Policy Gradient for Coherent Risk Measures

Aviv Tamar
Electrical Engineering Department, The Technion - Israel Institute of Technology, Haifa, Israel 32000

Yinlam Chow
Institute for Computational & Mathematical Engineering (ICME), Stanford University, Stanford CA, USA 94305

Mohammad Ghavamzadeh
Adobe Research, 345 Park Avenue, San Jose CA, USA 95110

Shie Mannor
Electrical Engineering Department, The Technion - Israel Institute of Technology, Haifa, Israel 32000

Abstract

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

1. Introduction

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

There are various approaches to quantifying the risk of a random cost, such as the celebrated Markowitz mean-variance model [Markowitz 1959], or the more recent Value at Risk (VaR) and Conditional Value at Risk (CVaR) [Rockafellar & Uryasev 2000]. The preference of one risk measure over another depends on factors such as sensitivity to rare events, ease of estimation from data, and computational tractability of the optimization problem, and in general, there is no single choice that dominates over the rest. However, the highly influential paper of [Artzner et al. 1999] 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.

When the optimization problem is sequential, such as when solving a Markov decision problem (MDP), another desirable property of a risk measure is time consistency. A time-consistent risk measure satisfies a “dynamic programming” style property: if a strategy is risk-optimal for an n-stage problem, then the component of the policy from the t-th time until the end (where t < n) is also risk-optimal (see principle of optimality in [Bertsekas 2005]). The recently proposed class of dynamic Markov coherent risk measures [Ruszczynski 2010] satisfies both the coherence and time consistency properties.

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

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

where $C$ is a random cost, controlled by tunable parameter vector $\theta$, and $\rho$ is a coherent risk measure. We consider both the time-consistent dynamic Markov coherent risk measures, and also the standard static coherent risk-measures without explicit temporal dependence.

For the static case and when the cost is of the form $C =$
Policy Gradient for Coherent Risk Measures

our contributions are:

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

Related Work

For the static-risk case, our approach is similar in spirit to policy gradient methods (Baxter & Bartlett 2001) a.k.a. the likelihood-ratio method in the simulation literature, (Glynn 1990), and may be seen as an extension of these methods to coherent risk objectives. Optimization of coherent risk measures was thoroughly investigated by Ruszczyński & Shapiro (2000) see also Shapiro et al. (2009) for the case discussed above, in which does not control the distribution of . For the case of MDPs and dynamic risk, Ruszczyński (2010) proposed a dynamic programming approach. This approach does not scale-up to large MDPs, due to the “curse of dimensionality”. The work of Tamar et al. (2014) on robust MDPs is relevant since an MDP with a dynamic coherent risk objective is essentially a robust MDP. Tamar et al. (2014) considered approximation only in the value function. For many problems, approximation in the policy space is more suitable (see, e.g., Marbach & Tsitsiklis 1998). 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 Tamar et al. (2014) in a part of our approach.

Risk-sensitive optimization in RL for specific risk functions has been studied recently by several authors. Borkar (2001) studied exponential utility functions, Tamar et al. (2012) and Prashanth & Ghaemmaghami (2013) studied mean-variance models, Chow & Ghaemmaghami (2014) and Tamar et al. (2015) studied CVaR in the static setting, and Petrik & Subramanian (2012) and Chow & Pavone (2014) 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. In particular, for the special case of CVaR, we obtain similar results to Tamar et al. (2015), but under weaker assumptions and simpler derivations.

2. Preliminaries

Consider a probability space (, , ) where is the set of outcomes (sample space), is a -algebra over representing the set of events we are interested in, and is a probability measure over the space of probability distributions, is a probability measure over parameterized by some tunable parameter . In the following, we suppress the notation of in -dependent quantities. To ease the technical exposition, in this paper we restrict our attention to finite probability spaces, i.e., has a finite number of elements. Our results can be extended to the -normed spaces without loss of generality, but the details are omitted for brevity. Denote by the space of random variables : , defined over the probability space (, , ). In this paper a random variable is interpreted as a cost, i.e., the smaller the realized value of , the better. For point-wise partial order, i.e., for all , we denote by the point-wise partial order, i.e., for all , we denote by the point-wise partial order, i.e., for all , we denote by the point-wise partial order, i.e., for all .

A MDP is a tuple (, , , , , ) where and are the state and action spaces; is a bounded deterministic cost; is the transition probability distribution; is a discount factor; and is the initial state. Actions are chosen accord-

\footnote{Our results may easily be extended to random costs and random initial states.}
ing to a $\theta$-parameterized stationary Markov 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 $\mu_\theta$ in the MDP.

2.1. Coherent Risk Measures

A risk measure is a function $\rho : Z \to \mathbb{R}$ 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) $\min_{\nu \in \mathcal{R}} \left\{ \nu + \frac{1}{1-\alpha} \mathbb{E}[Z - \nu]^+ \right\}$. A risk measure is called coherent [Artzner et al. (1999)] if it satisfies the following conditions for all $Z, W \in Z$:

- **A1** Convexity: $\forall \lambda \in [0, 1], \rho(\lambda Z + (1-\lambda)W) \leq \lambda \rho(Z) + (1-\lambda)\rho(W)$;
- **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 $\rho(\lambda Z) = \lambda \rho(Z)$.

These conditions ensure the “rationality” of single-period risk assessments; the reader is referred to Artzner et al. (1999) for a detailed motivation. The following representation theorem (Shapiro et al. 2009) shows an important property of the coherent risk measures that is fundamental to our gradient-based approaches.

**Theorem 2.1.** A risk measure $\rho : Z \to \mathbb{R}$ is coherent if and only if there exists a convex bounded and closed set $\mathcal{U} \subset \mathcal{B}$ such that

$$\rho(Z) = \max_{\xi : \xi_{P_0} \in \mathcal{U}(P_0)} \mathbb{E}_{\xi}[Z].$$

(1)

The result essentially states that any coherent risk measure is an expectation w.r.t. a worst-case density function $\xi_{P_0}$, chosen adversarially from a suitable set of test density functions $\mathcal{U}(P_0)$, referred to as risk envelope.

In this paper, we assume that the risk envelope $\mathcal{U}(P_0)$ takes on the following form.

**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

$$\mathcal{U}(P_0) = \left\{ \xi_{P_0} : \begin{array}{l}
g_e(\xi, P_0) = 0, \quad \forall e \in \mathcal{E}, \\
f_i(\xi, P_0) \leq 0, \quad \forall i \in \mathcal{I}, \\
\sum_{\omega \in \Omega} \xi(\omega)P_0(\omega) = 1, \quad \xi(\omega) \geq 0 \end{array} \right\},$$

where each constraint $g_e(\xi, P_0)$ is an affine function in $\xi$, each constraint $f_i(\xi, P_0)$ is a convex function in $\xi$, and there exists a strictly feasible point $\tilde{\xi}$. Furthermore, for any given $\xi \in \mathcal{B}$, $f_i(\xi, P_0)$ and $g_e(\xi, P_0)$ are twice differentiable in $p$, and there exists a $M > 0$ such that

$$\max \left\{ \max_{\xi \in \mathcal{E}} \frac{df_i(\xi, P_0)}{dp(\omega)}, \max_{\xi \in \mathcal{E}} \frac{dg_e(\xi, P_0)}{dp(\omega)} \right\} \leq M, \forall \omega \in \Omega.$$

Assumption 2.2 implies that the risk envelope $\mathcal{U}(P_0)$ is known in an explicit form. From Theorem 6.6 of Shapiro et al. (2009), $\rho$ is a coherent risk if and only if $\mathcal{U}(P_0)$ is a convex and weakly* compact set. In the case of a finite probability space, by Theorem 3.18 of Rudin (1991), we conclude that if $\rho$ is a coherent risk, $\mathcal{U}(P_0)$ is convex and compact. 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.

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. Since in this paper we are also interested in dynamic and time-consistent risk measures, we need to provide a multi-period generalization of the concepts presented in Section 2.1. Here we closely follow the discussion in Ruszczyński (2010).

Consider a probability space $(\Omega, \mathcal{F}, P_0)$, 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 $\mathcal{F}_0 = \{\Omega, 0\}$, 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 $(\Omega, \mathcal{F}_t, P_0)$, and also let $\mathcal{Z}_{0:T} := \mathcal{Z}_0 \times \cdots \times \mathcal{Z}_T$ 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 $\theta$, i.e., $Z_{0:T} = (Z_0 = \gamma^0C(x_0, a_0), \ldots, Z_T = \gamma^TC(x_T, a_T)) \in \mathcal{Z}_{0:T}$.

**Time Consistency** A very important class of dynamic risk measures are those that are time consistent. Example 2.1 in [Lanctu et al. (2011)] 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 and we refer the readers to Ruszczyński (2010) and Lanctu et al. (2011) for further insights.

To evaluate risk consistently in a dynamic setting, we no longer construct a single risk metric, but rather a sequence of risk metrics $\rho_{i,T} : \mathcal{Z}_{i,T} \to \mathbb{R}$, mapping a future stream of random costs into a risk metric/assessment at time $T$ for $t \in \{0, \ldots, T\}$. The sequence of metrics $\{\rho_{i,T}\}_{i=0}^T$ should
be consistent over time [Shapiro, 2009, i2c et al., 2011]. A widely accepted notion of time-consistency is as follows (Ruszczyński, 2010): 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 \( t \). Remarkably, Theorem 1 in Ruszczyński (2010) shows that the “multi-stage composition of risk”, i.e., for each \( t \in \{0, \ldots, T\} \), the mapping \( \rho_{t: T} : Z_t \rightarrow Z_t \) is defined as

\[
\rho_{t: T}(Z) = Z_t + \rho_t \left( Z_{t+1} + \rho_{t+1}(Z_{t+2} + \ldots + \rho_{T-1}(Z_T)) \right),
\]

with each \( \rho_t \) being a static risk measure, is a necessary and sufficient condition for time consistency.

**Markov Coherent Risk Measures** are a very useful class of dynamic time-consistent risk measures that are particularly important for our study of risk in MDPs. Following Eq. 3 for an MDP \( \mathcal{M} \), we define the Markov coherent risk measure \( \rho_T(\mathcal{M}) \) as

\[
\rho_T(\mathcal{M}) = C(x_0, a_0) + \gamma \rho \left( C(x_1, a_1) + \gamma \rho \left( C(x_2, a_2) + \ldots + \gamma \rho \left( C(x_{T-1}, a_{T-1}) + \gamma \rho \left( C(x_T, a_T) \right) \right) \right) \right),
\]

where \( \rho \) is a static coherent risk measure that satisfies Assumption 2.2 and \( x_0, a_0, \ldots, x_T, a_T \) is a trajectory drawn from the MDP \( \mathcal{M} \) under policy \( \mu_0 \). It is important to note that the static coherent risk at state \( x \in X \) is induced by the transition probability \( P_\theta(\cdot|x) = \sum_{a \in A} P(x'|x, a)\mu_\theta(a|x) \). We also define \( \rho_\infty(\mathcal{M}) = \lim_{T \rightarrow \infty} \rho_T(\mathcal{M}) \), which is well-defined since \( \gamma < 1 \) and the cost is bounded. We further assume that \( \rho \) is a Markov risk measure, i.e., the evaluation of each static coherent risk measure \( \rho \) is not allowed to depend on the whole past. Explicitly, for any \( t \geq 0 \) and state dependent random variable \( Z(x_{t+1}) \in Z_{t+1} \), the risk evaluation is given by

\[
\rho(Z(x_{t+1})) = \max_{\xi : \xi P_\theta(\cdot|x_0) \subseteq \mathcal{U}(x_t, P_\theta(\cdot|x_t))} \mathbb{E}_\xi [Z(x_{t+1})],
\]

where we let \( \mathcal{U}(x_t, P_\theta(\cdot|x_t)) \) denote the risk-envelope (2) with \( P_\theta \) 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. More details can be found in Section 5.

