Some contributions to Markov decision processes

Thesis submitted in accordance with the requirements of
the University of Liverpool for the degree of Doctor in Philosophy
by

Shanyun Chu

Supervisors: Dr. Yi Zhang and Dr. Alexey Piunovskiy

Department of Mathematical Sciences
University of Liverpool

July 2015

Abstract

In a nutshell, this thesis studies discrete-time Markov decision processes (MDPs) on Borel Spaces, with possibly unbounded costs, and both expected (discounted) total cost and long-run expected average cost criteria.

In Chapter 2, we systematically investigate a constrained absorbing MDP with expected total cost criterion and possibly unbounded (from both above and below) cost functions. We apply the convex analytic approach to derive the optimality and duality results, along with the existence of an optimal finite mixing policy. We also provide mild conditions under which a general constrained MDP model with state-action-dependent discount factors can be equivalently transformed into an absorbing MDP model. Chapter 3 treats a more constrained absorbing MDP, as compared with that in Chapter 2. The dynamic programming approach is applied to a reformulated unconstrained MDP model and the optimality results are obtained. In addition, the correspondence between policies in the original model and the reformulated one is illustrated.

In Chapter 4, we attempt to extend the dynamic programming approach for standard MDPs with expected total cost criterion to the case, where the (iterated) coherent risk measure of the cost is taken as the performance measure to be minimized. The cost function under our consideration is allowed to be unbounded from the below, and possibly arbitrarily unbounded from the above. Under a fairly weak version of continuity-compactness conditions, we derive the optimality results for both the finite and infinite horizon cases, and establish value iteration as well as policy iteration algorithms. The standard MDP and the iterated conditional value-at-risk of the cost function are illustrated as two examples.

Chapter 5 and 6 tackle MDPs with long-run expected average cost criterion. In Chapter 5, we consider a constrained MDP with possibly unbounded (from both above and below) cost functions. Under Lyapunov-like conditions, we show the sufficiency of stable policies to the concerned constrained problem. Furthermore, we introduce the corresponding space of performance vectors and manage to characterize each of its extreme points with a deterministic stationary policy. Finally, the existence of an optimal finite mixing policy is justified. Chapter 6 concerns an unconstrained MDP with the cost functions unbounded from the below and possibly arbitrarily unbounded from the above. We provide a detailed discussion on the issue of sufficient policies in the denumerable case, establish the average cost optimality inequality (ACOI) and show the existence of an optimal deterministic stationary policy.

In Chapter 7, an inventory-production system is taken as an example of real-world applications to illustrate the main results in Chapter 2 and 5.

Acknowledgement

I would like to express my sincere gratitude to my primary supervisor, Dr. Yi Zhang, for his invaluable suggestions and consistent support during my research. I would also like to thank my secondary supervisor, Dr. Alexey Piunovskiy, for his active arrangement of a series of seminars, which indeed broadened my horizon and helped me better understand topics in related areas. I am grateful to the financial support offered by Department of Mathematical Science over the past four years. Finally, I am indebted to my parents and anonymous friends, without whom this thesis is unlikely to come into being.

Notations

$\square$	end of a proof
$:=$	equality of definition
$\mathbb{R}$	the set of real numbers
$\mathbb{K}$	the set of feasible state-action pairs
$\delta_x(\cdot)$	a dirac measure concentrated on the point x
$\mathbf{1}_{\{\cdot\}}$	the indicator function
U^H	the set of history-dependent policies
U^M	the set of (randomized) Markov policies
U^S	the set of (randomized) stationary policies
U^{DS}	the set of deterministic stationary policies
$\mathcal{B}(S)$	Borel σ-algebra of subsets of S
$\mathbf{B}(S)$	the space of measurable bounded functions on S
$\mathbf{C}(S)$	the space of continuous bounded functions on S
$\mathbf{B}_w(S)$	the space of w-bounded measurable functions on S
$\mathcal{M}(S)$	the space of finite measures on $\mathcal{B}(S)$
$\mathcal{M}_w(S)$	the space of w-bounded finite measures on $\mathcal{B}(S)$
$\mathcal{P}(S)$	the space of probability measures on S
$\mathcal{M}$	the model of a Markov decision process

Contents

Chapter 1

General Introduction

1.1 Introduction

The thesis deals with discrete-time Markov decision processes, which are frequently and will be referred to as MDPs for short throughout the dissertation, in Borel spaces, with unbounded costs. The criteria to be optimized are expected total cost and long-run expected average cost. Incidentally, a MDP model with discounted expected total cost criterion is regarded as the special case of the one with (undiscounted) expected total cost criterion. These problems form an important class of stochastic control problems with various applications to telecommunication, inventory management, finance and so on.

Roughly speaking, the thesis can be divided into two parts, where the first half consists of Chapter 2, 3 and 4 with the objective of minimizing expected total cost, and the other half is made up of Chapter 5 and 6 with the objective of minimizing long-run expected average cost. Either part concerns both constrained and unconstrained problems. It is worth mentioning that Chapter 4 introduces a new concept called *risk measures*, which makes it possible for one to consider a risk-sensitive MDP. In the remainder of this section, we give a general introduction of standard MDPs with expected total cost criterion, the interplay between a risk-sensitive MDP model and the notion of *iterated coherent risk measures*, MDPs with long-run expected average cost criterion, and at last some contributions we make to the current literature.

Standard MDPs with the objective of minimizing the expected total cost have been intensively studied at least since the 1950s partially due to the fact that the theory can be technically subtle and involving (especially when the state space is a general Borel space) and partially for its rich and various applications. The results on the general theory of MDPs, i.e., the existence of optimal policies and the establishment of the optimality equations, etc., can be found, e.g., in the monographs and textbooks [8, 47, 50, 75], and the applications of MDPs to finance and insurance are well demonstrated in [8].

In regard to a constrained MDP model, there are three main approaches to deal with it. The most popular and easiest way is *direct method* or *convex-analytic approach* (see [2, 20, 34, 72]), which rewrites the original controlled problem as a linear one over the set of *occupation measures*. The main techniques used by this approach are *convex analysis* and *Lagrange multipliers*. The second similar approach widely adopted is called *linear programming* (see [2, 47, 48, 50]). It combines together the primal functional and source of constraints to form both the primal linear program (PLP) and the corresponding dual linear problem (DLP). In addition to (primal) optimality results obtained via the *convex-analytic approach*, the duality results are derived under proper conditions as well. Finally, as the same with an unconstrained model, the popular *dynamic programming approach* is employed; see [65]. However, this approach does not automatically fit well with constrained models and is thus fairly restrictive in its applications. On the other hand, it possesses the advantage that the previous two approaches do not enjoy. To be specific, it allows one to characterize the optimal policy explicitly whenever the initial distribution is given. In connection with the material presented in this dissertation, *direct method* is adopted in Chapter 2 and 5; *linear programming* is used in Chapter 2; *dynamic programming approach* is employed in Chapter 3, in which we consider a more constrained problem.

For the concerned problem in Chapter 2, to the best of our knowledge, the constrained absorbing MDP model in Borel spaces, with unbounded (from both above and below) cost functions and total cost criterion, is not yet investigated in the current literature. The recent advances in this topic include [2, 26, 60] and [50, Chap.9], where [50, 60] consider an unconstrained absorbing MDP model and follow the *dynamic programming approach*, [2] studies a constrained absorbing MDP model on a countable state space, and [26] treats a constrained absorbing MDP model in Borel spaces but with cost functions bounded from one side. Therefore, it is natural that we investigate systematically such an absorbing MDP model, which is of interest in its own right. To this end, we follow the *convex analytic approach*, which is well demonstrated in [18, 72] for different models.

It is well known that a standard discounted MDP model can be equivalently viewed as an undiscounted MDP model. The same assertion also holds if we consider a non-standard and more general discounted MDP model with a state-action-dependent factor. Should the transformed undiscounted model be absorbing, the results for the discounted MDP model would immediately follow from those of the absorbing MDP model. In the present work we provide reasonably verifiable conditions, which, on the one hand, guarantee the transformed undiscounted model to be absorbing, and on the other hand, also allow the state-action-dependent discount factor not necessarily separated from one. To the best of our knowledge, there is only limited literature on discounted MDP models with non-constant discount factors, see [42, 85], both of which consider an unconstrained model, and follow the *dynamic programming approach*.

Chapter 3 concerns a more constrained absorbing MDP model, which can be seen as a natural extension of the one treated in Chapter 2 and [65]. Technically, the more constrained problem is automatically raised if we allow the cost functions to be unbounded from below (of course, controlled by a weight function), see the reformulation in Chapter 3. With a similar reformulation introduced by [65], the establishment of corresponding optimality equation and an optimal deterministic stationary policies is shown by the main optimality results derived in Chapter 4, where a risk-sensitive MDP model is under our consideration.

In view of the form of the concerned problem, constraints are required to be satisfied for expected total cost over the infinite horizon.

In contrast, constraints are further required to be satisfied over every finite horizon in addition to the infinite case in Chapter 2. In terms of methods applied to deal with corresponding problems, the *dynamic programming approach* is employed here as distinct from *convex analytic approach* in Chapter 2. To implement *dynamic programming approach* to constrained problems, the values of all constrained functionals up to every time step is recorded by a specific vector, which is in turn incorporated into the definition of new cost function. Indeed, the idea is similar to the penalty function method. It is worth mentioning that the model considered here differs from that in the above reference in two respects. Firstly, the costs are not necessarily lower-bounded, but rather controlled by a weight function so as to allow unboundedness from both the above and below. Secondly, the criterion under investigation is to minimize total expected cost, where [65] considers a discounted total cost problem. As is mentioned in Remark 2.1(b) and Section 2.6, the latter issue can be in some way addressed by reformulating a discounted MDP as a transient one. Finally, we mention that it is allowed to consider a denumerable model and a Borel but finite one, where the technicality problem remains unresolved in the more general Borel case.

The material presented in Chapter 4 is from [23]. As compared with the standard MDP, where the decision makers are risk neutral, i.e., the expectation of the cost function is the only performance measure, the more recent development is in the direction of incorporating into the MDP model the concept of risk measures. One popular way of doing so is to consider the expected value of the risk measure of the total cost, see [22, 61], where an entropy risk mapping (also known as the exponential utility) is to be minimized; see also [7, 15], where other risk measures are considered. Under some conditions such problems can be transformed to equivalent standard MDPs, see [9]. Instead of considering the risk of the aggregated cost, another way of incorporating risk measures in MDPs is to consider the aggregated (or say iterated) risk; see [70, 82, 90]. This is also what Chapter 4 attempts to do. We point out that for a multistage problem, it is demonstrated in [70] (see also [69]) that optimizing with respect to the expected risk measure (or say utility) is subject to limitations in representing a rational decision making, which can be overcome by optimizing with respect to the aggregated (iterated) risk. Both [69, 70] briefly consider finite horizon problems in finite or countable spaces.In Chapter 4, we focus on the coherent risk measures, whose definition is given in Section 4.2 below. The axiomatic definitions of coherent risk measures are first given in [5], and one popular example of coherent risk measures is the conditional value-at-risk; see [71, 77, 78], Example 4.2 and the appendix below. A single stage optimization problem of the conditional value-at-risk is considered in [77], see also [78].

