A CONVEX ANALYTIC APPROACH TO RISK-AWARE MARKOV DECISION PROCESSES *

WILLIAM B. HASKELL AND RAHUL JAIN †

Abstract. In classical Markov decision process (MDP) theory, we search for a policy that say, minimizes the expected infinite horizon discounted cost. Expectation is of course, a risk neutral measure, which does not suffice in many applications, particularly in finance. We replace the expectation with a general risk functional, and call such models risk-aware MDP models. We consider minimization of such risk functionals in two cases, the expected utility framework, and Conditional Value-at-Risk, a popular coherent risk measure. Later, we consider risk-aware MDPs wherein the risk is expressed in the constraints. This includes stochastic dominance constraints, and the classical chance-constrained optimization problems. In each case, we develop a convex analytic approach to solve such risk-aware MDPs. In most cases, we show that the problem can be formulated as an infinite dimensional linear program in occupation measures when we augment the state space. We provide a discretization method and finite approximations for solving the resulting LPs. A striking result is that the chance-constrained MDP problem can be posed as a linear program via the convex analytic method.

Keywords: Markov decision processes, Stochastic optimization, risk measures, Conditional value-at-risk, Stochastic dominance constraints, Convex analytic approach.

AMS Subject Classification: 90C40, 91B30, 90C34, 49N15.

1. Introduction.

Consider a Markov decision process on a state space $\mathbb{S}$, action space $\mathbb{A}$, cost function $c(s,a)$, discount factor $\gamma \in (0,1)$ and a transition kernel Q. Typically, we want to find an optimal policy π that solves $\inf_\pi \mathbb{E}^\pi \left(\sum_{t=0}^\infty \gamma^t c(s_t,a_t)\right)$. Of course, as we well know this problem can be solved by dynamic programming. The key is that one can show that there exists an optimal policy that is stationary and Markovian.

However, dynamic programming is not the only method to solve the problem. Alternatives include posing the optimization problem as a linear program as in the 'convex analytic approach'. The convex analytic approach is developed for finite state and action spaces in [6, 13, 25, 31], and for Borel state and action spaces in [22, 23, 34]. In this approach, we introduce an occupation measure $\mu_\pi(s,a)$, which for a fixed stationary policy π can be interpreted as the discounted empirical frequency of visiting state s and taking action a. Thus, we can pose a linear program

$$\inf_\mu \left\{ \sum_{(s,a)\in\mathbb{S}\times\mathbb{A}} \mu(s,a)c(s,a) \text{ s.t. linear constraints} \right\}$$

whose solution gives the optimal occupation measure from which the optimal policy can be derived.

Often, however, in multi-stage decision-making, the risk-neutral objective is not enough. A decision maker may be risk-averse and may want to explicitly model his risk aversion. Thus, we are faced with a different optimization problem:

$$\inf_\pi \rho\left(\sum_{t=0}^\infty \gamma^t c(s_t,a_t)\right), \tag{1.1}$$

*This research is supported by the Office of Naval Research Young Investigator Award #N000141210766 and the National Science Foundation CAREER Award #0954116.

†William Haskell is an assistant professor in the ISE department at National University of Singapore. Rahul Jain is an associate professor and the Kenneth C. Dahlberg Early Career Chair in the Departments of EE, CS and ISE at the University of Southern California.

where ρ is a coherent risk measure such as conditional value-at-risk (CVaR). Problem (1.1) presents significant challenges since it need not be convex (depending on the risk measure) and the optimal policy will in general depend on history (rather than just the current state). It follows that the 'principle of optimality' used in writing down the dynamic programming equations does not hold, and thus Problem (1.1) cannot be solved via the dynamic programming method.

Problem (1.1) is not artificial or abstract. There is a tremendous interest in taking risk into account in sequential decision-making in areas like finance and insurance, operations management, smart-grid power networks, and even robotics.

In finance, for example, the portfolio optimization is really a sequential decision-making problem. Furthermore, investors are rarely risk-neutral. The mean-variance approach of Markowitz [32, 42] has been popular for risk-averse investors. Yet, the financial crisis of 2008 has raised the need to consider other direct measures of risk. The problem of single-stage portfolio optimization with such risk measures (the conditional value-at-risk) was first addressed in [37]. How to do dynamic portfolio optimization with such risk measures has been an open problem. Our convex analytic framework provides a solution.

As mentioned above, Problem (1.1) does not admit an optimal stationary policy, nor a dynamic programming solution. On the other hand, introducing an occupation measure via the convex analytic method ([6, 7]), leads to static optimization problems in a space of occupation measures. We develop this approach for Problem (1.1) and other related problems. This opens up a variety of new modeling options for decision makers seeking risk-aware policies.

We make the following main contributions in this paper. We treat four different risk-aware sequential stochastic optimization formulations. Of these, the expected utility formulation, the stochastic dominance and chance-constrained formulations lead to linear programming problems. We also give a treatment for optimization of risk functionals such as conditional value-at-risk (CVaR), and briefly discuss mean deviation and mean semi-deviation risk measures. These lead to non-convex optimization problems, albeit over convex feasible regions. In this case, we give a sequence of linear programming approximations that asymptotically yield an optimal solution. Thus, in each case the convex analytic methodology that we use yields a tractable solution. This is based on a state space augmentation approach akin to that used in [4, 5] and finite approximation methods for infinite-dimensional linear programs adapted from [21, 29, 30]. This, when combined with a discretization scheme, allows us to provide a convergence rate.

A striking observation we make is that, unlike the single-stage chance-constrained optimization problem which is generally non-convex and very difficult to solve [35], the sequential chance-constrained optimization problem can actually be reformulated as a linear programming problem via the convex analytic approach.

Throughout this paper, we encounter a dichotomy between sequential and single-stage risk management. In the static setting, we optimize over a *random variable* while in the sequential setting, we have to optimize over a *measure*. This leads to serious technical difficulties in the sequential setting as compared to the static setting.

Related Literature. There is a substantial body of work on risk measures in the static setting, [17, 28, 39]. Risk measures have also been considered in [26, 27], where the expected utility of countable state finite action MDPs is minimized. In [15], variance penalties for MDPs are considered. In [41], the mean-variance trade-off in MDPs is further explored. [14] shows how to solve the Bellman equation for

risk-sensitive control of MDPs. In [5], a finite horizon MDP with a conditional value-at-risk constraint on total cost is considered. Both an offline and online stochastic approximation algorithm are developed.

The most closely related works to this paper are the following. In [38], the class of Markov risk measures is proposed. This class of risk measures leads to tractable dynamic programming formulations. However, we note that most common risk measures, e.g., CVaR are not Markov. [3] shows how to minimize the average value-at-risk of costs in MDPs. [18] applies stochastic dominance constraints to the long-run average and infinite horizon discounted reward distributions in MDPs. [43] minimizes the conditional value-at-risk of discounted total cost and provides a static nonlinear programming formulation. Existence of a solution to such a problem is established though no tractable method is given for its solution. [4] considers minimization of the certainty equivalent of MDP costs, where the classical risk-sensitive MDP is a special case. In [9], numerical methods are developed for risk-aware MDPs with one-stop conditional risk measures. Specifically, value and policy iteration algorithms are developed and their convergence is established. Our work is distinct from the above in considering a range of risk-aware sequential optimization formulations, using a convex analytic approach, and in most cases yielding a linear programming problem, or at least approximation via a sequence of linear programs.

2. Preliminaries. This section reviews standard notation for MDPs and then discusses risk-aware MDPs.

2.1. Discounted MDPs. Consider a discrete time MDP given by the 5-tuple $(\mathbb{S},\, \mathbb{A},\, \{A(s) : s \in \mathbb{S}\},\, Q,\, c)$. The state space $\mathbb{S}$ and the action space $\mathbb{A}$ are Borel spaces, measurable subsets of complete and separable metric spaces, with corresponding Borel $\sigma-$algebras $\mathcal{B}(\mathbb{S})$ and $\mathcal{B}(\mathbb{A})$. We use $\mathcal{P}(\mathbb{S})$ to denote the space of probability measures over $\mathbb{S}$ with respect to $\mathcal{B}(\mathbb{S})$, and we define $\mathcal{P}(\mathbb{A})$ analogously. For each state $s \in \mathbb{S}$, the set $A(s) \subset \mathbb{A}$ is a measurable set in $\mathcal{B}(\mathbb{A})$ and it indicates the set of feasible actions available in state s. We assume that the multifunction $s \to A(s)$ permits a measurable map $\phi : S \to A$ such that $\phi(s) \in A(s)$.

The set of feasible state-action pairs is given by $\mathbb{K} = \{(s,a) \in \mathbb{S} \times \mathbb{A} : a \in A(s)\}$, with corresponding Borel $\sigma-$algebra $\mathcal{B}(\mathbb{K})$. The transition law Q governs the system evolution. For $B \in \mathcal{B}(\mathbb{S})$, $Q(B \mid s,\, a)$ is the probability of next visiting the set B given the current state-action pair (s,a). Finally, $c : \mathbb{K} \to \mathbb{R}$ is a measurable cost function that depends on state-action pairs. We will emphasize cost throughout (rather than reward) because risk functions are typically defined for losses.

We now describe two classes of policies for MDPs. Let $\mathcal{H}_t$ be the set of *histories* at time t, $\mathcal{H}_0 = \mathbb{S}$, $\mathcal{H}_1 = \mathbb{K} \times \mathbb{S}$, and $\mathcal{H}_t = \mathbb{K}^t \times \mathbb{S}$ for all $t \geq 2$. A specific history $h_t \in \mathcal{H}_t$ records the state-action pairs visited at times $0, 1, \ldots, t-1$ as well as the current state s_t. Define Π to be the set of all *history-dependent randomized policies*: collections of mappings $\pi = (\pi_t)_{t \geq 0}$ where $\pi_t : \mathcal{H}_t \to \mathcal{P}(\mathbb{A})$ for all $t \geq 0$. Given a history $h_t \in \mathcal{H}_t$ and a set $B \in \mathcal{B}(\mathbb{A})$, $\pi_t(B \mid h_t)$ is the probability of selecting an action in B. Define Π_0 to be the class of *stationary randomized Markov policies*: mappings $\pi : \mathbb{S} \to \mathcal{P}(\mathbb{A})$ which only depend on history through the current state. For a policy $\pi \in \Pi_0$, a given state $s \in \mathbb{S}$, and a set $B \in \mathcal{B}(\mathbb{A})$, $\pi(B \mid s)$ is the probability of choosing an action in B. The class Π_0 is a subset of Π and we explicitly assume that Π and Π_0 only include feasible policies that respect the constraints $\mathbb{K}$.

The canonical measurable space of MDP trajectories is $(\Omega,\, \mathcal{B}) = (\mathbb{K}^\infty,\, \mathcal{B}(\mathbb{K})^\infty)$, and specific trajectories are written as $\omega \in \Omega$. The state and action at time t are denoted s_t and a_t, respectively. Formally, for a trajectory $\omega \in \Omega$, $s_t(\omega)$ and $a_t(\omega)$

are the state and action at time t along this trajectory. With respect to an initial distribution $\nu \in \mathcal{P}(\mathbb{S})$, any policy $\pi \in \Pi$ determines a probability measure P_ν^π on $(\Omega, \mathcal{B})$ and a corresponding stochastic process $\{(s_t, a_t)\}_{t\geq 0}$. The resulting probability space is $(\Omega, \mathcal{B}, P_\nu^\pi)$. The expectation operator with respect to P_ν^π is denoted $\mathbb{E}_\nu^\pi[\cdot]$.

For discount factor $\gamma \in (0,1)$, consider: $C(\pi,\nu) = \mathbb{E}_\nu^\pi[\sum_{t=0}^\infty \gamma^t c(s_t,a_t)]$. We assume that costs c are bounded above and below throughout this paper, which is one way of ensuring that $C(\pi,\nu)$ is well-defined.

Assumption 2.1. *There exist $\underline{c}$ and $\bar{c}$ such that $0 < \underline{c} \leq c(s,a) \leq \bar{c} < \infty$ for all $(s,a) \in \mathbb{K}$.*

This assumption streamlines our presentation and is reasonable in practice. For example, in a real newsvendor problem there are limits on how much we can order in any one period, and thus the maximum possible order cost is bounded in each period. Under Assumption 2.1, the inequalities $0 < \underline{c}/(1-\gamma) \leq \sum_{t=0}^\infty \gamma^t c(s_t,a_t) \leq \bar{c}/(1-\gamma) < \infty$ hold for all trajectories $\omega \in \Omega$. We use $\mathcal{Y} := [0, \bar{c}/(1-\gamma)]$ to denote the interval in which the running costs $\sum_{t=0}^T \gamma^t c(s_t,a_t)$ lie for all finite horizons T. This interval will appear quite often.

The classical *infinite horizon discounted cost minimization problem* is

$$\inf_{\pi\in\Pi} C(\pi,\nu). \tag{2.1}$$

It is well known that a stationary policy in Π_0 is optimal for Problem (2.1) under suitable conditions (for example, this result is found in [36] for finite and countable state spaces, and in [22, 23] for general Borel state and action spaces).

2.2. Risk-aware MDPs. Problem (2.1) has a risk-neutral objective but sometimes decision makers may be risk-averse or risk-seeking. We now introduce risk-aware extensions of Problem (2.1). It is convenient to introduce a fixed reference probability space, since the underlying probability space $(\Omega, \mathcal{B}, P_\nu^\pi)$ for MDP trajectories is changing as π varies. Consider the probability space $([0,1], \mathcal{B}([0,1]), \mathbb{P})$ with primitive uncertainties denoted by $\xi \in [0,1]$, where $\mathcal{B}([0,1])$ is the Borel $\sigma-$algebra on $[0,1]$ and $\mathbb{P}$ is the uniform distribution on $[0,1]$. We define $\mathcal{L} = \mathcal{L}_\infty([0,1], \mathcal{B}([0,1]), \mathbb{P})$ to be the space of essentially bounded random variables on $([0,1], \mathcal{B}([0,1]), \mathbb{P})$. Random variables defined on $[0,1]$ with support in $\mathcal{Y}$, such as the infinite horizon discounted cost $\sum_{t=0}^\infty \gamma^t c(s_t,a_t)$, are included in $\mathcal{L}$. Recall that two random variables X and Y are equal in distribution, written $X =_d Y$, if $\Pr\{X \leq \eta\} = \Pr\{Y \leq \eta\}$ for all $\eta \in \mathbb{R}$. Let $C_\nu^\pi \in \mathcal{L}$ be defined by

$$\Pr\{C_\nu^\pi \leq \eta\} = P_\nu^\pi\left\{\sum_{t=0}^\infty \gamma^t c(s_t,a_t) \leq \eta\right\}, \forall \eta \in \mathbb{R},$$

i.e., C_ν^π is a random variable that is equal in distribution to $\sum_{t=0}^\infty \gamma^t c(s_t,a_t)$ on $\mathbb{R}$ when the underlying probability distribution on trajectories is P_ν^π. All C_ν^π have support contained in the interval $\mathcal{Y}$.

Risk functions ρ will be mappings $\rho : \mathcal{L} \to \mathbb{R}$. Now, consider a fixed risk function, a mapping $\rho : \mathcal{L} \to \mathbb{R}$ which associates a scalar with random variables in $\mathcal{L}$. We will make the following assumption about the risk measures.

Assumption 2.2. *Risk measures are law invariant, i.e., $\rho(X) = \rho(Y), \forall X =_d Y$.*

Most common risk measures, and all of the risk measures we consider in this paper, are law invariant. By assuming law invariance, we are simply saying that we only care about the distribution of costs on $\mathbb{R}$. We do not care about properties of the underlying probability space on which these random variables are defined.

Under this assumption the random variables $\{C_\nu^\pi\}_{\pi\in\Pi}$ can be used in place of the measurable mapping $\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)$, since a law invariant risk function ρ will not distinguish between them because they have the same distribution on $\mathbb{R}$ by definition. A natural risk-aware extension of Problem (2.1) is then

$$\inf_{\pi\in\Pi}\rho\left(C_\nu^\pi\right). \tag{2.2}$$

3. State space augmentation. This section describes our general approach for solving Problem (2.2) based on state space augmentation and convex analytic methods. Note that history dependence shows up in Problem (2.2) - it is to be expected that some information about the history h_t will have to be appended to the current state s_t at all times to get a near-optimal risk-aware policy. The idea behind state space augmentation is to partially or totally capture the history (see [3, 4, 5]), so that this information is available for decision-making.