### 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 \( \rho \) as defined in Section 2, the static risk problem (SRP) is given by

\[
\min_{\theta} \rho(Z).
\]

For an MDP \( \mathcal{M} \) and a dynamic Markov coherent risk measure \( \rho_T \) as defined by Eq. 4, the dynamic risk problem (DRP) is given by

\[
\min_{\theta} \rho_\infty(\mathcal{M}).
\]

Except for very limited cases, there is no reason to hope that neither the SRP in (6) nor the DRP in (7) should be tractable problems, since the dependence of the risk measure on \( \theta \) may be complex and non-convex. In this work, we aim towards a more modest goal and search for a locally optimal \( \theta \). 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

\[
\nabla_{\theta} \rho(Z) \quad \text{and} \quad \nabla_{\theta} \rho_\infty(\mathcal{M}).
\]

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 \( \theta \). 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 (6) and (7). From the structure of the dynamic risk in Eq. 4, one may think that a gradient estimator for \( \rho(Z) \) may help us to estimate the gradient \( \nabla_{\theta} \rho_\infty(\mathcal{M}) \). Indeed, this is the approach that we take in this paper. The rest of the paper is thus structured as follows. In Section 4, we propose an estimator for \( \nabla_{\theta} \rho(Z) \), which is used in Section 5 to estimate \( \nabla_{\theta} \rho_\infty(\mathcal{M}) \).

### 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 (Marbach & Tsitsiklis, 1998).

**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 and is satisfied in various applications such as queueing systems, inventory management, and financial engineering (see, e.g., the survey by Fu, 2006).
Using Theorem 2.1 and Assumption 2.2 for each θ, we have that ρ(Z) is the solution to the convex optimization problem (1) (for that value of θ). The Lagrangian function of (1), denoted by \( L_θ(ξ, λ^F, λ^E, λ^Z) \), may be written as

\[
L_θ(ξ, λ^F, λ^E, λ^Z) = \sum_{ω∈Ω} ξ(ω) P_θ(ω) Z(ω) - λ^F \left( \sum_{ω∈Ω} ξ(ω) P_θ(ω) - 1 \right) - \sum_{e∈E} λ^E(e) g_e(ξ, P_θ) - \sum_{i∈I} λ^Z(i) f_i(ξ, P_θ).
\]

The convexity of (1) and its strictly feasibility due to Assumption 2.2 implies that \( L_θ(ξ, λ^F, λ^E, λ^Z) \) has a non-empty set of saddle points \( S \). The next theorem presents a formula for the gradient \( ∇θρ(Z) \). As we shall subsequently show, this formula is particularly convenient for devising sampling based estimators for \( ∇θρ(Z) \).

**Theorem 4.2.** Let Assumptions 2.2 and 4.1 hold. For any saddle point \((ξ_θ^*, λ_θ^*, λ_θ^E, λ_θ^Z) ∈ S \) of (8), we have that

\[
∇θρ(Z) = E_θ \left[ ∇θ \log P(ω)(Z - λ_θ^*) \right] - \sum_{e∈E} λ_θ^E(e) \nablaθg_e(ξ_θ^*; P_θ) - \sum_{i∈I} λ_θ^Z(i) \nablaθf_i(ξ_θ^*; P_θ).
\]

The proof of this theorem, given in the supplementary material, involves an application of the Envelope theorem (Milgrom & Segal 2002) and a standard ‘likelihood-ratio’ trick. We now demonstrate the utility of Theorem 4.2 with several examples. The full details for deriving these results are in the supplementary material.

### 4.1. Example 1: CVaR

The α-CVaR risk measure (Rockafellar & Uryasev 2000) is defined by ρCVaR(Z; α) = \[ \inf_{P_θ} \{ t + α^{-1} E \{ Z - t \} \} \]. The risk envelope for CVaR is known to be \( U = \{ ξ P_θ : ξ(ω) ∈ [0, α^{-1}] \}, \ \sum_{ω∈Ω} ξ(ω) P_θ(ω) = 1 \} \) (Shapiro et al. 2009). Furthermore, Shapiro et al. (2009) show that \( ξ_θ^*(ω) = α^{-1} \) when \( Z(ω) > λ_θ^* \), and \( ξ_θ^*(ω) = 0 \) when \( Z(ω) < λ_θ^* \), where \( λ_θ^* \) is any \( (1-α) \)-quantile of \( Z \). Let \( q_α \) denote such a quantile. Using Theorem 4.2 we can easily show that

\[
∇θρ_{CVaR}(Z; α) = E \left[ ∇θ \log P(ω)(Z - q_α) \right] Z(ω) > q_α.
\]

This formula was recently proved by Tamar et al. (2015) for the case of continuous distributions under additional technical assumptions, and used to devise a sampling based estimator by replacing the conditional expectation with a sample average. Here we show that it holds regardless of these assumptions and in the discrete case as well.

### 4.2. Example 2: Mean-Semideviation

Let \( SD[Z] = \left( E \left[ (Z - E[Z])^2 \right] \right)^{1/2} \) denote the semideviation of \( Z \). The mean-semideviation risk measure is defined as \( ρ_{MSD}(Z; α) = E[Z] + α SD[Z] \) and shown to be coherent for \( α ∈ [0, 1] \) (Shapiro et al. 2009). We have the following result for this risk measure.

**Proposition 4.3.** Under Assumption 4.1 we have

\[
∇θρ_{MSD}(Z; α) = \nablaθE[Z] + α E[Z - E[Z]] \nablaθE[P(ω)(Z - E[Z])] - \nablaθE[Z].
\]

This proposition can be used to devise a sampling based estimator for \( ∇θρ_{MSD}(Z; α) \) by replacing all the expectations with sample averages. The algorithm along with the proof of the proposition are in the supplementary material.

### 4.3. General Gradient Estimation Algorithm

We now consider a general coherent risk \( ρ(Z) \), for which, in contrast to the CVaR and mean-semideviation cases, the Lagrangian saddle-point is not known analytically. We show that \( ∇θρ(Z) \) may be estimated using a sample average approximation (SAA) (Shapiro et al. 2009) of the formula in Theorem 4.2.

Assume that we are given \( N \) i.i.d. samples \( ω_i ∼ P_θ \), \( i = 1, ..., N \), and let \( P_{θ,N}(ω) = \frac{1}{N} \sum_{i=1}^{N} I{ω_i = ω} \) denote the corresponding empirical distribution. Also, let the sample risk envelope \( U(P_{θ,N}) \) be defined according to Eq. 2 with \( P_θ \) replaced by \( P_{θ,N} \). Consider the following SAA version of the optimization in Eq. 1

\[
ρ_N(Z) = \max_{ξ ∈ U(P_{θ,N})} \sum_{i=1}^{N} P_{θ,N}(ω_i) ξ(ω_i) Z(ω_i).
\]

(9)

Note that (9) defines a convex optimization problem. In the following, we assume that a solution to (9) may be computed efficiently using standard convex optimization tools such as interior point methods (Boyd & Vandenberghe 2009). Let \( ξ_{θ,N} \) denote a solution to (9) and \( λ_θ^{F,N}, \lambda_θ^E,N, λ_θ^Z,N \) denote the corresponding KKT multipliers. We propose the following estimator for the gradient, based on Theorem 4.2

\[
∇θρ_{N}(Z) = \sum_{i=1}^{N} P_{θ,N}(ω_i) ξ_{θ,N}^N(ω_i) ∇θ \log P(ω_i)(Z(ω_i) - λ_{θ,N}^*) - \sum_{e∈E} λ_{θ,N}^E(e) \nablaθg_e(ξ_{θ,N}^N; P_{θ,N}) - \sum_{i∈I} λ_{θ,N}^Z(i) \nablaθf_i(ξ_{θ,N}^N; P_{θ,N}).
\]

(10)
In the following we show that under some conditions on the set $\mathcal{U}(P_0)$, $\nabla_{\theta,N}\phi(Z)$ is a consistent estimator of $\nabla_{\theta}\phi(Z)$. 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 $\tilde{S} \subset C$ is non-empty and bounded. (II) The functions $f_e(\xi, P_0)$ for all $e \in \mathcal{E}$ and $f_i(\xi, P_0)$ for all $i \in \mathcal{I}$ are finite-valued and continuous (in $\xi$) on $C_\xi$. (III) For $N$ large enough the set $S_N$ is non-empty and $S_N \subset C$ a.e. 1. Further assume that: (IV) If $\xi_N P_{\theta,N} \in \mathcal{U}(P_{\theta,N})$ and $\xi_N$ converges a.e. 1 to a point $\xi$, then $\xi P_0 \in \mathcal{U}(P_0)$. We then have that $\lim_{N \to \infty} \rho_N(Z) = \rho(Z)$ and $\lim_{N \to \infty} \nabla_{\theta,N}\rho(Z) = \nabla_{\theta}\phi(Z)$ a.e. 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 and 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 $P_0$ uniformly on $\Omega$, (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 measures.

To summarize this section, we have seen that by exploiting the special structure of coherent risk measures in Theorem 2.4 and by the envelope-theorem style result of Theorem 4.2, we were able to derive sampling-based, likelihood-ratio-style algorithms for estimating the policy gradient $\nabla_{\theta}\phi(Z)$ of coherent static risk measures. The gradient estimation algorithms developed here for static risk measures will be used as a subroutine in our subsequent treatment of dynamic risk measures.

## 5. Gradient Formula for Dynamic Risk

In this section, we first derive a new formula for the gradient of a general Markov-coherent dynamic risk measure $\nabla_{\theta}\rho_{\infty}(\mathcal{M})$ that involves the value function of the risk objective $\rho_{\infty}(\mathcal{M})$ (e.g., the value function proposed by Ruszczyński [2010]). This formula extends the well-known "policy gradient theorem" (Sutton et al. 2000; Konda & Tsitsiklis 2000) 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 $\theta$, calculate the risk-sensitive value function of $\rho_{\infty}(\mathcal{M})$ (see Section 5.2), and

**Actor:** Using the critic’s value function, estimate $\nabla_{\theta}\rho_{\infty}(\mathcal{M})$ by sampling (see Section 5.3).

The value function proposed by Ruszczyński (2010) 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 Ruszczyński [2010]) 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 (Bertsekas & Tsitsiklis, 1996). In particular, temporal difference (TD) learning methods (Sutton & Barto, 1998) are popular for this purpose, which have been recently extended to robust MDPs by Tamar et al. (2014). We use their (robust) TD algorithm and show how our critical use it to approximate the risk-sensitive value function. We then discuss how the error introduced by this approximation affects the gradient estimate of the actor.

### 5.1. 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 $C_0(\mathcal{X}) = \{\sum a_\mu C(x,a)\mu(a|x)\}$ be the stage-wise cost function induced by policy $\mu_\theta$. We now define the risk sensitive Bellman operator $T_\theta[V] := B(\mathcal{X}) \mapsto B(\mathcal{X})$ as

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

According to Theorem 1 in Ruszczyński (2010), the operator $T_\theta$ has a unique fixed-point $V_\theta$, 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 $\theta$, 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.