One of the contributions in Chapter 4 is to allow cost functions being defined in a more relaxed way. As far as applications are concerned, the reason for considering $+\infty$-valued cost functions in economics is explained in [62, 63], where examples involving a $+\infty$-valued utility function popular in economics are presented, see e.g., Example 2 of [62] or Example 4 of [63], where some relevant references in the economics literature can be also found.

In Chapter 5, in addition to the conventional optimality results, we study the existence of a mixing optimal policy to a constrained average optimality problem. One way of establishing such a result is through characterizing the extreme points of the space of occupation or stable measures with those generated by deterministic stationary policies. In particular, [3, 2] considered the characterization of extreme points of stable measures (state-action frequencies) with respect to an average problem under the unichainedness assumption and moment-like conditions in the denumerable case; [20] improved the above results by removing the moment-like conditions, but instead applying one of the geometric results, i.e., Dubin's Lemma (see [30, Main result]). [72] studied the space of occupation measures with respect to discounted problems when the state space is a general Borel space, whereas [26] investigated an absorbing problem with total cost criterion and show the same result by the similar reasoning applied in [72]. It should be pointed out that among the above literature, especially [20], there is no straightforward way of extending the obtained results from the denumerable case to the general Borel one. Owing to this difficulty, we show the existence of an mixing optimal policy, instead by studying the geometric properties of the space of performance vectors.

The study of the space of performance vectors was initiated by [33, 34] as compared with the space of occupation or stable measures [3, 18, 20]. The merit of this object lies in that the space of performance vectors is indeed a finite-dimensional Euclidean space, which is more convenient to deal with and enjoys more well-established results than the space of measures does. Besides, it should be mentioned that the space of performance vectors is more powerful in dealing with multi-objective problems. Indeed, our proof of the main results in Chapter 5 follows in a similar manner as in [34] by recursively constructing sub-models so as to reducing the dimension of the space of performance vectors. The difference lies in that [34] considered a weighted discounted problem, whereas the average optimality problem is considered by us.

In the present work, we attempt to show that the occupation measure optimal to the concerned problem can be represented by convex combination of at most $M+1$ occupation measures, each of which is generated by a deterministic stationary policy, where M denotes the number of constraints. It should be emphasized that the problem studied in Chapter 5 differs from that in [41, 43], which focused on the space of randomized strategies in the appropriate topology. They seem to resemble each other, but indeed turns out to be two distinctive problems. The difference is roughly explained in [72, p.89]. Concretely, mixing policies randomize only before the first step by randomly selecting a specific deterministic policy, which is implemented throughout the evolution of the process. In contrast, randomized policies considered in [41, 43] randomize at each step as the process evolves. In addition, it should be mentioned that the problem considered here is essentially different from the sequence of articles initiated by [53] aiming at finding a "minimum pair" $(\gamma, \hat{\pi})$. In comparison, here we consider an average optimal problem with the initial distribution γ being fixed. As explained in [53, Remark 2.2], the approach adopted in the above reference can be in some sense extended to the present problem by imposing certain type of ergodic hypothesis. However, as will be seen in Assumption 5.3 and 5.4, the property of ergodicity will be imposed merely on the class of deterministic stationary polices rather than “stable policies” introduced in [53]. The objective of seeking for an optimal mixing policy implies the intrinsic necessity of restriction to the family of deterministic stationary polices, essentially as distinct from [53].

Chapter 6 treats an unconstrained average optimality problem under the strong continuity-compactness condition. A detailed summary of the early development of the theory of MDPs with average cost criterion can be found in [4]. We only show a sketched and selected version. Derman [27] first showed the existence of a global optimal deterministic stationary policy in the case of finite state and action space. One of the tricky points regarding average problems lies in that even an *epsilon*-optimal policy may not exist if either the state or action space is infinite; see [6, 79, 80], [31, Chap.7] and [32, Sect.5] for counter-examples. A milestone of the theory of average problems was the *vanishing discount factor approach* established by Blackwell [14]. He observed that there is a close relationship between discount and average problems, i.e., the latter one can be viewed as an approximation of a series of discount problems as the discount factors increase to 1. If the cost function is bounded, Derman [28] studied the average cost optimality equation (ACOE), which provides the optimal value function as a constant and induces the optimal deterministic stationary policy. Hordijk [57] extended the results in [28] from countable-state-finite-action case to countable-state-compact-action one. Sennott [87, 88] considered problems with cost functions bounded from the below. One can apply the well-celebrated Abelian (Tauberian) Theorem (see Theorem C.1) and the conventional diagonal argument to establish the average cost optimality inequality (ACOI). Again, She showed there exists an optimal deterministic stationary policy. Cavazos-Cadena [21] provided an important example to illustrate that the ACOE fails to hold whereas ACOI holds. [46] and [86] considered the general Borel state space, and studied the conditions under which the “relative difference” is bounded. Another direction along which the theory of MDPs with average criterion developed is the *convex analytic approach* having been introduced previously. He relaxed the assumption of uniform ergodicity on the class of deterministic stationary polices when implementing *vanishing discount factor approach*, and investigated "stable policies" only requiring positive Harris recurrence under the assumption of unichainedness. [50] and [44] considered the w-bounded cost functions, where w is a weight function, under strong continuity-compactness conditions.

To the best of our knowledge, we first consider an MDP with average criterion and one-sided w-bounded cost functions, and provide verifiable conditions for the establishement of optimality results. Our methods combine both the *vanishing discount factor approach* and imposing the functional ergodic property on negative part of the cost. We give a detailed discussion on the notion of sufficiency of stationary policies in the denumerable case, which indeed requires moment-like conditions. The ACOI is established and the existence of an optimal deterministic stationary policy to the concerned problem is justified. Finally, we provide an illustrative example in the denumerable case.

In Chapter 7, an inventory-production problem is considered as an example of applications to illustrate the main results obtained in Chapter 2 and 5. Moreover, some of the preliminary, well-known and frequently referred results are collected in the appendix for the convenience of the readers.

Having said the above, the main contributions of this dissertation together with the closely related literature can be summarized as follows:

- For a constrained absorbing MDP model in Borel spaces with possibly unbounded (from both above and below) cost functions and total cost criteria, we derive the optimality and duality results, together with the existence of an optimal mixing policy. The obtained results to various extent, complement [60] and [50, Chap.9] by considering constrained models, and [2] by considering models in Borel spaces, [72] by studying models with undiscounted total cost criteria, and [26] by considering unbounded (both from above and below) cost functions and undertaking the underlying duality analysis.

- We provide mild conditions to guarantee that a constrained discounted MDP model in Borel spaces with a state-action-dependent discount factor possibly not separated from one to be equivalently transformed into an absorbing MDP model. The obtained results complement [65, 85] by considering constrained models with state-action-dependent discount factors.

- We consider a more constrained absorbing MDP model (defined in Chapter 3) in both denumerable and Borel spaces with cost functions possibly unbounded from the below and arbitrarily unbounded from the above, derive the optimality results, establish the existence of a randomized Markov policy. The obtained results differ from [65] by considering a problem in the more constrained context with total cost criterion, and allowing the cost functions unbounded from the below.

- We establish the optimality equation for the reformulated more constrained problem, the existence of an optimal deterministic stationary policy to the problem in the reformulated model, and the correspondence between policies of the original problem and the reformulated one.

- For a MDP in Borel state and action spaces, where the aggregated coherent risk measure is minimized, we establish the optimality equation as well as the value iteration and policy iteration algorithms, and prove the existence of an optimal deterministic stationary policy under quite general conditions. The obtained results complement but differ from the article [90] at least in the following aspects. Firstly, we allow more general cost functions into consideration, that is, the growth (in both directions) of the cost function must be bounded by a specific weight function in [90], whereas the cost here, not only being allowed to be unbounded from the both directions, can be arbitrarily unbounded from the above, and possibly $+\infty$-valued (also cf. [82] where the bounded cost is considered). Note that when one studies constrained MDP problems using the penalty cost method, the $+\infty$-valued cost function appears automatically; see [65] and the material in Chapter 3. Secondly, the analysis in [90] is based on a contraction argument, whereas here we follow a (weakly) monotone convergence argument. On the other hand, more general (not necessarily coherent) risk measures are covered in [90].

- For a constrained MDP model in Borel spaces with possibly unbounded (from both above and below) cost functions and long-run expected average criterion, we show that any extreme point of the space of performance vectors corresponding to stable measures can be generated by a deterministic stationary policy, establish the existence of an optimal mixing policy to the constrained problem over no more than $M+1$ deterministic stationary policies (M is number of constraints). The obtained results complement [20] by considering a general Borel space, differ from [72] and [3] by studying the long-run average cost criterion.

- For an unconstrained average optimality problem, we provide verifiable conditions for the sufficiency of stationary policies in the denumerable case, complementing [3] by considering cost functions unbounded from both directions without a common weight function, and thus rewrite the original problem as a linear one over the space of invariant measures.

- We establish the average cost optimal inequality (ACOI) and the existence of an optimal deterministic stationary policy for an average optimality problem in Borel spaces, extending all the current literature, e.g., [44, 46, 47, 49, 50, 86, 87, 88, 96], by allowing cost functions unbounded from the below while keeping their positive part not necessarily bounded by a prescribed weight function.

1.2 Preliminaries

1.2.1 Markov decision processes

This subsection is devoted to the background knowledge in regard to an MDP model. The material in this section is quite standard and we refer the reader to the books [47, 50] for greater details.

A model of Markov Decision Processes (MDP) $\mathcal{M}$ is defined be a five-tuple

$$\{S, A, (A(x) : x \in S), Q(dy|x, a), c(x, a)\} \tag{1.1}$$

consisting of

- a Borel space S endowed with Borel σ-algebra $\mathcal{B}(S)$ is called the *state space* and elements of which are viewed as states;
- a Borel space A endowed with Borel σ-algebra $\mathcal{B}(A)$ is called the *action space* and elements of which are viewed as actions;
- $A(\cdot)$ is a set-valued mapping which assigns to each $x \in S$ the nonempty set of admissible actions $A(x) \in \mathcal{B}(A)$. It is assumed that graph $A(\cdot)$, denoted by $\mathbb{K} := \{(x, a);\ x \in S,\ a \in A(x)\}$, is a product Borel-measurable subset of $S \times A$ and contains the graph of a measurable function $f : S \to A$;
- $Q(dy|x, a)$ is a stochastic kernel from $\mathbb{K}$ to S, which is called the transition probability or the transition law. $Q(dy|x, a)$ can be viewed as a Borel-measurable function from $\mathbb{K}$ to $[0, 1]$ for each $\Gamma_S \in \mathcal{B}(S)$, and $Q(\cdot|x, a)$ as a probability measure on $(S, \mathcal{B}(S))$ for each $(x, a) \in \mathbb{K}$;
- The one-stage cost function $c(x, a)$ is a Borel-measurable function from $\mathbb{K}$ to $[-\infty, +\infty]$.

Remark 1.1 *(a) The assumption of the existence of a measurable selector contained in the graph $\mathbb{K}$ ensures the existence of at least one deterministic stationary policy, and is also a standing one throughout the*

dissertation.
(b) Unless specified otherwise, we use the term "measurability" instead of "Borel measurability" throughout.
(c) In subsequent chapters, the definition of cost functions will vary from case to case, and could sometimes be more restrictive than the aforementioned one; moreover, the Model $\mathcal{M}$ is best understood as a prototype, so that more components are allowed to be introduced for specific problems, and we shall mention the corresponding modifications.