In this section, we first augment the state space to keep track of the running cost over the entire time horizon. Then, we argue that the infinite horizon problem can be approximated with arbitrary precision by a finite horizon problem. Furthermore, the finite horizon problem can be solved exactly with convex analytic methods. In particular, we will estimate C_ν^π for a given π with another random variable that is distributed according to an occupation measure. With this method we can, in principle, approximate any instance of Problem (2.2) with a static optimization problem.

In [4], $\mathbb{S}$ is augmented with two new state variables to keep track of the running cost and the discounting. We borrow the augmented state space from [4] and denote it as $\tilde{\mathbb{S}}:=\mathbb{S}\times\mathcal{Y}\times(0,1]$, where the first component is the state in the original MDP, the second component keeps track of the running cost (recall $\mathcal{Y}=[0,\,\bar{c}/(1-\gamma)]$ is the support of the running cost $y_t=\sum_{i=0}^{t}\gamma^i c(s_i,a_i)$ for all $t\geq 0$), and the third component adjusts for the discounting. A state in $\tilde{\mathbb{S}}$ looks like (s_t,y_t,z_t) where we continue to use s_t to denote the original state at time t, y_t is the running cost at the beginning of time t, and z_t is the discount factor for time t. The set of feasible actions in state $(s,y,z)\in\tilde{\mathbb{S}}$ only depends on s, so we define the augmented set of feasible state-action pairs to be $\tilde{\mathbb{K}}:=\left\{(s,y,z,a)\in\tilde{\mathbb{S}}\times\mathbb{A}:\,a\in A(s)\right\}$, along with its Borel $\sigma-$algebra $\mathcal{B}\left(\tilde{\mathbb{K}}\right)$, and we assume $\tilde{\mathbb{K}}$ is closed in $\tilde{\mathbb{S}}\times\mathbb{A}$. Occupation measures for the solution of Problem (2.2) will be defined on $\tilde{\mathbb{K}}$.

The evolution of the augmented state $\{(s_t,y_t,z_t)\}_{t\geq 0}$ is as follows: $\{s_t\}_{t\geq 0}$ still evolves as per the original transition kernel Q: $s_{t+1}\sim Q(\cdot\,|\,s_t,a_t)$, $\forall t\geq 0$, and its evolution does not depend on the running costs or the discounting. The running costs $\{y_t\}_{t\geq 0}$ evolve deterministically according to $y_{t+1}=z_t c(s_t,a_t)+y_t$, for all $t\geq 0$. The discount factors $\{z_t\}_{t\geq 0}$ also evolve deterministically according to $z_{t+1}=\gamma\,z_t$, for all $t\geq 0$. The state variable $\{z_t\}_{t\geq 0}$ is just a geometric series in place to make sure the running costs are updated correctly. We initialize $y_0=0$ since no costs have been assessed before time $t=0$. Also, we initialize $z_0=1$ so that the costs $c(s_0,\,a_0)$ at time $t=0$ are not discounted. Note then that $y_t=\sum_{i=0}^{t-1}\gamma^i c(s_i,a_i)$ for all $t\geq 1$. We emphasize that the augmented state variables y_t and z_t are completely determined by the history $h_t\in\mathcal{H}_t$. We let $\tilde{Q}$ be the transition kernel for the augmented state $\{(s_t,y_t,z_t)\}_{t\geq 0}$, and defined by

$$\tilde{Q}\left(B,\,z\,c(s,a)+y,\,\gamma\,z\,|\,s,y,z,a\right)=Q\left(B\,|\,s,a\right),\ \forall B\in\mathcal{B}(\mathbb{S}),\ \forall\,(s,y,z,a)\in\tilde{\mathbb{K}}.$$

We emphasize again that the augmented state variables (y_{t+1},z_{t+1}) are deterministic functions of (s_t,y_t,z_t,a_t) for all $t\geq 0$.

Now we describe a new class of policies for use in tandem with the augmented state space. Let Π_1 be the class of *augmented stationary randomized Markov policies*: mappings $\pi : \tilde{\mathbb{S}} \to \mathcal{P}(\mathbb{A})$ which depend on history only through the current augmented state (s, y, z). Policies $\pi \in \Pi_1$ are allowed to use the running cost and the discount level to make decisions. Because the running cost and discount factor are functions of the history h_t, we consider Π_1 to be a subset of the set of all policies Π and we see that $\Pi_0 \subset \Pi_1 \subset \Pi$, where we view the set of stationary Markov policies on the original state space Π_0 as a subset of Π_1.

On a trajectory $\omega \in \Omega$, $y_t(\omega)$ and $z_t(\omega)$ are the running cost and discount factor at time t. Our earlier initial state distribution ν used to define Problem (2.1) can be extended to an initial state distribution on the augmented state space. We denote this initial distribution on the augmented state space by ν as well, since the initial conditions on y_0 and z_0 are deterministic, i.e. $\nu(\mathbb{S} \times \{0\} \times \{1\}) = 1$, corresponding to $y_0 = 0$ and $z_0 = 1$. Along with the (augmented) initial distribution ν, a policy $\pi \in \Pi$ gives a probability measure P_ν^π on $(\Omega, \mathcal{B})$ that determines a corresponding stochastic process $\{(s_t, y_t, z_t, a_t)\}_{t \geq 0}$ on the augmented set of state-action pairs.

Now that we have the augmented state space in place, we will discuss a general method for approximating Problem (2.2). Our approximation scheme is based on the intuition, confirmed in the following lemma, that the running cost y_t is a good estimate of the infinite horizon cost $\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)$ on every trajectory $\omega \in \Omega$ for large enough t. The following result uses boundedness of costs.

Lemma 3.1. *For any $\epsilon > 0$, there is a $T = T(\epsilon)$ such that*

$$\| y_t(\omega) - \sum_{t=0}^{\infty} \gamma^t c(s_t(\omega), a_t(\omega)) \| < \epsilon, \, \forall \omega \in \Omega,$$

for all $t \geq T$.

Lemma 3.1 justifies our interest in the running cost $y_T = \sum_{t=0}^{T} \gamma^t c(s_t, a_t)$ at time T for large enough T. We now compare the risk of the finite horizon cost y_T versus the infinite horizon cost $\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)$ in terms of ρ. In the next assumption, we will abuse notation and write $\rho(y_T)$ and $\rho(\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t))$ to denote the risk function ρ from Problem (2.2) evaluated at y_T and $\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)$ when the underlying probability space is $(\Omega, \mathcal{B}, P_\nu^\pi)$, i.e., the underlying policy is implicit.

Assumption 3.2. *For any $\epsilon > 0$, there is a T such that*

$$|\rho(y_T) - \rho\left(\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)\right)| \leq \epsilon, \quad \forall \pi \in \Pi.$$

Assumption 3.2 amounts to uniform continuity of ρ across policies in the running cost. It means that when T is large and y_T is close to $\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)$ almost surely across policies $\pi \in \Pi$, then the risk $\rho(y_T)$ is close to $\rho(\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t))$ across policies $\pi \in \Pi$. The key in Assumption 3.2 is that the error guarantee does not depend on the policy $\pi \in \Pi$. We will show that it is easy to guarantee Assumption 3.2 holds for the specific risk-aware MDPs that we consider in Section 5.

Under Assumption 3.2, we can justify working with a truncation $\sum_{t=0}^{T} \gamma^t c(s_t, a_t)$ of the infinite horizon discounted cost. In analogy with C_ν^π, we define $C_{\nu,T}^\pi \in \mathcal{L}$ to satisfy $\Pr\{C_{\nu,T}^\pi \leq \eta\} = P_\nu^\pi\left\{\sum_{t=0}^{T} \gamma^t c(s_t, a_t) \leq \eta\right\}$, for all $\eta \in \mathbb{R}$, so that $C_{\nu,T}^\pi$ has the same distribution on $\mathbb{R}$ as the finite horizon cost $\sum_{t=0}^{T} \gamma^t c(s_t, a_t)$ at time T with respect to P_ν^π. The next lemma considers the quality of $\rho\left(C_{\nu,T}^\pi\right)$ versus $\rho(C_\nu^\pi)$, it follows immediately from law invariance and Assumption 3.2.

Lemma 3.3. *Suppose Assumption 3.2 holds. Then, for any $\epsilon > 0$, there is a T such that $|\rho\left(C^{\pi}_{\nu,\,T}\right) - \rho\left(C^{\pi}_{\nu}\right)| \leq \epsilon$ for all $\pi \in \Pi$.*

The preceding Lemma 3.3 shows how we are using Assumption 3.2 and why we assumed law invariance of ρ. We have defined $C^{\pi}_{\nu,\,T}$ to be equal in distribution to y_T when the underlying probability distribution is P^{π}_{ν}. Lemma 3.3 confirms that the risk of $C^{\pi}_{\nu,\,T}$, $\rho\left(C^{\pi}_{\nu,\,T}\right)$, is close to the risk of the infinite horizon discounted cost C^{π}_{ν}, $\rho\left(C^{\pi}_{\nu}\right)$, uniformly across policies. We needed the error guarantee on $|\rho\left(C^{\pi}_{\nu,\,T}\right) - \rho\left(C^{\pi}_{\nu}\right)|$ to be independent of $\pi \in \Pi$.

Since $C^{\pi}_{\nu,\,T}$ approximates C^{π}_{ν} well for large T, we can approximate Problem (2.2) with the truncated problem

$$\inf_{\pi \in \Pi} \rho\left(C^{\pi}_{\nu,\,T}\right). \tag{3.1}$$

Since we have truncated the planning horizon, Problem (3.1) turns out to be exactly solvable with convex analytic methods since we can compute the distribution of y_T exactly for any finite T, as we will show. We can use a solution of Problem (3.1) to get a near-optimal solution for Problem (2.2), as confirmed in the following lemma. However, the best we can hope for is a near-optimal policy in Π_1, but not in Π_0 (which is defined over the unaugmentated state space), for general ρ. Denote $\rho^* := \inf_{\pi \in \Pi} \rho\left(C^{\pi}_{\nu}\right)$.

Lemma 3.4. *Choose any $\epsilon > 0$. Then, there is a T such that:*

(i) $\inf_{\pi \in \Pi} \rho\left(C^{\pi}_{\nu,\,T}\right) < \rho^* + \epsilon$;

(ii) For $\hat{\pi}$ with $\rho\left(C^{\hat{\pi}}_{\nu,\,T}\right) \leq \inf_{\pi \in \Pi} \rho\left(C^{\pi}_{\nu,\,T}\right) + \epsilon$, we have $\rho\left(C^{\hat{\pi}}_{\nu}\right) < \rho^* + 3\,\epsilon$.

Proof. (i) Choose T such that $|\rho\left(C^{\pi}_{\nu,\,T}\right) - \rho\left(C^{\pi}_{\nu}\right)| \leq \epsilon/2$ for all $\pi \in \Pi$, and then choose π' such that $\rho\left(C^{\pi'}_{\nu}\right) < \rho^* + \epsilon/2$. We are guaranteed that such a π' exists by the definition of infimum. It follows that

$$\rho\left(C^{\pi'}_{\nu,\,T}\right) \leq \rho\left(C^{\pi'}_{\nu}\right) + |\rho\left(C^{\pi'}_{\nu,\,T}\right) - \rho\left(C^{\pi'}_{\nu}\right)| \leq \rho^* + \epsilon.$$

(ii) Now, for $\hat{\pi}$ with $\rho\left(C^{\hat{\pi}}_{\nu,\,T}\right) \leq \inf_{\pi \in \Pi} \rho\left(C^{\pi}_{\nu,\,T}\right) + \epsilon$, it follows

$$\rho\left(C^{\hat{\pi}}_{\nu}\right) \leq \rho\left(C^{\hat{\pi}}_{\nu,\,T}\right) + |\rho\left(C^{\hat{\pi}}_{\nu}\right) - \rho\left(C^{\hat{\pi}}_{\nu,\,T}\right)| \leq \inf_{\pi \in \Pi} \rho\left(C^{\pi}_{\nu,\,T}\right) + 2\,\epsilon \leq \rho^* + 3\,\epsilon,$$

where the second inequality uses part (i). ∎

Next we show that the distribution of $C^{\pi}_{\nu,\,T}$ can be computed exactly with convex analytic methods for any T. The idea is to modify the transition kernel so that it no longer updates the running cost after time T, i.e., $y_{t+1} = y_t$ for all $t \geq T$ and thus $y_t = y_T$ for all $t \geq T$. Then, we use the conditional distribution of μ^{π}_{ν} on $\{t \geq T\}$ to compute the distribution of $C^{\pi}_{\nu,\,T}$ on $\mathbb{R}$. Note that we can use the augmented state variable $\{z_t\}_{t \geq 0}$ to know when we have reached time T, i.e., when $z_t = \gamma^T$, so that we do not have to explicitly keep track of time and augment the state even further. The modified transition kernel is

$$\tilde{Q}_T\left(B,\, y',\, z' \,|\, s, y, z, a\right) = \begin{cases} Q\left(B \,|\, s, a\right), & y' = z\, c\left(s, a\right) + y,\; z' = \gamma\, z,\; z > \gamma^T, \\ Q\left(B \,|\, s, a\right), & y' = y,\; z' = z,\; z = \gamma^T, \\ 0, & \text{otherwise}, \end{cases},$$

for any $B \in \mathcal{B}\left(\mathbb{S}\right)$. The transition kernel $\tilde{Q}_T$ is very close to $\tilde{Q}$, however it stops updating the running cost and the discount factor once $z_t = \gamma^T$ is reached at time T.

We now introduce the machinery for occupation measures in order to compute the distribution of $C^{\pi}_{\nu,\,T}$, starting with the necessary functional spaces. Let $\mathcal{M}\left(\tilde{\mathbb{K}}\right)$ be the space of finite signed measures on $\left(\tilde{\mathbb{K}},\,\mathcal{B}\left(\tilde{\mathbb{K}}\right)\right)$ in the total variation norm $\|\mu\|_{\mathcal{M}\left(\tilde{\mathbb{K}}\right)} = \int_{\tilde{\mathbb{K}}} |\mu|\left(d\left(s,y,z,a\right)\right)$, and let $\mathcal{M}_{+}\left(\tilde{\mathbb{K}}\right)$ be the set of positive measures. The upcoming occupation measures for our convex analytic approach will belong to $\mathcal{M}\left(\tilde{\mathbb{K}}\right)$. Additionally, we will need the space dual to $\mathcal{M}\left(\tilde{\mathbb{K}}\right)$: let $\mathcal{F}\left(\tilde{\mathbb{K}}\right)$ be the space of bounded measurable functions $f:\tilde{\mathbb{K}}\to\mathbb{R}$ in the supremum norm $\|f\|_{\mathcal{F}\left(\tilde{\mathbb{K}}\right)} = \sup_{(s,y,z,a)\in\tilde{\mathbb{K}}} |f\left(s,y,z,a\right)|$. For a measure $\mu\in\mathcal{M}\left(\tilde{\mathbb{K}}\right)$ and a function $f\in\mathcal{F}\left(\tilde{\mathbb{K}}\right)$, we define the duality pairing $\langle\mu,\,f\rangle = \int_{\tilde{\mathbb{K}}} f\left(s,y,z,a\right)\mu\left(d\left(s,y,z,a\right)\right)$, the integral of f with respect to the measure μ. When $\mu\in\mathcal{M}\left(\tilde{\mathbb{K}}\right)$ is a probability measure and $f\in\mathcal{F}\left(\tilde{\mathbb{K}}\right)$, then $\langle\mu,\,f\rangle$ can be interpreted as an expectation.

Now, let I_B be the indicator function of a measurable set $B\in\mathcal{B}\left(\tilde{\mathbb{K}}\right)$. We can define the augmented *infinite horizon discounted occupation measure* $\mu^{\pi}_{\nu}\in\mathcal{M}_{+}\left(\tilde{\mathbb{K}}\right)$ of a policy π on $\tilde{\mathbb{K}}$ as

$$\mu^{\pi}_{\nu}\left(B\right) = \sum_{t=0}^{\infty}\gamma^{t}\mathbb{E}^{\pi}_{\nu}\left[I_B\left(s_t,y_t,z_t,a_t\right)\right] = \sum_{t=0}^{\infty}\gamma^{t}P^{\pi}_{\nu}\left\{\left(s_t,y_t,z_t,a_t\right)\in B\right\},\ \forall B\in\mathcal{B}\left(\tilde{\mathbb{K}}\right).$$

We can interpret $\mu^{\pi}_{\nu}\left(B\right)$ as the expected discounted number of visits to state-action pairs in the set $B\subset\tilde{\mathbb{K}}$ when following the policy π.

The next theorem expresses the distribution of $C^{\pi}_{\nu,\,T}$ in terms of μ^{π}_{ν}. This theorem is the foundation of the convex analytic approach for Problem (3.1).