### 5.2. Value Function Approximation

Consider the linear approximation of the risk-sensitive value function $V_\theta(x) \approx v^\top \phi(x)$, where $\phi(\cdot) \in \mathbb{R}^{d_2}$ is the $d_2$-dimensional state-dependent feature vector. Thus, the approximate value function belongs to the low dimensional sub-space $\mathcal{V} = \{\phi|v \in \mathbb{R}^{d_2}\}$, where $\Phi : \mathcal{X} \mapsto \mathbb{R}^{d_2}$ is a function mapping such that $\Phi(x) = \phi(x)$. The goal of our critic is to find a good approximation of $V_\theta$ from simulated trajectories of the MDP. In order to have a well-defined approximation scheme, we first impose the following standard assumption (Bertsekas & Tsitsiklis, 1996).

**Assumption 5.1.** The mapping $\Phi$ has full column rank.

For a function $\gamma : \mathcal{X} \mapsto \mathbb{R}$, we define its weighted (by $d$) $l_2$-norm as $\|\gamma\|_d = \sqrt{\sum_x d(x'|x)^2 \gamma(x')^2}$, where $d$ is a distribution over $\mathcal{X}$. Using this, we define $\Pi : \mathcal{X} \mapsto \mathcal{V}$, 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 $V_\theta$ with the fixed-point of the joint operator $\Pi T_\theta$.  

Policy Gradient for Coherent Risk Measures
i.e., \( \hat{V}_\theta(x) = v_\theta^T \phi(x) \), such that

\[
\forall x \in X, \quad \hat{V}_\theta(x) = \Pi T_\theta(\hat{V}_\theta(x)).
\]  

(12)

From Eq. [5] that has been derived from Theorem 2.1 for dynamic risks, it is easy to see that the risk-sensitive Bellman equation (11) is a robust Bellman equation [Nili and El Ghaoui 2005] with uncertainty set \( U(x, P_\theta(\cdot|x)) \). Thus, we may use the TD approximation of the robust Bellman equation proposed by Tamar et al. [2014] to find an approximation of \( V_\theta \). We will need the following assumption analogous to Assumption 2 in Tamar et al. [2014].

**Assumption 5.2.** There exists \( \kappa \in (0, 1) \) such that \( \xi(x') \leq \kappa/\gamma \), for all \( \xi(\cdot)P_\theta(\cdot|x) \in U(x, P_\theta(\cdot|x)) \) and all \( x, x' \in X \).

Given Assumption 5.2, Proposition 3 in Tamar et al. [2014] guarantees that the projected risk-sensitive Bellman operator \( \Pi T_\theta \) is a contraction w.r.t. \( d_0 \)-norm. Therefore, Eq. [12] has a unique fixed-point solution \( v_\theta(x) = v_\theta^* \phi(x) \). This means that \( v_\theta \in \mathbb{R}^{\kappa_x} \) satisfies \( v_\theta^* \in \arg \min \|T_\theta(\Phi v) - \Phi v\|_{d_0}^2 \). By the projection theorem on Hilbert spaces, the orthogonality condition for \( v_\theta^* \) becomes

\[
\sum_{x \in X} d_\theta(x|x_0)\phi(x)\phi(x)^T v_\theta = \sum_{x \in X} d_\theta(x|x_0)\phi(x)C_\theta(x) + \gamma \sum_{x \in X} d_\theta(x|x_0)\phi(x) \max_{\xi: \xi P_\theta(\cdot|x) \in U(x, P_\theta(\cdot|x))} \mathbb{E}[\Phi v]_\theta.
\]

(13)

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 \( \theta \), we may estimate the fixed-point solution \( v_\theta^* \) using the projected risk sensitive value iteration (PRSVI) algorithm with the update rule

\[
v_{k+1} = \left( \frac{1}{N} \sum_{t=0}^{N-1} \phi(x_t)\phi(x_t)^T \right)^{-1} \left[ \frac{1}{N} \sum_{t=0}^{N-1} \phi(x_t)C_\theta(x_t) \right] + \gamma \frac{1}{N} \sum_{t=0}^{N-1} \phi(x_t) \max_{\xi: \xi P_\theta(\cdot|x) \in U(x, P_\theta(\cdot|x))} \mathbb{E}[\Phi v_k]_\theta.
\]

(13)

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

In order to implement the iterative algorithm [13], one must repeatedly solve the inner optimization problem \( \max_{\xi: \xi P_\theta(\cdot|x) \in U(x, P_\theta(\cdot|x))} \mathbb{E}[\Phi v]_\theta \). When the state space \( X \) is large, solving this optimization problem is often computationally expensive or even intractable. Similar to Section 3.4 of Tamar et al. [2014], 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 \( P_N(x'|x, a) = \frac{\sum_{t=0}^{N-1} 1\{x_t=x, a_t=a, x_{t+1}=x'\}}{\sum_{t=0}^{N-1} 1\{x_t=x, a_t=a\}} \) and \( P_{\theta,N}(x'|x) = \sum_{a \in A} P_N(x'|x, a)\mu_\theta(a|x) \). Consider the following \( \ell_2 \)-regularized empirical robust optimization problem

\[
\rho_N(\Phi v) = \max_{\xi: \xi P_{\theta,N}(\cdot|x) \in U(x, P_{\theta,N}(\cdot|x))} \mathbb{E}[\Phi v]_\theta + \frac{1}{2N} \|P_{\theta,N}(x'|x)\xi(x')\|^2.
\]

(14)

As in Meng and Xu [2006], the \( \ell_2 \)-regularization term in this optimization problem guarantees convergence of optimizers \( \xi^* \) 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 \( \lambda_{\theta,x;N}^*, \lambda_{\theta,x;N}^{\infty} \), \( \lambda_{\theta,x;N}^\gamma \), the corresponding KKT multipliers.

We obtain the empirical PRSVI algorithm by replacing the inner optimization \( \max_{\xi: \xi P_{\theta,N}(\cdot|x) \in U(x, P_{\theta,N}(\cdot|x))} \mathbb{E}[\Phi v]_\theta \) in Eq. [13] with \( \rho_N(\Phi v) \) from Eq. [14]. Similarly, as both \( N \) and \( k \) tend to infinity, \( v_k \) converges w.p. 1 to \( v_\theta^* \). More details can be found in the supplementary material.

### 5.3. Gradient Estimation

In Section 5.2, we showed that we may effectively approximate the value function of a fixed policy \( \theta \) using the (empirical) PRSVI algorithm in Eq. [13]. In this section, we first derive a formula for the gradient of the Markov-coherent dynamic risk measure \( \rho_{\infty}(M) \), and then propose a SAA algorithm for estimating this gradient, in which we use the SAAs approximation of value function from Section 5.2. As described in Section 5.1, \( \rho_{\infty}(M) = V_\theta(x_0) \), and thus, we shall first derive a formula for \( \nabla \theta V_\theta(x_0) \).

Let \( (\xi_{\theta,x}^*, \lambda_{\theta,x}^*, \lambda_{\theta,x}^{\infty}, \lambda_{\theta,x}^\gamma) \) be the saddle point of (8) corresponding to the state \( x \in X \). 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 \( (\lambda_{\theta,x}^*, \lambda_{\theta,x}^{\infty}, \lambda_{\theta,x}^\gamma) \). 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 5.3.** The likelihood ratio \( \nabla \theta \log \mu_\theta(a|x) \) is well-defined and bounded for all \( x \in X \) and \( a \in A \).

As in Theorem 4.2 for the static case, we may use the envelope theorem and the risk-sensitive Bellman equation, \( V_\theta(x) = C_\theta(x) + \gamma \max_{\xi: \xi P_\theta(\cdot|x) \in U(x, P_\theta(\cdot|x))} \mathbb{E}[\Phi v]_\theta \), to derive a formula for \( \nabla \theta V_\theta(x) \). We report this result in Theorem 4.3. In the case when these spaces are continuous, the empirical transition probability can be found by kernel density estimation.

---

3In 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.
rem [5.4] which is analogous to the risk-neutral policy gradient theorem (Sutton et al. 2000 [Konda & Tsitsiklis 2000 Bhatnagar et al. 2009]. The proof is in the supplementary material.

Theorem 5.4. Under Assumptions 2.2 we have

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

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

$$h_\theta(x, a) = C(x, a) + \sum_{x' \in I} P(x'|x, a) \xi(x') \left[ \gamma V_\theta(x') - \lambda_\theta^P \right] - \sum_{x' \in I} \lambda_\theta^P \left( \frac{dP_i(\xi(x'), x')}{dP(x')} - \sum_{x' \in I} \lambda_\theta^P \left( \frac{dP_i(\xi(x'), x')}{dP(x')} \right) \right].$$

(15)

Theorem 5.4 indicates that the policy gradient of the Markov-coherent dynamic risk measure $\rho_\infty(M)$, i.e., $\nabla \rho_\infty(M) = \nabla V_\theta$, is equivalent to the risk-neutral value function $\theta$ in a MDP with the stage-wise cost function $\nabla \log \mu_\theta(a|x) h_\theta(x, a)$ (which is well-defined and bounded), and transition probability $P_\theta(\cdot|x) \xi(x)$.

Thus, when the saddle points are known and the state space $X$ is not too large, we can compute $\nabla V_\theta$ using a policy evaluation algorithm. However, when the state space is large, exact calculation of $\nabla V_\theta$ by policy evaluation becomes impossible, and our goal would be to derive a sampling method to estimate $\nabla V_\theta$. Unfortunately, since the risk envelop depends on the policy parameter $\theta$, unlike the risk-neutral case, the risk sensitive (or robust) Bellman equation $T_\theta[V_\theta(x)]$ in (11) is nonlinear in the stationary Markov policy $\mu_\theta$. Therefore $h_\theta$ cannot be considered using the action-value function (Q-function) of the robust MDP. Therefore, even if the exact formulation of the value function $V_\theta$ 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 $V_\theta$ 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 $V_\theta^*$. To address the sampling issues, we propose the following two-phase sampling procedure for estimating $\nabla V_\theta$.

(1) Generate $N$ trajectories $\{x_t^{(n)}, a_t^{(n)}, x_t^{(n)}| a_t^{(n)} \}_{j=1}^N$ from the Markov chain induced by policy $\theta$ and transition probabilities $P_\theta^*(\cdot|x) := \xi(x)$.

(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 $h_\theta(x, a)$

$$h_{\theta, N}(x, a) := C(x, a) + \frac{1}{N} \sum_{k=1}^N \xi^{(y(k))} \left[ \gamma \phi(y(k))-\lambda_\theta^P \right]$$

$$-\sum_{e \in E} \lambda_\theta^P \left( \frac{dP_i(\xi^{(y(k))}, x)}{dP(y(k))} - \sum_{e \in E} \lambda_\theta^P \left( \frac{dP_i(\xi^{(y(k))}, x)}{dP(y(k))} \right) \right].$$

(3) Calculate an estimate of $\nabla V_\theta$ using the following average over all the samples:

$$\frac{1}{N} \sum_{j=1}^N \sum_{t=0}^{\infty} \gamma^t \log \mu_\theta(a_t^{(j)}|x_t^{(j)}) h_{\theta, N}(x_t^{(j)}, a_t^{(j)})$$

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 $(\xi(x), \lambda_\theta^P, \lambda_\theta^E, \lambda_\theta^I)$ do not have a closed-form solution, we may follow the SAA procedure of Section 5.2 and replace them and the transition probabilities $P_N(x'|x, a)$ with their sample estimates $(\xi(x), \lambda_\theta^P, \lambda_\theta^E, \lambda_\theta^I)$ and $P_N(x'|x, a)$ respectively.