For every $t = 0, 1, 2, \dots$, let H_t denote the space of admissible trajectories up to time t. To put it precisely, $H_0 := S$, and $H_t := \mathbb{K} \times H_{t-1}$ when $t \geq 1$.

Generally speaking, an MDP is a discrete-time Markov decision process. At each time step, an action is selected by the decision maker based on possibly all the past information. Depending on the current state and action, some costs are incurred and the transition probability, i.e., the conditional distribution of the state for the next time step, is determined. A policy is defined as a sequence of actions to be taken at each time step, and it can be categorized as follows.

- A history-dependent policy π is a sequence of stochastic kernels $(\pi_t(da|h_t))_{t=0,1,2,\dots}$ concentrated on $A(x_t)$, i.e.,

  $$\pi_t(A(x_t)|h_t) = 1 \tag{1.2}$$

  where $h_t = (x_0, a_0, \dots, x_{t-1}, a_{t-1}, x_t)$ is the observed history up to time t;

- The policy is called (randomized) Markov if the stochastic kernels π_t only depend on the current state and time so that we may write $\pi_t(da|h_t) = \pi_t(da|x_t)$;

- The policy is called (randomized) stationary if the stochastic kernels π_t only depend on the current state so that we can write $\pi_t(da|x_0, a_0, \dots, x_{t-1}, a_{t-1}, x_t) = \pi(da|x_t)$, which is often referred to as φ throughout the dissertation;

- The policy is called deterministic stationary, if for a stationary policy there exists a measurable mapping $f : S \to A$, $f(x) \in A(x), \forall\ x \in S$ such that the stochastic kernel can be written as $\pi(da|x) = \delta_{f(x)}(da)$, where $\delta_x(\cdot)$ denotes the dirac measure concentrated on $x \in S$.

We denote by U^H (respectively, U^M, U^S, U^{DS}) the class of history-dependent (respectively, randomized Markov, randomized stationary, deterministic stationary) policies. It is obvious that $U^{DS} \subseteq U^S \subseteq U^M \subseteq U^H \neq \emptyset$.

Let $(\Omega, \mathcal{F})$ be the canonical measurable space where $\Omega := (S \times A)^\infty$ denotes the space of trajectories over infinite horizon and $\mathcal{F}$ is its product measurable σ-algebra.

Let $\pi \in U^H$ be an arbitrary policy and $\gamma \in \mathcal{P}(S)$ an initial distribution, where $\mathcal{P}(S)$ denotes the space of probability measures on S equipped with the usual weak topology. Then, by the well-known Ionescu-Tulcea Theorem (see [47, Prop.C.10 and Rem.11]), there exists a unique probability measure P^π_γ constructed on the canonical space $(\Omega, \mathcal{F})$, which is concentrated on $H_\infty := \mathbb{K}^\infty$ by (1.2), that is, $P^\pi_\gamma(H_\infty) = 1$. Furthermore, for every $\Gamma_S \in \mathcal{B}(S)$, $\Gamma_A \in \mathcal{B}(A)$, and $h_t \in H_t$, $t = 0, 1, 2, \ldots,$

$$\begin{aligned} P^\pi_\gamma(x_0 \in \Gamma_S) &= \gamma(\Gamma_S), \\ P^\pi_\gamma(a_t \in \Gamma_A | h_t) &= \pi_t(\Gamma_A | h_t), \\ P^\pi_\gamma(x_{t+1} \in \Gamma_S | h_t, a_t) &= Q(\Gamma_S | x_t, a_t) \end{aligned}$$

Formally, the controlled stochastic process $(x_t)_{t=0,1,2,\ldots}$ on the probability triple $(\Omega, \mathcal{F}, P^\pi_\gamma)$ is called a discrete-time Markov decision process (MDP), which is also known as a Markov control process.

The corresponding expectation is denoted by E^π_γ. When $\gamma(dx)$ is concentrated at a point, say $x \in S$, we use the simplified notations P^π_x and E^π_x.

Let $\varphi \in U^S$ (respectively, $f \in U^{DS}$) be an arbitrary randomized stationary (respectively, deterministic stationary) policy, and $c(x,a)$ and $Q(dy|x,a)$ be the cost function and stochastic kernels coming from (1.1), we define, for each $x \in S$,

$$c(x,\varphi) := \int_A c(x,a)\varphi(da|x) \tag{1.3}$$

and

$$Q_\varphi(\Gamma_S|x) := Q(\Gamma_S|x,\varphi) = \int_A Q(\Gamma_S|x,a)\varphi(da|x) \quad \forall\, \Gamma_S \in \mathcal{B}(S).$$

In particular, for the specified deterministic stationary policy (measurable selector) f, (1.3) and (1.4) reduce to

$$c(x,f) := c(x,f(x)) \text{ and } Q_f(\Gamma_S|x) := Q(\Gamma_S|x,f) = Q(\Gamma_S|x,f(x))$$

Note that each of the above functions are measurable on S.

Proposition 1.1 *If $\varphi \in U^S$ is a randomized stationary policy, then $(x_t)_{t=0,1,2,\ldots}$ is a homogeneous Markov chain with $Q_\varphi(dy|x)$ being the stochastic transition kernel, that is, for each $\Gamma_S \in \mathcal{B}(S)$ and $t = 0,1,2,\ldots,$*

$$\begin{aligned} P_x^\varphi(x_{t+1} \in \Gamma_S|x_0,\ldots,x_t) &= P_x^\varphi(x_{t+1} \in \Gamma_S|x_t) \\ &= Q_\varphi(\Gamma_S|x_t) \end{aligned}$$

One of the consequences of the above proposition is that the t-stage transition probabilities can be denoted by $Q_\varphi^t(dy|x)$, $t = 0,1,2,\ldots$, that is

$$Q_\varphi^t(\Gamma_S|x) := P_x^\varphi(x_t \in \Gamma_S) \quad \forall\, x \in S,\ \Gamma_S \in \mathcal{B}(S),$$

with $Q_\varphi^1(\Gamma_S|x) := Q_\varphi(\Gamma_S|x)$ and $Q_\varphi^0(\Gamma_S|x) := \mathbf{1}_{\{x\in\Gamma_S\}}$, where $\mathbf{1}_{\{\}}$ stands for the indicator function. That is, we can write $Q_\varphi^t(dy|x)$ recursively as

$$\begin{aligned} Q_\varphi^t(\Gamma_S|x) &= \int_S Q_\varphi(\Gamma_S|y)Q_\varphi^{t-1}(dy|x) \\ &= \int_S Q_\varphi^{t-1}(\Gamma_S|y)Q_\varphi(dy|x) \end{aligned}$$

Given an initial distribution $\gamma \in \mathcal{P}(S)$, the objectives of our concern in this dissertation are to minimize the discounted expected total cost , i.e.,

$$V_{discount}(\pi, \gamma) := E_\gamma^\pi \left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t) \right] \to \min_{\pi \in U^H},$$

the expected total cost, i.e.,

$$V_{total}(\pi, \gamma) := E_\gamma^\pi \left[\sum_{t=0}^{\infty} c(x_t, a_t) \right] \to \min_{\pi \in U^H},$$

and the long-run expected average cost, i.e.,

$$V_{average}(\pi, \gamma) := \overline{\lim_{n \to \infty}} \frac{1}{n} E_\gamma^\pi \left[\sum_{t=0}^{n-1} c(x_t, a_t) \right] \to \min_{\pi \in U^H}$$

respectively, provided each of the above is well defined in some sense.

If no initial distribution is fixed (excluding the dirac measure $\delta_x(\cdot)$ as well), the objectives to be minimized are represented by

$$V_{discount}(\pi, x) := E_x^\pi \left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t) \right] \to \min_{\pi \in U^H}$$

for the discounted expected total cost,

$$V_{total}(\pi, x) := E_x^\pi \left[\sum_{t=0}^{\infty} c(x_t, a_t) \right] \to \min_{\pi \in U^H}$$

for the expected total cost, and

$$V_{average}(\pi, x) := \overline{\lim_{n \to \infty}} \frac{1}{n} E_x^\pi \left[\sum_{t=0}^{n-1} c(x_t, a_t) \right] \to \min_{\pi \in U^H}$$

for the long-run expected average cost, respectively.

Let $V_{criterion}(\pi, \gamma)$ denote any of $V_{discount}(\pi, \gamma)$, $V_{total}(\pi, \gamma)$ and $V_{average}(\pi, \gamma)$, and accordingly, $V_{criterion}(\pi, x)$ denote any of $V_{discount}(\pi, x)$, $V_{total}(\pi, x)$ and $V_{average}(\pi, x)$, We formally define the optimal policies for the concerned problems.

Definition 1.1 *(a) A policy $\pi^* \in U^H$ is called optimal for the problem with a given initial distribution $\gamma \in \mathcal{P}(S)$*

$$V_{criterion}(\pi, \gamma) \to \min_{\pi \in U^H}$$

if

$$V_{criterion}(\pi^*, \gamma) \leq V_{criterion}(\pi, \gamma)$$

for each policy $\pi \in U^H$;
(b) A policy $\pi^* \in U^H$ is called optimal for the problem

$$V_{criterion}(\pi, x) \to \min_{\pi \in U^H}$$

if

$$V_{criterion}(\pi^*, x) \leq V_{criterion}(\pi, x)$$

for each policy $\pi \in U^H$ and for each initial state $x \in S$.

We finish this subsection with a useful result, which is often referred to as Derman-Strauch Lemma (see [29] for the original version) in the literature.

Lemma 1.1 *Let $\gamma \in \mathcal{P}(S)$ be an arbitrarily fixed initial distribution. For each policy $\pi = (\pi_t)_{t=0,1,2,\ldots} \in U^H$, there is a (randomized) Markov policy $\pi^M = (\pi^M_t)_{t=0,1,2,\ldots} \in U^M$ such that*

$$P^{\pi}_{\gamma}(x_t \in \Gamma_S, a_t \in \Gamma_A) = P^{\pi^M}_{\gamma}(x_t \in \Gamma_S, a_t \in \Gamma_A) \quad \forall\ \Gamma_S \in \mathcal{B}(S),\ \Gamma_A \in \mathcal{B}(A)$$

for each $t = 0, 1, 2, \ldots$. Here for each $t = 0, 1, 2, \ldots,$ one can obtain π^M_t as the stochastic kernel from S to A such that

$$P^{\pi}_{\gamma}(x_t \in dx, a_t \in da) = P^{\pi}_{\gamma}(x_t \in dx)\pi^M_t(da|x) \tag{1.4}$$

The Derman-Strauch Lemma states that given an initial distribution $\gamma \in \mathcal{P}(S)$, and for each history-dependent policy π, there exists a Markov policy π^M such that

$$E_\gamma^\pi\left[\sum_{t=0}^{n-1} c(x_t, a_t)\right] = E_\gamma^{\pi^M}\left[\sum_{t=0}^{n-1} c(x_t, a_t)\right]$$

for each $n = 1, 2, \ldots,$ and

$$E_\gamma^\pi\left[\sum_{t=0}^{\infty} c(x_t, a_t)\right] = E_\gamma^{\pi^M}\left[\sum_{t=0}^{\infty} c(x_t, a_t)\right].$$

It is standard to fix an initial distribution $\gamma \in \mathcal{P}(S)$ when dealing with constrained MDPs, and the Derman-Strauch Lemma asserts that we can always restrict our attention to the class of Markov policies U^M. The same argument remains valid even in the more constrained context, which is introduced and studied in Chapter 3.