Theorem 3.5. *For any $\eta\in\mathbb{R}$,*

$$Pr\left\{C^{\pi}_{\nu,\,T}\leq\eta\right\} = \frac{1-\gamma}{\gamma^{T}}\int_{\tilde{\mathbb{S}}} I\left\{y\leq\eta,\,z=\gamma^{T}\right\}\mu^{\pi}_{\nu}\left(d\left(s,y,z,a\right)\right).$$

Proof. Compute

$$\Pr\left\{C^{\pi}_{\nu,\,T}\leq\eta\right\} = \frac{1-\gamma}{\gamma^{T}}\mathbb{E}^{\pi}_{\nu}\left[\sum_{t=T}^{\infty}\gamma^{t}I\left\{y_t\leq\eta\right\}\right],$$

using the fact that $I\left\{y_T\leq\eta\right\} = I\left\{y_t\leq\eta\right\}$ for all $t\geq T$. Finally

$$\mathbb{E}^{\pi}_{\nu}\left[\sum_{t=T}^{\infty}\gamma^{t}I\left\{y_t\leq\eta\right\}\right] = \int_{\tilde{\mathbb{K}}} I\left\{y_t\leq\eta,\,z_t=\gamma^{T}\right\}\mu^{\pi}_{\nu}\left(d\left(s,y,z,a\right)\right).$$

□

Based on the preceding theorem, we construct a random variable in $\mathcal{L}$ that is determined by μ^{π}_{ν} and that is equal in distribution to $C^{\pi}_{\nu,\,T}$. For a measure $\mu\in\mathcal{M}\left(\tilde{\mathbb{K}}\right)$, define the random variable $X\left(\mu\right)$ in $\mathcal{L}$ to have the distribution

$$\Pr\left\{X\left(\mu\right)\leq\eta\right\} = \frac{1-\gamma}{\gamma^{T}}\int_{\tilde{\mathbb{K}}} I\left\{y\leq\eta,\,z=\gamma^{T}\right\}\mu\left(d\left(s,y,z,a\right)\right),\ \forall\eta\in\mathbb{R}.$$

By Theorem 3.5, $X\left(\mu_{\nu}^{\pi}\right)$ is equal in distribution to $C_{\nu,\,T}^{\pi}$.

Occupation measures μ_{ν}^{π} for policies $\pi \in \Pi$ have special properties that can be conveniently expressed as a linear mapping. Define $\mathcal{M}\left(\tilde{\mathbb{S}}\right)$ to be the space of finite signed measures on $\left(\tilde{\mathbb{S}},\, \mathcal{B}\left(\tilde{\mathbb{S}}\right)\right)$ in the total variation norm. Introduce the linear mapping $L_{0,\,T} : \mathcal{M}\left(\tilde{\mathbb{K}}\right) \rightarrow \mathcal{M}\left(\tilde{\mathbb{S}}\right)$ defined by

$$\left[L_{0,\,T}\mu\right](B) := \hat{\mu}\left(B\right) - \gamma \int_{\tilde{\mathbb{K}}} \tilde{Q}_{T}\left(B \,|\, s,y,z,a\right) \mu\left(d\left(s,y,z,a\right)\right), \forall B \in \mathcal{B}\left(\tilde{\mathbb{S}}\right), \tag{3.2}$$

where $\hat{\mu}\left(B\right) = \int_{B\times\mathbb{A}} \mu\left(d\left(s,y,z,a\right)\right)$, for all $B \in \mathcal{B}\left(\tilde{\mathbb{S}}\right)$ is the marginal distribution of the measure μ on $\tilde{\mathbb{S}}$. With this notation in place, we can write the convex analytic form of Problem (3.1),

$$\inf_{\mu\in\mathcal{M}\left(\tilde{\mathbb{K}}\right)} \left\{\rho\left(X\left(\mu\right)\right) \,:\, L_{0,\,T}\mu = \nu\right\}. \tag{3.3}$$

It is worth mentioning that when $\rho\left(X\left(\mu\right)\right)$ is concave in μ (in particular, when it is linear in μ), then Problem (3.3) has an optimal solution at an extreme feasible measure μ (see [34, Theorem 19]). Since extreme feasible measures correspond to deterministic policies, it follows that randomized policies are not needed.

We formally justify the equivalence between Problem (3.1) and Problem (3.3) in the next lemma. Specifically, we show that an optimal solution for either problem can be used to construct an optimal solution for the other problem. The intuition is that the constraint $L_{0,\,T}\mu = \nu$ defines all feasible occupation measures that can be produced by policies in Π. Further, since we are using the modified transition kernel $\tilde{Q}_{T}$, we know that $\rho\left(X\left(\mu\right)\right)$ is equal to $\rho\left(C_{\nu,\,T}^{\pi}\right)$ for some policy π based on Theorem 3.5.

We note that given an occupation measure $\mu \in \mathcal{M}\left(\tilde{\mathbb{K}}\right)$, we get a policy $\pi_{\mu} \in \Pi$ defined by the conditional distribution of μ on $\mathbb{A}$, $\pi_{\mu}\left(B \,|\, s,y,z\right) = \mu\left(B \,|\, s,y,z\right)$, for all $B \in \mathcal{B}\left(\mathbb{A}\right)$, for each state $(s,y,z) \in \tilde{\mathbb{S}}$.

Lemma 3.6. *If π is $\epsilon-$optimal for Problem (3.1), then μ_{ν}^{π} is $\epsilon-$optimal for Problem (3.3). Conversely, if μ is $\epsilon-$optimal for Problem (3.3), then π_{μ} is $\epsilon-$optimal for Problem (3.1).*

Proof. All occupation measures μ_{ν}^{π} for $\pi \in \Pi$ must satisfy $L_{0,\,T}\mu_{\nu}^{\pi} = \nu$. If μ satisfies the equality $L_{0,\,T}\mu = \nu$, then $\pi_{\mu} \in \Pi$. Thus, any feasible μ for Problem (3.3) gives a feasible policy $\pi \in \Pi$, and vice versa.

We know that $\inf_{\pi\in\Pi} \rho\left(C_{\nu,\,T}^{\pi}\right) = \inf_{\pi\in\Pi} \rho\left(X\left(\mu_{\nu}^{\pi}\right)\right)$, since $\rho\left(X\left(\mu_{\nu}^{\pi}\right)\right) = \rho\left(C_{\nu,\,T}^{\pi}\right)$ by construction of $X\left(\mu_{\nu}^{\pi}\right)$ and by Assumption 2.2. Further $\inf_{\pi\in\Pi} \rho\left(X\left(\mu_{\nu}^{\pi}\right)\right) = \inf_{\mu\in\mathcal{M}\left(\tilde{\mathbb{K}}\right)} \left\{\rho\left(X\left(\mu\right)\right) \,:\, L_{0,\,T}\mu = \nu\right\}$, since μ_{ν}^{π} is feasible for Problem (3.3) for any $\pi \in \Pi$, and $\pi_{\mu} \in \Pi$ for any μ feasible to Problem (3.3). So, the optimal values of Problem (3.1) and Problem (3.3) are equal. Note $\rho\left(X\left(\mu\right)\right) = \rho\left(C_{\nu,\,T}^{\pi_{\mu}}\right)$ again by Assumption 2.2 and the definition of π_{μ} to get the desired result. ∎

To get cleaner notation throughout the paper, we introduce the additional linear mapping $L_{1,\,T} : \mathcal{M}\left(\tilde{\mathbb{K}}\right) \rightarrow \mathcal{M}\left(\mathcal{Y}\right)$, where $\mathcal{M}\left(\mathcal{Y}\right)$ is the space of finite signed measures on $\mathcal{Y} = [0,\, \bar{c}/\left(1-\gamma\right)]$, defined by

$$\left[L_{1,\,T}\mu\right](B) := \frac{1-\gamma}{\gamma^{T}} \int_{\tilde{\mathbb{K}}} I\left\{y \in B,\, z = \gamma^{T}\right\} \mu\left(d\left(s,y,z,a\right)\right), \forall B \in \mathcal{B}\left(\mathcal{Y}\right). \tag{3.4}$$

Note that $\theta = L_{1,T}\mu$ is proportional to the marginal distribution of μ on $\mathcal{Y}$ conditioned on the event $\left\{z = \gamma^T\right\}$ that appears in the statement of Theorem 3.5. Equivalently, $\theta = L_{1,T}\mu$ is the marginal distribution of μ on $\mathcal{Y}$ conditioned on the event $\{t \geq T\}$, i.e., we have passed the time T and the running costs are no longer updated. We are introducing this shorthand to get a cleaner statement of Problem (3.3), since it is difficult to write out $\rho\left(X\left(\mu\right)\right)$ directly for specific forms of ρ.

Given a distribution $\theta \in \mathcal{M}\left(\mathcal{Y}\right)$, we abuse notation and let $X\left(\theta\right)$ be the random variable in $\mathcal{L}$ defined by $\Pr\left\{X\left(\theta\right) \leq \eta\right\} = \theta\left\{y \leq \eta\right\}$, for all $\eta \in \mathbb{R}$. Under this definition, $X\left(\theta\right)$ is equal in distribution to $X\left(\mu\right)$ where $\theta = L_{1,T}\mu$. Problem (3.3) is then equivalent to

$$\rho^* \triangleq \inf_{\mu \in \mathcal{M}\left(\tilde{\mathbb{K}}\right),\, \theta \in \mathcal{M}\left(\mathcal{Y}\right)} \left\{\rho\left(X\left(\theta\right)\right) \,:\, L_{0,T}\mu = \nu,\, L_{1,T}\mu = \theta\right\}, \tag{3.5}$$

using the fact that $\rho\left(X\left(\theta\right)\right) = \rho\left(X\left(\mu\right)\right)$ when $\theta = L_{1,T}\mu$. The additional measure θ helps us write the objective $\rho\left(X\left(\theta\right)\right)$ more cleanly once we choose specific functional forms for ρ. For this reason, we will focus on Problem (3.5) rather than Problem (3.3) in the remainder of the paper.

4. Finite approximations. Problem (3.5) is generally an infinite-dimensional optimization problem, it has infinitely many variables and constraints. Such problems are extremely hard to solve directly, but they can be approximated by finite-dimensional optimization problems, i.e., problems with finitely many variables and constraints. In this section we develop two approaches for making finite approximations of Problem (3.5). First, we explain how the aggregation-relaxation-inner approximation method can give finite approximations for Problem (3.5). This method works in full generality, and has been studied for infinite-dimensional linear programming problems (see [21, 23]). Second, we develop a discretization scheme for the augmented state variables. This discretization scheme is intuitive and leads easily to a convergence rate analysis.

4.1. Aggregation-relaxation-inner approximation. We now elucidate finite approximations for the infinite-dimensional linear programming problems that arise in the classical convex analytic approach for risk-neutral MDPs. These finite approximations are based on an aggregation of the constraints, a relaxation of the aggregate constraints, and then an inner approximation of the decision variables. We note similar developments in [21, 23] for infinite-dimensional linear programming problems. In our case, we have to take extra care because Problem (3.5) has a nonlinear objective in general. We make the following assumption about the objective.

Assumption 4.1. *$\rho\left(X\left(\cdot\right)\right) \,:\, \mathcal{M}\left(\mathcal{Y}\right) \to \mathbb{R}$ is weakly continuous.*

Assumption 4.1 is needed to establish an asymptotic convergence of our approximation. In earlier work on approximation of infinite-dimensional linear programs [21, 23], there was no need for Assumption 4.1 because the objective function was linear. It was possible to use Fatou's lemma to show that a sequence of approximate solutions converges to an optimal solution asymptotically. Our Problem (3.5) has a nonlinear objective and thus needs a special consideration. We will show that this assumption is met for all of the risk-aware optimization formulations that we study in the next section.

We also make the following assumption about the augmented state space, and the augmented set of state-action pairs.

Assumption 4.2. *$\tilde{\mathbb{S}}$ and $\tilde{\mathbb{K}}$ are locally compact separable metric spaces.*

This assumption is met under many circumstances, for instance if $\tilde{\mathbb{S}}$ and $\tilde{\mathbb{K}}$ are Euclidean spaces - which is usually the case in practice. Assumption 4.2 is needed so that we can approximate probability measures in $\mathcal{M}\left(\tilde{\mathbb{K}}\right)$ and $\mathcal{M}\left(\tilde{\mathbb{S}}\right)$ with probability measures that have finite support.

We begin by describing aggregation of the constraints. Let $\mathcal{C}\left(\tilde{\mathbb{S}}\right)$ be the space of continuous functions on $\tilde{\mathbb{S}}$. The constraint $L_{0,\,T}\mu = \nu$ is equivalent to $\langle L_{0,T}\mu - \nu,\, f\rangle = 0,\, \forall f \in \mathcal{C}_0\left(\tilde{\mathbb{S}}\right)$, where $\mathcal{C}_0\left(\tilde{\mathbb{S}}\right) \subset \mathcal{C}\left(\tilde{\mathbb{S}}\right)$ is any countable dense subset of $\mathcal{C}\left(\tilde{\mathbb{S}}\right)$, by [23, Lemma 12.5.2]. Similarly, let $\mathcal{C}(\mathcal{Y})$ be the space of continuous functions on $\mathcal{Y}$, and let $\mathcal{C}_1(\mathcal{Y})$ be a countable dense subset of $\mathcal{C}(\mathcal{Y})$. The constraint $L_{1,\,T}\mu = \theta$ is equivalent to $\langle L_{1,T}\mu - \theta,\, f\rangle = 0,\, \forall f \in \mathcal{C}_1(\mathcal{Y})$. We will now approximate these two new representations of the equality constraints in Problem (3.5). Let $\{C_{0,\,k}\}_{k\geq 0}$ be an increasing sequence of finite sets with $C_{0,\,k} \uparrow \mathcal{C}_0\left(\tilde{\mathbb{S}}\right)$ and $\{C_{1,\,k}\}_{k\geq 0}$ be an increasing sequence of finite sets with $C_{1,\,k} \uparrow \mathcal{C}_1(\mathcal{Y})$. Now we discuss the inner approximation of the infinite-dimensional decision variables μ and θ. Let $\mathbb{S}' \subset \mathbb{S}$, $\mathbb{Y}' \subset \mathcal{Y}$, $\mathbb{Z}' \subset (0,1]$, and $\mathbb{A}' \subset \mathbb{A}$ be countable dense sets, with increasing sequences $\{S_k\}_{k\geq 0}$, $\{Y_k\}_{k\geq 0}$, $\{Z_k\}_{k\geq 0}$, and $\{A_k\}_{k\geq 0}$ with $S_k \uparrow \mathbb{S}'$, $Y_k \uparrow \mathbb{Y}'$, $Z_k \uparrow \mathbb{Z}'$, and $A_k \uparrow \mathbb{A}'$.

Finally, take $\Delta_{0,k} = \mathcal{P}(S_k \times Y_k \times Z_k \times A_k)$ and $\Delta_{1,k} = \mathcal{P}(Y_k)$. The sets $\Delta_0 = \cup_{k=1}^{\infty}\Delta_{0,k}$ and $\Delta_1 = \cup_{k=1}^{\infty}\Delta_{1,k}$ are then dense in the spaces of probability measures on $\tilde{\mathbb{K}}$ and $\mathcal{Y}$ respectively, under Assumption 4.2.

The resulting finite approximation of Problem (3.5) is then

$$\begin{aligned}\boldsymbol{P}(C_{0,\,k},\, C_{1,\,k},\, \epsilon_k,\, \Delta_{0,\,l},\, \Delta_{1,\,l}) \;:\; \inf_{\mu\in\mathcal{M}\left(\tilde{\mathbb{K}}\right),\,\theta\in\mathcal{M}(\mathcal{Y})} & \rho(X(\theta)) \\ \text{s.t. } & |\langle L_{0,\,T}\mu - \nu,\, f\rangle| \leq \epsilon_k, \quad \forall f \in C_{0,\,k}, \\ & |\langle L_{1,\,T}\mu - \theta,\, f\rangle| \leq \epsilon_k, \quad \forall f \in C_{1,\,k}, \\ & \mu \in \Delta_{0,\,l},\, \theta \in \Delta_{1,\,l}.\end{aligned}$$

Problem $\boldsymbol{P}(C_{0,\,k},\, C_{1,\,k},\, \epsilon_k,\, \Delta_{0,\,l},\, \Delta_{1,\,l})$ has finitely many constraints indexed by $C_{0,\,k}$ and $C_{1,\,k}$, and finitely many variables since we have restricted the support of μ and θ. We continue to view $\mu \in \Delta_{0,\,l}$ and $\theta \in \Delta_{1,\,l}$ as elements of $\mathcal{M}\left(\tilde{\mathbb{K}}\right)$ and $\mathcal{M}(\mathcal{Y})$, respectively. Note that the constraints in $\boldsymbol{P}(C_{0,\,k},\, C_{1,\,k},\, \epsilon_k,\, \Delta_{0,\,l},\, \Delta_{1,\,l})$ include an error tolerance of ϵ_k, since we cannot expect to satisfy them exactly with discretized decision variables. The next theorem considers the behavior of $\boldsymbol{P}(C_{0,\,k},\, C_{1,\,k},\, \epsilon_k,\, \Delta_{0,\,l},\, \Delta_{1,\,l})$ as $k,\, l \to \infty$, and its proof is similar to the preceding two theorems.