At the end, we show the convergence of the above two-phase sampling procedure. Let $d_{P_\theta^*, x}^N(x_0)$ and $\pi_{P_\theta^*, x}^N(x_0)$ be the state and state-action occupancy measure induced by the transition probability function $P_\theta^*(\cdot|x)$, respectively. Similarly, let $d_{P_{\theta, N}^*, x}^N(x_0)$ and $\pi_{P_{\theta, N}^*, x}^N(x_0)$ be the state and state-action occupancy measure induced by the estimated transition probability function $P_{\theta, N}^*(\cdot|x) := \xi(x) P_{\theta, N}^*(\cdot|x)$. 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

$$\frac{1}{N} \sum_{j=1}^N \sum_{t=0}^{\infty} \gamma^t \mathbf{1}\{x_t^{(j)} = x, a_t^{(j)} = a\} = \pi_{P_{\theta, N}^*, x}^N(x_0, a_0).$$

Based on the strongly convex property of the $\ell_2$-regularized objective function in the inner robust optimization problem $\min_{\Pi_\theta} \Phi_{\Pi_\theta}$, we can show that both the state-action occupancy measure $\pi_{P_{\theta, N}^*, x}^N(x_0, a_0)$ and the stage-wise cost $h_{\theta, N}(x, a)$ converge to their true values within a value function approximation error bound $\Delta = \|\Phi_{\Pi_\theta}^* - V_\theta\|_\infty$. We refer the readers to the supplementary materials for these technical results. These results together with Theorem 5.4 imply the consistency of the policy gradient estimation.

Theorem 5.5. For any $x_0 \in X$, the following expression holds with probability 1:

$$\lim_{N \to \infty} \frac{1}{N} \sum_{j=1}^N \sum_{t=0}^{\infty} \gamma^t \log \mu_\theta(a_t^{(j)}|x_t^{(j)}) h_{\theta, N}(x_t^{(j)}, a_t^{(j)})$$

$$- \nabla V_\theta(x_0) = 0.$$

Thm. 5.5 guarantees that as the value function approximation error decreases and the number of samples increases, the sampled gradient converges to the true gradient.
6. Discussion and Conclusion

We presented sampling-based algorithms for estimating the gradient of both static and dynamic coherent risk measures, using two new ‘policy gradient’ style formulas. In contrast to previous approaches, our algorithms apply when the tunable parameters control the distribution of the random variable, and thus extend coherent risk optimization to new domains such as reinforcement learning.

In this work we have not discussed the important question of which coherent risk measure to prefer for a given problem. In the supplementary material, we report on experiments that we conducted using our algorithms on the American options domain (Tamar et al., 2014), a standard benchmark for risk-sensitive RL. In these experiments, optimizing the dynamic risk resulted in very conservative policies compared to the static risk. An in-depth empirical study is left to future work.

References

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

Baxter, J. and Bartlett, P. Infinite-horizon policy-gradient estimation. JAIR, 15:319–350, 2001.

Bertsekas, D. Dynamic Programming and Optimal Control, Vol I. Athena Scientific, 3rd edition, 2005.

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

Bertsekas, D. and Tsitsiklis, J. Neuro-Dynamic Programming. Athena Scientific, 1996.

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

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

Boyd, S. and Vandenberghe, L. Convex optimization. Cambridge university press, 2009.

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

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

Fiacco, A. Introduction to sensitivity and stability analysis in nonlinear programming. Elsevier, 1983.

Fu, M. Gradient estimation. In Simulation, volume 13 of Handbooks in Operations Research and Management Science, pp. 575 – 616. Elsevier, 2006.

Glynn, P. Likelihood ratio gradient estimation for stochastic systems. Communications of the ACM, 33(10):75–84, 1990.

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

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

Konda, V. and Tsitsiklis, J. Actor-critic algorithms. In Advances in Neural Information Processing Systems, pp. 1008–1014, 2000.

Li, Y., Szepesvari, Cs., and Schuurmans, D. Learning exercise policies for american options. In Proc. of the Twelfth International Conference on Artificial Intelligence and Statistics, JMLR: W&CP, volume 5, pp. 352–359, 2009.

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

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

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

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

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

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

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

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

Rockafellar, R., Wets, R., and Wets, M. Variational analysis, volume 317. Springer, 1998.

Rudin, W. Functional analysis. international series in pure and applied mathematics, 1991.
Ruszczyński, A. Risk-averse dynamic programming for Markov decision processes. Mathematical Programming, 125(2):235–261, 2010.

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

Shapiro, A. On a time consistency concept in risk averse multistage stochastic programming. Operations Research Letters, 37(3):143–147, 2009.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. Lectures on Stochastic Programming, chapter 6, pp. 253–332. SIAM, 2009.

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

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

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

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

Tamar, A., Glassner, Y., and Mannor, S. Optimizing the CVaR via sampling. In AAAI, 2015.
A. Gradient Results for Static Mean-Semideviation

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

$$\rho_{\text{MSD}}(Z) = \mathbb{E}[Z] + c \left( \mathbb{E} \left[ (Z - \mathbb{E}[Z])^2 \right] \right)^{1/2}, \quad (16)$$

Following the derivation in [Shapiro et al., 2009], note that $(\mathbb{E} \left[ |Z|^2 \right])^{1/2} = \|Z\|_2$, where $\| \cdot \|_2$ denotes the $L_2$ norm of the space $L_2(\Omega, \mathcal{F}, P_0)$. The norm may also be written as:

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

and hence

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

It follows that Eq. (1) holds with

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

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

$$\rho_{\text{MSD}}(Z) = \sup_{\|\xi\|_2 \leq 1, \xi \geq 0} \langle 1 + c\xi - c\mathbb{E} [\xi], Z \rangle. \quad (17)$$

Let $\bar{\xi}$ denote an optimal solution for (17). In [Shapiro et al., 2009] it is shown that $\bar{\xi}$ is a contact point of $(Z - \mathbb{E}[Z])_+$, that is

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

and we have that

$$\bar{\xi} = \frac{(Z - \mathbb{E}[Z])_+}{\| (Z - \mathbb{E}[Z])_+ \|_2} = \frac{(Z - \mathbb{E}[Z])_+}{\text{SD}(Z)}. \quad (18)$$

Let $\xi$ be a probability distribution, but for $c \in [0, 1]$, it can be shown [Shapiro et al., 2009] that $1 + c\xi - c\mathbb{E} [\xi]$ always is.

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

**Proposition A.1.** Under Assumption (17) we have that

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

and, according to the standard likelihood-ratio method,

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

**Proof.** Note that in Eq. (17) the constraints do not depend on $\theta$. Therefore, using the envelope theorem we obtain that

$$\nabla_\theta \rho(Z) = \nabla_\theta \langle 1 + c\xi - c\mathbb{E} [\xi], Z \rangle$$

$$= \nabla_\theta \langle 1, Z \rangle + c \nabla_\theta \langle \xi, Z \rangle - c \nabla_\theta \langle \mathbb{E} [\xi], Z \rangle. \quad (19)$$

We now write each of the terms in Eq. (19) as an expectation. We start with the following standard likelihood-ratio result:

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

Also, we have that

$$\langle \mathbb{E} [\xi], Z \rangle = \mathbb{E} \left[ \mathbb{E} [\xi] \mathbb{E}[Z] \right].$$
therefore, by the derivative of a product rule:
\[ \nabla_\theta (\mathbb{E} \left[ \xi \right], Z) = \nabla_\theta \mathbb{E} \left[ \xi \mathbb{E} [ Z ] \right] + \mathbb{E} \left[ \xi \right] \nabla_\theta \mathbb{E} [ Z ] . \]

By the likelihood-ratio trick and Eq. (18) we have that
\[ \nabla_\theta \mathbb{E} \left[ \xi \right] = \frac{1}{\text{SD}(Z)} \mathbb{E} \left[ \nabla_\theta \log P(\omega)(Z - \mathbb{E} [ Z ] )_+ \right] . \]

Also, by the likelihood-ratio trick
\[ \nabla_\theta \mathbb{E} \left[ \xi Z \right] = \mathbb{E} \left[ \nabla_\theta \log P(\omega) \xi Z \right] . \]

Plugging these terms back in Eq. (19), we have that
\[ \nabla_\theta \rho \left( Z \right) = \nabla_\theta \mathbb{E} [ Z ] + c \mathbb{E} \left[ \nabla_\theta \log P(\omega)(Z - \mathbb{E} [ Z ] )_+ \left( \nabla_\theta \log P(\omega)(Z - \mathbb{E} [ Z ] ) - \nabla_\theta \mathbb{E} [ Z ] \right) \right] . \]

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

**Algorithm 1 GMSD**

1: Given:
   - Risk level \( c \)
   - An i.i.d. sequence \( z_1, \ldots, z_N \sim P_\omega \).
2: Set \( \hat{\mathbb{E}} [ Z ] = \frac{1}{N} \sum_{i=1}^{N} z_i \).
3: Set \( \hat{\text{SD}}(Z) = \left( \frac{1}{N} \sum_{i=1}^{N} (z_i - \hat{\mathbb{E}} [ Z ])_+^2 \right)^{1/2} \).
4: Set \( \hat{\nabla_\theta \mathbb{E} [ Z ]} = \frac{1}{N} \sum_{i=1}^{N} \nabla_\theta \log P(z_i) z_i \).
5: Return:
\[ \nabla_\theta \hat{\rho}(Z) = \hat{\nabla_\theta \mathbb{E} [ Z ]} + \frac{c}{\hat{\text{SD}}(Z)} \frac{1}{N} \sum_{i=1}^{N} \left( z_i - \hat{\mathbb{E}} [ Z ] \right)_+ \left( \nabla_\theta \log P(z_i)(z_i - \hat{\mathbb{E}} [ Z ] ) - \hat{\nabla_\theta \mathbb{E} [ Z ]} \right) . \]

**B. Proof of Theorem 4.2**

First note from Assumption 2.2 that

(i) Slater’s condition holds in the primal optimization problem (1).
Policy Gradient for Coherent Risk Measures

(ii) \( L_0(\xi, \lambda^P, \lambda^E, \lambda^Z) \) is convex in \( \xi \) and concave in \((\lambda^P, \lambda^E, \lambda^Z)\).