1.2.2 Weighted-norm spaces

Let S be a Borel space, and $\mathbf{B}(S)$ be the space of real-valued measurable bounded functions f on S, with the supremum norm

$$\|f\| := \sup_{x\in S} |f(x)|.$$

We denote by $\mathbf{C}(S)$ the subspace of $\mathbf{B}(S)$ consisting of all continuous measurable functions on S.

We assume that $w(\cdot) : S \to [1, +\infty)$ is a given measurable function, and often referred to as a weight function. For each real-valued measurable function f on S, we define its w-norm as

$$\|f\|_w := \sup_{x\in S} \frac{|f(x)|}{w(x)},$$

note that $w(\cdot) \equiv 1$ makes w-norm reduce the supremum norm.

We define

$$\mathbf{B}_w(S) := \{f(\cdot) : f(\cdot) \text{ defined on } S \text{ is measurable, and } \|f\|_w < \infty\},$$

and accordingly, denote by $\mathbf{C}_w(S)$ the subspace of $\mathbf{B}_w(S)$ consisting of all continuous measurable functions on S.

Let $\mathcal{M}(S)$ be the space of finite measures on S such that

$$\sup_{\mu\in\mathcal{M}(S)} \mu(S) < \infty.$$

We equip $\mathcal{M}(S)$ with the usual weak topology generated by all the elements in $\mathbf{B}(S)$, denoted by $\tau_{usual}(\mathcal{M}(S))$, which is metrizable. Note that the space of probability measures, denoted by $\mathcal{P}(S)$, is a subset of $\mathcal{M}(S)$ with $P(S) = 1$ for each $P \in \mathcal{P}(S)$. Let $\mathcal{M}_w(S)$ be the space of finite measures on S, where w is the prescribed weight function, such that

$$\sup_{\mu\in\mathcal{M}_w(S)} \int_S w(x)\mu(dx) < \infty.$$

Note that $\mathcal{M}_w(S)$ is a subset of $\mathcal{M}(S)$ because of $w(\cdot) \geq 1$.

In the sequel, we assume that the weight function w is continuous on S. Indeed, there exists a one-to-one correspondence between $\mathcal{M}_w(S)$ and $\mathcal{M}(S)$. For each $\mu \in \mathcal{M}_w(S)$, one can define $\tilde{\mu} \in \mathcal{M}(S)$ by

$$\tilde{\mu}(dx) := \mu(dx)w(x), \tag{1.5}$$

and for each $\tilde{\mu} \in \mathcal{M}(S)$, one can reproduce $\mu \in \mathcal{M}_w(S)$ by

$$\mu(dx) := \frac{\tilde{\mu}(dx)}{w(x)}. \tag{1.6}$$

This correspondence defines the topology $\tau(\mathcal{M}_w(S))$ on $\mathcal{M}_w(S)$ as the image of $\tau_{usual}(\mathcal{M}(S))$. Now $(\mathcal{M}_w(S), \tau(\mathcal{M}_w(S)))$ and $(\mathcal{M}(S), \tau_{usual}(\mathcal{M}(S)))$ are homeomorphic. The convergence in $\tau(\mathcal{M}_w(S))$ is called the w-weak convergence, and is denoted by $\overset{w}{\to}$. Since $\tau_{usual}(\mathcal{M}(S))$ is metrizable, $\tau(\mathcal{M}_w(S))$ is metrizable, too, and $\mu_n \overset{w}{\to} \mu$ as $n \to \infty$, where $\mu_n, \mu \in \mathcal{M}_w(S)$, if and only if $\lim_{n\to\infty} \int_S f(x)\mu_n(dx) = \int_S f(x)\mu(dx)$ for each $f \in \mathbf{C}_w(S)$.

We summarize the above discussions in the following remark for future reference.

Remark 1.2 *The topology $\tau(\mathcal{M}_w(S))$ on $\mathcal{M}_w(S)$ is metrizable, and it is indeed the weak topology on $\mathcal{M}_w(S)$ generated by $\mathbf{C}_w(S)$.*

Chapter 2

Constrained absorbing MDP with total cost criterion

2.1 Introduction

This chapter is organized as follows: Section 2.2 is about a constrained absorbing MDP model. We present the properties of occupation measures, show the closedness and compactness of the space of occupation measures in a proper topology in Section 2.3. In Section 2.4 and 2.5, we reformulate the original problem as a primal linear program (PLP) in the space of occupation measures, derive the existence of a stationary optimal policy, and prove the absence of the duality gap between the PLP and its dual linear program (DLP). Section 2.6 is about a (non-standard) discounted MDP model, where we firstly present some conditions that guarantee the (non-standard) discounted MDP model to be equivalently transformed into an absorbing MDP model.

2.2 Problem formulation and assumptions

A constrained MDP model in Borel spaces with total cost criteria is a 7-tuple

$$\{S, A, A(x), Q(dy|x,a), c_0(x,a), (c_n(x,a), d_n)_{n=1,\dots,N}, \gamma(dy)\},$$

In addition to the first five components introduced in Chapter 1.3.1, we make the following remarks. $c_0(x,a)$ is a measurable function in $\mathbb{K}$ representing the key cost function. $c_n(x,a)$ are measurable functions on $\mathbb{K}$, and $d_n \in \mathbb{R}$, where N is the number of constraints, representing the sources of constraints. $\gamma(dy)$ is a probability measure on $(S, \mathcal{B}(S))$ representing a predetermined initial distribution.

In this chapter we are concerned with an absorbing MDP model, which is defined similarly to [50, Def.9.6.1] as follows.

Definition 2.1 *A constrained MDP model in Borel spaces with total cost criteria is called absorbing if the following conditions are satisfied.*
(a) The state space can be written as $S = \mathbf{S} \bigcup \{\Delta_S\}$*, where* $\mathbf{S}$ *is a Borel space, and* Δ_S *is an isolated point. The action space can be written as* $A = \mathbf{A} \bigcup \{\Delta_A\}$*, where* $\mathbf{A}$ *is a Borel space, and* Δ_A *is an isolated point. Furthermore,* $A(\Delta_S) := \{\Delta_A\}$*, and* $\forall\ x \in \mathbf{S}, A(x) \subseteq \mathbf{A}$.
(b) $Q(\{\Delta_S\}|\Delta_S, \Delta_A) = 1$ *and* $\forall\ x \in \mathbf{S}, a \in A(x)$*,* $Q(\{\Delta_S\}|x,a) = 1 - Q(\mathbf{S}|x,a)$.
(c) $\forall\ n = 0, \ldots, N, c_n(\Delta_S, \Delta_A) = 0$.
(d) There exists a constant $k \geq 0$ *and a measurable function* w *on* S : $w(x) \geq 1$ *on* $\mathbf{S}$*,* $w(\Delta_S) = 0$ *such that*

$$\sup_{\pi \in U^H, x \in \mathbf{S}} \frac{\sum_{t=0}^{\infty} E_x^\pi \left[w(x_t)\right]}{w(x)} \leq k.$$

We define the measurable set $\mathbf{K} := \mathbb{K} \bigcap (\mathbf{S} \times \mathbf{A})$.

Remark 2.1 *(a) On the one hand, the absorbing MDP model in the sense of Definition 2.1 is a special case of the one considered in [26] (see Sec.2 therein). However, the presence of the "weight" function* w *in Definition 2.1 allows one to consider the cost functions* $c_n(x,a), n = 0, 1, \ldots, N$ *unbounded from both above and below (see Chapter 1), which is not covered in [26]. On the other hand, it is also a special case of the transient MDP model defined in [50, Def.9.6.1], where the authors consider only an unconstrained problem.*
(b) In our absorbing model we consider only a single absorbing point Δ_S*. There are two reasons. Firstly, one can transform a discounted MDP*

model with a state-action-dependent discounted factor into an absorbing model by adding to the original state space an absorbing point. Secondly, an absorbing set can often be compressively viewed and treated as an absorbing point.

(c) In what follows, the fixed function $w(\cdot)$ *comes from Definition 2.1.*

(d) A constrained MDP model with total cost criteria is absorbing if it holds that

$$\int_{\mathbf{S}} w(y)Q(dy|x,a) \leq \beta w(x) + bl(x), \forall\ x \in \mathbf{S},$$

where $0 \leq \beta < 1$ *and* $b \geq 0$ *are two constants,* $l(\cdot)$ *is a measurable function on* S *such that* $l(\Delta_S) = 0$, $0 \leq l(x) \leq 1$ *on* $\mathbf{S}$*, and there exists a constant* $0 < \hat{l} < 1$ *satisfying* $E_x^\pi\left[l(x_t)\right] \leq \hat{l}^t, \forall\ t = 0, 1, \ldots, \pi \in U^H$*. This follows from the reasoning presented in [50, Ex.9.6.7]. We also remind the reader of the fact that this condition can be satisfied only if* $Q(dy|x,a)$ *is substochastic, see [50, p.55].*

Below in this chapter the MDP model under our consideration is absorbing in the sense of Definition 2.1.

Assumption 2.1 *(a) There exists a constant* $\hat{c} \geq 0$ *such that*

$$\sup_{a \in A(x)} |c_n(x,a)| \leq \hat{c}w(x), n = 0, 1, \ldots, N.$$

(b) $\int_{\mathbf{S}} w(x)\gamma(dx) < \infty$*, and* $\gamma(\{\Delta_S\}) = 0$*, where* $\gamma(dx)$ *is the initial distribution.*

(c) $\forall\ u \in \boldsymbol{B}_w(\mathbf{S})$*, it holds that* $\sup_{x \in \mathbf{S}} \frac{\sup_{a \in A(x)} \left|\int_{\mathbf{S}} Q(dy|x,a)u(y)\right|}{w(x)} < \infty$.

Assumption 2.1(c) is needed for technical reasons. Assumption 2.1(a,b) ensures the following optimization problem of our interest to be well defined in the sense that all the expected total costs are finite (see also

[50, Prop.9.6.4]):

$$W_0(\pi) := E_\gamma^\pi \left[\sum_{t=0}^{\infty} c_0(x_t, a_t)\right] \to \min_{\pi \in U^H} \tag{2.1}$$

s.t.

$$W_n(\pi) := E_\gamma^\pi \left[\sum_{t=0}^{\infty} c_n(x_t, a_t)\right] \leq d_n, n = 1, 2, \ldots, N.$$

We denote by $U^{feasible} := \{\pi \in U^H : W_n(\pi) \leq d_n, n = 1, \ldots, N\}$ the set of feasible policies for Problem (2.1).

Remark 2.2 *It is a standing assumption in this chapter that $U^{feasible} \neq \emptyset$.*

In this chapter we are interested in the existence of an optimal solution (together with its form) to Problem (2.1). For this purpose, it is convenient to rewrite Problem (2.1) as a linear program over the space of occupation measures (that is, we follow the convex analytic approach), which are introduced in the next section.