Theorem 4.3. *Let ρ^*_{kl} be the optimal value of Problem $\boldsymbol{P}(C_{0,\,k},\, C_{1,\,k},\, \epsilon_k,\, \Delta_{0,\,l},\, \Delta_{1,\,l})$, and let $\{(\mu_{kl},\, \theta_{kl})\}_{k,\,l\geq 0}$ be a sequence of solutions of it. Then,*

(i) Problem $\boldsymbol{P}(C_{0,\,k},\, C_{1,\,k},\, \epsilon_k,\, \Delta_{0,\,l},\, \Delta_{1,\,l})$ is solvable for each k, for all sufficiently large l.

(ii) $\rho^*_{kl} \to \rho^*$ as $k \to \infty$ and $l \to \infty$. Every weak accumulation point of $\{(\mu_{kl},\, \theta_{kl})\}_{k,\,l\geq 0}$ is an optimal solution of Problem (3.5).

Proof. This result is similar to the proof of [23, Theorem 12.5.3]. Only now, we use weak-continuity of the objective $\rho(X(\cdot))$ in θ to get $\liminf_{i\to\infty}\rho(X(\theta_i)) \geq \rho(X(\theta))$, whenever $\theta_i \to \theta$ in the weak topology. In [23, Theorem 12.5.3], Fatou's lemma was used to establish $\liminf_{i\to\infty}\langle c,\, \mu_i\rangle \geq \langle c,\, \mu\rangle$, when $\mu_i \to \mu$ in the weak topology. □

Because we worked in a general setting (we have made no assumptions on the state and action spaces other than Assumption 4.2), the convergence results in this

subsection are asymptotic. Next we look at a finite approximation scheme that has convergence rate guarantees when $\mathbb{S}$ and $\mathbb{A}$ are finite.

4.2. Discretization. In this subsection, we consider the special case when the original state and action spaces are finite. Here, we only need to discretize the augmenting state variables defined on $\mathcal{Y} \times (0,1]$. Although the running cost can only take finitely many values when $\mathbb{S}$ and $\mathbb{A}$ are finite, we still discretize $\mathcal{Y}$ because the number of possible values for the running cost can be quite large. We propose a natural discretization scheme for the running cost and the discounting that readily leads to convergence rate estimates. We will use $\mathbb{Y} \subset \mathcal{Y}$ to denote a general finite discretization of $\mathcal{Y}$. The choice of a discretization $\mathbb{Z} \subset (0,1]$ is automatic once T has been fixed, specifically it is $\mathbb{Z} = \{\gamma^t\}_{t=0}^{T}$.

We will discretize $\mathcal{Y}$ into the finite set $\mathbb{Y} \subset \mathcal{Y}$, where the granularity of $\mathbb{Y}$ is $\sup_{y \in \mathcal{Y}} \inf_{\hat{y} \in \mathbb{Y}} |y - \hat{y}|$, the set distance between $\mathcal{Y}$ and $\mathbb{Y}$. Given $\epsilon > 0$, a set $\mathbb{Y}$ with granularity exists with not more than $\lceil \bar{c}/(\epsilon(1-\gamma)) \rceil$ elements. We introduce a new stochastic process $\{\bar{y}_t\}_{t \geq 0}$ for the discretized running costs on $\mathbb{Y}$ to differentiate from the original continuous running costs $\{y_t\}_{t \geq 0}$. The system dynamic for $\{\bar{y}_t\}_{t \geq 0}$ is

$$\bar{y}_{t+1} = \arg\min_{y \in \mathbb{Y}} |y - (z_t\, c(s_t, a_t) + \bar{y}_t)|, \, \forall t \geq 0,$$

which simply assigns $\bar{y}_{t+1}$ to be the closest point in $\mathbb{Y}$ to the original update given by $z_t\, c(s_t, a_t) + \bar{y}_t$. We will show that the error of this scheme increases linearly with time. We will also introduce $\{\bar{z}_t\}_{t \geq 0}$ to denote the discretized discounting process whose system dynamic is given by

$$\bar{z}_{t+1} = \begin{cases} \gamma\, \bar{z}_t, & \bar{z}_t > \gamma^T, \\ \bar{z}_t, & \bar{z}_t = \gamma^T. \end{cases}$$

The process $\{\bar{z}_t\}_{t \geq 0}$ is the same as $\{z_t\}_{t \geq 0}$ up until time T.

The stochastic processes $\{\bar{y}_t\}_{t \geq 0}$ and $\{\bar{z}_t\}_{t \geq 0}$ are defined on $(\Omega, \mathcal{F})$ along with $\{(s_t, a_t)\}_{t \geq 0}$. By construction, $\{\bar{y}_t\}_{t \geq 0}$ is a function of $\{(s_t, a_t)\}_{t \geq 0}$ and $\{\bar{z}_t\}_{t \geq 0}$ is a deterministic process. We denote the discretized augmented state space as $\bar{\mathbb{S}} = \mathbb{S} \times \mathbb{Y} \times \mathbb{Z}$, and the corresponding discretized set of state-action pairs is $\bar{\mathbb{K}} = \{(s, y, z, a) \in \mathbb{S} \times \mathbb{Y} \times \mathbb{Z} \times \mathbb{A} : a \in A(s)\}$. Corresponding to the stochastic process $\{(s_t, \bar{y}_t, \bar{z}_t, a_t)\}_{t \geq 0}$, we introduce the transition kernel $\bar{Q}_T$ which accounts for the system dynamics of $\{\bar{y}_t\}_{t \geq 0}$ and $\{\bar{z}_t\}_{t \geq 0}$. We let $\bar{L}_{0,\,T}$ be the linear operator $L_{0,\,T}$ with $\bar{Q}_T$ in place of $\tilde{Q}_T$,

$$\left[\bar{L}_{0,\,T}\mu\right](B) := \hat{\mu}(j) - \gamma \sum_{(s,y,z,a) \in \bar{\mathbb{K}}} \bar{Q}_T(j \mid s, y, z, a)\, \mu(s, y, z, a), \forall j \in \bar{\mathbb{S}}, \tag{4.1}$$

where $\hat{\mu}(j) = \sum_{a \in \mathbb{A}} \mu(s, y, z, a)$, for all $(s, y, z) \in \bar{\mathbb{S}}$. Equation (4.1) is the discretized analog of $L_{0,\,T}$ defined in equation (3.2). Similarly, we define $\bar{L}_{1,\,T}$ to be the linear operator $L_{1,\,T}$ suitably modified for use on $\mathbb{Y}$,

$$\left[\bar{L}_{1,\,T}\mu\right](y) := \frac{1-\gamma}{\gamma^T} \sum_{(s,y,z,a) \in \bar{\mathbb{K}}} I\left\{y = j,\, z = \gamma^T\right\} \mu(s, y, z, a), \, \forall j \in \mathbb{Y}. \tag{4.2}$$

Again, equation (4.2) is the discretized analog of $L_{0,\,T}$ defined in equation (3.4). For a measure $\theta \in \mathcal{M}(\mathbb{Y})$, define the random variable $\bar{X}(\theta)$ in $\mathcal{L}$ to have the distribution $\Pr\left\{\bar{X}(\mu) = j\right\} = \theta(\eta), \ \forall \eta \in \mathbb{Y}$.

The convex analytic formulation for the discretized MDP is then

$$\inf_{\mu\in\mathbb{R}^{|\bar{\mathbb{K}}|},\,\theta\in\mathbb{R}^{|\mathbb{Y}|}} \left\{\rho\left(\bar{X}\left(\theta\right)\right) \,:\, \bar{L}_{0,\,T}\mu=\nu,\,\bar{L}_{1,\,T}\mu=\theta\right\}. \tag{4.3}$$

Problem (4.3) has finitely many variables and constraints by construction. We now want to compare solutions of the discretized Problem (4.3) and the original Problem (3.5).

Theorem 4.4. *Let $\rho^*=\inf_{\pi\in\Pi}\rho\left(C_\nu^\pi\right)$ and choose any $\epsilon>0$ and $T=T\left(\epsilon\right)$. If the granularity of $\mathbb{Y}$ is smaller than ϵ/T, then:*

(i) $|y_T\left(\omega\right)-\bar{y}_T\left(\omega\right)|<\epsilon$ for all $\omega\in\Omega$;

(ii) Under Assumption 3.2, $|\rho\left(y_T\right)-\rho\left(\bar{y}_T\right)|\leq\epsilon$;

(iii) For $\mu\in\mathcal{M}\left(\bar{\mathbb{K}}\right)$ with

$$\rho\left(\bar{X}\left(\theta\right)\right)\leq\inf_{\mu\in\mathcal{M}\left(\bar{\mathbb{K}}\right)}\left\{\rho\left(\bar{X}\left(\mu\right)\right)\,:\,\bar{L}_{0,\,T}\mu=\nu,\,\bar{L}_{1,\,T}\mu=\theta\right\}+\epsilon$$

we have $\rho\left(C_\nu^{\hat{\pi}}\right)<\rho^*+3\,\epsilon$ where $\hat{\pi}$ is the policy generated by μ.

Proof. (i) By definition, $\bar{y}_0=y_0=c\left(s_0,a_0\right)$ so $|y_0-\bar{y}_0|=0$. Now, $|y_1-\bar{y}_1|<\epsilon$ by construction of $\{\bar{y}_t\}_{t\geq0}$ and the fact that $\mathbb{Y}$ has granularity ϵ. For the inductive step, suppose $|y_t-\bar{y}_t|<t\,\epsilon$. Then, $y_{t+1}=z_t\,c\left(s_t,a_t\right)+y_t$ and $\bar{y}_{t+1}=\arg\min_{y\in\mathbb{Y}}|y-\left(z_t\,c\left(s_t,a_t\right)+\bar{y}_t\right)|$. Then,

$$\begin{aligned}|y_{t+1}-\bar{y}_{t+1}|&\leq|y_{t+1}-\left(z_t\,c\left(s_t,a_t\right)+\bar{y}_t\right)|+|\left(z_t\,c\left(s_t,a_t\right)+\bar{y}_t\right)-\bar{y}_{t+1}|\\&=|y_t-\bar{y}_t|+|\left(z_t\,c\left(s_t,a_t\right)+\bar{y}_t\right)-\bar{y}_{t+1}|\\&<\left(t+1\right)\epsilon,\end{aligned}$$

using the update for $\bar{y}_{t+1}$ and the induction hypothesis.

(ii) Follows immediately from part (i).

(iii) For $\pi\in\Pi$, we have that $\rho\left(\bar{X}\left(\mu_\nu^\pi\right)\right)=\rho\left(\bar{y}_T\right)$ by construction. ∎

Problem (4.3) can be solved exactly, at least in principle, because it is a finite-dimensional optimization problem. When there is additional problem structure, we can extend this discretization scheme to solve Problem (4.3) when $\mathbb{S}$ and $\mathbb{A}$ are infinite. Such a situation occurs in the dynamic risk-averse newsvendor problem which we will report in the future.

5. Optimization of risk functionals. Section 3 shows how to cast risk-aware MDPs as static optimization problems and Section 4 shows how to construct tractable approximations. In this section, we apply our general methodology to two specific risk-aware MDPs. First, we minimize expected utility and then we minimize CVaR. Notably, the expected utility minimizing MDPs lead to linear programming problems in occupation measures. The CVaR minimizing MDPs lead to nonconvex problems in occupation measures, but these problems can be solved with a sequence of linear programming problems.

5.1. Expected utility risk functional. Utility functions can be used to express a decision maker's risk preferences. Often, decision makers have increasing marginal costs. Hence, we focus on utility functions that are increasing and convex. For a fixed increasing, convex, and continuous utility function $u:\mathbb{R}\to\mathbb{R}$, we can replace the risk-neutral expectation $\mathbb{E}_\nu^\pi\left[\cdot\right]$ with the expected utility $\mathbb{E}_\nu^\pi\left[u\left(\cdot\right)\right]$. The resulting risk-aware MDP is

$$\inf_{\pi\in\Pi}\mathbb{E}\left[u\left(C_\nu^\pi\right)\right]=\inf_{\pi\in\Pi}\mathbb{E}_\nu^\pi\left[u\left(\sum_{t=0}^{\infty}\gamma^t c\left(s_t,a_t\right)\right)\right]. \tag{5.1}$$

Since we are focusing on costs rather than rewards, we prefer lower expected utility to higher expected utility. It would be more correct to refer to u as a "disutility" function since it measures costs, but we continue to use the more common term "utility". Problem (5.1) has been studied with state space augmentation in [4] for concave and convex utility functions u. In [4], it is shown that Problem (5.1) can be solved with value iteration and policy iteration on the augmented state space.

We first approximate Problem (5.1) with a finite horizon problem, and then we formulate the resulting problem as a static optimization problem in occupation measures. The resulting static optimization problem turns out to be a linear programming problem. Next, we use linear programming duality to reveal the dual problem in value functions - from which we can recover dynamic programming equations.

Next we confirm that the expected utility objective $\mathbb{E}\left[u\left(\cdot\right)\right]$ for Problem (5.1) satisfies Assumption 3.2, so that our earlier error bounds from Lemma 3.4 apply.

Lemma 5.1. *The risk function $\rho\left(\cdot\right) = \mathbb{E}\left[u\left(\cdot\right)\right]$ satisfies Assumption 3.2.*

Proof. Since u is increasing, convex, and continuous on $\mathcal{Y}$, it is also Lipschitz continuous on this interval. Without loss of generality, we can take the Lipschitz constant to be one by appropriately scaling u by a positive constant. Now compute

$$\begin{aligned}
&|\mathbb{E}_\nu^\pi\left[u\left(\sum_{t=0}^{T}\gamma^t c\left(s_t,a_t\right)\right)\right] - \mathbb{E}_\nu^\pi\left[u\left(\sum_{t=0}^{\infty}\gamma^t c\left(s_t,a_t\right)\right)\right]| \\
&\leq \mathbb{E}_\nu^\pi\left[|u\left(\sum_{t=0}^{T}\gamma^t c\left(s_t,a_t\right)\right) - u\left(\sum_{t=0}^{\infty}\gamma^t c\left(s_t,a_t\right)\right)|\right] \\
&\leq \mathbb{E}_\nu^\pi\left[|\sum_{t=T+1}^{\infty}\gamma^t c\left(s_t,a_t\right)|\right],
\end{aligned}$$

where the last term can be made arbitrarily small by taking $T \to \infty$. In the second inequality, we are using Lipschitz continuity, i.e. $|u\left(x\right) - u\left(y\right)| \leq |x-y|$ for all x and y. ∎

As for the general Problem (2.2), we fix the time horizon T and consider the truncated problem

$$\inf_{\pi\in\Pi}\mathbb{E}\left[u\left(C_{\nu,\,T}^{\pi}\right)\right], \tag{5.2}$$

which can be solved exactly with convex analytic methods. Problem (5.2) is equivalent to the following optimization problem:

$$\inf_{\mu\in\mathcal{M}\left(\tilde{\mathbb{K}}\right),\,\theta\in\mathcal{M}\left(\mathcal{Y}\right)}\left\{\mathbb{E}\left[u\left(X\left(\theta\right)\right)\right] \,:\, L_{0,\,T}\mu = \nu,\, L_{1,\,T}\mu = \theta\right\}, \tag{5.3}$$

where $L_{0,\,T}$ is defined in (3.2) and $L_{1,\,T}$ is defined in (3.4). By definition of $X\left(\theta\right)$, the objective $\mathbb{E}\left[u\left(X\left(\theta\right)\right)\right] = \int u\left(y\right)\theta\left(dy\right) = \langle\theta,\, u\rangle$ is linear in θ. Thus, we see that Problem (5.3) is actually a linear programming problem. We remind the reader that Problem (5.3) gives rise to a deterministic optimal policy by [34, Theorem 19] because it is linear.

