Figure 2: Learning curves (θ values vs. training episodes) of the analytical-solution based policy gradient (labeled CVaR), and the general sampling based algorithm (labeled CVaRS), for different values of N - the sampling budget.

E Proof of Theorem 5.2

Similar to the proof of Theorem 4.2, 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.,

$$\begin{aligned}\max_{\xi\,:\,\xi P_\theta(\cdot|x)\in\mathcal{U}(x,P_\theta(\cdot|x))} \sum_{x'\in\mathcal{X}} \xi(x')P_\theta(x'|x)V_\theta(x') &= \max_{\xi\geq 0} \min_{\lambda^{\mathcal{P}},\lambda^{\mathcal{I}}\geq 0,\lambda^{\mathcal{E}}} L_{\theta,x}(\xi,\lambda^{\mathcal{P}},\lambda^{\mathcal{E}},\lambda^{\mathcal{I}}) \\ &= \min_{\lambda^{\mathcal{P}},\lambda^{\mathcal{I}}\geq 0,\lambda^{\mathcal{E}}} \max_{\xi\geq 0} L_{\theta,x}(\xi,\lambda^{\mathcal{P}},\lambda^{\mathcal{E}},\lambda^{\mathcal{I}}).\end{aligned}$$

the gradient formula in (10) can be written as

$$\begin{aligned}\nabla_\theta V_\theta(x) &= \nabla_\theta \left[C_\theta(x) + \gamma \max_{\xi\,:\,\xi P_\theta(\cdot|x)\in\mathcal{U}(x,P_\theta(\cdot|x))} \mathbb{E}_\xi[V_\theta]\right] \\ &= \gamma \sum_{x'\in\mathcal{X}} \xi^*_{\theta,x}(x')P_\theta(x'|x)\nabla_\theta V_\theta(x') + \sum_{a\in\mathcal{A}} \mu_\theta(a|x)\nabla_\theta \log \mu_\theta(a|x) h_\theta(x,a),\end{aligned}$$

where the stage-wise cost function $h_\theta(x,a)$ is defined in (27). By defining $\widehat{h}_\theta(x) = \sum_{a\in\mathcal{A}} \mu_\theta(a|x)\nabla_\theta \log \mu_\theta(a|x) h_\theta(x,a)$ and unfolding the recursion, the above expression implies

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

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 Gradient Formula for Dynamic Risk - Full Results

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 [30]). This formula extends the well-known "policy gradient theorem" [34, 17] 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 F.3), and
Actor: Using the critic's value function, estimate $\nabla_\theta \rho_\infty(\mathcal{M})$ by sampling (see Section F.4).

The value function proposed by [30] 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 [30]) 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 [6]. In particular, *temporal difference* (TD) learning methods [33] are popular for this purpose, which have been recently extended to robust MDPs by [37]. 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.

F.1 Dynamic Risk

We provide a multi-period generalization of the concepts presented in Section 2.1. Here we closely follow the discussion in [30].

Consider a probability space $(\Omega, \mathcal{F}, P_\theta)$, a filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \cdots \subset \mathcal{F}_T \subset \mathcal{F}$, and an adapted sequence of real-valued random variables Z_t, $t \in \{0, \ldots, T\}$. We assume that $\mathcal{F}_0 = \{\Omega, \emptyset\}$, i.e., Z_0 is deterministic. For each $t \in \{0, \ldots, T\}$, we denote by $\mathcal{Z}_t$ the space of random variables defined over the probability space $(\Omega, \mathcal{F}_t, P_\theta)$, and also let $\mathcal{Z}_{t,T} := \mathcal{Z}_t \times \cdots \times \mathcal{Z}_T$ be a sequence of these spaces. The sequence of random variables Z_t can be interpreted as the stage-wise costs observed along a trajectory generated by an MDP parameterized by a parameter θ, i.e., $Z_{0,T} \doteq \big(Z_0 = \gamma^0 C(x_0, a_0), \ldots, Z_T = \gamma^T C(x_T, a_T)\big) \in \mathcal{Z}_{0,T}$.

In particular, we are interested in the sequence of random variables induced by the trajectories from a Markov decision process (MDP) parameterized by parameter θ.