2.3 Properties of occupation measures

For an absorbing MDP model (see Definition 2.1) we define under Assumption 2.1(b) the occupation measures in a similar way to [50, (9.4.4)].

Definition 2.2 *The occupation measure $\nu^\pi(dx \times da)$ of a policy $\pi \in U^H$ for an absorbing MDP model is a measure on $\mathcal{B}(\mathbf{S} \times \mathbf{A})$ defined by*

$$\nu^\pi(\Gamma_S \times \Gamma_A) := \sum_{t=0}^{\infty} P_\gamma^\pi(x_t \in \Gamma_S, a_t \in \Gamma_A), \forall\ \Gamma_S \in \mathcal{B}(\mathbf{S}), \Gamma_A \in \mathcal{B}(\mathbf{A}). \tag{2.2}$$

The projection (or marginal) of $\nu^\pi(\cdot)$ on $\mathbf{S}$ is written as $\nu^\pi(\cdot \times A)$, and the space of occupation measures is denoted by $\mathcal{D}$. It is evident that under Assumption 2.1(b) every occupation measure is finite and $\mathcal{D}$ is uniformly bounded. Indeed,

$$\begin{aligned} & \sup_{\nu^\pi \in \mathcal{D}} \nu^\pi(\mathbf{S}\times\mathbf{A}) \leq \sup_{\pi\in U^H} \sum_{t=0}^{\infty} E_\gamma^\pi\left[w(x_t)\right] \\ = & \sup_{\pi\in U^H} \sum_{t=0}^{\infty} \int_{\mathbf{S}} \gamma(dx) E_x^\pi\left[w(x_t)\right] \leq k \int_{\mathbf{S}} \gamma(dx) w(x) < \infty. \end{aligned} \tag{2.3}$$

It is also easy to understand that every occupation measure is concentrated on $\mathbf{K}$ (see [50, Rem.9.4.2(b)] or [31, Thm.1, p.88]). Moreover, the occupation measures defined by Definition 2.2 are essentially the same as those defined by [26, (2.2)].

The next proposition is from [26], and we sketch its proof immediately after the statement.

Proposition 2.1 *For an absorbing MDP model, suppose Assumption 2.1(b,c) is satisfied. Then the following assertions hold.*
(a) $\forall\ \varphi \in U^S$, $\nu^\varphi(\Gamma_S\times\Gamma_A) = \int_{\Gamma_S} \nu^\varphi(dx\times A)\varphi(\Gamma_A|x), \forall\ \Gamma_S \in \mathcal{B}(\mathbf{S}), \Gamma_A \in \mathcal{B}(\mathbf{A})$.
(b) $\forall\ \pi \in U^H$, $\exists\ \varphi \in U^S$: $\nu^\pi(\Gamma_S\times\Gamma_A) = \int_{\Gamma_S} \nu^\pi(dx\times A)\varphi(\Gamma_A|x) = \nu^\varphi(\Gamma_S\times\Gamma_A), \forall\ \Gamma_S\in\mathcal{B}(\mathbf{S}), \Gamma_A\in\mathcal{B}(\mathbf{A})$. *Furthermore, given such a* $\varphi \in U^S$, *a policy* $\varphi' \in U^S$ *generates* ν^φ *if and only if* $\forall\ \Gamma_A\in\mathcal{B}(\mathbf{A})$, $\varphi(\Gamma_A|x) = \varphi'(\Gamma_A|x)$ $\nu^\varphi(\cdot\times A)$*-a.s.*
(c) A finite measure $\nu(dx\times da)$ *on* $\mathbf{S}\times\mathbf{A}$ *concentrated on* $\mathbf{K}$ *is an occupation measure for a policy* π *if and only if*

$$\int_{\mathbf{S}\times\mathbf{A}} \nu(dx\times da) w(x) < \infty, \tag{2.4}$$

and

$$\nu(\Gamma_S\times A) = \gamma(\Gamma_S) + \int_{\mathbf{S}\times\mathbf{A}} Q(\Gamma_S|y,a)\nu(dy\times da), \tag{2.5}$$

for each $\Gamma_S \in \mathcal{B}(S)$.
(d) $\mathcal{D}$ *is convex.*
(e) If $\nu\in\mathcal{D}$ *is an extreme point, then there exists a deterministic stationary policy* $f\in U^{DS}$ *such that* $\nu(dx\times da) = \nu^f(dx\times da)$.

Proof. (a) We have the following observation

$$\begin{aligned}
P_\gamma^\varphi(x_t \in \Gamma_S, a_t \in \Gamma_A) &= E_\gamma^\varphi \mathbf{1}_{\{x_t \in \Gamma_S, a_t \in \Gamma_A\}} \\
&= E_\gamma^\varphi E_\gamma^\varphi \left[\mathbf{1}_{\{x_t \in \Gamma_S, a_t \in \Gamma_A\}} | x_t\right] \\
&= \int_{\Gamma_S} P_\gamma^\varphi(a_t \in \Gamma_A | x_t = x) P_\gamma^\varphi(x_t \in dx) \\
&= \int_{\Gamma_S} \int_{\Gamma_A} \varphi(da|x) P_\gamma^\varphi(x_t \in dx)
\end{aligned}$$

The second equality comes from the tower property of conditional expectation, and the last line follows from the canonical construction of a MDP along with the fact that $\varphi \in U^S$. Taking summation on both sides of the above equality with respect to t over 0 to ∞ justifies part (a).

(c) We first show the easier "only if" part. Observe that (2.4) is a direct consequence of (2.3). (2.5) holds true due to the following calculation,

$$\begin{aligned}
\hat{\nu}^\pi(\Gamma_S) &= \sum_{t=0}^{\infty} P_\gamma^\pi(x_t \in dx) \\
&= P_\gamma^\pi(x_0 \in \Gamma_S) + \sum_{t=1}^{\infty} P_\gamma^\pi(x_t \in \Gamma_S) \\
&= \gamma(\Gamma_S) + \sum_{t=1}^{\infty} E_\gamma^\pi[P_\gamma^\pi(x_t \in \Gamma_S | x_{t-1}, a_{t-1})] \\
&= \gamma(\Gamma_S) + \sum_{t=1}^{\infty} \int_{\mathbf{S}\times\mathbf{A}} P_\gamma^\pi(x_t \in \Gamma_S | y, a) P_\gamma^\pi(x_{t-1} \in dy, a_{t-1} \in da) \\
&= \gamma(\Gamma_S) + \int_{\mathbf{S}\times\mathbf{A}} Q(\Gamma_S | y, a) \nu^\pi(dy, da)
\end{aligned}$$

which coincides with the expression of (2.5).

The "if" part involves slightly more complications. Let $\nu(dx \times da)$ be a finite measure that satisfies both (2.4) and (2.5), which is allowed to be disintegrated as $\nu(dx \times da) = \nu(dx \times A)\varphi(da|x)$ (see [47, Prop.D.8(a)]). For the obtained stationary policy $\varphi \in U^S$, we generate the corresponding occupation measure ν^φ and complete our proof by showing that ν and ν^φ agree with each other.

Let $u \in \mathbf{B}_w(\mathbf{S} \times \mathbf{A})$ be an arbitrary function with $u(\Delta_S, \Delta_A) = 0$, and define $J_u^\varphi(x) := E_x^\varphi[\sum_{t=0}^\infty u(x_t, a_t)]$. It is easy to show that $J_u^\varphi(x) \in \mathbf{B}_w(\mathbf{S})$ by Definition 2.1(d), and is a solution of the following equation

$$\begin{aligned} h(x) &= \int_{\mathbf{A}} \left[u(x,a) + \int_{\mathbf{S}} h(y) Q(dy|x,a) \right] \varphi(da|x) \\ &= u(x,\varphi) + \int_{\mathbf{S}} h(y) Q(dy|x,\varphi) \quad \forall\ x \in S. \end{aligned} \tag{2.6}$$

With the above observation in mind, we have

$$\begin{aligned} & \int_{\mathbf{S}\times\mathbf{A}} u(x,a)\nu(dx \times da) \\ = & \int_{\mathbf{S}\times\mathbf{A}} J_u^\varphi(x)\nu(dx \times da) - \int_{\mathbf{S}\times\mathbf{A}} \int_{\mathbf{S}} J_u^\varphi(y) Q(dy|x,a)\nu(dx \times da) \\ = & \int_{\mathbf{S}} J_u^\varphi(x)\gamma(dx) + \int_{\mathbf{S}} J_u^\varphi(x) \int_{\mathbf{S}\times\mathbf{A}} Q(dx|y,a)\nu(dy,da) \\ & - \int_{\mathbf{S}\times\mathbf{A}} \int_{\mathbf{S}} J_u^\varphi(y) Q(dy|x,a)\nu(dx \times da) \\ = & E_\gamma^\varphi[\sum_{t=0}^\infty u(x_t,a_t)] \\ = & \int_{\mathbf{S}\times\mathbf{A}} u(x,a)\nu^\varphi(dx \times da) \end{aligned}$$

The first equality comes from taking expectation on both sides of (2.6) with respect to $\nu(dx \times A)$ with h being replaced by J_u^φ, and $\int_{\mathbf{S}} J_u^\varphi(y) Q(dy|x,a) \in \mathbf{B}_w(\mathbf{S})$ is due to Assumption 2.1(c).

Part (d) is trivial. Parts (b,e) come from [26, Lem.4.2, Lem.4.7], whose proofs are similar to those of [72, Lem.25, Thm.19], where Assumption 2.1(c) is needed, see [72, p.308]. □

Remark 2.3 *Since $\nu \in \mathcal{D}$ is concentrated on $\mathbf{K}$, there is no loss of generality that below we regard occupation measures ν as measures on $\mathbf{K}$. Notations such as $\int_{\mathbf{S}\times\mathbf{A}} f(x,a)\nu(dx \times da)$, which are still in use, should be accordingly understood.*

Assumption 2.2 *(a) The function $w(x)$ is continuous on $\mathbf{S}$.*
(b) For any bounded continuous function $u \in \mathbf{C}(\mathbf{S})$, $\int_{\mathbf{S}} u(x)Q(dx|y,a)$ is continuous in $(y,a) \in \mathbf{K}$.

(c) There exists a moment (see Appendix B.1) $v(x,a)$ on $\mathbf{K}$ satisfying

$$\sup_{\nu\in\mathcal{D}}\int_{\mathbf{K}} v(x,a)w(x)\nu(dx\times da)<\infty.$$

Assumption 2.2 and Assumption 2.3 formulated below are "compactness-continuity" conditions, which, in various forms, are commonly assumed to derive optimality results. In particular, our condition is similar to those imposed in [73] and [47, Con.5.7.4]. See more discussions on this in Remark 2.4 below.

The definition below follows the material presented in Chapter 1 under Assumption 2.2(a).