Remark 5.2. The discretized finite-dimensional version of the expected utility mini-

mizing MDP is

$$\begin{aligned}
&\inf_{\mu,\,\theta} \sum_{y\in\mathbb{Y}} u\left(y\right)\theta\left(y\right) && (5.4)\\
&\text{s.t. } \nu\left(j\right) = \sum_{a\in\mathbb{A}} \mu\left(j,\,a\right) - \gamma \sum_{(s,y,z,a)\in\bar{\mathbb{K}}} \bar{Q}_T\left(j \mid s,y,z,a\right)\mu\left(s,y,z,a\right), && \forall j\in\bar{\mathbb{S}},\\
&\qquad \theta\left(\xi\right) = \frac{1-\gamma}{\gamma^T} \sum_{(s,y,z,a)\in\bar{\mathbb{K}}} I\left\{y=\xi,\,z=\gamma^T\right\}\mu\left(s,y,z,a\right), && \forall \xi\in\mathbb{Y},
\end{aligned}$$

where we are using

$$\hat{\mu}\left(j\right) = \sum_{a\in\mathbb{A}} \mu\left(j,a\right)$$

for all $j\in\bar{\mathbb{S}}$.

The next step in our analysis is to take the dual of Problem (5.3). Duality is helpful here in two ways. First, it enhances our understanding of Problem (5.3) by providing a certificate of optimality. Second, it reveals a linear programming problem in value functions for Problem (5.3). To proceed, we define the adjoint of the linear operator $L_{1,\,T}$. Let $\mathcal{F}\left(\mathbb{R}\right)$ be the space of bounded measurable functions $f:\mathbb{R}\to\mathbb{R}$.

Lemma 5.3. *The adjoint of $L_{1,\,T}$ is $L^*_{1,\,T}:\mathcal{F}\left(\mathbb{R}\right)\to\mathcal{F}\left(\tilde{\mathbb{K}}\right)$ defined by*

$$\left[L^*_{1,\,T}f\right]\left(s,y,z,a\right) := \frac{1-\gamma}{\gamma^T} f\left(y\right) I\left\{z\le\gamma^T\right\},\ \forall\left(s,y,z,a\right)\in\tilde{\mathbb{K}}.$$

Proof. The adjoint is defined by
$\langle f,\,L_{1,T}\mu\rangle = \langle L^*_{1,T}f,\,\mu\rangle$, so we compute

$$\begin{aligned}
\langle f,\,L_{1,\,T}\mu\rangle &= \int_{\mathbb{R}} f\left(y\right)\left[L_{1,\,T}\mu\right]\left(dy\right)\\
&= \int_{\mathbb{R}} f\left(y'\right)\left[\frac{1-\gamma}{\gamma^T}\int_{\tilde{\mathbb{K}}} I\left\{y=y',\,z\le\gamma^T\right\}\mu\left(d\left(s,y,z,a\right)\right)\right]dy'\\
&= \int_{\tilde{\mathbb{K}}}\left[\int_{\mathbb{R}} \frac{1-\gamma}{\gamma^T} f\left(y'\right) I\left\{y=y',\,z\le\gamma^T\right\}dy'\right]\mu\left(d\left(s,y,z,a\right)\right)\\
&= \int_{\tilde{\mathbb{K}}}\left[\frac{1-\gamma}{\gamma^T} f\left(y\right) I\left\{z\le\gamma^T\right\}\right]\mu\left(d\left(s,y,z,a\right)\right).
\end{aligned}$$

□

We are now ready to compute the Lagrangian dual of Problem (5.3). This derivation follows from the infinite dimensional linear programming theory (see [2]). We are assured that the Lagrangian dual will be a linear programming problem because the primal Problem (5.3) is a linear programming problem. Define $\mathcal{F}\left(\tilde{\mathbb{S}}\right)$ to be the space of bounded measurable functions $f:\tilde{\mathbb{S}}\to\mathbb{R}$. The value function for Problem (5.2) exists in the space $\mathcal{F}\left(\tilde{\mathbb{S}}\right)$, and the following dual problem is an optimization problem in value functions.

Theorem 5.4. *The dual to Problem (5.3) is*

$$\sup_{v \in \mathcal{F}\left(\tilde{\mathbb{S}}\right)} \langle v,\, \nu \rangle \tag{5.5}$$

$$\begin{aligned} s.t.\ \ v\left(s,y,z\right) &\leq \gamma \int_{\tilde{\mathbb{S}}} v\left(\xi\right) \tilde{Q}_T\left(d\xi \mid s,y,z,a\right) \\ &+ \frac{1-\gamma}{\gamma^T} u\left(y\right) I\left\{z \leq \gamma^T\right\}, \qquad \forall \left(s,y,z,a\right) \in \tilde{\mathbb{K}}. \end{aligned} \tag{5.6}$$

Proof. Let the Lagrange multiplier for constraint $L_{0,\,T}\mu = \nu$ be $v \in \mathcal{F}\left(\tilde{\mathbb{S}}\right)$ and let the Lagrange multiplier for constraint $L_{1,\,T}\mu = \theta$ be $w \in \mathcal{F}\left(\mathbb{R}\right)$, then the Lagrangian for Problem (5.3) is

$$\Psi\left(\mu,\, \theta,\, v,\, w\right) = \mathbb{E}\left[u\left(X\left(\theta\right)\right)\right] + \langle v,\, L_{0,\,T}\mu - \nu \rangle + \langle w,\, L_{1,\,T}\mu - \theta \rangle.$$

Problem (5.3) is equivalent to

$$\inf_{\theta,\, \mu \geq 0} \sup_{v,\, w} \Psi\left(\mu,\, \theta,\, v,\, w\right),$$

so the dual problem is defined to be

$$\sup_{v,\, w} \inf_{\theta,\, \mu \geq 0} \Psi\left(\mu,\, \theta,\, v,\, w\right).$$

Rearranging the Lagrangian

$$\begin{aligned} \Psi\left(\mu,\, \theta,\, v,\, w\right) &= \langle \theta,\, u \rangle + \langle v,\, L_{0,\,T}\mu - \nu \rangle + \langle w,\, L_{1,\,T}\mu - \theta \rangle \\ &= \langle \theta,\, u - w \rangle + \langle \mu,\, L^*_{0,\,T} v + L^*_{1,\,T} w \rangle - \langle v,\, \nu \rangle \end{aligned}$$

we see that the dual problem is

$$\begin{aligned} &\sup_{v,\, w} \ -\langle v,\, \nu \rangle \\ &\text{s.t.}\, u - w \geq 0, \\ &\qquad L^*_{0,\,T} v + L^*_{1,\,T} w \geq 0. \end{aligned}$$

The adjoint of $L_{0,\,T}$ is $L^*_{0,\,T} : \mathcal{F}\left(\tilde{\mathbb{S}}\right) \to \mathcal{F}\left(\tilde{\mathbb{K}}\right)$ defined by

$$\left[L^*_{0\,T} h\right]\left(s,y,z,a\right) := h\left(s,y,z\right) - \gamma \int_{\tilde{\mathbb{S}}} h\left(\xi\right) \tilde{Q}_T\left(d\xi \mid s,y,z,a\right),$$

(see [19] for example.) We have the adjoint of $L_{1,\,T}$ from Lemma 5.3, which gives the form of the dual stated above after switching the sign of the unrestricted value function $v = -v$. ∎

Problem (5.5) - (5.6) is a maximization problem which drives the components of v to be as large as possible. Thus, constraint (5.6) must be binding at optimality for some action $a \in \mathbb{A}$ for every state $(s,y,z) \in \tilde{\mathbb{S}}$, or else we could increase v further. We then see that the dynamic programming equations on the augmented state space $\tilde{\mathbb{S}}$ are

$$v\left(s,y,z\right) = \inf_{a \in A(s)} \left\{ \frac{1-\gamma}{\gamma^T} u\left(y\right) I\left\{z \leq \gamma^T\right\} + \gamma \int_{\tilde{\mathbb{S}}} v\left(\xi\right) \widetilde{Q}_T\left(d\xi \mid s,a\right) \right\}, \forall \left(s,y,z\right) \in \tilde{\mathbb{S}}. \tag{5.7}$$

The term $\frac{1-\gamma}{\gamma^T}u(y)\,I\left\{z\leq\gamma^T\right\}$ appears as a cost function on the augmented state space, the original cost function c is absorbed into the transition kernel. We emphasize that the preceding Bellman equations are stationary on the augmented state space.

Using the dynamic programming equations (5.7), we can write the value function for Problem (5.2) as

$$v^{\pi}(s,0,1)=\inf_{\pi\in\Pi}\mathbb{E}_{\nu}^{\pi}\left[\sum_{t=0}^{\infty}\frac{1-\gamma}{\gamma^T}u(y_t)\,I\left\{z_t\leq\gamma^T\right\}\right],\ \forall s\in\mathbb{S}.$$

On the augmented state space, we have a cost function $\frac{1-\gamma}{\gamma^T}u(y)\,I\left\{z\leq\gamma^T\right\}$ that depends on y and z but not s and a, since the original cost function was absorbed into the transition kernel. The new cost function $\frac{1-\gamma}{\gamma^T}u(y)\,I\left\{z\leq\gamma^T\right\}$ is always zero up until time T, after which it accounts for the running cost.

Recall that a problem is 'solvable' when its optimal value is attained, there is 'no duality gap' between a primal and its dual when both problems have the same optimal value, and 'strong duality' holds between a primal and its dual when both problems are solvable. To conclude this section, we turn to the issues of solvability and strong duality for Problems (5.3) and (5.5) - (5.6). The following technical conditions ensure that Problem (5.3) is solvable, and that strong duality holds.

Assumption 5.5. *(i) The cost function* $\frac{1-\gamma}{\gamma^T}u(y)\,I\left\{z\leq\gamma^T\right\}:\tilde{\mathbb{K}}\to\mathbb{R}$ *is inf-compact, i.e., the level sets*

$$\left\{(s,y,z,a)\in\tilde{\mathbb{K}}:\ \frac{1-\gamma}{\gamma^T}u(y)\,I\left\{z\leq\gamma^T\right\}\leq\epsilon\right\}$$

are compact for all $\epsilon\geq 0$.

(ii) The transition law $\tilde{Q}_T$ *is weakly continuous, i.e.*

$$(s,y,z,a)\to\int_{\mathbb{S}}v(s)\,\tilde{Q}_T(ds\,|\,s,y,z,a)\in\mathcal{C}_b\left(\tilde{\mathbb{K}}\right),\ \forall v\in\mathcal{C}_b\left(\tilde{\mathbb{S}}\right),$$

where $\mathcal{C}_b\left(\tilde{\mathbb{K}}\right)$ *and* $\mathcal{C}_b\left(\tilde{\mathbb{S}}\right)$ *denote the space of continuous and uniformly bounded functions on* $\tilde{\mathbb{K}}$ *and* $\tilde{\mathbb{S}}$*, respectively.*

(iii) There exists a uniformly bounded minimizing sequence $\left\{v^i\right\}_{i\geq 0}\subset\mathcal{F}\left(\tilde{\mathbb{S}}\right)$ *for Problem (5.5) - (5.6).*

The preceding assumptions are standard in the literature on the convex analytic approach to MDPs (see the monographs [22, 23] for a summary). The next theorem summarizes well-known solvability and duality results for infinite-dimensional LPs, as applied to Problems (5.3) and (5.5) - (5.6). As a reminder, a primal and its dual have no duality gap when their optimal values are equal, and strong duality holds when both optimal values are attained.

Theorem 5.6. *(i) ([19, Theorem 3.2]) Under Assumptions 2.1 and 5.5(i)(ii), Problem (5.3) is solvable.*

(ii) [19, Theorem 4.6] Under Assumptions 2.1 and 5.5(i)(ii), there is no duality gap between Problem (5.3) and Problem (5.5) - (5.6).

(iii) [2, Theorem 3.9] Under Assumptions 2.1 and 5.5(i)(ii)(iii), strong duality holds between Problems (5.3) and (5.5) - (5.6).

When strong duality holds, the optimal values of Problems (5.3) and (5.5) - (5.6) are equal. In this situation, we can recover an optimal policy for Problem (5.2) by

solving either the primal problem in occupation measures or the dual problem in value functions. If μ^* is an optimal solution to Problem (5.3), then $\pi^* \in \Pi$ defined by

$$\pi^*(B \mid s, y, z) = \mu^*(B \mid s, y, z), \forall B \in \mathcal{B}(\mathbb{A})$$

is an optimal policy for Problem (5.2). Conversely, if v^* is an optimal solution to Problem (5.5) - (5.6), then a greedy policy with respect to v^*,

$$\pi^*(s, y, z) \in \arg\min_{a \in A(s)} \left\{ \frac{1-\gamma}{\gamma^T} u(y) I\left\{z \leq \gamma^T\right\} + \gamma \int_{\tilde{\mathbb{S}}} v^*(\xi) \tilde{Q}_T(d\xi \mid s, a) \right\},$$

is an optimal policy for Problem (5.2).

Remark 5.7. As we mentioned at the beginning of this subsection, Problem (5.1) is studied with dynamic programming methods in [4]. Our results here are of a very different flavor. First, we solve Problem (5.1) with the convex analytic approach while [4] develops value iteration and policy iteration algorithms for Problem (5.1). Our present paper and [4] share the same type of history augmentation, which necessitates an uncountable state space. However, we are able to provide a finite approximation and error guarantees while the algorithms in [4] are purely conceptual.

5.2. Conditional value-at-risk. CVaR is among the most popular risk functions, the CVaR-minimizing policy at level $\beta \in (0,1)$ solves

$$\begin{aligned} &\inf_{\pi \in \Pi} \inf_{\eta \in \mathbb{R}} \left\{ \eta + \frac{1}{1-\beta} \mathbb{E}\left[(C_\nu^\pi - \eta)_+\right] \right\} \\ &= \inf_{\pi \in \Pi} \inf_{\eta \in \mathbb{R}} \left\{ \eta + \frac{1}{1-\beta} \mathbb{E}_\nu^\pi \left[\left(\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) - \eta \right)_+ \right] \right\}. \end{aligned} \tag{5.8}$$

The dependence on π enters through the expectation inside the minimization problem. Problem (5.8) is solved with state space augmentation and dynamic programming algorithms in [3]. In contrast, we provide a solution via the convex analytic approach.

Problem (5.1) naturally leads to a linear programming problem in occupation measures. Here we will see that the convex analytic formulation for Problem (5.8) is nonconvex. This fact is in contrast to stochastic optimization with CVaR, which gives convex optimization problems.

The next lemma shows that we do not have to minimize over all $\eta \in \mathbb{R}$ when evaluating the CVaR of $C_{\nu,T}^\pi$, we only need to minimize over $\eta \in \mathcal{Y}$.

Lemma 5.8. *For any $X \in \mathcal{L}$ with support contained in $\mathcal{Y}$,*

$$\inf_{\eta \in \mathbb{R}} \left\{ \eta + \frac{1}{1-\beta} \mathbb{E}\left[(X-\eta)_+\right] \right\} = \inf_{\eta \in \mathcal{Y}} \left\{ \eta + \frac{1}{1-\beta} \mathbb{E}\left[(X-\eta)_+\right] \right\}.$$

Proof. For $\eta < 0$, we have

$$\eta + \frac{1}{1-\beta} \mathbb{E}\left[(X-\eta)_+\right] = \eta + \frac{1}{1-\beta} \mathbb{E}[X-\eta] = \left(\frac{-\beta}{1-\beta}\right)\eta + \frac{1}{1-\beta}\mathbb{E}[X],$$

which is increasing as $\eta \to -\infty$. For $\eta > \bar{c}/(1-\gamma)$, we have

$$\eta + \frac{1}{1-\beta} \mathbb{E}\left[(X-\eta)_+\right] = \eta,$$

which is increasing as $n \to \infty$. ∎

Now we check Assumption 3.2 for CVaR to get error bounds.