Thus by the duality result in convex optimization \cite{Boyd&Vandenberghe2009}, the above conditions imply strong duality and we have \( \rho(Z) = \max_{\xi \in \Omega} \min_{\lambda^P, \lambda^E, \lambda^Z} \max_{\xi} L_0(\xi, \lambda^P, \lambda^E, \lambda^Z). \) From Assumption 2.2, one can also see that the family of functions \( \{ L_0(\xi, \lambda^P, \lambda^E, \lambda^Z) \}_{(\xi, \lambda^P, \lambda^E, \lambda^Z) \in \mathbb{R}^{|\Omega|} \times \mathbb{R}^{|E|} \times \mathbb{R}^{|Z|}} \) is equi-differentiable in \( \theta \), \( L_0(\xi, \lambda^P, \lambda^E, \lambda^Z) \) is Lipschitz, as a result, an absolutely continuous function in \( \theta \), and thus, \( \nabla_\theta L_0(\xi, \lambda^P, \lambda^E, \lambda^Z) \) is continuous and bounded at each \( (\xi, \lambda^P, \lambda^E, \lambda^Z) \). Then for every selection of saddle point \((\hat{\xi}, \hat{\lambda}^P, \hat{\lambda}^E, \hat{\lambda}^Z) \in S \) of \( (8) \), using the Envelope theorem for saddle-point problems (see Theorem 4 of Milgrom & Segal2002), we have

\[
\nabla_\theta \max_{\xi \geq 0} \min_{\lambda^P, \lambda^E, \lambda^Z} L_0(\xi, \lambda^P, \lambda^E, \lambda^Z) = \nabla_\theta L_0(\xi, \lambda^P, \lambda^E, \lambda^Z) \big|_{(\xi^*, \lambda^*_0, \lambda^*_0, \lambda^*_0)} .
\]

The result follows by writing the gradient in \((20)\) explicitly, and using the likelihood-ratio trick:

\[
\sum_{\omega \in \Omega} \xi(\omega) \nabla_\theta P_\theta(\omega) Z(\omega) - \lambda^P \sum_{\omega \in \Omega} \xi(\omega) \nabla_\theta P_\theta(\omega) = \sum_{\omega \in \Omega} \xi(\omega) P(\omega) \nabla_\theta \log P(\omega) (Z(\omega) - \lambda^P),
\]

where the last equality is justified by Assumption 4.1.

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

\[
L_{\theta; N}(\xi, \lambda^P, \lambda^E, \lambda^Z) = \sum_{\omega \in \Omega} \xi(\omega) P_{\theta; N}(\omega) Z(\omega) - \lambda^P \left( \sum_{\omega \in \Omega} \xi(\omega) P_{\theta; N}(\omega) - 1 \right) - \sum_{e \in E} \lambda^E(e) f_e(\xi, P_{\theta; N}) - \sum_{i \in I} \lambda^Z(i) f_i(\xi, P_{\theta; N}).
\]

Recall that \( S \subset \mathbb{R}^{|\Omega|} \times \mathbb{R} \times \mathbb{R}^{|E|} \times \mathbb{R}^{|Z|} \) denotes the set of saddle points of the true Lagrangian \((8)\). Let \( S_N \subset \mathbb{R}^{|\Omega|} \times \mathbb{R} \times \mathbb{R}^{|E|} \times \mathbb{R}^{|Z|} \) denote the set of SAA Lagrangian \((21)\) saddle points.

Suppose that there exists a compact set \( C \equiv C_\xi \times C_\lambda \), where \( C_\xi \subset \mathbb{R}^{|\Omega|} \) and \( C_\lambda \subset \mathbb{R} \times \mathbb{R}^{|E|} \times \mathbb{R}^{|Z|} \) such that:

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

(ii) The functions \( f_e(\xi, P_\theta) \) for all \( e \in E \) and \( f_i(\xi, P_\theta) \) for all \( i \in I \) are finite valued and continuous (in \( \xi \)) on \( C_\xi \).

(iii) For \( N \) large enough the set \( S_N \) is non-empty and \( S_N \subset C \) w.p. 1.

Recall from Assumption 2.2 that for each fixed \( \xi \in B \), 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 \( P_{\theta; N} \rightarrow P_\theta \) w.p. 1. From the definition of the Lagrangian function and continuity of constraint functions, one can easily see that for each \( (\xi, \lambda^P, \lambda^E, \lambda^Z) \in \mathbb{R}^{|\Omega|} \times \mathbb{R} \times \mathbb{R}^{|E|} \times \mathbb{R}^{|Z|} \), \( L_{\theta; N}(\xi, \lambda^P, \lambda^E, \lambda^Z) \rightarrow L_0(\xi, \lambda^P, \lambda^E, \lambda^Z) \) 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 \( \xi_N \) converges w.p. 1 to a point \( \xi \), then \( \xi \in \mathcal{U}(P_\theta) \).

According to the discussion in Page 161 of Shapiro et al. \cite{Shapiro2009}, 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 \( \xi_N \rightarrow \xi \) w.p. 1.
and from Theorem 6.6 in [Shapiro et al. 2009], we know that both sets \( \mathcal{U}(P_0;N) \) and \( \mathcal{U}(P_0) \) are convex and compact. Furthermore, note that we have

(vi) The objective function on (1) is linear, finite valued and continuous in \( \xi \) on \( C_\xi \) (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 \( \xi \).

From (i,iv,v,vi,vii), and under the same lines of proof as in Theorem 5.5 of [Shapiro et al. 2009], we have that

\[
\rho_N(Z) \rightarrow \rho(Z) \text{ w.p. 1 as } N \rightarrow \infty, \tag{22}
\]

\[
\mathbb{D}\{P_N, P\} \rightarrow 0 \text{ w.p. 1 as } N \rightarrow \infty, \tag{23}
\]

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 [Shapiro et al. 2009], that \( L_0;N(\xi_0^*, \lambda_0^*, P, \lambda_0^*, \lambda_0^*; P_0; N, \lambda_0^*; P_0) \rightarrow L_0(\xi_0^*, \lambda_0^*, P, \lambda_0^*, \lambda_0^*; P_0) \) w.p. 1 and \( \mathbb{D}\{S_N, S\} \rightarrow 0 \text{ w.p. 1 as } N \rightarrow \infty. \) Based on the definition of the deviation of sets, the limit point of any element in \( S_N \) is also an element in \( S. \)

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

**Part 1**

We first show that \( L_0;N(\xi_0^*, \lambda_0^*, P, \lambda_0^*, \lambda_0^*; P_0; N, \lambda_0^*; P_0) \) converges to \( L_0(\xi_0^*, \lambda_0^*, P, \lambda_0^*, \lambda_0^*; P_0) \) w.p. 1 as \( N \rightarrow \infty. \)

For each fixed \( (\lambda^P, \lambda^E, \lambda^T) \in C_{\lambda} \), the function \( L_0(\xi, \lambda^P, \lambda^E, \lambda^T) \) is convex and continuous in \( \xi \). Together with the point-wise S.L.L.N. property, Theorem 7.49 of Shapiro et al. [2009] implies that \( L_0;N(\xi, \lambda^P, \lambda^E, \lambda^T) \rightarrow L_0(\xi, \lambda^P, \lambda^E, \lambda^T) \) as \( N \rightarrow \infty. \) where \( \rightarrow \) denotes epi-convergence. Furthermore, since the objective and constraint functions are convex in \( \xi \) and are finite valued on \( C_{\xi} \), the set \( \text{dom}L_0(\cdot, \lambda^P, \lambda^E, \lambda^T) \) has non-empty interior. It follows from Theorem 7.27 of Shapiro et al. [2009] that epi-convergence of \( L_0;N \) to \( L_0 \) implies uniform convergence on \( C_{\xi} \), i.e., \( \sup_{\xi \in C_{\xi}} |L_0;N(\xi, \lambda^P, \lambda^E, \lambda^T) - L_0(\xi, \lambda^P, \lambda^E, \lambda^T)| \leq \epsilon. \) On the other hand, for each fixed \( \xi \in C_{\xi} \), the function \( L_0(\xi, \lambda^P, \lambda^E, \lambda^T) \) is linear and thus continuous in \( (\lambda^P, \lambda^E, \lambda^T) \) and \( \text{dom}L_0(\xi, \cdot, \cdot, \cdot) \) has non-empty interior. It follows from analogous arguments that \( \sup_{(\lambda^P, \lambda^E, \lambda^T) \in C_{\lambda}} |L_0;N(\xi, \lambda^P, \lambda^E, \lambda^T) - L_0(\xi, \lambda^P, \lambda^E, \lambda^T)| \leq \epsilon. \) Combining these results implies that for any \( \epsilon > 0 \) and a.e. \( \omega_{SA} \in \Omega_{SA} \) there is a \( N^*(\epsilon, \omega_{SA}) \) such that

\[
\sup_{(\xi, \lambda^P, \lambda^E, \lambda^T) \in C_{\lambda}} |L_0;N(\xi, \lambda^P, \lambda^E, \lambda^T) - L_0(\xi, \lambda^P, \lambda^E, \lambda^T)| \leq \epsilon. \tag{24}
\]

Now, assume by contradiction that for some \( N > N^*(\epsilon, \omega_{SA}) \) we have \( L_0;N(\xi_0^*, \lambda_0^*, P_0; N, \lambda_0^*; P_0) \) greater than \( N^*(\epsilon, \omega_{SA}) \) then by definition of the saddle points

\[
L_0;N(\xi_0^*, \lambda_0^*, \lambda_0^*, \lambda_0^*; P_0; N, \lambda_0^*; P_0) \geq L_0;N(\xi_0^*, \lambda_0^*, \lambda_0^*, \lambda_0^*; N, \lambda_0^*; P_0) > L_0(\xi_0^*, \lambda_0^*, \lambda_0^*, \lambda_0^*; P_0) + \epsilon \Rightarrow L_0(\xi_0^*, \lambda_0^*, \lambda_0^*, \lambda_0^*; P_0) + \epsilon,
\]

contradicting (24).

Similarly, assuming by contradiction that \( L_0(\xi_0^*, \lambda_0^*, \lambda_0^*, \lambda_0^*; P_0) > \epsilon \) gives

\[
L_0(\xi_0^*, \lambda_0^*, \lambda_0^*, \lambda_0^*; P_0) \geq L_0(\xi_0^*, \lambda_0^*, \lambda_0^*, \lambda_0^*; P_0) > L_0;N(\xi_0^*, \lambda_0^*, \lambda_0^*, \lambda_0^*; N, \lambda_0^*; P_0) + \epsilon \Rightarrow L_0;N(\xi_0^*, \lambda_0^*, \lambda_0^*, \lambda_0^*; P_0) + \epsilon,
\]

also contradicting (24).

It follows that \( |L_0;N(\xi_0^*, \lambda_0^*, P_0, \lambda_0^*, \lambda_0^*, \lambda_0^*; N, \lambda_0^*; P_0) - L_0(\xi_0^*, \lambda_0^*, \lambda_0^*, \lambda_0^*; P_0)| \leq \epsilon \) for all \( N > N^*(\epsilon, \omega_{SA}) \), and therefore

\[
\lim_{N \to \infty} L_0;N(\xi_0^*, \lambda_0^*, P_0, \lambda_0^*, \lambda_0^*, \lambda_0^*; N, \lambda_0^*; P_0) = L_0(\xi_0^*, \lambda_0^*, P_0, \lambda_0^*, \lambda_0^*, \lambda_0^*), \tag{25}
\]

w.p. 1.
Part 2  Let us now show that \( D \{ S_N, S \} \to 0 \). We argue by contradiction. Suppose that \( D \{ S_N, S \} \not\to 0 \). Since \( C \) is compact, we can assume that there exists a sequence \((\xi_{\theta;0}^*, \lambda_{\theta;0}^p, \lambda_{\theta;0}^e, \lambda_{\theta;0}^T) \in S_N \) that converges to a point \((\tilde{\xi}^*, \tilde{\lambda}^p, \tilde{\lambda}^e, \tilde{\lambda}^T) \in C \) and \((\tilde{\xi}^*, \tilde{\lambda}^p, \tilde{\lambda}^e, \tilde{\lambda}^T) \notin S \). However, from (23) we must have that \( \tilde{\xi}^* \in \mathcal{P} \). Therefore, we must have that
\[
L_\theta(\tilde{\xi}^*, \tilde{\lambda}^p, \tilde{\lambda}^e, \tilde{\lambda}^T) > L_\theta(\xi_{\theta;0}^*, \lambda_{\theta;0}^p, \lambda_{\theta;0}^e, \lambda_{\theta;0}^T),
\]
by definition of the saddle point set.