Let $\mathcal{M}(\mathbf{K})$ be the set of finite measures on $\mathbf{K}$ such that

$$\sup_{M\in\mathcal{M}(\mathbf{K})} M(\mathbf{K})<\infty.$$

We equip $\mathcal{M}(\mathbf{K})$ with the usual weak topology generated by the set of bounded continuous functions on $\mathbf{K}$, denoted by $\tau_{usual}(\mathcal{M}(\mathbf{K}))$, which is metrizable (see [74, 95]). We call a measurable function $f(x,a)$ on $\mathbf{K}$ w-bounded if

$$\sup_{x\in\mathbf{S}}\frac{\sup_{a\in A(x)}|f(x,a)|}{w(x)}<\infty,$$

and the set of such functions is denoted by $\mathbf{B}_w(\mathbf{K})$. Let $\mathcal{M}_w(\mathbf{K})$ denote the set of finite measures on $\mathbf{K}$ such that

$$\sup_{M\in\mathcal{M}_w(\mathbf{K})}\int_{\mathbf{K}} M(dx\times da)w(x)<\infty.$$

Now we reveal some topological properties of $\mathcal{D}$ as a subset of $\mathcal{M}_w(\mathbf{K})$.

Theorem 2.1 *For an absorbing MDP model, suppose Assumption 2.1(b,c) and Assumption 2.2(a,b) are satisfied. Then the following assertions hold.*
(a) $\mathcal{D}$ is a closed subset of the topological space $(\mathcal{M}_w(\mathbf{K}),\tau(\mathcal{M}_w(\mathbf{K})))$.
(b) If in addition Assumption 2.2(c) is also satisfied, then $\mathcal{D}$ is compact in $(\mathcal{M}_w(\mathbf{K}),\tau(\mathcal{M}_w(\mathbf{K})))$.

Proof. (a) According to Remark 1.2, it suffices to consider the convergence of sequences. So we take a sequence of elements $\nu_n \in \mathcal{D}$ and $\nu \in \mathcal{M}_w(\mathbf{K})$ such that $\nu_n \overset{w}{\to} \nu$. Below we prove that $\nu \in \mathcal{D}$. To this end, by Proposition 2.1(c) we only need verify the validities of (2.4) and (2.5). Inequality (2.4) obviously holds because of the second inequality in (2.3) and $\int_{\mathbf{S}\times\mathbf{A}} w(x)\nu(dx \times da) = \lim_{n\to\infty} \int_{\mathbf{S}\times\mathbf{A}} w(x)\nu_n(dx \times da)$ by the supposition. It remains to verify the validity of (2.5) as follows. Let us define a measure on $\mathbf{S}$ by $\tilde{\nu}(dx) := \gamma(dx) + \int_{\mathbf{K}} Q(dx|y,a)\nu(dy \times da)$. Then on the one hand, $\nu_n(dx \times A) \to \tilde{\nu}(dx)$ in the corresponding usual weak topology, which is metrizable (see [74, 95]). Indeed, for any fixed bounded continuous function $u(x)$ on $\mathbf{S}$ we have

$$\begin{aligned}
& \lim_{n\to\infty} \int_{\mathbf{S}\times\mathbf{A}} u(x)\nu_n(dx \times da) \\
= \; & \int_{\mathbf{S}} u(x)\gamma(dx) + \lim_{n\to\infty} \int_{\mathbf{S}} u(x) \int_{\mathbf{K}} Q(dx|y,a)\nu_n(dy \times da) \\
= \; & \int_{\mathbf{S}} u(x)\gamma(dx) + \lim_{n\to\infty} \int_{\mathbf{K}} \int_{\mathbf{S}} u(x)Q(dx|y,a)\nu_n(dy \times da) \\
= \; & \int_{\mathbf{S}} u(x)\gamma(dx) + \int_{\mathbf{K}} \int_{\mathbf{S}} u(x)Q(dx|y,a)\nu(dy \times da) = \int_{\mathbf{S}} u(x)\tilde{\nu}(dx),
\end{aligned}$$

where the first equality is by Proposition 2.1(c), and the third equality follows from Assumption 2.2(b). On the other hand, $\nu_n(dx \times A) \to \nu(dx \times A)$ simply because $\nu_n(dx \times A) \overset{w}{\to} \nu(dx \times A)$ in the corresponding w-weak topology, and the w-weak topology is at least as strong as the usual weak topology (see [40, Exer.7, p.127]). From this and the uniqueness of the usual weak limit, we conclude $\nu(\cdot \times A) = \tilde{\nu}(\cdot)$, i.e., (2.5) holds for $\nu(\cdot \times A)$, as required.

(b) By part (a) of this theorem, it suffices to prove that $\mathcal{D}$ is precompact in the topology $(\mathcal{M}_w(\mathbf{K}), \tau(\mathcal{M}_w(\mathbf{K})))$. In accordance with Remark 1.2, this is the same as to prove that $\tilde{\mathcal{D}}$ is precompact in the topology $(\mathcal{M}(\mathbf{K}), \tau_{usual}(\mathcal{M}(\mathbf{K})))$, where $\tilde{\mathcal{D}}$ is the image of $\mathcal{D}$ via (1.5), and in correspondence the elements of $\mathcal{D}$ (respectively, $\tilde{\mathcal{D}}$) are denoted by ν (respectively, $\tilde{\nu}$). It can be very easily verified based on Definition B.1 that if there is a moment $v(x,a)$ on $\mathbf{K}$ such that $\sup_{\tilde{\nu}\in\tilde{\mathcal{D}}} \int_{\mathbf{K}} v(x,a)\tilde{\nu}(dx \times da) < \infty$, then the family $\tilde{\mathcal{D}}$ is tight (see also [47, Prop.E.8], where this simple observation is formulated for a more restrictive case). Hence, Assumption 2.2(c) implies the existence of such a moment $v(x, a)$ and thus the tightness of the family $\tilde{\mathcal{D}}$. The tightness together with that $\tilde{\mathcal{D}}$ is uniformly bounded in the sense of (2.3) means that $\tilde{\mathcal{D}}$ is precompact because of Prohorov's theorem, see Theorem B.1. Thus the proof is completed. □

We remark on Assumption 2.2(c) and Theorem 2.1 as follows.

Remark 2.4 *(a) Assumption 2.2(c) is satisfied if* $\forall\ x \in \mathbf{S}, A(x) \equiv \mathbf{A}$, *and the spaces* $\mathbf{A}$ *and* $\mathbf{S}$ *are both compact. In this case, we may take* $v(x, a) \equiv 1$ *as a moment (see Definition B.1(a)). This is the assumption imposed in [72].*

(b) According to [47, Rem.5.7.5] and (2.3), Assumption 2.2(c) also holds if the following are met:

(i) $\mathbf{S}$ *and* $\mathbf{A}$ *are* σ*-compact.*

(ii) The multifunction $x \to A(x)$ *is compact-valued and upper semicontinuous, i.e., for any* F *closed in* A, $\{x \in \mathbf{S} : A(x) \bigcap F \neq \emptyset\}$ *is closed in* $\mathbf{S}$ *(see the appendix attached or [47, Appendix D] for more details).*

(iii) There exists a nonnegative measurable function $w'(x)$ *on* S *such that*

(1) The requirement of Definition 2.1 and Assumption 2.1(b) are also satisfied with the function $w(x)$ *being replaced by* $w(x)w'(x)$.

(2) $\forall\ \epsilon \geq 0$, $\exists$ *a compact set* $S_\epsilon \subseteq \mathbf{S} : w'(x) \geq \epsilon$, $\forall\ x \notin S_\epsilon$.

In this case, the function w' *is a desired moment.*

(c) In case $w(x) \equiv 1$ *on* $\mathbf{S}$, $(\mathcal{M}_w(\mathbf{K}), \tau(\mathcal{M}_w(\mathbf{K})))$ *and* $(\mathcal{M}(\mathbf{K}), \tau_{usual}(\mathcal{M}(\mathbf{K})))$ *coincide. Then by [26, Lem.4.8], the statement of Theorem 2.1(b) also holds if we replace Assumption 2.2(c) with the condition (b,ii) formulated above in this remark. Indeed, Assumption 2.2(c) is only needed to prove Theorem 2.1(b), which in turn is only used in the proof of Theorem 2.2(a) below. Therefore, in case we can take* $w(x) \equiv 1$ *on* $\mathbf{S}$*, Assumption 2.2(c) can be simply replaced everywhere with the condition (b,ii) formulated above in this remark.*

(d) Assumption 2.2(c) is stronger than the compactness of $A(x), \forall\ x \in \boldsymbol{S}$, *as observed in [73, Lem.3.10].*

2.4 Optimality results

For the concerned absorbing MDP model, suppose Assumption 2.1(a,b) is satisfied. Then Problem (2.1) can be rewritten in the form of a well-defined linear program as follows:

$$\begin{aligned}
&\int_{\mathbf{K}} c_0(x,a)\nu(dx\times da) \to \min_{\nu} &&(2.7)\\
s.t.&\\
&\int_{\mathbf{K}} c_n(x,a)\nu(dx\times da) \le d_n, n=1,2,\dots,N,\\
&\nu\in\mathcal{D}.
\end{aligned}$$

That is why we are interested in the properties of occupation measures as presented in the above.

Assumption 2.3 *The functions $c_n(x,a), n = 0,1,\dots,N$ are all lower semicontinuous on $\mathbf{K}$.*

This assumption together with Assumption 2.2 validates the generalized Weierstrass' theorem, which leads to the existence of an optimal solution to Problem (2.7), see the proof of Theorem 2.2(a) below.

Assumption 2.4 *(a) There exists a policy $\hat{\pi}$ such that the inequalities in Problem (2.1) are all strict, i.e., $W_n(\hat{\pi}) < d_n, n = 1,2,\dots,N$.*
(b) The functions $c_n(x,a), n = 1,2,\dots,N$ are all continuous on $\mathbf{K}$.

Assumption 2.4(a) is known as Slater's condition, which validates Khun Tucker's theorem stated in the proof of part (b) of the next theorem.

Theorem 2.2 *For an absorbing MDP model, suppose Assumption 2.1, Assumption 2.2 and Assumption 2.3 are satisfied. Then the following assertions hold.*
(a) Problem (2.7) is solvable, and there is a stationary optimal policy to Problem (2.1).
(b) If in addition Assumption 2.4 is also satisfied, then there exist constants $\lambda_n^*, n = 1,2,\dots,N+1$ and occupation measures $\nu_n^*, n = 1,\dots,N+1$ such that $\lambda_n^* \ge 0$, $\sum_{n=1}^{N+1}\lambda_n^* = 1$, $\nu_n^*, n = 1,2,\dots,N+1$ are generated by

deterministic stationary policies, say f_n, and the occupation measure defined by $\nu^{Opt} := \sum_{n=1}^{N+1} \lambda_n^* \nu_n^*$ solves Problem (2.7). Here N is the number of constraints (inequalities) in Problem (2.7).

Proof. (a) Firstly, we prove a preliminary result under the conditions of this theorem.

Preliminary result: $\forall\ n = 0, 1, \ldots, N$, $\exists$ a sequence of w-bounded continuous functions $c_n^m(x,a), m = 1, 2, \cdots : \forall\ (x,a) \in \mathbf{K}, c_n^m(x,a) \uparrow c_n(x,a)$ as $m \uparrow \infty$. Moreover, $\forall\ n = 0, 1, \ldots, N, \exists\ R \geq 0 : |c_n^m(x,a)| \leq Rw(x), \forall\ m = 1, 2, \ldots.$

Proof. Throughout this proof, let $n = 0, 1, \ldots, N$ be arbitrarily fixed.