Lemma 5.9. *[16, Lemma 4.3] The risk function* $\rho(\cdot) = \inf_{\eta\in\mathbb{R}} \left\{\eta + \frac{1}{1-\beta}\mathbb{E}\left[(\cdot-\eta)_+\right]\right\}$ *satisfies Assumption 3.2.*

For a fixed time horizon T, the truncated problem is

$$\inf_{\pi\in\Pi}\inf_{\eta\in\mathbb{R}}\left\{\eta + \frac{1}{1-\beta}\mathbb{E}\left[\left(C^{\pi}_{\nu,T}-\eta\right)_+\right]\right\}. \tag{5.9}$$

Problem (5.9) can be expressed as a static problem in occupation measures:

$$\inf_{\mu\in\mathcal{M}\left(\tilde{\mathbb{K}}\right),\,\theta\in\mathcal{M}(\mathcal{Y})}\left\{\inf_{\eta\in\mathbb{R}}\left\{\eta + \frac{1}{1-\beta}\mathbb{E}\left[(X(\theta)-\eta)_+\right]\right\} \,:\, L_{0,T}\mu=\nu,\, L_{1,T}\mu=\theta\right\}, \tag{5.10}$$

by Theorem 3.5 and Lemma 3.6. By definition, $\mathbb{E}\left[(X(\theta)-\eta)_+\right] = \int_{\mathcal{Y}}(y-\eta)_+\,\theta(dy)$, so we define

$$g(\theta) := \inf_{\eta\in\mathbb{R}}\left\{\eta + \frac{1}{1-\beta}\int_{\mathcal{Y}}(y-\eta)_+\,\theta(dy)\right\}$$

to be the CVaR objective. Then Problem (5.10) can be written simply as

$$\inf_{\mu\in\mathcal{M}\left(\tilde{\mathbb{K}}\right),\,\theta\in\mathcal{M}(\mathcal{Y})}\left\{g(\theta) \,:\, L_{0,T}\mu=\nu,\, L_{1,T}\mu=\theta\right\}.$$

Problem (5.10) is a convex analytic formulation for Problem (5.8), but it is not a convex optimization problem - it has a convex feasible region but a concave minimization objective. However, since the objective of Problem (5.10) is concave, we are still guaranteed the existence of a deterministic optimal policy by [34, Theorem 19].

Remark 5.10. When $\mathbb{S}$ and $\mathbb{A}$ are finite, we can use the discretization from Subsection 4.2 to approximate Problem (5.9) with

$$\begin{aligned}
&\inf_{\mu,\,\theta}\inf_{\eta\in\mathcal{Y}}\left\{\eta + \frac{1}{1-\beta}\sum_{y\in\mathbb{Y}}(y-\eta)_+\,\theta(y)\right\} && (5.11)\\
&\text{s.t. } \nu(j) = \hat{\mu}(j) - \gamma\sum_{(s,y,z,a)\in\bar{\mathbb{K}}}\bar{Q}_T(j\,|\,s,y,z,a)\,\mu(s,y,z,a), && \forall j\in\bar{\mathbb{S}},\\
&\theta(\xi) = \frac{1-\gamma}{\gamma^T}\sum_{(s,y,z,a)\in\bar{\mathbb{K}}} I\left\{y=\xi,\, z=\gamma^T\right\}\mu(s,y,z,a), && \forall \xi\in\mathbb{Y}.
\end{aligned}$$

Since the preceding problem is finite-dimensional, we can apply the algorithm from [30] directly, rather than applying the algorithm to each finite approximation in problems $\boldsymbol{P}(C_{0,k},\, C_{1,k},\, \epsilon_k,\, \Delta_{0,l},\, \Delta_{1,l})$ and taking $k,\, l\to\infty$.

The next lemma concerns properties of g, its concavity and its continuity with respect to the weak topology on $\mathcal{M}(\mathcal{Y})$. Part (i) is by definition, the proof of part (ii) follows by an approximation argument with a finite set $N\subset\mathcal{Y}$.

Lemma 5.11. *(i)* $g(\cdot)$ *is concave in* θ.

(ii) $g(\cdot)$ *is weak-continuous.*

In this case, Problem $\boldsymbol{P}(C_{0,k},\, C_{1,k},\, \epsilon_k,\, \Delta_{0,l},\, \Delta_{1,l})$ takes the form:

$$\inf_{\mu\in\mathcal{M}\left(\tilde{\mathbb{K}}\right),\,\theta\in\mathcal{M}(\mathcal{Y})} g\left(\theta\right) \tag{5.12}$$

$$\text{s.t.}\,\left|\langle L_{0,\,T}\mu-\tilde{\nu},\,f\rangle\right|\leq\epsilon_k, \qquad \forall f\in C_{0,k}, \tag{5.13}$$

$$\left|\langle L_{1,\,T}\mu-\theta,\,f\rangle\right|\leq\epsilon_k, \qquad \forall f\in C_{1,k}, \tag{5.14}$$

$$\mu\in\Delta_{0,l},\,\theta\in\Delta_{1,l}, \tag{5.15}$$

where $\Delta_{0,l}\subset\mathcal{M}\left(\tilde{\mathbb{K}}\right)$ and $\Delta_{1,l}\subset\mathcal{M}\left(\mathcal{Y}\right)$ are the sets of probability measures with finite support defined in Subsection 4.1. The preceding problem has finitely many decision variables and finitely many constraints. We can use the successive linear approximation method to solve Problem (5.12) - (5.15) exactly, see [30]. In the finite setting, the subgradient is

$$\partial g\left(\hat{\theta}\right)=\operatorname{conv}\left\{\left\{\left(y-\eta\right)_+\right\}_{y\in\Delta_{1,k}}\,:\,g\left(\hat{\theta}\right)=\eta+\frac{1}{1-\beta}\sum_{y\in\Delta_{1,k}}\left(y-\eta\right)_+\hat{\theta}\left(y\right)\right\}$$

$\subset\mathbb{R}^{|\Delta_{1,k}|}$, where we are viewing g as a function on probability distributions on $\Delta_{1,k}$. The idea is that at each candidate point, we will linearize the objective of Problem (5.12) - (5.15) and then solve the resulting LP to get the next candidate point. This procedure is justified since we know the optimal solution of a concave minimization problem will be found at an extreme point of the feasible region. We emphasize that successive linear approximation is being applied to the discretized Problem (5.12) - (5.15), not the infinite-dimensional Problem (5.10). Let $\left(\mu^i,\,\theta^i\right)$ be the i^{th} candidate solution to Problem (5.12) - (5.15), then the linearization of Problem (5.12) - (5.15) at $\left(\mu^i,\,\theta^i\right)$ is

$$\inf_{\mu\in\mathcal{M}\left(\tilde{\mathbb{K}}\right),\,\theta\in\mathcal{M}(\mathcal{Y})} \langle s^i,\,\theta-\theta^i\rangle$$

$$\begin{aligned}\text{s.t.}\,&\left|\langle L_{0,\,T}\mu-\nu,\,f\rangle\right|\leq\epsilon_k, && \forall f\in C_{0,k},\\ &\left|\langle L_{1,\,T}\mu-\theta,\,f\rangle\right|\leq\epsilon_k, && \forall f\in C_{1,k},\\ &\mu\in\Delta_{0,k},\,\theta\in\Delta_{1,k},\end{aligned}$$

where $s^i\in\partial g\left(\theta^i\right)$. The solution of the preceding problem becomes the next candidate solution $\left(\mu^{i+1},\,\theta^{i+1}\right)$. It is shown in [30] that this procedure will converge to the optimal solution of Problem (5.12) - (5.15).

Remark 5.12. Instead of minimizing the CVaR of costs, we could maximize the CVaR of reward. For a reward function $r:\mathbb{K}\to\mathbb{R}$, we can let $y_t=\sum_{i=0}^{t}\gamma^i r\left(s_i,a_i\right)$ be the running reward. Using our same truncation argument, we get the optimization problem

$$\sup_{\mu\in\mathcal{M}\left(\tilde{\mathbb{K}}\right),\,\theta\in\mathcal{M}(\mathcal{Y})}\left\{\inf_{\eta\in\mathbb{R}}\left\{\eta+\frac{1}{1-\beta}\int\left(y-\eta\right)_+\theta\left(dy\right)\right\}\,:\,L_{0,\,T}\mu=\nu,\,L_{1,\,T}\mu=\theta\right\}.$$

We have already established that CVaR is a concave function of θ, thus this problem is automatically convex since it is maximizing a concave function.

Remark 5.13. We can treat mean-deviation and mean-semideviation in the same manner as above. The resulting static problems in occupation measures are nonconvex.

The situation is mitigated somewhat by the fact that the feasible regions are determined by linear constraints, and thus the feasible region is convex. The only nonconvexity is in the objectives.

6. Risk constrained optimization. We have so far addressed MDP models with minimization of some risk function of the infinite horizon discounted cost. In this section, we extend our development to risk-constrained MDPs. For an additional risk function $\vartheta : \mathcal{L} \to \mathbb{R}$ and a constant κ, we can add a constraint to Problem (2.2) to get

$$\inf_{\pi \in \Pi} \left\{ \rho\left(C_\nu^\pi\right) \, : \, \vartheta\left(C_\nu^\pi\right) \leq \kappa \right\}. \tag{6.1}$$

We will study two specific instances of Problem (6.1) in this section: one based on stochastic dominance and the other on chance constraints.

6.1. Stochastic dominance constraints. Stochastic dominance relations (or stochastic orders) are partial orders on the space of random variables (see [33, 40]). They have major relevance to risk management because they allow us to express risk preferences for an entire *class* of decision makers, as opposed to Problem (5.1) which represents only a single decision maker. Optimization with stochastic dominance constraints was addressed in [10, 11, 12, 18].

Definition 6.1. For random variables $X, Y \in \mathbb{R}$, X is *dominated* by Y in the *increasing convex stochastic order*, written $X \leq_{icx} Y$, if $\mathbb{E}\left[u\left(X\right)\right] \leq \mathbb{E}\left[u\left(Y\right)\right]$ for all increasing convex functions $u : \mathbb{R} \to \mathbb{R}$ such that both expectations exist.

If $X \leq_{icx} Y$, then any risk-averse decision maker with an increasing convex utility function would prefer the random variable X to Y. We will assume that there is a reference random variable Y that serves as a benchmark. The benchmark Y expresses the user's desiderata regarding the properties of a favorable cost distribution.

Fortunately, a computationally tractable representation of $\leq_{icx}$ is available. Denote $(x)_+ = \max\{x, 0\}$. It is known (see [10]) that in this case $X \leq_{icx} Y$ is equivalent to

$$\mathbb{E}\left[(X - \eta)_+\right] \leq \mathbb{E}\left[(Y - \eta)_+\right], \forall \eta \in \mathbb{R}.$$

The above parametric representation of $\leq_{icx}$ is easier to implement than its original definition, since it reduces to a continuum of constraints indexed by a single parameter, whereas the original definition is a continuum of constraints indexed by a function space. We can also write $X \leq_{icx} Y$ as the system

$$\mathbb{E}\left[(X - \eta)_+\right] \leq \mathbb{E}\left[(Y - \eta)_+\right], \forall \eta \in \operatorname{supp} Y,$$

where $\operatorname{supp} Y$ is the support of Y. Furthermore, when $\operatorname{supp} Y = \{\eta_i\}_{i \in I}$ for a finite index I, then $X \leq_{icx} Y$ is equivalent to

$$\mathbb{E}\left[(X - \eta_i)_+\right] \leq \mathbb{E}\left[(Y - \eta_i)_+\right], \forall i \in I.$$

The next assumption about the boundedness of the benchmark streamlines our analysis, and is not unreasonable in practice.

Assumption 6.2. *The support of the benchmark Y is contained in $[\eta_1, \eta_2]$.*

With these ingredients, we propose the stochastic dominance-constrained MDP:

$$\begin{aligned} &\inf_{\pi\in\Pi} \mathbb{E}_\nu^\pi \left[\sum_{t=0}^{\infty} \gamma^t c\left(s_t, a_t\right)\right] \\ &\text{s.t. } \mathbb{E}_\nu^\pi \left[\left(\sum_{t=0}^{\infty} \gamma^t c\left(s_t, a_t\right) - \eta\right)_+\right] \leq \mathbb{E}\left[\left(Y-\eta\right)_+\right], \forall \eta \in [\eta_1, \eta_2]. \end{aligned} \tag{6.2}$$

Equivalently,

$$\inf_{\pi\in\Pi} \left\{\mathbb{E}\left[C_\nu^\pi\right] \,:\, \mathbb{E}\left[\left(C_\nu^\pi - \eta\right)_+\right] \leq \mathbb{E}\left[\left(Y-\eta\right)_+\right], \forall \eta \in [\eta_1, \eta_2]\right\}.$$

In our earlier work [18], we applied stochastic dominance constraints to the steady state in the classic convex analytic formulation. Problem (6.2) differs substantially because it has stochastic dominance constraints on the discounted cost over the entire time horizon.

We will see that some of the results and proofs for Problem (6.2) in this section are similar to those for Problem (5.1). The connection between Problem (6.2) and Problem (5.1) is natural: stochastic dominance constraints are defined in terms of a continuum of constraints on the expected utility of cost, while Problem (5.1) minimizes the expected utility for a single utility function. Our overall plan is the same, we will approximate Problem (6.2) with a truncated problem that can be solved exactly with convex analytic methods.

Problem (6.2) is constrained, so we cannot apply Lemma 3.4 to get error bounds. We also need to consider satisfaction of the constraints. The following result is immediate since $(x-\eta)_+$ is Lipschitz continuous for all η. Recall $y_T = \sum_{t=0}^{T} \gamma^t c\left(s_t, a_t\right)$ is the running cost at time T.

Lemma 6.3. *For any $\epsilon > 0$, there is a T such that*

$$\left|\left(y_T - \eta\right)_+ - \left(\sum_{t=0}^{\infty} \gamma^t c\left(s_t, a_t\right) - \eta\right)_+\right| \leq \epsilon, \forall \eta \in [\eta_1, \eta_2],$$

for all $\pi \in \Pi$.

We will approximate Problem (6.2) with the truncation

$$\inf_{\pi\in\Pi} \left\{\mathbb{E}\left[C_{\nu,T}^\pi\right] \,:\, \mathbb{E}\left[\left(C_{\nu,T}^\pi - \eta\right)_+\right] \leq \mathbb{E}\left[\left(Y-\eta\right)_+\right], \forall \eta \in [\eta_1, \eta_2]\right\}. \tag{6.3}$$

We use this to get an estimate on the quality (both in terms of optimality and feasibility) of a solution to Problem (6.3) versus Problem (6.2).

The next lemma states that a near optimal solution for Problem (6.3) will be nearly optimal and nearly feasible for Problem (6.2). Let Δ_{SD} and $\Delta_{SD,T}$ denote the feasible regions of Problems (6.2) and (6.3), respectively. The next proof is immediate.

Lemma 6.4. *Choose T as in the statement of Lemma 6.3, then:*

(i) $\left|\mathbb{E}\left[\left(C_{\nu,T}^\pi - \eta\right)_+\right] - \mathbb{E}\left[\left(C_\nu^\pi - \eta\right)_+\right]\right| < \epsilon, \forall \eta \in [\eta_1, \eta_2]$*;*

(ii) For $\hat{\pi} \in \Delta_{SD,T}$ with $\mathbb{E}\left[C_{\nu,T}^{\hat{\pi}}\right] \leq \inf_{\pi\in\Delta_{SD,T}} \mathbb{E}\left[C_{\nu,T}^\pi\right] + \epsilon$ we have $\mathbb{E}\left[C_\nu^{\hat{\pi}}\right] < \inf_{\pi\in\Delta_{SD}} \mathbb{E}\left[C_\nu^\pi\right] + 3\,\epsilon$ and $\mathbb{E}\left[\left(C_\nu^{\hat{\pi}} - \eta\right)_+\right] \leq \mathbb{E}\left[\left(Y-\eta\right)_+\right] + \epsilon$ for all $\eta \in [\eta_1, \eta_2]$.

Note that Lemma 6.4 guarantees near-feasibility of an optimal solution to Problem (6.3) with respect to Problem (6.2), rather than exact feasibility.

To succinctly express the stochastic dominance constraint, let $\mathcal{C}([\eta_1, \eta_2])$ be the space of continuous functions on $[\eta_1, \eta_2]$ in the supremum norm. We define a new linear operator $L_2 : \mathcal{M}(\mathcal{Y}) \to \mathcal{C}([\eta_1, \eta_2])$ via

$$[L_2\theta](\eta) := \int_{\mathcal{Y}} (y-\eta)_+ \, \theta(dy), \, \forall \eta \in [\eta_1, \eta_2],$$

which gives the vector of expected utilities $\mathbb{E}\left[(X(\theta)-\eta)_+\right]$ as η ranges over $[\eta_1, \eta_2]$. Finally, let $g \in \mathcal{C}([\eta_1, \eta_2])$ be defined by

$$g(\eta) = \mathbb{E}\left[(Y-\eta)_+\right], \, \forall \eta \in [\eta_1, \eta_2],$$

the vector of expected utilities of the benchmark over $\eta \in [\eta_1, \eta_2]$. Using L_2, we can write Problem (6.3) in convex analytic terms as

$$\inf_{\mu \in \mathcal{M}(\bar{\mathbb{K}}), \, \theta \in \mathcal{M}(\mathcal{Y})} \{\mathbb{E}[X(\theta)] \, : \, L_{0,T}\mu = \nu, \, L_{1,T}\mu = \theta, \, L_2\theta \geq g\}. \tag{6.4}$$

By the same reasoning as for Problem (5.3), Problem (6.4) is a linear programming problem.