Explicitly, for any $t \geq 0$ and state dependent random variable $Z(x_{t+1}) \in \mathcal{Z}_{t+1}$, the risk evaluation is given by

$$\rho\big(Z(x_{t+1})\big) = \max_{\xi \,:\, \xi P_\theta(\cdot|x_t) \in \mathcal{U}(x_t, P_\theta(\cdot|x_t))} \mathbb{E}_\xi\big[Z(x_{t+1})\big], \tag{22}$$

where we let $\mathcal{U}(x_t, P_\theta(\cdot|x_t))$ denote the risk-envelope (2) with P_θ replaced with $P_\theta(\cdot|x_t)$. The Markovian assumption on the risk measure $\rho_T(\mathcal{M})$ allows us to optimize it using dynamic programming techniques.

F.2 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 $\mathcal{X}$ and $C_\theta(x) = \sum_{a \in \mathcal{A}} C(x, a)\mu_\theta(a|x)$ be the stage-wise cost function induced by policy μ_θ. We now define the risk sensitive Bellman operator $T_\theta[V] : B(\mathcal{X}) \mapsto B(\mathcal{X})$ as

$$T_\theta[V](x) := C_\theta(x) + \gamma \max_{\xi P_\theta(\cdot|x) \in \mathcal{U}(x, P_\theta(\cdot|x))} \mathbb{E}_\xi[V]. \tag{23}$$

According to Theorem 1 in [30], 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.

F.3 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 sub-space $\mathcal{V} = \{\Phi v | v \in \mathbb{R}^{\kappa_2}\}$, 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 [6].

Assumption F.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 joint operator ΠT_θ, i.e., $\tilde{V}_\theta(x) = v_\theta^{*\top} \phi(x)$, such that

$$\forall x \in \mathcal{X}, \qquad \tilde{V}_\theta(x) = \Pi T_\theta[\tilde{V}_\theta](x). \tag{24}$$

From Eq. 22 that has been derived from Theorem 2.1 for dynamic risks, it is easy to see that the risk-sensitive Bellman equation (23) is a robust Bellman equation [23] with uncertainty set $\mathcal{U}(x, P_\theta(\cdot|x))$. Thus, we may use the TD approximation of the robust Bellman equation proposed by [37] to find an approximation of V_θ. We will need the following assumption analogous to Assumption 2 in [37].

Assumption F.2. *There exists $\kappa \in (0, 1)$ such that $\xi(x') \leq \kappa/\gamma$, for all $\xi(\cdot) P_\theta(\cdot|x) \in \mathcal{U}(x, P_\theta(\cdot|x))$ and all $x, x' \in \mathcal{X}$.*

Given Assumption F.2, Proposition 3 in [37] guarantees that the projected risk-sensitive Bellman operator ΠT_θ is a contraction w.r.t. d_θ-norm. Therefore, Eq. 24 has a unique fixed-point solution $\tilde{V}_\theta(x) = v_\theta^{*\top} \phi(x)$. 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

$$\begin{aligned}\sum_{x \in \mathcal{X}} d_\theta(x|x_0) \phi(x) \phi(x)^\top v_\theta^* = \sum_{x \in \mathcal{X}} d_\theta(x|x_0) \phi(x) C_\theta(x) \\ + \gamma \sum_{x \in \mathcal{X}} d_\theta(x|x_0) \phi(x) \max_{\xi \,:\, \xi P_\theta(\cdot|x) \in \mathcal{U}(x, P_\theta(\cdot|x))} \mathbb{E}_\xi[\Phi v_\theta^*].\end{aligned}$$

As a result, given a long enough trajectory $x_0, a_0, x_1, a_1, \ldots, x_{N-1}, a_{N-1}$ 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

$$\begin{aligned}v_{k+1} = \left(\frac{1}{N} \sum_{t=0}^{N-1} \phi(x_t) \phi(x_t)^\top \right)^{-1} \Bigg[\frac{1}{N} \sum_{t=0}^{N-1} \phi(x_t) C_\theta(x_t) \\ + \gamma \frac{1}{N} \sum_{t=0}^{N-1} \phi(x_t) \max_{\xi P_\theta(\cdot|x_t) \in \mathcal{U}(x_t, P_\theta(\cdot|x_t))} \mathbb{E}_\xi[\Phi v_k] \Bigg].\end{aligned} \tag{25}$$