For the first assertion, we argue as follows. Since $c_n(x,a)$ is w-bounded and lower semicontinuous on $\mathbf{K}$, and $w(x)$ is continuous on $\mathbf{S}$ (and thus on $\mathbf{K}$), we have that the function $\overline{c}_n(x,a) := \frac{c_n(x,a)}{w(x)}$ is bounded and lower semicontinuous on $\mathbf{K}$. Then by Proposition A.1 or [72, Thm.A1.14] one can take a sequence of bounded continuous functions $\overline{c}_n^m(x,a), m = 1, 2, \ldots$ such that $\overline{c}_n^m(x,a) \uparrow \overline{c}_n(x,a)$ as $m \uparrow \infty$, i.e., $\overline{c}_n^m(x,a)w(x) \uparrow c_n(x,a)$ as $m \uparrow \infty$. Let us define

$$c_n^m(x,a) := \overline{c}_n^m(x,a)w(x), \forall\ m = 1, 2, \ldots,$$

which is the desired sequence.

Now since $c_n(x,a)$ and $c_n^1(x,a)$ are both w-bounded, there exists

$$R_n \geq 0 : \sup_{a \in A(x)} |c_n(x,a)| \leq R_n w(x), \sup_{a \in A(x)} |c_n^1(x,a)| \leq R_n w(x), \forall\ x \in \mathbf{S}.$$

The fact that $\forall\ (x,a) \in \mathbf{K}, c_n^m(x,a) \uparrow c_n(x,a)$ as $m \to \infty$ implies that $\forall\ m = 1, 2, \ldots,$

$$\begin{aligned} |c_n^m(x,a)| &\leq \max\{|c_n(x,a)|, |c_n^1(x,a)|\} \\ &\leq \max\{\sup_{a \in A(x)} |c_n(x,a)|, \sup_{a \in A(x)} |c_n^1(x,a)|\} \leq R_n w(x). \end{aligned}$$

Since $n = 0, 1, \ldots, N$ and $m = 1, 2, \ldots$ are both arbitrarily fixed, it remains to take $R := \max_{n=0,\ldots,N}(R_n)$ for the second assertion. □ Note that the above preliminary result also holds for measurable functions on

S that are bounded in the w-norm form the below.

Consider the space of feasible occupation measures defined by

$$\mathcal{D}_{feasible} := \left\{ \nu \in \mathcal{D} : \int_{\mathbf{K}} c_n(x,a)\nu(dx \times da) \leq d_n, n = 1, \dots, N \right\}.$$

Firstly, let us prove that $\mathcal{D}_{feasible}$ is w-weakly compact in $\mathcal{D}$. With Remark 1.2 in mind, suppose $\nu_j \overset{w}{\to} \nu$ as $j \to \infty$, where $\nu_j \in \mathcal{D}_{feasible}$, and $\nu \in \mathcal{D}$. By Lebesgue's dominated convergence theorem, the preliminary result just established in the above and using the notations therein, we have that $\forall\ n = 1, 2, \dots, N$,

$$\begin{aligned}
\int_{\mathbf{K}} \nu(dx \times da) c_n(x,a) &= \int_{\mathbf{K}} \nu(dx \times da) \lim_{m \to \infty} c_n^m(x,a) \\
&= \lim_{m \to \infty} \int_{\mathbf{K}} \nu(dx \times da) c_n^m(x,a) \\
&= \lim_{m \to \infty} \lim_{j \to \infty} \int_{\mathbf{K}} \nu_j(dx \times da) c_n^m(x,a) \leq d_n,
\end{aligned}$$

i.e., $\nu \in \mathcal{D}_{feasible}$. Therefore, $\mathcal{D}_{feasible}$ is w-weakly closed in $\mathcal{D}$. This and Theorem 2.1(b) imply that $\mathcal{D}_{feasible}$ is w-weakly compact in $\mathcal{D}$.

Secondly, let us prove that the functional on $\mathcal{D}$ defined by $\mathcal{D}_{feasible} \ni \nu :\to \int_{\mathbf{K}} c_0(x,a)\nu(dx \times da)$ is lower semicontinuous on $\mathcal{D}_{feasible}$ equipped with the w-weak topology. But an argument similar to the one used above would result in that $\left\{\nu \in \mathcal{D}_{feasible} : \int_{\mathbf{K}} c_0(x,a)\nu(dx \times da) \leq r\right\}$ is w-weakly closed $\forall\ r \in \mathbb{R}$, which is equivalent to the lower semicontinuity (in the w-weak topology) of $\int_{\mathbf{K}} c_0(x,a)\nu(dx \times da)$, see [1, p.43].

Now Problem (2.7) is solvable by the generalized Weierstrass' theorem (see [1, Thm.2.43]). The last statement of this theorem follows from this and Proposition 2.1. □

(b) We firstly recall from [72, Thm.A2.1] a version of Kuhn Tucker's theorem, then present some preliminary results, and finally give the main proof of this part of the theorem.

Khun Tucker's theorem: For linear program (2.7), we define the Lagrange function $\underline{L}(\nu, \vec{Y})$ on $\mathcal{D} \times (\mathbb{R}^0_+)^N$ by

$$\underline{L}(\nu, \vec{Y}) := \int_{\mathbf{K}} c_0(x,a)\nu(dx \times da) + \sum_{n=1}^{N} Y_n \left(\int_{\mathbf{K}} c_n(x,a)\nu(dx \times da) - d_n \right),$$

where and also in the sequel, we often use the generic notation $\vec{Y} = (Y_1, Y_2, \dots, Y_N)$, and the primal functional $G(\vec{Y})$ on $\mathbb{R}^N$ by

$$\begin{aligned} G(\vec{Y}) \ := \ \inf \Bigg\{ & \int_{\mathbf{K}} c_0(x,a)\nu(dx \times da) : \nu \in \mathcal{D} \\ & \int_{\mathbb{K}} c_n(x,a)\nu(dx \times da) - d_n \leq Y_n, n = 1, 2, \dots, N \Bigg\}. \end{aligned}$$

Then an occupation measure ν^{Opt} solves Problem (2.7) if and only if there exists a $\vec{Y}^* = (Y_1^*, Y_2^*, \dots, Y_N^*) \in (\mathbb{R}^0_+)^N$ such that

$$\int_{\mathbf{K}} c_n(x,a)\nu^{Opt}(dx \times da) - d_n \leq 0, \forall \ n = 1, 2, \dots, N, \tag{2.8}$$

$$\underline{L}(\nu^{Opt}, \vec{Y}^*) = \inf_{\nu \in \mathcal{D}} \underline{L}(\nu, \vec{Y}^*) \tag{2.9}$$

and

$$\sum_{n=1}^{N} Y_n^* \left(\int_{\mathbf{K}} c_n(x,a)\nu^{Opt}(dx \times da) - d_n \right) = 0. \tag{2.10}$$

Indeed, under Assumption 2.1(a,b), $G(\vec{Y}) > -\infty$, $G(\vec{0}) < \infty$. These two facts about $G(\cdot)$ together with Slater's condition (i.e., Assumption 2.4(a)) satisfy the condition of [72, Thm.A2.1] (see also [72, p.298-299] for more details), from which we infer for the result formulated above.

Preliminary observation 1: The set $\underline{\mathcal{D}}_{\vec{Y}^*} := \{\nu \in \mathcal{D} : \underline{L}(\nu, \vec{Y}^*) = \inf_{\nu \in \mathcal{D}} \underline{L}(\nu, \vec{Y}^*)\}$ is non-empty and convex.

Indeed, by part (a) of this theorem and the aforementioned Kuhn Tucker's theorem, there exists $\nu^* \in \mathcal{D}$ and $\vec{Y}^* = (Y_1^*, Y_2^*, \dots, Y_N^*) \in$

$(\mathbb{R}^0_+)^N$ such that

$$\begin{aligned}
&\int_{\mathbf{K}} c_n(x,a)\nu^*(dx \times da) - d_n \le 0, n = 1,2,\dots,N,\\
&\underline{L}(\nu^*, \vec{Y}^*) = \inf_{\nu \in \mathcal{D}} \underline{L}(\nu, \vec{Y}^*),\\
&\sum_{n=1}^{N} Y_n^* \left(\int_{\mathbf{K}} c_n(x,a)\nu^*(dx \times da) - d_n \right) = 0.
\end{aligned}$$

Hence, $\underline{\mathcal{D}}_{\vec{Y}^*}$ is nonempty. Its convexity is obvious.

Remark 2.5 *Below in this proof* $\vec{Y}^*$ *and* ν^* *come from Preliminary observation 1 and are fixed.*

Preliminary observation 2: If $\underline{\nu}$ is an extreme point of $\underline{\mathcal{D}}_{\vec{Y}^*}$, then it is also an extreme point of $\mathcal{D}$.

Indeed, if $\underline{\nu} = \lambda\nu_1 + (1-\lambda)\nu_2$, where $\lambda \in (0,1), \nu_1, \nu_2 \in \mathcal{D}$, then $\underline{L}(\underline{\nu}, \vec{Y}^*) = \inf_{\nu \in \mathcal{D}} \underline{L}(\nu, \vec{Y}^*) = \lambda \underline{L}(\nu_1, \vec{Y}^*) + (1-\lambda)\underline{L}(\nu_2, \vec{Y}^*)$ because of the linearity of the Lagrange function $\underline{L}(\cdot, \vec{Y}^*)$. However, this further leads to that $\nu_1, \nu_2 \in \underline{\mathcal{D}}_{\vec{Y}^*}$. Since $\underline{\nu}$ is an extreme point of $\underline{\mathcal{D}}_{\vec{Y}^*}$, it must be that $\nu_1 = \nu_2$. Hence, $\underline{\nu}$ is also an extreme point of $\mathcal{D}$.

Preliminary observation 3: The set $\underline{\mathcal{D}}_{\vec{Y}^*} \subseteq \mathcal{D}$ is w-weakly compact.

Since $\mathcal{D}$ is w-weakly compact (see Theorem 2.1), it suffices to show that $\underline{\mathcal{D}}_{\vec{Y}^*}$ is w-weakly closed in $\mathcal{D}$. We observe that the function $\underline{L}(\cdot, \vec{Y}^*)$ is (w-weakly) lower semicontinuous in $\mathcal{D}$, which follows from the lower semicontinuity of the functions $c_n(x,a), n = 0,1,\dots,N$. Now consider a sequence of $\underline{\nu}_l \in \underline{\mathcal{D}}_{\vec{Y}^*}$ and $\underline{\nu} \in \mathcal{D}$ such that $\underline{\nu}_l \overset{w}{\to} \underline{\nu}$. Then by the ($w$-weak) lower semicontinuity of $\underline{L}(\cdot, \vec{Y}^*)$ and [1, Lem.2.42], we have $\inf_{\nu \in \mathcal{D}} \underline{L}(\nu, \vec{Y}^*) \le \underline{L}(\underline{\nu}, \vec{Y}^*) \le \underline{\lim}_{l \to \infty} \underline{L}(\underline{\nu}_l, \vec{Y}^*) = \inf_{\nu \in \mathcal{D}} \underline{L}(\nu, \vec{Y}^*)$, i.e., $\underline{\nu} \in \underline{\mathcal{D}}_{\vec{Y}^*}$. Thus, $\underline{\mathcal{D}}_{\vec{Y}^*}$ is w-weakly closed in $\mathcal{D}$.