Now,
\[
L_{\theta;N}(\xi_{\theta;0}^*, \lambda_{\theta;0}^p, \lambda_{\theta;0}^e, \lambda_{\theta;0}^T) - L_\theta(\xi_{\theta;0}^*, \lambda_{\theta;0}^p, \lambda_{\theta;0}^e, \lambda_{\theta;0}^T) = \left[ L_{\theta;N}(\xi_{\theta;0}^*, \lambda_{\theta;0}^p, \lambda_{\theta;0}^e, \lambda_{\theta;0}^T) - L_\theta(\xi_{\theta;0}^*, \lambda_{\theta;0}^p, \lambda_{\theta;0}^e, \lambda_{\theta;0}^T) \right] + \left[ L_\theta(\xi_{\theta;0}^*, \lambda_{\theta;0}^p, \lambda_{\theta;0}^e, \lambda_{\theta;0}^T) - L_\theta(\xi_{\theta;0}^*, \lambda_{\theta;0}^p, \lambda_{\theta;0}^e, \lambda_{\theta;0}^T) \right].
\]

The first term in the r.h.s. of (26) tends to zero, using the argument from (24), and the second by continuity of \( L_\theta \) guaranteed by (ii). We thus obtain that \( L_{\theta;N}(\xi_{\theta;0}^*, \lambda_{\theta;0}^p, \lambda_{\theta;0}^e, \lambda_{\theta;0}^T) \) tends to \( L_\theta(\xi_{\theta;0}^*, \lambda_{\theta;0}^p, \lambda_{\theta;0}^e, \lambda_{\theta;0}^T) \), which is a contradiction to (25).

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

Consider Eq. (10). Since \( \nabla_\theta \log P(\cdot) \) is bounded by Assumption 4.1 and \( \nabla_\theta f_1(\cdot; P_\theta) \) and \( \nabla_\theta g_c(\cdot; P_\theta) \) are bounded by Assumption 2.2 and using our previous result \( D \{ S_N, S \} \to 0 \), we have that for a.e. \( \omega_{SAA} \in \Omega_{SAA} \)
\[
\lim_{N \to \infty} \nabla_{\theta;N} \rho(Z) = \sum_{\omega \in \Omega} P_\theta(\omega)\xi_{\theta}^p(\omega) \nabla_\theta \log P(\omega)(Z(\omega) - \lambda_{\theta;0}^p) \]
\[
- \sum_{\ell \in \mathcal{E}} \lambda_{\theta;0}^e(\ell) \nabla_\theta g_c(\xi_{\theta}^p; P_\theta)
- \sum_{i \in \mathcal{I}} \lambda_{\theta;0}^T(i) \nabla_\theta f_1(\xi_{\theta}^p; P_\theta)
= \nabla_\theta \rho(Z),
\]
where the first equality is obtained from the Envelop theorem (see Theorem 4.2) with \((\xi_{\theta}, \lambda_{\theta}^p, \lambda_{\theta}^e, \lambda_{\theta}^T) \in S_N \cap S\) is the limit point of the converging sequence \(\{(\xi_{\theta;0}^*, \lambda_{\theta;0}^p, \lambda_{\theta;0}^e, \lambda_{\theta;0}^T)\}_{N\in\mathbb{N}}\).

D. Convergence Analysis of Empirical PRSVI

Lemma D.1 (Technical Lemma). Let \( P(\cdot|\cdot) \) and \( \tilde{P}(\cdot|\cdot) \) 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 \( \xi : \xi \tilde{P}(\cdot|x) \in \mathcal{U}(x, \tilde{P}(\cdot|x)) \),
\[
\sum_{x' \in \mathcal{X}} |\xi(x') - \tilde{\xi}(x')| \leq M_\xi \sum_{x' \in \mathcal{X}} \left| P(x'|x) - \tilde{P}(x'|x) \right|.
\]

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 \( \text{dom}(\mathcal{U}) \) and the graph of \( \mathcal{U}(x, P(\cdot|x)) \) is closed, by Theorem 2.7 in Rockafellar et al. (1998), \( \mathcal{U}(x, P(\cdot|x)) \) is a Lipschitz set-valued mapping with respect to the Hausdorff distance. Thus, for any \( \xi : \xi \mathcal{P}(\cdot|x) \in \mathcal{U}(x, \mathcal{P}(\cdot|x)) \), the following expression holds for some \( M_\xi > 0 \):
\[
\inf_{\xi \in \mathcal{U}(x, \tilde{P}(\cdot|x))} \sum_{x' \in \mathcal{X}} |\xi(x') - \tilde{\xi}(x')| \leq M_\xi \sum_{x' \in \mathcal{X}} \left| P(x'|x) - \tilde{P}(x'|x) \right|.
\]

Next, we want to show that the infimum of the left side is attained. Since the objective function is convex, and \( \mathcal{U}(x, \tilde{P}(\cdot|x)) \) is a convex compact set, there exists \( \tilde{\xi} : \tilde{\xi} \tilde{P}(\cdot|x) \in \mathcal{U}(x, \tilde{P}(\cdot|x)) \) such that infimum is attained.
Lemma D.2 (Strong Law of Large Number). Consider the sampling based PRSVI algorithm with update sequence \( \{ \hat{v}_k \} \). Then as both \( N \) and \( k \) tend to \( \infty \), \( \hat{v}_k \) converges with probability 1 to \( v^*_\theta \), the unique solution of projected risk sensitive fixed point equation \( \Pi T_{\hat{v}} \Phi v = \Phi v \).

**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.,

\[
\frac{1}{N} \sum_{i=0}^{N-1} 1_{\{ x_t = x \}} \to d_\theta(x|x_0), \text{ and } \hat{P}(x'|x,a) \to P(x'|x,a), \forall x, x' \in \mathcal{X}, a \in \mathcal{A}.
\]

Therefore with probability 1,

\[
\frac{1}{N} \sum_{i=0}^{N-1} \phi(x_t) \phi(x_t)^\top \to \sum_x d_\theta(x|x_0) \cdot \phi(x) \phi(x)^\top(x),
\]

\[
\frac{1}{N} \sum_{i=0}^{N-1} \phi(x_t) C_\theta(x_t) \to \sum_x d_\theta(x|x_0) \cdot \phi(x) C_\theta(x).
\]

Now we show that following expression holds with probability 1,

\[
\max_{\xi: \xi \in \mathcal{U}(x_t, P_{\theta,N}(x_t))} \sum_{x' \in \mathcal{X}} \xi(x') P_{\theta,N}(x'|x_t) v^\top(x') + \frac{1}{2N} (\xi(x') P_{\theta,N}(x'|x_t))^2 \to \max_{\xi: \xi \in \mathcal{U}(x_t, P_{\theta,N}(x_t))} \sum_{x' \in \mathcal{X}} \xi(x') P_{\theta,N}(x'|x_t) v^\top(x').
\] (27)

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

\[
\max_{\xi: \xi \in \mathcal{U}(x_t, P_{\theta,N}(x_t))} \sum_{x' \in \mathcal{X}} \xi(x') P_{\theta,N}(x'|x_t) v^\top(x') + \frac{1}{2N} (\xi(x') P_{\theta,N}(x'|x_t))^2 - \max_{\xi: \xi \in \mathcal{U}(x_t, P_{\theta,N}(x_t))} \sum_{x' \in \mathcal{X}} \xi(x') P_{\theta,N}(x'|x_t) v^\top(x') \leq \| \Phi v \|_\infty \left( M_{\xi^\ast_{\theta,x_t,N}} + \max_{x' \in \mathcal{A}} | \xi^\ast_{\theta,x_t,N}(x') | \right) \sum_{x' \in \mathcal{X}} | P_{\theta}(x'|x_t) - P_{\theta,N}(x'|x_t) | + \frac{1}{2N}.
\]

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

\[
\max_{\xi: \xi \in \mathcal{U}(x_t, P_{\theta,N}(x_t))} \sum_{x' \in \mathcal{X}} \xi(x') P_{\theta,N}(x'|x_t) v^\top(x') + \frac{1}{2N} (\xi(x') P_{\theta,N}(x'|x_t))^2 - \max_{\xi: \xi \in \mathcal{U}(x_t, P_{\theta,N}(x_t))} \sum_{x' \in \mathcal{X}} \xi(x') P_{\theta,N}(x'|x_t) v^\top(x') \leq \| \Phi v \|_\infty \max \left\{ \left( M_{\xi^\ast} + \max_{x' \in \mathcal{X}} | \xi^\ast(x') | \right), \left( M_{\xi^\ast_{\theta,x_t,N}} + \max_{x' \in \mathcal{A}} | \xi^\ast_{\theta,x_t,N}(x') | \right) \right\} \sum_{x' \in \mathcal{X}} | P_{\theta}(x'|x_t) - P_{\theta,N}(x'|x_t) | + \frac{1}{2N}.
\]

Therefore, the claim in expression (27) 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,

\[
\frac{1}{N} \sum_{i=0}^{N-1} \phi(x_t) \rho(\Phi v_t) \to d_\theta(x|x_0) \phi(x), \max_{\xi: \xi \in \mathcal{U}(x_t, P_{\theta,N}(x_t))} \sum_{x' \in \mathcal{X}} \xi(x') P_{\theta,N}(x'|x_t) v^\top_\theta(x')
\]
Combining the above arguments implies
\[
\frac{1}{N} \sum_{t=0}^{N-1} \phi(x_t) P_N(\Psi v_k) \rightarrow d_\theta(x|x_0) \phi(x) \to
\max_{\xi: \xi P_\theta(x) \in \mathcal{U}(x, P_\theta(x))} \sum_{x' \in \mathcal{X}} \xi(x') P_\theta(x'|x) v_\theta^\top(\phi(x')).
\]

As \( N \to \infty \), the above arguments imply that \( v_k - \hat{v}_k \to 0 \). On the other hand, Proposition 1 in [Tamar et al. 2014] implies that the projected risk sensitive Bellman operator \( P_\theta V_k \) is a contraction, it follows that from the analysis in Section 6.3 in [Bertsekas 2012] that the sequence \( \{\hat{v}_k\} \) generated by projected value iteration converges to the unique fixed point \( \Phi v_\theta^* \).