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 (25), one must repeatedly solve the inner optimization problem $\max_{\xi P_\theta(\cdot|x) \in \mathcal{U}(x, P_\theta(\cdot|x))} \mathbb{E}_\xi[\Phi v]$. 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 [37], we propose the following SAA approach to solve this problem. For the trajectory, $x_0, a_0, x_1, a_1, \ldots, x_{N-1}, a_{N-1}$, we define the empirical transition probability $P_N(x'|x, a) \doteq$

$\frac{\sum_{t=0}^{N-1} \mathbf{1}\{x_t=x,\, a_t=a,\, x_{t+1}=x'\}}{\sum_{t=0}^{N-1} \mathbf{1}\{x_t=x,\, a_t=a\}}$[4] and $P_{\theta;N}(x'|x) = \sum_{a\in\mathcal{A}} P_N(x'|x,a)\mu_\theta(a|x)$. Consider the following ℓ_2-regularized empirical robust optimization problem[5]

$$\rho_N(\Phi v) = \max_{\xi:\xi P_{\theta;N}\in\mathcal{U}(x,P_{\theta;N})} \sum_{x'\in\mathcal{X}} P_{\theta;N}(x'|x)\xi(x')\phi^\top(x')v + \frac{1}{2N}\big[P_{\theta;N}(x'|x)\xi(x')\big]^2. \tag{26}$$

As in [20], the ℓ_2-regularization term in this optimization problem guarantees convergence of 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 $\lambda^{*,\mathcal{P}}_{\theta,x;N}$, $\lambda^{*,\mathcal{E}}_{\theta,x;N}$, $\lambda^{*,\mathcal{I}}_{\theta,x;N}$, the corresponding KKT multipliers.

We obtain the empirical PRSVI algorithm by replacing the inner optimization $\max_{\xi P_\theta(\cdot|x_t)\in\mathcal{U}(x_t,P_\theta(\cdot|x_t))} \mathbb{E}_\xi[\Phi v^*_\theta]$ in Eq. 25 with $\rho_N(\Phi v)$ from Eq. 26. Similarly, as both N and k tend to infinity, v_k converges w.p. 1 to v^*_θ. More details can be found in the supplementary material.

F.4 Gradient Estimation

In Section F.3, we showed that we may effectively approximate the value function of a fixed policy θ using the (empirical) PRSVI algorithm in Eq. 25. 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 F.3. As described in Section F.2, $\rho_\infty(\mathcal{M}) = V_\theta(x_0)$, 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 4.1 of the static case.

Assumption F.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 4.2 for the static case, we may use the envelope theorem and the risk-sensitive Bellman equation, $V_\theta(x) = C_\theta(x) + \gamma \max_{\xi P_\theta(\cdot|x)\in\mathcal{U}(x,P_\theta(\cdot|x))} \mathbb{E}_\xi[V_\theta]$, to derive a formula for $\nabla_\theta V_\theta(x)$. We report this result in Theorem F.4, which is analogous to the risk-neutral policy gradient theorem [34, 17, 7]. The proof is in the supplementary material.

Theorem F.4. *Under Assumptions 2.2, we have*

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

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

$$h_\theta(x,a) = C(x,a) + \sum_{x'\in\mathcal{X}} P(x'|x,a)\xi^*_{\theta,x}(x')\Big[\gamma V_\theta(x') - \lambda^{*,\mathcal{P}}_{\theta,x} - \sum_{i\in\mathcal{I}} \lambda^{*,\mathcal{I}}_{\theta,x}(i)\frac{df_i(\xi^*_{\theta,x},p)}{dp(x')} - \sum_{e\in\mathcal{E}} \lambda^{*,\mathcal{E}}_{\theta,x}(e)\frac{dg_e(\xi^*_{\theta,x},p)}{dp(x')}\Big]. \tag{27}$$

[4] 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.

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

Theorem F.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), and transition probability $P_\theta(\cdot|x)\xi^*_{\theta,x}(\cdot)$. 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 (23) is nonlinear in the stationary Markov policy μ_θ. Therefore 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 many applications the value function V_θ is not known in advance, which further complicates gradient estimation. 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 $\{x_0^{(j)}, a_0^{(j)}, x_1^{(j)}, a_1^{(j)}, \ldots\}_{j=1}^N$ from the Markov chain induced by policy θ and transition probabilities $P^\xi_\theta(\cdot|x) := \xi^*_{\theta,x}(\cdot) P_\theta(\cdot|x)$.