Remark 6.5. The discretized finite-dimensional version of the stochastic dominance-constrained MDP is given by

$$\begin{aligned}
&\inf_{\mu, \, \theta} \sum_{y \in \mathbb{Y}} y \, \theta(y) && (6.5)\\
&\text{s.t.} \, \nu(j) = \sum_{a \in \mathbb{A}} \mu(j, \, a) - \gamma \sum_{(s,y,z,a) \in \bar{\mathbb{K}}} \bar{Q}_T(j \mid s, y, z, a) \, \mu(s, y, z, a), && \forall j \in \bar{\mathbb{S}},\\
&\quad \theta(\xi) = \frac{1-\gamma}{\gamma^T} \sum_{(s,y,z,a) \in \bar{\mathbb{K}}} I\left\{y = \xi, \, z = \gamma^T\right\} \mu(s, y, z, a), && \forall \xi \in \mathbb{Y},\\
&\quad \sum_{y \in \mathbb{Y}} (y - \eta_i)_+ \, \theta(y) \leq \mathbb{E}\left[(Y - \eta_i)_+\right], && \forall i \in I,
\end{aligned}$$

where we are assuming the benchmark has finite support indexed by I.

The following lemma connects Problem (6.4) with Problem (6.3).

Lemma 6.6. *If π is optimal for Problem (6.3), then μ_ν^π is optimal for Problem (6.4). Conversely, if μ is optimal for Problem (6.4), then π_μ is optimal for Problem (6.3).*

Proof. We use the fact that $\mathbb{E}\left[\left(C_{\nu,T}^{\pi} - \eta\right)_+\right] = \mathbb{E}\left[(X(\mu_\nu^\pi) - \eta)_+\right]$ and also that $\mathbb{E}\left[\left(C_{\nu,T}^{\pi_\mu} - \eta\right)_+\right] = \mathbb{E}\left[(X(\mu) - \eta)_+\right]$ for all $\eta \in [\eta_1, \eta_2]$. Since,

$$\inf_{\pi \in \Delta_{SD,\,T}} \mathbb{E}\left[C_{\nu,T}^{\pi}\right] = \inf_{\pi \in \Delta_{SD,\,T}} \mathbb{E}\left[X(\mu_\nu^\pi)\right],$$

it follows that

$$\inf_{\pi \in \Delta_{SD,\,T}} \mathbb{E}\left[X(\mu_\nu^\pi)\right] = \inf_{\mu \in \mathcal{M}(\bar{\mathbb{K}}), \, \theta \in \mathcal{M}(\mathcal{Y})} \{\mathbb{E}[X(\theta)] \, : \, L_{0,T}\mu = \nu, \, L_{1,T}\mu = \theta, \, L_2\theta \geq g\},$$

since μ_ν^π is feasible for Problem (6.4) for any $\pi \in \Delta_{SD,T}$, and $\pi_\mu \in \Delta_{SD,T}$ for any μ feasible to Problem (6.4). So, the optimal values of Problem (6.3) and Problem (6.4) are equal. ∎

We now derive the dual to Problem (6.4). As for Problem (5.3), duality can be used to get a certificate of optimality for Problem (6.4). Additionally, dynamic programming equations will emerge, although there is a significant difference between the dynamic programming equations for the unconstrained Problem (5.2) versus the constrained Problem (6.4). We first compute the adjoint of L_2 since it will appear in the dual to Problem (6.4).

Lemma 6.7. *The adjoint of L_2 is $L_2^* : \mathcal{M}([\eta_1, \eta_2]) \to \mathcal{F}(\mathbb{R})$ defined by*

$$[L_2^*\Lambda](\eta) := \int_{\eta_1}^{\eta_2} (y-\eta)_+ \Lambda(d\eta), \ \forall \eta \in \mathbb{R}.$$

Proof. Take $\Lambda \in \mathcal{M}([\eta_1, \eta_2])$, then

$$\langle \Lambda, L_2\theta \rangle = \int_{\eta_1}^{\eta_2} \left[\int_{\mathcal{Y}} (y-\eta)_+ \theta(dy) \right] \Lambda(d\eta) = \int_{\mathcal{Y}} \left[\int_{\eta_1}^{\eta_2} (y-\eta)_+ \Lambda(d\eta) \right] \theta(dy),$$

by Fubini's theorem. ∎

We report the dual of Problem (6.4) in the next theorem. It is an optimization problem in value functions, and it will have some similarities to Problem (5.5) - (5.6). However, the Lagrange multiplier of the stochastic dominance constraint will now appear in the dual.

Theorem 6.8. *The dual to Problem (6.4) is*

$$\sup_{v \in \mathcal{F}\left(\tilde{\mathbb{S}}\right),\, \Lambda \in \mathcal{M}([\eta_1, \eta_2])} \langle v, \nu \rangle - \langle \Lambda, g \rangle \tag{6.6}$$

$$\begin{aligned} s.t.\, v(s) \leq \gamma \int_{\tilde{\mathbb{S}}} v(\xi) \tilde{Q}_T(d\xi \mid s, a) + \frac{1-\gamma}{\gamma^T} y\, I\left\{z \leq \gamma^T\right\} \\ + \frac{1-\gamma}{\gamma^T} \int_{\eta_1}^{\eta_2} (y-\eta)_+ \Lambda(d\eta)\, I\left\{z \leq \gamma^T\right\}, \\ \forall (s, y, z, a) \in \tilde{\mathbb{K}}. \end{aligned} \tag{6.7}$$

Proof. Let the Lagrange multiplier for constraint $L_{0,T}\mu = \nu$ be $v \in \mathcal{F}\left(\tilde{\mathbb{S}}\right)$, the Lagrange multiplier for constraint $L_{1,T}\mu = \theta$ be $w \in \mathcal{F}(\mathbb{R})$, and the Lagrange multiplier for constraint $L_2\theta \geq g$ be $\Lambda \in \mathcal{M}([\eta_1, \eta_2])$. The Lagrangian for Problem (6.4) is then

$$\Psi(\mu, \theta, v, w, \Lambda) = \mathbb{E}[X(\theta)] + \langle v, L_{0,T}\mu - \nu \rangle + \langle w, L_{1,T}\mu - \theta \rangle + \langle \Lambda, L_2\theta - g \rangle.$$

Problem (6.4) is then equivalent to

$$\inf_{\theta,\, \mu \geq 0} \sup_{v,\, w,\, \Lambda} \{\Psi(\mu, \theta, v, w, \Lambda) \,:\, \Lambda \leq 0\},$$

so the dual problem is defined to be

$$\sup_{v,\, w,\, \Lambda} \left\{ \inf_{\theta,\, \mu \geq 0} \Psi(\mu, \theta, v, w, \Lambda) \,:\, \Lambda \leq 0 \right\}.$$

Rearranging the Lagrangian

$$\begin{aligned} \Psi(\mu, \theta, v, w, \Lambda) &= \langle \theta, y \rangle + \langle v, L_{0,T}\mu - \nu \rangle + \langle w, L_{1,T}\mu - \theta \rangle + \langle \Lambda, L_2\theta - g \rangle \\ &= \langle \theta, y - w + L_2^*\Lambda \rangle + \langle \mu, L_0^* v + L_1^* w \rangle - \langle v, \nu \rangle - \langle \Lambda, g \rangle \end{aligned}$$

we see that the dual problem is

$$\begin{aligned} \sup_{v,w,\Lambda} \; & -\langle v,\, \nu\rangle - \langle \Lambda,\, g\rangle \\ \text{s.t. } & y - w + L_2^*\Lambda \geq 0, \\ & L_0^* v + L_1^* w \geq 0. \end{aligned}$$

Switch the sign of v, and take $w(y) = y + \int_{\eta_1}^{\eta_2} (y-\eta)_+ \Lambda(d\eta)$ to get the desired form. □

Remark 6.9. The dual Problem (6.6) - (6.7) is naturally a linear programming problem, since the primal Problem (6.4) is an LP. Problem (6.6) - (6.7) is significant for two reasons. First, it reveals the role that utility functions play as the Lagrange multipliers of the stochastic dominance constraints. This result has already been discovered in stochastic optimization in [10, 11], and it is unsurprising that it holds for MDPs as well. Basically, the Lagrange multiplier Λ induces an increasing convex function of y through the expression

$$\int_{\eta_1}^{\eta_2} (y-\eta)_+ \Lambda(d\eta)\, I\left\{z \leq \gamma^T\right\}.$$

Each $(y-\eta)_+$ is convex in y, Λ is a nonnegative measure, the sum of increasing convex functions is increasing and convex, and we view z as fixed so $I\left\{z \leq \gamma^T\right\}$ is a constant. Second, Problem (6.6) - (6.7) reveals a new form of dynamic programming equations.

The resulting dynamic programming equations for Problem (6.3) are $v(s,y,z) =$

$$\begin{aligned} &\inf_{a\in A(s)} \left\{ \frac{1-\gamma}{\gamma^T} y\, I\left\{z \leq \gamma^T\right\} + \int_{\eta_1}^{\eta_2} (y-\eta)_+ \Lambda(d\eta) + \gamma \int_{\tilde{\mathbb{S}}} v(\xi)\, \tilde{Q}_T(d\xi \mid s,a) \right\}, \\ &\forall\, (s,y,z) \in \tilde{\mathbb{S}}. \end{aligned} \tag{6.8}$$

Notice that the dual variable Λ appears in the preceding objective function: Λ is here because the original problem was constrained. The entire expression

$$\frac{1-\gamma}{\gamma^T} y\, I\left\{z \leq \gamma^T\right\} + \int_{\eta_1}^{\eta_2} (y-\eta)_+ \Lambda(d\eta)$$

acts as a cost function on the augmented state space. Equation (6.3) thus does not represent Bellman iteration in the traditional sense, since there is an external tuning parameter Λ that is not determined by value iteration. The appearance of such an external tuning parameter is typical for constrained MDPs (see [1]): in general, constraints in optimization and control problems cause Lagrange multipliers to appear.

We make the following assumptions to guarantee solvability and strong duality.

Assumption 6.10. *There exists a bounded minimizing sequence* $\left\{\left(v^i,\, \Lambda^i\right)\right\}_{i\geq 0}$ *in Problem (6.6) - (6.7).*

This type of assumption is common in the literature on infinite-dimensional LPs (see [2, 21, 19, 20], for example). It ensures that the dual optimal value is attained by giving a minimizing sequence that attains this value. The next theorem summarizes solvability and strong duality results for Problem (6.3), and is based on linear programming duality.

Theorem 6.11. *Suppose Assumptions 2.1, 5.5, and 6.10 hold. Also suppose Problem (6.3) is feasible. Then:*

(i) Problem (6.4) is solvable;

(ii) There is no duality gap between Problems (6.4) and (6.6) - (6.7);

(iii) Strong duality holds between Problems (6.4) and (6.6) - (6.7).

Remark 6.12. Our development in this section extends to multivariate stochastic dominance constraints with minor modifications. Suppose we have a vector-valued cost function $d : \mathbb{K} \to \mathbb{R}^n$ in addition to c. We are interested in constraining the distribution of the vector valued random variable $\sum_{t=0}^{\infty} \gamma^t d(s_t, a_t)$. However, for $n \geq 2$ there is no parametric representation of $\geq_{icv}$, see [24, 8]. In general, we will take $\mathcal{U}$ to be a collection of increasing concave functions from $\mathbb{R}^n$ to $\mathbb{R}$, and the associated relaxed dominance constraints

$$\mathbb{E}_{\nu}^{\pi}\left[u\left(\sum_{t=0}^{\infty}\gamma^t d(s_t, a_t)\right)\right] \geq \mathbb{E}\left[u(Y)\right], \forall u \in \mathcal{U}.$$

The development for this case is largely similar to the one in this section, at the expense of more complicated notation and some further technical assumptions on d and $\mathcal{U}$.

6.2. Chance constraints. We now consider chance-constrained MDPs in this subsection. Chance constrained optimization problems are usually non-convex and very difficult to solve. Fortunately, in our framework, they lead to linear programming problems because we are optimizing over measures rather than random variables. The development here is actually quite similar to the one in the preceding subsection.

If we view the indicator function as a type of utility step function, then the probability

$$P_{\nu}^{\pi}\left\{\sum_{t=0}^{\infty}\gamma^t c(s_t, a_t) \leq \eta\right\} = \mathbb{E}_{\nu}^{\pi}\left[I\left(\sum_{t=0}^{\infty}\gamma^t c(s_t, a_t) \leq \eta\right)\right]$$

is an expected utility. A collection of chance constraints is thus similar to a collection of constraints on expected utilities, like stochastic dominance constraints. The general chance-constrained MDP is

$$\begin{aligned} &\inf_{\pi\in\Pi}\left\{\mathbb{E}\left[C_{\nu}^{\pi}\right] : \Pr\left\{C_{\nu}^{\pi} \leq \eta_i\right\} \geq 1-\delta_i,\, i \in I\right\} \qquad (6.9)\\ =\ &\inf_{\pi\in\Pi}\left\{\mathbb{E}_{\nu}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t c(s_t, a_t)\right] : P_{\nu}^{\pi}\left\{\sum_{t=0}^{\infty}\gamma^t c(s_t, a_t) \leq \eta_i\right\} \geq 1-\delta_i,\, i \in I\right\}, \end{aligned}$$

where $\eta_i \geq 0$ and $\delta_i \in (0, 1)$ for all $i \in I$ are given constants.

We use the next lemma to justify truncation of Problem (6.9).

Lemma 6.13. *For any $\eta \geq 0$ and $\epsilon > 0$, there is a T such that*

$$|P_{\nu}^{\pi}\left\{\sum_{t=0}^{T}\gamma^t c(s_t, a_t) \leq \eta\right\} - P_{\nu}^{\pi}\left\{\sum_{t=0}^{\infty}\gamma^t c(s_t, a_t) \leq \eta\right\}| \leq \epsilon$$

for all $\pi \in \Pi$.

Proof. We already know y_T converges to $\sum_{t=0}^{\infty}\gamma^t c(s_t, a_t)$ on all trajectories as $T \to \infty$, and thus

$$P_{\nu}^{\pi}\left\{\sum_{t=0}^{T}\gamma^t c(s_t, a_t) \leq \eta\right\} \to P_{\nu}^{\pi}\left\{\sum_{t=0}^{\infty}\gamma^t c(s_t, a_t) \leq \eta\right\}$$

as $T \to \infty$, since almost sure convergence implies convergence in distribution. □

We now consider the truncation

$$\inf_{\pi \in \Pi} \left\{ \mathbb{E}\left[C^{\pi}_{\nu,\,T}\right] \,:\, \Pr\left\{C^{\pi}_{\nu,\,T} \leq \epsilon_i\right\} \geq 1-\delta_i,\, i \in I \right\}. \tag{6.10}$$

Since we have the distribution of $C^{\pi}_{\nu,\,T}$, we can compute $\Pr\left\{C^{\pi}_{\nu,\,T} \leq \epsilon_i\right\}$ with Theorem 3.5 to get

$$\Pr\left\{C^{\pi}_{\nu,\,T} \leq \eta_i\right\} = \frac{1-\gamma}{\gamma^T} \int_{\tilde{\mathbb{S}}} I\left\{y \leq \eta_i,\, z = \gamma^T\right\} \mu^{\pi}_{\nu}\left(d\left(s,y,z,a\right)\right) = \int I\left\{y \leq \eta_i\right\} \theta\left(dy\right)$$

where $\theta = L_{1,\,T}\mu^{\pi}_{\nu}$. In particular, $\Pr\left\{C^{\pi}_{\nu,\,T} \leq \eta_i\right\}$ is a linear function of μ^{π}_{ν}, and thus it is a linear function of θ. We introduce a linear operator $L_3 : \mathcal{M}\left(\mathbb{R}\right) \to \mathbb{R}^{|I|}$ defined by

$$\left[L_3\theta\right]_i := \int_{\mathbb{R}} I\left\{y \leq \eta_i\right\} \theta\left(dy\right),\, \forall i \in I.$$

We also define $g = \left(1-\delta_i\right)_{i \in I}$ to get

$$\inf_{\mu \in \mathcal{M}\left(\tilde{\mathbb{K}}\right),\, \theta \in \mathcal{M}(\mathcal{Y})} \left\{\mathbb{E}\left[X\left(\theta\right)\right] \,:\, L_{0,\,T}\mu = \nu,\, L_{1,\,T}\mu = \theta,\, L_3\theta \leq g\right\}. \tag{6.11}$$

Problem (6.11) is a linear programming problem. We will compute its dual to obtain dynamic programming equations, which are similar to those for Problem (6.3). First, we derive the adjoint of L_3.