This in turns implies that the sequence \( \{\Phi v_k\} \) converges to \( \Phi v_\theta^* \).

\[ \square \]

**E. Proof of Theorem 5.4**

Similar to the proof of Theorem 4.2, recall the saddle point definition of \( (\xi^*_\theta, \lambda^*_\theta, P^*_\theta, \phi^*_{\theta,x}, \lambda^*_\theta, \lambda^*_\theta) \in S \) and strong duality result, i.e.,
\[
\max_{\xi: \xi P_\theta(x) \in \mathcal{U}(x, P_\theta(x))} \sum_{x' \in \mathcal{X}} \xi(x') P_\theta(x'|x) V_\theta(x') = \min_{\lambda \geq 0} \max_{\lambda^\top \geq 0} L_{\theta, x}(\xi, \lambda^\top, \lambda^\top, \lambda^\top) = \min_{\lambda^\top \geq 0} \max_{\xi \geq 0} L_{\theta, x}(\xi, \lambda^\top, \lambda^\top).
\]

the gradient formula in (20) can be written as
\[
\nabla_\theta V_\theta(x) = \nabla_\theta \left[ C_\theta(x) + \gamma \max_{\xi: \xi P_\theta(x) \in \mathcal{U}(x, P_\theta(x))} \mathbb{E}_\xi [V_\theta] \right] = \gamma \sum_{x' \in \mathcal{X}} \xi^*_\theta(x') P_\theta(x'|x) \nabla_\theta V_\theta(x') + \sum_{a \in A} \mu_\theta(a|x) \nabla_\theta \log \mu_\theta(a|x) h_\theta(x, a),
\]

where the stage-wise cost function \( h_\theta(x, a) \) is defined in (15). By defining \( \hat{h}_\theta(x) = \sum_{a \in A} \mu_\theta(a|x) \nabla_\theta \log \mu_\theta(a|x) h_\theta(x, a) \) and unfolding the recursion, the above expression implies
\[
\nabla_\theta V_\theta(x) = \hat{h}_\theta(x_0) + \gamma \sum_{x_1 \in \mathcal{X}} P_\theta(x_1|x_0) \xi^*_\theta(x_1) \left[ \hat{h}_\theta(x_1) + \gamma \sum_{x_2 \in \mathcal{X}} P_\theta(x_2|x_1) \xi^*_\theta(x_2) \nabla_\theta V_\theta(x_2) \right].
\]

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, \( \lim_{t \to \infty} \rho(\gamma^t V_\theta(x_1)) = 0 \), when \( x_0 = x \) the above expression implies the result of this theorem.

**F. Technical Results in Section 5.3**

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

where the solution of the above empirical problem is \( \xi^*_\theta(x) \) and the corresponding KKT multipliers are \((\lambda^*_\theta, P^*_\theta, \phi^*_{\theta,x}, \lambda^*_\theta, \lambda^*_\theta)\). Comparing to the optimization problem for \( \rho_N(\Phi v) \), i.e.,
\[
\rho_N(\Phi v) = \max_{\xi: \xi P_\theta(x) \in \mathcal{U}(x, P_\theta(x))} \sum_{x' \in \mathcal{X}} P_\theta(x'|x) \xi(x') \phi^\top(x') v + \frac{1}{2N} (\xi(x') P_\theta(x'|x))^2,
\]

where the solution of the above empirical problem is \( \xi^*_\theta(x) \) and the corresponding KKT multipliers are \((\lambda^*_\theta, P^*_\theta, \phi^*_{\theta,x}, \lambda^*_\theta, \lambda^*_\theta)\), the optimization problem in (28) can be viewed as having a skewed objective function of the problem in (29), within the deviation of magnitude \( \Delta + 1/2N \) where \( \Delta = \|v_\theta^* - V_\theta\|_\infty \). Before getting into the main analysis, we have the following observations.
Lemma F.2.

(i) Without loss of generality, we can also assume \((\xi^*_{\theta,x;N}, \lambda^*_{\theta,x;N}, \lambda^*_{\theta,x;N}, \lambda^*_{\theta,x;N})\) follows the strict complementary slackness condition. \(^5\)

(ii) Recall from Assumption 2.2 that the functions \(f_i(\xi, p)\) and \(g_i(\xi, p)\) are twice differentiable in \(\xi\) 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 (29) 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 \((\lambda^*_{\theta,x;N}, \lambda^*_{\theta,x;N}, \lambda^*_{\theta,x;N})\). Furthermore, define the Lagrangian function

\[
\hat{L}_{\theta,N}(\xi, \lambda^P, \lambda^E, \lambda^T) \doteq \sum_{x' \in \mathcal{X}} P_{\theta,N}(x'|x)\xi(x')\phi^\top(x')\nu + \frac{1}{2N}(\xi(x')P_{\theta,N}(x'|x)\xi(x') - 1)
\]

\[-\sum_{i \in I} \lambda^E(i) f_i(\xi, P_{\theta,N}(\cdot|x)) + \sum_{i \in I} \lambda^T(i) f_i(\xi, P_{\theta,N}(\cdot|x)).\]

One can easily conclude that \(\nabla^2 \hat{L}_{\theta,N}(\xi, \lambda^P, \lambda^E, \lambda^T) = -P_{\theta,N}(\cdot|x)^\top P_{\theta,N}(\cdot|x)/N - \sum_{i \in I} \lambda^T(i) \nabla^2 f_i(\xi, P_{\theta,N}(\cdot|x))\) such that for any vector \(\nu \neq 0\),

\[\nu^\top \nabla^2 \hat{L}_{\theta,N}(\xi^*_{\theta,x;N}, \lambda^*_{\theta,x;N}, \lambda^*_{\theta,x;N}, \lambda^*_{\theta,x;N}) \nu < 0,\]

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

Based on all the above analysis, we have the following sensitivity result from Corollary 3.2.4 in Fiacco [1983], derived based on Implicit Function Theorem.

**Proposition F.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 (28) and (29) are bounded as follows:

\[
\begin{bmatrix}
\xi^*_{\theta,x;N} \\
\lambda^P_{\theta,x;N} \\
\lambda^E_{\theta,x;N} \\
\lambda^T_{\theta,x;N}
\end{bmatrix}
= 
\begin{bmatrix}
\bar{\xi}_{\theta,x;N} \\
\bar{\lambda}^P_{\theta,x;N} \\
\bar{\lambda}^E_{\theta,x;N} \\
\bar{\lambda}^T_{\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 \((\bar{\lambda}^P_{\theta,x;N}, \bar{\lambda}^E_{\theta,x;N}, \bar{\lambda}^T_{\theta,x;N}) \to (\lambda^P_{\theta,x}, \lambda^E_{\theta,x}, \lambda^T_{\theta,x})\) 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 \(\hat{h}_{\theta,N}(x,a)\) and \(\gamma\)-visiting distribution \(\pi_N(x,a)\).

**Lemma F.2.** There exists a constant \(M_h > 0\) such that \(\max_{x \in \mathcal{X}, a \in \mathcal{A}} |h_{\theta}(x,a) - \lim_{N \to \infty} \hat{h}_{\theta,N}(x,a)| \leq M_h \Delta\).

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

\[
\left| \hat{h}_{\theta,N}(x,a) - h_{\theta}(x,a) \right| \leq M \sum_{i \in I} \left| \lambda^T_{\theta,x;N}(i) - \lambda^T_{\theta,x;N}(i) \right| + M \sum_{e \in E} \left| \lambda^E_{\theta,x;N}(e) - \lambda^E_{\theta,x;N}(e) \right| + \left\| \lambda^P_{\theta,x;N} - \lambda^P_{\theta,x;N} \right\| + \gamma \| V_0 \|_\infty |\xi^*_{\theta,x;N} - \xi^*_{\theta,x;N}| + 1 + \gamma \| V_0 \|_\infty \| \Phi_{\theta} \|_\infty
\]

\[+ \gamma \| V_0 \|_\infty \max \{ \| \xi^*_{\theta,x;N} \|_\infty, \| \xi^*_{\theta,x;N} \|_\infty \} \| P(\cdot|x,a) - P_N(\cdot|x,a) \|_1. \]

\(^5\)The existence of strict complementary slackness condition 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.
Note that at \( N \to \infty \), \( \| P(\cdot|a) - P_N(\cdot|a) \|_1 \to 0 \) 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 Propositions \[4.4, F.1\] the proof of this Lemma is completed by letting \( N \to \infty \) and defining

\[
M_h(x) = \max\{1, M, \frac{\gamma C_{\max}}{1-\gamma} \| \xi_{\theta,x}^* \|_1 \} + \gamma \Delta
\]

\[
\leq \left( \max\{1, M, \frac{\gamma C_{\max}}{1-\gamma} \| \Phi_{\theta,x}^{-1} \xi_{\theta,x} \|_1 + \gamma \} \right) \Delta.
\]

\[\Box\]

**Lemma F.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:

\[
\gamma \sum_{x'\in\mathcal{X}} d_{\rho_\theta}(x'|x) P_{\theta}^\xi(x|x') = d_{\rho_\theta}(x) - (1-\gamma)1\{x_0 = x\},
\]

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

\[
(I - \gamma P_{\theta}^\xi)^\top d_{\rho_\theta}(\cdot|x) = 1\{x_0 = x\}, \quad \forall x \in \mathcal{X}.
\]

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

\[
(I - \gamma P_{\theta;N}^\xi)^\top d_{\rho_{\theta;N}} = 1\{x_0 = z\} \in \mathcal{X}.
\]

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

\[
d_{\rho_\theta} - d_{\rho_{\theta;N}} - \gamma \left( (P_{\theta}^\xi)^\top d_{\rho_\theta} - (P_{\theta;N}^\xi)^\top d_{\rho_{\theta;N}} \right) = 0,
\]

which further implies

\[
(I - \gamma P_{\theta}^\xi)^\top \left( d_{\rho_\theta} - d_{\rho_{\theta;N}} \right) = \gamma \left( P_{\theta}^\xi - P_{\theta;N}^\xi \right)^\top d_{\rho_{\theta;N}}
\]

\[
\iff \quad \left( d_{\rho_\theta} - d_{\rho_{\theta;N}} \right) = (I - \gamma P_{\theta}^\xi)^\top \gamma \left( P_{\theta}^\xi - P_{\theta;N}^\xi \right)^\top d_{\rho_{\theta;N}}.
\]

Notice that with transition probability matrix \( P_{\theta}^\xi(\cdot|x) \), we have \( (I - \gamma P_{\theta}^\xi)^{-1} = \sum_{i=0}^{\infty} (\gamma P_{\theta}^\xi)^i < \infty \). The series is summable because by Perron-Frobenius theorem, the maximum eigenvalue of \( P_{\theta}^\xi \) is less than or equal to 1 and \( I - \gamma P_{\theta}^\xi \) is invertible. On the other hand, for every given \( x_0 \in \mathcal{X} \),

\[
\left\{ (P_{\theta}^\xi - P_{\theta;N}^\xi)^\top d_{\rho_{\theta;N}} \right\} (z') = \sum_{x \in \mathcal{X}} \sum_{k=0}^{\infty} \gamma^k (1-\gamma) \mathbb{E}_{P_{\theta;N}^\xi}(x_k = x|x_0) \left( P_{\theta}^\xi(z'|x_k) - P_{\theta;N}^\xi(z'|x_k) \right), \quad \forall z' \in \mathcal{X}
\]

\[
= \mathbb{E}_{P_{\theta;N}^\xi} \left( \sum_{k=0}^{\infty} \gamma^k (1-\gamma) \left( P_{\theta}^\xi(z'|x_k) - P_{\theta;N}^\xi(z'|x_k) \right) | x_0 \right), \quad \forall z' \in \mathcal{X}
\]

\[
\leq \mathbb{E}_{P_{\theta;N}^\xi} \left( \sum_{k=0}^{\infty} \gamma^k (1-\gamma) \left| P_{\theta}^\xi(z'|x_k) - P_{\theta;N}^\xi(z'|x_k) \right| | x_0 \right), \quad \forall z' \in \mathcal{X}
\]

\[
= Q(z'), \quad \forall z' \in \mathcal{X}.
\]
Note that every element in matrix \((I - \gamma P^\xi_\theta)^{-1} = \sum_{t=0}^{\infty} \left(\gamma P^\xi_\theta\right)^t\) is non-negative. This implies for any \(z \in \mathcal{X}\),

\[
\left\{d P^\xi_\theta - d P^\xi_{\theta,N}\right\}(z) = \left\{\left(I - \gamma P^\xi_\theta\right)^{-T} \gamma \left(P^\xi_\theta - P^\xi_{\theta,N}\right)^T d P^\xi_{\theta,N}\right\}(z),
\]

\[
\leq \left\{\left(I - \gamma P^\xi_\theta\right)^{-T} \gamma Q\right\}(z) = \left\{\left(I - \gamma P^\xi_\theta\right)^{-T} \gamma Q\right\}(z).
\]