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

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

(3) Calculate an estimate of ∇V_θ using the following average over all the samples: $\frac{1}{N}\sum_{j=1}^N \sum_{t=0}^\infty \gamma^t \nabla_\theta \log \mu_\theta(a_t^{(j)}|x_t^{(j)}) h_{\theta,N}(x_t^{(j)}, a_t^{(j)})$.

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 $P_N(x'|x,a)$.

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 F.3 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.

At the end, we show the convergence of the above two-phase sampling procedure. Let $d_{P^\xi_\theta}(x|x_0)$ and $\pi_{P^\xi_\theta}(x,a|x_0)$ be the state and state-action occupancy measure induced by the transition probability function $P^\xi_\theta(\cdot|x)$, respectively. Similarly, let $d_{P^\xi_{\theta;N}}(x|x_0)$ and $\pi_{P^\xi_{\theta;N}}(x,a|x_0)$ be the state and state-action occupancy measure induced by the estimated transition probability function $P^\xi_{\theta;N}(\cdot|x) := \xi^*_{\theta,x;N}(\cdot) P_{\theta;N}(\cdot|x)$. 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 $\frac{1}{N}\sum_{j=1}^N \sum_{t=0}^\infty \gamma^t \mathbf{1}\{x_t^{(j)} = x, a_t^{(j)} = a\} = \pi_{P^\xi_{\theta;N}}(x,a|x_0)$. 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 $\pi_{P^\xi_{\theta;N}}(x,a|x_0)$ 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 the supplementary materials for these technical results. These results together with Theorem F.4 imply the consistency of the policy gradient estimation.

Theorem F.5. *For any $x_0 \in \mathcal{X}$, the following expression holds with probability 1:*

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

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

G Convergence Analysis of Empirical PRSVI

Lemma G.1 (Technical Lemma)**.** *Let $P(\cdot|\cdot)$ and $\widetilde{P}(\cdot|\cdot)$ be two arbitrary transition probability matrices. At state $x \in \mathcal{X}$, for any $\xi : \xi P(\cdot|x) \in \mathcal{U}(x, P(\cdot|x))$, there exists a $M_\xi > 0$ such that for some $\tilde{\xi} : \tilde{\xi}\widetilde{P}(\cdot|x) \in \mathcal{U}(x, \widetilde{P}(\cdot|x))$,*

$$\sum_{x'\in\mathcal{X}} |\xi(x') - \tilde{\xi}(x')| \le M_\xi \sum_{x'\in\mathcal{X}} \left| P(x'|x) - \widetilde{P}(x'|x) \right|.$$

Proof. From Theorem 2.1, we know that $\mathcal{U}(x, P(\cdot|x))$ 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 $\mathcal{U}(x, P(\cdot|x))$ is closed, by Theorem 2.7 in [29], $\mathcal{U}(x, P(\cdot|x))$ is a Lipschitz set-valued mapping with respect to the Hausdorff distance. Thus, for any $\xi : \xi P(\cdot|x) \in \mathcal{U}(x, P(\cdot|x))$, the following expression holds for some $M_\xi > 0$:

$$\inf_{\hat{\xi}\in\mathcal{U}(x,\widetilde{P}(\cdot|x))} \sum_{x'\in\mathcal{X}} |\xi(x') - \hat{\xi}(x')| \le M_\xi \sum_{x'\in\mathcal{X}} \left| P(x'|x) - \widetilde{P}(x'|x) \right|.$$

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

Lemma G.2 (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.,

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

Therefore with probability 1,

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

Now we show that following expression holds with probability 1:

$$\begin{aligned} &\max_{\xi\,:\,\xi P_{\theta;N}(\cdot|x_t)\in\mathcal{U}(x_t, P_{\theta;N}(\cdot|x_t))} \sum_{x'\in\mathcal{X}} \xi(x') P_{\theta;N}(x'|x_t) v^\top \phi\,(x') + \frac{1}{2N} (\xi(x') P_{\theta;N}(x'|x_t))^2 \\ \to &\max_{\xi\,:\,\xi P_\theta(\cdot|x_t)\in\mathcal{U}(x_t, P_\theta(\cdot|x_t))} \sum_{x'\in\mathcal{X}} \xi(x') P_\theta(x'|x_t) v^\top \phi\,(x'). \end{aligned} \tag{28}$$