Lemma 6.14. *The adjoint of L_3 is $L^*_3 : \mathbb{R}^{|I|} \to \mathcal{F}\left(\mathbb{R}\right)$ defined by*

$$\left[L^*_3\lambda\right]_i := \sum_{i \in I} \lambda_i I\left\{y \leq \eta_i\right\},\, \forall i \in I.$$

Proof. Take $\lambda \in \mathbb{R}^{|I|}$, then

$$\left\langle \lambda,\, L_3\theta \right\rangle = \sum_{i \in I} \lambda_i \int I\left\{y \leq \eta_i\right\} \theta\left(dy\right) = \int \left[\sum_{i \in I} \lambda_i I\left\{y \leq \eta_i\right\}\right] \theta\left(dy\right).$$

□

The computation of the dual of Problem (6.11) is similar to the one for the dual of Problem (6.4), and we omit the detailed computation. Furthermore, strong duality holds under similar sufficient conditions.

Theorem 6.15. *The dual to Problem (6.11) is*

$$\sup_{v \in \mathcal{F}\left(\tilde{\mathbb{S}}\right),\, \lambda \in \mathbb{R}^{|I|}} \left\langle v,\, \nu \right\rangle - \left\langle \lambda,\, g \right\rangle \tag{6.12}$$

$$\begin{aligned} s.t.\;\; v\left(s,y,z\right) \leq\; & \gamma \int_{\tilde{\mathbb{S}}} v\left(\xi\right) \tilde{Q}_T\left(d\xi \,|\, s,a\right) + \frac{1-\gamma}{\gamma^T} y\, I\left\{z \leq \gamma^T\right\} \\ & + \frac{1-\gamma}{\gamma^T} \sum_{i \in I} \lambda_i I\left\{y \leq \eta_i\right\} I\left\{z \leq \gamma^T\right\},\, \forall \left(s,y,z,a\right) \in \tilde{\mathbb{K}}. \end{aligned} \tag{6.13}$$

As in the case of the stochatic dominance-constrained MDPs, we can infer dynamic programming equations for Problem (6.10) as well from the dual Problem (6.12) - (6.13). such equations have an external tuning parameter λ, and are thus not true dynamic programming equations. Instead, it is better to solve Problem (6.11) and its dual by linear programming.

Remark 6.16. We now consider $\mathbb{S}$ and $\mathbb{A}$ to be finite so that we can apply the discretization technique from Subsection 4.2 to get finite-dimensional versions of Problems (6.10) and (6.12) - (6.13). This explicityly brings out the linear programming structure of the chance-constrained problem.

First, the discretized primal problem in occupation measures is

$$\begin{aligned}
&\inf_{\mu,\,\theta} \sum_{y\in\mathbb{Y}} y\,\theta\left(y\right) && (6.14)\\
&\text{s.t. } \nu\left(j\right) = \hat{\mu}\left(j\right) - \gamma \sum_{(s,y,z,a)\in\bar{\mathbb{K}}} \bar{Q}_T\left(j \mid s,y,z,a\right)\mu\left(s,y,z,a\right), && \forall j\in\bar{\mathbb{S}},\\
&\qquad \theta\left(\xi\right) = \frac{1-\gamma}{\gamma^T} \sum_{(s,y,z,a)\in\bar{\mathbb{K}}} I\left\{y=\xi,\, z=\gamma^T\right\}\mu\left(s,y,z,a\right), && \forall \xi\in\mathbb{Y},\\
&\qquad \sum_{y\in\mathbb{Y}} I\left\{y\le\eta_i\right\}\theta\left(y\right) \ge 1-\delta_i, && \forall i\in I.
\end{aligned}$$

The dual problem is now simply

$$\begin{aligned}
&\sup_{v,\,\lambda} \sum_{j\in\bar{\mathbb{S}}} v\left(j\right)\nu\left(j\right) - \sum_{i\in I}\lambda_i\left(1-\delta_i\right)\\
&\text{s.t. } v\left(s,y,z\right) \le \gamma\sum_{\xi\in\bar{\mathbb{S}}} v\left(\xi\right)\bar{Q}_T\left(\xi\mid s,a\right) + \frac{1-\gamma}{\gamma^T} y\, I\left\{z\le\gamma^T\right\}\\
&\qquad + \frac{1-\gamma}{\gamma^T}\sum_{i\in I}\lambda_i I\left\{y\le\eta_i\right\} I\left\{z\le\gamma^T\right\},\ \forall\left(s,y,z,a\right)\in\bar{\mathbb{K}}.
\end{aligned}$$

Both of these problems are finite-dimensional linear programming problems. When either problem is feasible, the other is feasible and we automatically have strong duality.

Remark 6.17. It is striking that our chance-constrained MDPs give rise to linear programming problems in occupation measures, since chance-constrained stochastic optimization problems are typically nonconvex. This observation is due to the fact that we are optimizing over measures rather than random-variable-valued mappings. Specifically, the probability of the event $\left\{\sum_{t=0}^{T}\gamma^t c\left(s_t,a_t\right)\le\epsilon\right\}$ is a linear function of the occupation measure.

7. Numerical illustration. We now illustrate the dependence of the size of the discretized static optimization problems on the desired accuracy. We fix an error tolerance $\epsilon>0$ throughout this section and compute the truncation $T=T\left(\epsilon\right)$. In particular, if we choose

$$T\left(\epsilon\right) = \left\lceil -\frac{\log\left(\epsilon\right)+\log\left(1-\gamma\right)-\log\left(\bar{c}\right)}{\log\left(\gamma\right)}\right\rceil, \tag{7.1}$$

it ensures that $\sum_{t=T+1}^{\infty}\left|\gamma^t c\left(s_t,a_t\right)\right|<\epsilon$. Then, for the expected utility minimizing MDP, if the utility function $u\left(\cdot\right)$ is Lipschitz continuous with constant 1, then for such a $T(\epsilon)$, we have

$$\left|\mathbb{E}_\nu^\pi\left[u\left(\sum_{t=0}^{T}\gamma^t c\left(s_t,a_t\right)\right)\right] - \mathbb{E}_\nu^\pi\left[u\left(\sum_{t=0}^{\infty}\gamma^t c\left(s_t,a_t\right)\right)\right]\right| < \epsilon.$$

	# of variables	# of constraints
Expected utility	484,022	48,422
Stochastic dominance	484,022	48,522
Chance constraints	484,022	48,522
CVaR	484,022	48,422

FIGURE 7.1. *Size of optimization problems for various discretized optimization problems.*

Similar calculations for the stochastic-dominance constrained and the chance-constrained MDPs show that truncation at $T(\epsilon)$ will result in an error of at most ϵ. The CVaR minimizing MDP differs slightly, and here $T(\epsilon)$ will result in error of at most $\epsilon(1-\beta)$. Once the threshold $T(\epsilon)$ is chosen we discretize $\mathcal{Y} = [0, \bar{c}]$ with granularity $\epsilon/T(\epsilon)$ to obtain $\mathbb{Y}$ such that

$$|\mathbb{Y}| = \left\lceil \frac{\bar{c}}{(\epsilon/T(\epsilon))} \right\rceil = \left\lceil \frac{\bar{c}\,T(\epsilon)}{\epsilon} \right\rceil,$$

and we also obtain $|\mathbb{Z}| = T(\epsilon)$. Each of the discretized primal problems (Problems 5.4, 5.11, 6.5 and 6.14) then has

$$|\overline{\mathbb{K}}| + |\mathbb{Y}| = |\mathbb{S}|\,|\mathbb{Y}|\,|\mathbb{Z}|\,|\mathbb{A}| + |\mathbb{Y}| = |\mathbb{Y}|\,(|\mathbb{S}|\,|\mathbb{Z}|\,|\mathbb{A}| + 1) = \left\lceil \frac{\bar{c}\,T(\epsilon)}{\epsilon} \right\rceil (|\mathbb{S}|\,|\mathbb{A}|\,T(\epsilon) + 1)$$

variables. Expected utility and CVaR both have

$$|\overline{\mathbb{S}}| + |\mathbb{Y}| = |\mathbb{S}|\,|\mathbb{Y}|\,|\mathbb{Z}| + |\mathbb{Y}| = |\mathbb{Y}|\,(|\mathbb{S}|\,|\mathbb{Z}| + 1) = \left\lceil \frac{\bar{c}\,T(\epsilon)}{\epsilon} \right\rceil (|\mathbb{S}|\,T(\epsilon) + 1)$$

constraints. The stochastic dominance and chance-constrained MDPs both have

$$|\overline{\mathbb{S}}| + |\mathbb{Y}| + |I| = |\mathbb{Y}|\,(|\mathbb{S}|\,|\mathbb{Z}| + 1) + |I| = \left\lceil \frac{\bar{c}\,T(\epsilon)}{\epsilon} \right\rceil (|\mathbb{S}|\,T(\epsilon) + 1) + |I|$$

constraints. We point out that CVaR is a nonlinear optimization problem, while all the others are all linear programming problems.

Let us now consider what these numbers look like with a particular MDP. Suppose the state and action space sizes are $|\mathbb{S}| = 100$ and $|\mathbb{A}| = 10$, the discount factor $\gamma = 0.9$, and the upper bound on the costs is $\bar{c} = 1$. Let us take $I = 100$. For CVaR, we use $\beta = 0.9$, and take $\epsilon = 1$. With this, $T(\epsilon) = 22$.

From table in Figure 7, we see that these are large linear programs that are solvable on standard laptops available today.

8. Conclusion. In this paper, we introduce a framework of risk-aware MDPs that is a generalization of the classical MDP and constrained MDP models. Therein, expectation of infinite horizon discounted costs is considered. We replace the expectation, a risk neutral measure, with a general risk functional. This framework encompasses many popular ways of expressing risk: expected (dis-)utility models, coherent risk measures such as Conditional Value-at-Risk, stochastic dominance and chance constraints. Prior attempts have focused on developing dynamic programming algorithms, albeit with limited success. This is because in such problems optimal policies are not stationary, and by the nature of the problem are going to be history dependent. In contrast, dynamic programming methods are successful when the underlying

stochastic processes are Markovian. We thus develop a convex analytic approach. We augment the state space, introduce an occupation measure on it, which then yields optimization formulations that are in most cases linear programs. These are indeed infinite dimensional LPs. Hence, we give methods for successive finite approximation of such LPs. A striking result here is that, unlike in static optimization, the chance-constrained MDP can be solved via a linear program. This is very promising. The methods and techniques we have developed are quite general and can be used to solve other risk-aware MDPs beyond those treated here.

REFERENCES

[1] Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

[2] Edward J. Anderson and Peter Nash. *Linear Programming in Infinite-Dimensional Spaces*. John Wiley & Sons, 1987.

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

[4] Nicole Bäuerle and Ulrich Rieder. More risk-sensitive markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2014.

[5] V. Borkar and R. Jain. Risk-constrained markov decision processes. In *Proc. of the IEEE Control and Decision Conference*, 2010.

[6] Vivek S Borkar. A convex analytic approach to markov decision processes. *Probability Theory and Related Fields*, 78(4):583–602, 1988.

[7] Vivek S. Borkar. Convex analytic methods in markov decision processes. In Eugene A. Feinberg, Adam Shwartz, and Frederick S. Hillier, editors, *Handbook of Markov Decision Processes*, volume 40 of *International Series in Operations Research & Management Science*, pages 347–375. Springer US, 2002.

[8] E. M. Bronshtein. Extremal convex functions. *Sibirskii Matematicheskii Zhurnal*, 19(1):10–18, 1978.

[9] Özlem Çavus and Andrzej Ruszczynski. Computational methods for risk-averse undiscounted transient markov models. *Operations Research*, 2014.

[10] Darinka Dentcheva and Andrzej Ruszczyński. Optimization with stochastic dominance constraints. *Society of Industrial and Applied Mathematics Journal of Optimization*, 14(2):548–566, 2003.

[11] Darinka Dentcheva and Andrzej Ruszczyński. Optimality and duality theory for stochastic optimization problems with nonlinear dominance constraints. *Mathematical Programming*, 99:329–350, 2004.

[12] Darinka Dentcheva and Andrzej Ruszczyński. Optimization with multivariate stochastic dominance constraints. *Mathematical Programming*, 117:111–127, 2009.

[13] Cyrus Derman. *Finite State Markovian Decision Processes*. Academic Press, Inc., Orlando, FL, USA, 1970.

[14] G. Di Masi and L. Stettner. Risk-sensitive control of discrete-time markov processes with infinite horizon. *SIAM Journal on Control and Optimization*, 38(1):61–78, 1999.

[15] Jerzy A. Filar, L. C. M. Kallenberg, and Huey-Miin Lee. Variance-penalized markov decision processes. *Mathematics of Operations Research*, 14(1):147–161, 1989.

[16] Hans Follmer and Alexander Schied. *Stochastic Finance: An Introduction in Discrete Time*. Walter de Gruyter, 2004.

[17] M. Frittelli, M. Maggis, and I. Peri. Risk measures on p(r) and value at risk with probability/loss function. *To appear in Mathematical Finance*, 2012.

[18] William B. Haskell and Rahul Jain. Stochastic dominance-constrained markov decision processes. *Society of Industrial and Applied Mathematics Journal on Control and Optimization*, 51(1):273–303, 2013.

[19] Onésimo Hernández-Lerma and Juan González-Hernández. Constrained Markov control processes in Borel spaces: the discounted case. *Mathematical Methods of Operations Research*, 52:271–285, 2000.

[20] Onésimo Hernández-Lerma, Juan González-Hernández, and Raquiel R. López-Martínez. Constrained average cost Markov control processes in Borel spaces. *SIAM J. Control Optim.*, 42(2):442–468, 2003.

[21] Onésimo Hernández-Lerma and Jean B Lasserre. Approximation schemes for infinite linear programs. *SIAM Journal on Optimization*, 8(4):973–988, 1998.

[22] Onesimo Hernandez-Lerma and Jean Bernard Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer-Verlag New York, Inc., 1996.
[23] Onesimo Hernandez-Lerma and Jean Bernard Lasserre. *Further Topics On Discrete-Time Markov Control Processes*. Springer-Verlag New York, Inc., 1999.
[24] Soren Johansen. The extremal convex functions. *Math. Scand.*, 34:61–68, 1974.
[25] Lodewijk Cornelis Maria Kallenberg. Linear programming and finite markovian control problems. *MC Tracts*, 148:1–245, 1983.
[26] David M Kreps. Decision problems with expected utility critera, i: upper and lower convergent utility. *Mathematics of Operations Research*, 2(1):45–53, 1977.
[27] David M Kreps. Decision problems with expected utility criteria, ii: Stationarity. *Mathematics of Operations Research*, 2(3):266–274, 1977.
[28] Shigeo Kusuoka. On law invariant coherent risk measures. *Advances in mathematical economics*, 3(1):83–95, 2001.
[29] Jean-Bernard Lasserre. *Moments, positive polynomials and their applications*, volume 1. World Scientific, 2009.
[30] Olvi L Mangasarian. Solution of general linear complementarity problems via nondifferentiable concave minimization. *Acta Mathematica Vietnamica*, 22(1):199–205, 1997.
[31] Alan S Manne. Linear programming and sequential decisions. *Management Science*, 6(3):259–267, 1960.
[32] Harry Markowitz. Portfolio selection*. *The journal of finance*, 7(1):77–91, 1952.
[33] Alfred Muller and Dietrich Stoyan. *Comparison Methods for Stochastic Models and Risks*. John Wiley and Sons, Inc., 2002.
[34] AB Piunovskiy. *Optimal control of random sequences in problems with constraints*, volume 410. Kluwer Academic Pub, 1997.
[35] Andras Prekopa. On probabilistic constrained programming. In *Proceedings of the Princeton symposium on mathematical programming*, pages 113–138. Princeton, New Jersey: Princeton University Press, 1970.
[36] Martin L. Puterman. *Markov Decision Processes Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2005.
[37] R Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.
[38] A. Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Mathematical programming*, 125(2):235–261, 2010.
[39] Andrzej Ruszczyński and Alexander Shapiro. Optimization of convex risk functions. *Mathematics of Operations Research*, 31(3):433–452, 2006.
[40] Moshe Shaked and J. George Shanthikumar. *Stochastic Orders*. Springer, 2007.
[41] Matthew J. Sobel. Mean-variance tradeoffs in an undiscounted mdp. *Operations Research*, 42(1):175–183, 1994.
[42] G Peter Todd. *Mean-variance analysis in portfolio choice and capital markets*, volume 66. John Wiley & Sons, 2000.
[43] Buheeerdun Yang. Conditional value-at-risk minimization in finite state markov decision processes: Continuity and compactness. *Journal of Uncertain Systems*, 7(1):50–57, 2013.