The last equality is due to the fact that every element in vector \(Q\) is non-negative. Combining the above results with Proposition 4.4 and F.1 and noting that

\[
(I - \gamma P^\xi_\theta)^{-1}e = \sum_{t=0}^{\infty} \left(\gamma P^\xi_\theta\right)^t e = \frac{1}{1 - \gamma}e,
\]

we further have that

\[
\|\pi - \pi_N\|_1 = \|d P^\xi_\theta - d P^\xi_{\theta,N}\|_1
\]

\[
\leq e^T \left(I - \gamma P^\xi_\theta\right)^{-T} \gamma Q
\]

\[
= \frac{\gamma}{1 - \gamma}e^T Q
\]

\[
\leq \frac{\gamma}{1 - \gamma} \max_{x \in \mathcal{X}} \left\|P^\xi_\theta(\cdot|x) - P^\xi_{\theta,N}(\cdot|x)\right\|_1
\]

\[
\leq \frac{\gamma}{1 - \gamma} \max_{x \in \mathcal{X}} \left\{||\xi_\theta(\cdot) - \xi_{\theta,N}(\cdot)||_1 + \|P_\theta(\cdot|x)\|_1\right. + \max\{\|\xi_{\theta,N}(\cdot)||_1, \|\xi_{\theta,N}(\cdot)||_1\} \|P(\cdot|x,a) - P_N(\cdot|x,a)\|_1\}.
\]

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 \(\|\xi_\theta(\cdot) - \xi_{\theta,N}(\cdot)||_1 \to 0\). We thus set the constant \(M_\pi\) as \(\gamma\|\Phi^{-1}_\theta \Psi_{\theta,x}\|_1/(1 - \gamma)\).

\[\square\]

G. Experiments

We empirically evaluate our algorithms on the American put option domain: a standard test-bed for risk-sensitive RL. [Li et al. 2009, Tamar et al. 2014, Chow & Ghavamzadeh 2014] that uses RL algorithms to maximize the revenue of executing an American option. In our setting, the state is continuous and represents the price of some stock. It evolves according to a geometric Brownian motion (GBM), i.e., \(x_{t+1}/x_t \sim N(\mu_t - \sigma_t^2/2, \sigma_t^2)\), where \(N\) is the log-normal distribution and \(\mu_t\) and \(\sigma_t\) are parameters. The action at each time \(t\) is binary. Action 0 stands for hold and action 1 stands for execute. An execution action generates reward \(R(x_t, 1) = \max\{0, K - x_t\}\), where \(K\) is fixed and known as the strike price, and terminates the episode: A hold action generates zero reward, i.e., \(R(x_t, 0) = 0\), and the price transitions to \(x_{t+1}\) as described above. Unless an execution occurred, the episode ends after \(T\) steps, with reward \(R(x_T, a_T) = \max\{0, K - x_T\}\), for any action \(a_T\). For the expected return, the optimal policy is a time dependent threshold policy that holds if \(x_t > \theta_t\) [Hull 2006], where \(\theta_t\) is the threshold, and executes otherwise. Accordingly, we search in the space of soft-threshold policies of the form

\[
\mu_\theta(a_t = 0|x_t) = \frac{1}{1 + \exp(-\beta(x_t - \theta_t))},
\]

for some softness parameter \(\beta > 0\). In this domain, it is common to introduce a discount factor to account for the risk-free interest rate that could have been earned had the option been executed (see, e.g., Li et al. 2009). Here, we chose a discount factor \(\gamma = 0.98\), i.e., the revenue random variable in the optimization problems is \(\mathcal{R}(x_0) = \sum_{t=0}^T \gamma^t R(x_t, a_t)\).

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

For the case of static optimization, we trained policies to optimize the following objective function:

\[
\rho\left(R(x_0, a_0) + \gamma R(x_1, a_1) + \ldots + \gamma^T R(x_T, a_T)\right),
\]

with
Policy Gradient for Coherent Risk Measures

1. Expectation, i.e., \( \rho(Z) = \mathbb{E}(Z) \),
2. CVaR(\( \alpha = 0.5 \)), i.e., \( \rho(Z) = \text{CVaR}_{0.5}(Z) \),
3. CVaR(\( \alpha = 0.7 \)), i.e., \( \rho(Z) = \text{CVaR}_{0.7}(Z) \),
4. 0.9 \cdot \text{Expectation} + 0.1 \cdot \text{CVaR}(\( \alpha = 0.5 \)), i.e., \( \rho(Z) = 0.9 \cdot \mathbb{E}(Z) + 0.1 \cdot \text{CVaR}_{0.5}(Z) \),
5. \( \text{Expectation} - 0.1 \cdot \text{Semi-deviation} \), i.e., \( \rho(Z) = \mathbb{E}(Z) - 0.1 \cdot \text{SD}(Z) \).

Using stochastic gradient descent. The gradient for the expectation was calculated using standard (episodic) policy gradient (Marbach & Tsitsiklis, 1998), the gradient for CVaR was calculated according to the GCVaR algorithm of Tamar et al. (2015), and the gradient for mean-semideviation was calculated using our GMSD algorithm of Section A. In Figure 1 we plot the histograms of the payoff of the different policies. The risk-averse nature of the policies trained with a risk-sensitive objective may be observed. In our experiments we set \( K = 1 \) and \( x_0 = 0.5 \). The numerical results of this experiment are reported in Table 1. Notice that when we set the objective function to CVaR(\( \alpha = 0.5 \)), the RL algorithm chooses a risk-averse policy and becomes very conservative. In order to avoid the potential loss due to market fluctuations, the induced policy executes the option almost immediately\(^6\) and receives a revenue of 0.5. Correspondingly, both objective functions CVaR(\( \alpha = 0.7 \)), 0.9 \cdot \text{Expectation} + 0.1 \cdot \text{CVaR}(\( \alpha = 0.5 \)), and Expectation – 0.1 \cdot \text{Semi-deviation} return less risk-averse policies, but still the revenues have lower return variability compared to the standard expectation counterpart.

For the case of dynamic optimization, we trained policies to optimize the following objective function:

\[
\rho \left( R(x_0, a_0) + \rho(\gamma R(x_1, a_1) + \ldots + \rho(\gamma^T R(x_T, a_T)) \ldots) \right),
\]

with

1. Expectation, i.e., \( \rho(Z) = \mathbb{E}(Z) \),
2. CVaR(\( \alpha = 0.5 \)), i.e., \( \rho(Z) = \text{CVaR}_{0.5}(Z) \),
3. 0.1 \cdot \text{Expectation} + 0.9 \cdot \text{CVaR}(\( \alpha = 0.7 \)), i.e., \( \rho(Z) = \mathbb{E}(Z) + 0.9 \cdot \text{CVaR}_{0.7}(Z) \),
4. 0.5 \cdot \text{Expectation} + 0.5 \cdot \text{CVaR}(\( \alpha = 0.7 \)), i.e., \( \rho(Z) = 0.5 \cdot \mathbb{E}(Z) + 0.5 \cdot \text{CVaR}_{0.7}(Z) \),
5. 0.9 \cdot \text{Expectation} + 0.1 \cdot \text{CVaR}(\( \alpha = 0.5 \)), i.e., \( \rho(Z) = 0.9 \cdot \mathbb{E}(Z) + 0.1 \cdot \text{CVaR}_{0.5}(Z) \),

using stochastic gradient descent, with the gradient calculated according to the algorithm of Section 5. We used RBF features for estimating the value function. Numerical results of this experiment are reported in Table 2 and the histograms of payoffs with respect to different objective functions are shown in Figure 2. The dynamic CVaR(\( \alpha = 0.5 \)) was very conservative, and chose to execute immediately. This also occurred with the case of 0.1 \cdot \text{Expectation} + 0.9 \cdot \text{CVaR}(\( \alpha = 0.7 \)). Compared to the static risk case, one can see that optimization with dynamic risk measure is more risk-averse. This is potentially due to the fact that for dynamic conditional value-at-risk, the tail risk compounds over time, for which the total tail risk CVaR(\( \alpha \)) \circ \ldots \circ CVaR(\( \alpha \)) becomes far more conservative than the static counterpart CVaR(\( \alpha \)). Practically,

| \( \text{Expectation} \) | \( \mathbb{E}(R(x_0)) \) | \( \sigma(R(x_0)) \) | \( \text{CVaR}_{0.5}(R(x_0)) \) | \( \text{CVaR}_{0.7}(R(x_0)) \) | \( \text{SD}(R(x_0)) \) |
|---|---|---|---|---|---|
| Expectation | 0.5548 | 0.2067 | 0.4061 | 0.4739 | 0.1746 |
| CVaR(\( \alpha = 0.5 \)) | 0.5000 | 0.0027 | 0.5000 | 0.5000 | 0.0010 |
| CVaR(\( \alpha = 0.7 \)) | 0.5004 | 0.1729 | 0.4304 | 0.4867 | 0.1601 |
| 0.9 \cdot \text{Expectation} + 0.1 \cdot \text{CVaR}(\( \alpha = 0.5 \)) | 0.5530 | 0.1931 | 0.4167 | 0.4804 | 0.1625 |
| Expectation – 0.1 \cdot \text{Semi-deviation} | 0.5551 | 0.1894 | 0.4208 | 0.4827 | 0.1597 |

Table 1. Performance comparison for the policies learned by different static risk-sensitive objective functions.

---

\(^6\)Out of 10000 Monte Carlo trials, there are only 7 trials with total reward not equal to 0.5. This is due to the system randomness of the Gaussian stock price model. In high precision, the corresponding results are \( \mathbb{E}(R(x_0)) = 0.500015840254391 \), \( \sigma(R(x_0)) = 0.002744831192968 \), \( \text{CVaR}_{0.5}(R(x_0)) = 0.499962977914256 \), \( \text{CVaR}_{0.7}(R(x_0)) = 0.499973555653040 \) and \( \text{SD}(R(x_0)) = 0.001028512000913 \).
one way to reduce conservatism in dynamic risk, while still maintaining control of reward variability, is to control the convex combination weightings between the combination of expectation and CVaR, as shown in the second plot of Figure 2. Unfortunately, the question of choosing an optimized pair of weighting and CVaR threshold is still quite challenging and is deferred for future research.

| Combination | $\mathbb{E}(R(x_0))$ | $\sigma(R(x_0))$ | $\text{CVaR}_{0.5}(R(x_0))$ | $\text{CVaR}_{0.7}(R(x_0))$ |
|-------------|----------------------|------------------|-----------------------------|-----------------------------|
| Expectation | 0.5437               | 0.1752           | 0.04157                     | 0.04611                     |
| CVaR($\alpha = 0.5$) | 0.5003 | 0.0165 | 0.4988 | 0.4992 |
| 0.1 · Expectation + 0.9 · CVaR($\alpha = 0.7$) | 0.5041 | 0.0538 | 0.4906 | 0.4933 |
| 0.5 · Expectation + 0.5 · CVaR($\alpha = 0.7$) | 0.5350 | 0.1440 | 0.4327 | 0.4735 |
| 0.9 · Expectation + 0.1 · CVaR($\alpha = 0.5$) | 0.5355 | 0.1571 | 0.4241 | 0.4712 |

Table 2. Performance comparison for the policies learned by different dynamic risk-sensitive objective functions.