Notice that for $\{\xi^*_{\theta,x_t;N}(x')\}_{x'\in\mathcal{X}} \in \arg\max_{\xi\,:\,\xi P_{\theta;N}(\cdot|x_t)\in\mathcal{U}(x_t,P_{\theta;N}(\cdot|x_t))} \sum_{x'\in\mathcal{X}} \xi(x')P_{\theta;N}(x'|x_t)v^\top\phi\,(x')$, Lemma G.1 implies

$$\begin{aligned}
&\max_{\xi\,:\,\xi P_{\theta;N}(\cdot|x_t)\in\mathcal{U}(x_t,P_{\theta;N}(\cdot|x_t))} \sum_{x'\in\mathcal{X}} \xi(x')P_{\theta;N}(x'|x_t)v^\top\phi\,(x') + \frac{1}{2N}(\xi(x')P_{\theta;N}(x'|x_t))^2 \\
&- \max_{\xi\,:\,\xi P_{\theta}(\cdot|x_t)\in\mathcal{U}(x_t,P_{\theta}(\cdot|x_t))} \sum_{x'\in\mathcal{X}} \xi(x')P_{\theta}(x'|x_t)v^\top\phi\,(x') \\
\leq &\|\Phi v\|_\infty \left(M_{\xi^*_{\theta,x_t;N}} + \max_{x\in\mathcal{X}} |\xi^*_{\theta,x_t;N}(x)|\right) \sum_{x'\in\mathcal{X}} |P_\theta(x'|x_t) - P_{\theta;N}(x'|x_t)| + \frac{1}{2N}.
\end{aligned}$$

The quantity $\max_{x\in\mathcal{X}} |\xi^*_{\theta,x_t;N}(x)|$ is bounded because $\mathcal{U}(x_t, P_{\theta;N}(\cdot|x_t))$ 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

$$\begin{aligned}
&\left|\max_{\xi\,:\,\xi P_{\theta;N}(\cdot|x_t)\in\mathcal{U}(x_t,P_{\theta;N}(\cdot|x_t))} \sum_{x'\in\mathcal{X}} \xi(x')P_{\theta;N}(x'|x_t)v^\top\phi\,(x') + \frac{1}{2N}(\xi(x')P_{\theta;N}(x'|x_t))^2\right. \\
&\left.- \max_{\xi\,:\,\xi P_{\theta}(\cdot|x_t)\in\mathcal{U}(x_t,P_{\theta}(\cdot|x_t))} \sum_{x'\in\mathcal{X}} \xi(x')P_{\theta}(x'|x_t)v^\top\phi\,(x')\right| \\
\leq &\|\Phi v\|_\infty \max\left\{\left(M_{\xi^*} + \max_{x\in\mathcal{X}} |\xi^*(x)|\right), \left(M_{\xi^*_{\theta,x_t;N}} + \max_{x\in\mathcal{X}} |\xi^*_{\theta,x_t;N}(x)|\right)\right\} \sum_{x'\in\mathcal{X}} |P_\theta(x'|x_t) - P_{\theta;N}(x'|x_t)| + \frac{1}{2N}.
\end{aligned}$$

Therefore, the claim in expression (28) holds when $N \to \infty$ and $\sum_{x'\in\mathcal{X}} |P_\theta(x'|x_t) - P_{\theta;N}(x'|x_t)| \to 0$. On the other hand, the strong law of large numbers also implies that with probability 1,

$$\frac{1}{N}\sum_{t=0}^{N-1} \phi(x_t)\rho(\Phi v_t) \to d_\theta(x|x_0)\phi(x) \max_{\xi\,:\,\xi P_\theta(\cdot|x)\in\mathcal{U}(x,P_\theta(\cdot|x))} \sum_{x'\in\mathcal{X}} \xi(x')P_\theta(x'|x){v^*_\theta}^\top\phi\,(x')\,.$$

Combining the above arguments implies

$$\frac{1}{N}\sum_{t=0}^{N-1} \phi(x_t)\rho_N(\Phi v_t) \to d_\theta(x|x_0)\phi(x) \max_{\xi\,:\,\xi P_\theta(\cdot|x)\in\mathcal{U}(x,P_\theta(\cdot|x))} \sum_{x'\in\mathcal{X}} \xi(x')P_\theta(x'|x){v^*_\theta}^\top\phi\,(x')\,.$$

As $N\to\infty$, the above arguments imply that $v_k - \widehat{v}_k \to 0$. On the other hand, Proposition 1 in [37] 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 [5] 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. □

H Technical Results

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

$$\max_{\xi\,:\,\xi P_{\theta;N}(\cdot|x)\in\mathcal{U}(x,P_{\theta;N}(\cdot|x))} \sum_{x'\in\mathcal{X}} P_{\theta;N}(x'|x)\xi(x')V_\theta(x'), \tag{29}$$

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 $\rho_N(\Phi v)$, i.e.,

$$\rho_N(\Phi v) = \max_{\xi\,:\,\xi P_{\theta;N}(\cdot|x)\in\mathcal{U}(x,P_{\theta;N}(\cdot|x))} \sum_{x'\in\mathcal{X}} P_{\theta;N}(x'|x)\xi(x')\phi^\top(x')v + \frac{1}{2N}(\xi(x')P_{\theta;N}(x'|x))^2, \tag{30}$$

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 (29) can be viewed as having a skewed objective function of the problem in (30), 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[6].

(ii) Recall from Assumption 2.2 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 2.2 implies the linear independence constraint qualification (LICQ).

(iv) Since optimization problem (30) has a convex objective function and convex/affine constraints in $\xi \in \mathbb{R}^{|\mathcal{X}|}$, 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

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

One can easily conclude that $\nabla^2 \widehat{L}_{\theta;N}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) = -P_{\theta;N}(\cdot|x)^\top P_{\theta;N}(\cdot|x)/N - \sum_{i \in \mathcal{I}} \lambda^{\mathcal{I}}(i) \nabla^2_\xi f_i(\xi, P_{\theta;N}(\cdot|x))$ 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 [13], derived based on Implicit Function Theorem.

Proposition H.1 (Basic Sensitivity Theorem)**.** *Under the Assumption 2.2, 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* (29) *and* (30) *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 4.4 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 H.2. *There exists a constant $M_h > 0$ such that* $\max_{x \in \mathcal{X}, a \in \mathcal{A}} |h_\theta(x,a) - \lim_{N \to \infty} \widehat{h}_{\theta;N}(x,a)| \leq M_h \Delta$.

[6]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 4.4 and H.1, the proof of this Lemma is completed by letting $N \to \infty$ and defining

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

□

Lemma H.3. *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:

$$\gamma \sum_{x'\in\mathcal{X}} d_{P^\xi_\theta}(x'|x) P^\xi_\theta(x|x') = d_{P^\xi_\theta}(x) - (1-\gamma)\mathbf{1}\{x_0 = x\}, \tag{31}$$

From here one easily notice this expression can be rewritten as follows:

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

On the other hand, by repeating the analysis with $P_{\theta;N}(\cdot|x)$, we can also write

$$\left(I - \gamma P^\xi_{\theta;N}\right)^\top d_{P^\xi_{\theta;N}} = \{\mathbf{1}\{x_0 = z\}\}_{z\in\mathcal{X}}.$$

Combining the above expressions implies for any $x \in \mathcal{X}$,

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

which further implies

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

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 $x_0 \in \mathcal{X}$,

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

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

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

The last equality is due to the fact that every element in vector $\mathcal{Q}$ is non-negative. Combining the above results with Proposition 4.4 and H.1, and noting that

$$(I-\gamma P_\theta^\xi)^{-1}e = \sum_{t=0}^{\infty}\left(\gamma P_\theta^\xi\right)^k e = \frac{1}{1-\gamma}e,$$

we further have that

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

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 $\|\xi_{\theta,x}^*(\cdot) - \xi_{\theta,x;N}^*(\cdot)\|_1 \to 0$. We thus set the constant M_π as $\gamma\|\Phi_{\theta,x}^{-1}\Psi_{\theta,x}\|_1/(1-\gamma)$. □