Preliminary observation 4: Define the mapping $\underline{Z} : \underline{\mathcal{D}}_{\vec{Y}^*} \to \mathbb{R}^N$ by

$$\begin{aligned}
\underline{Z}(\nu) \ := \ & \left(\int_{\mathbf{K}} c_1(x,a)\nu(dx \times da), \int_{\mathbf{K}} c_2(x,a)\nu(dx \times da), \dots, \right.\\
& \left. \int_{\mathbf{K}} c_N(x,a)\nu(dx \times da) \right).
\end{aligned}$$

Then the non-empty set $Q := \{\underline{Z}(\nu) : \nu \in \underline{\mathcal{D}}_{\vec{Y}^*}\}$ is convex compact in $\mathbb{R}^N$.

It is evident that $\underline{Z}(\cdot)$ is (w-weakly) continuous on $\underline{\mathcal{D}}_{\vec{Y}^*}$ (see Assumption 2.4(b)). Now the compactness of Q follows from the (w-weak) continuity of $\underline{Z}(\cdot)$, the (w-weak) compactness of $\underline{\mathcal{D}}_{\vec{Y}^*}$ (see Preliminary observation 3) and [1, Thm.2.34], while the convexity of Q is because of the linearity of $\underline{Z}(\cdot)$.

Main proof of part (b) of the theorem: By Krein-Milman's theorem (see Theorem C.2) and Preliminary observation 4, the set Q is the convex hull of its extreme points, which together with Carathéodory's convexity theorem (see Theorem C.3), implies that every point in Q can be represented as a convex combination of $N+1$ extreme points of Q. Thus,

$$\underline{Z}(\nu^*) = \sum_{n=1}^{N+1} \lambda_n^* g_n, \tag{2.11}$$

where ν^* solves Problem (2.1) and is defined earlier in this proof, $\forall\ n = 1, 2, \dots, N+1, \lambda_n^* \in [0,1]$, $\sum_{n=1}^{N+1} \lambda_n^* = 1$, and g_n is an extreme point of Q. The mapping $\underline{Z}$ from $\underline{\mathcal{D}}_{\vec{Y}^*}$ onto Q is (w-weakly) continuous and linear, and the sets $\underline{\mathcal{D}}_{\vec{Y}^*}$ and Q are both convex compact (see Preliminary observations 1,3,4), so that by [67, Chap.XI,T13], $\forall\ n = 1, 2, \dots, N+1$, $\exists$ an extreme point ν_n^* of $\underline{\mathcal{D}}_{\vec{Y}^*}$ satisfying

$$g_n = \underline{Z}(\nu_n^*). \tag{2.12}$$

Now we claim that $\nu^{Opt} := \sum_{n=1}^{N+1} \lambda_n^* \nu_n^*$ is the required solution to Problem (2.1). Indeed, according to the linearity of $\underline{Z}(\cdot)$, (2.11) and (2.12), we have $\underline{Z}(\nu^{Opt}) = \underline{Z}(\nu^*)$, which further implies that (2.8) and (2.10) are satisfied (see the definition of $\underline{Z}(\cdot)$ in Preliminary observation 4 and the definition of $\vec{Y}^*$ in Preliminary observation 1). The validity of (2.9) follows from the convexity of $\underline{\mathcal{D}}_{\vec{Y}^*}$ and $\nu^{Opt} \in \underline{\mathcal{D}}_{\vec{Y}^*}$.

Finally, since $\nu_n^*, n = 1, 2, \dots, N+1$ are extreme points of $\underline{\mathcal{D}}_{\vec{Y}^*}$ and thus extreme points of $\mathcal{D}$ (see Preliminary observation 2), it only remains to apply Proposition 2.1 for the existence of the corresponding deterministic stationary policies f_n generating ν_n^*, i.e., $\nu^{f_n} = \nu_n^*, n =$

$1, 2, \dots, N+1.$ □

Part (b) of the above theorem means that there exists an optimal mixing policy (in the sense of [2]) on the set of deterministic stationary policies $(f_n)_{n=1,2,\dots,N+1}$. Generally speaking, the class of deterministic stationary policies is not sufficient for constrained problems, see [39, Ex.1]. On the other hand, when the problem is unconstrained, i.e., $N = 0$, one can show the existence of an optimal deterministic stationary policy, based on the dynamic programming approach as done in [50, Chap.9], and some of their results are formulated in Lemma 2.1 below. Nevertheless, the presented convex analytic approach in this chapter is more powerful in dealing with constrained problems.

2.5 Duality results

Problem (2.7) is often referred to as the PLP of the underlying absorbing MDP model. In the sequel, under Assumption 2.1, Assumption 2.2(a,b) we formulate its DLP, and under some further conditions we prove the absence of the duality gap.

Following [50, Chap.12], we need the following objects:

- two dual pairs $(\mathcal{X}, \mathcal{Y})$ and $(\mathcal{Z}, \mathcal{V})$,
- a (weakly) continuous linear mapping from $\mathcal{X}$ to $\mathcal{Z}$,
- a positive cone Co in $\mathcal{X}$ and its dual cone Co^* in $\mathcal{Y}$,
- two fixed points, namely $B \in \mathcal{Z}$ and $C \in \mathcal{Y}$.

Below we denote by $\mathcal{M}_w^{\pm}(\mathbf{K})$ the set of signed measures on $\mathbf{K}$ with a finite w-norm, i.e., $\forall\ M \in \mathcal{M}_w^{\pm}(\mathbf{K})$, we have $\int_{\mathbf{K}} w(x)|M|(dx \times da) < \infty$, where $|M|$ denotes the total variation of M. Similarly, $\mathcal{M}_w^{\pm}(\mathbf{S})$ denotes the set of signed measures on $\mathbf{S}$ with a finite w-norm.

Two dual pairs $(\mathcal{X}, \mathcal{Y})$ **and** $(\mathcal{Z}, \mathcal{V})$ **:** We consider the following four linear spaces defined by

$$\begin{aligned}
\mathcal{X} &:= \mathcal{M}_w^{\pm}(\mathbf{K}) \times \mathbb{R}^N \\
&= \{X = (\nu, b_1, \ldots, b_N) : \nu \in \mathcal{M}_w^{\pm}(\mathbf{K}), b_n \in \mathbb{R}, n = 1, 2, \ldots, N\}, \\
\mathcal{Y} &:= \mathbf{B}_w(\mathbf{K}) \times \mathbb{R}^N \\
&= \{Y = (f, e_1, \ldots, e_N) : f \in \mathbf{B}_w(\mathbf{K}), e_n \in \mathbb{R}, n = 1, 2, \ldots, N\}, \\
\mathcal{Z} &:= \mathcal{M}_w^{\pm}(\mathbf{S}) \times \mathbb{R}^N \\
&= \{Z = (z_0, k_1, \ldots, k_N) : z_0 \in \mathcal{M}_w^{\pm}(\mathbf{S}), k_n \in \mathbb{R}, n = 1, 2, \ldots, N\}, \\
\mathcal{V} &:= \mathbf{B}_w(\mathbf{S}) \times \mathbb{R}^N \\
&= \{V = (v', j_1', \ldots, j_N') : v' \in \mathbf{B}_w(\mathbf{S}), j_n' \in \mathbb{R}, n = 1, 2, \ldots, N\}.
\end{aligned}$$

With the bilinear forms

$$\langle X, Y \rangle := \int_{\mathbf{K}} f(x, a)\nu(dx \times da) + \sum_{n=1}^{N} e_n b_n$$

and

$$\langle Z, V \rangle := \int_{\mathbf{S}} v'(x) z_0(dx) + \sum_{n=1}^{N} k_n j_n',$$

we finally have the promised two dual pairs, namely $(\mathcal{X}, \mathcal{Y})$ and $(\mathcal{Z}, \mathcal{V})$.

Remark 2.6 *(a) We equip $\mathcal{X}$ with the weak topology generated by all elements of $\mathcal{Y}$ when viewed as linear functionals on $\mathcal{X}$ through $\langle \cdot, Y \rangle$. We denote this weak topology by $\tau(\mathcal{X}, \mathcal{Y})$. By similarly understanding the notation of $\tau(\mathcal{Y}, \mathcal{X})$ and so on, we have four topological linear spaces: $(\mathcal{X}, \tau(\mathcal{X}, \mathcal{Y}))$, $(\mathcal{Y}, \tau(\mathcal{Y}, \mathcal{X}))$, $(\mathcal{Z}, \tau(\mathcal{Z}, \mathcal{V}))$ and $(\mathcal{V}, \tau(\mathcal{V}, \mathcal{Z}))$. These weak topologies are compatible with the underlying bilinear forms (see [1, p.211-215]).*
(b) Evidently, $\mathcal{M}_w(\mathbf{K})$ (defined earlier) is closed in $(\mathcal{M}_w^{\pm}(\mathbf{K}), \tau(\mathcal{M}_w^{\pm}(\mathbf{K})))$, where $\tau(\mathcal{M}_w^{\pm}(\mathbf{K}))$ denotes the weak topology on $\mathbf{K}$ generated by the set of all w-bounded continuous functions on $\mathbf{K}$. Under Assumption 2.1(b,c) and Assumption 2.2(a,b), $\mathcal{D}$ is closed in $(\mathcal{M}_w^{\pm}(\mathbf{K}), \tau(\mathcal{M}_w^{\pm}(\mathbf{K})))$ because of Theorem 2.1(a) and the aforementioned observation.

A continuous linear mapping from $(\mathcal{X}, \tau(\mathcal{X}, \mathcal{Y}))$ to $(\mathcal{Z}, \tau(\mathcal{Z}, \mathcal{V}))$:

Consider a linear mapping from $\mathcal{X}$ to $\mathcal{Z}$, namely U, defined by $\forall\ X = (\nu, \beta_1, \beta_2, \ldots, \beta_N) \in \mathcal{X}, U \circ X = Z = (z_0, k_1, k_2, \ldots, k_N)$, where $\forall\ \Gamma_S \in \mathcal{B}(\mathbf{S})$,

$$z_0(\Gamma_S) = \hat{\nu}(\Gamma_S) - \int_{\mathbf{K}} Q(\Gamma_S|y, a)\nu(dy \times da)$$

with $\hat{\nu}$ denoting the projection of ν on $\mathbf{S}$, and $\forall\ n = 1, \ldots, N, k_n = \int_{\mathbf{K}} c_n(x, a)\nu(dx \times da) + b_n$. Its adjoint mapping, namely U^*, is defined by $\forall\ V = (v', j_1', j_2', \ldots, j_N') \in \mathcal{V}, U^* \circ V = Y = (f, e_1, e_2, \ldots, e_N) \in \mathcal{Y}$, where

$$f(x, a) = v'(x) - \int_{\mathbf{S}} v'(y)Q(dy|x, a) + \sum_{n=1}^{N} j_n' c_n(x, a),$$

and $e_n = j_n', \forall\ n = 1, \ldots, N$. Indeed, the relation $\langle U \circ X, V\rangle = \langle X, U^* \circ V\rangle$ can be directly verified. Now one can infer from [50, Prop.12.2.5] for that U is the required continuous linear mapping from $(\mathcal{X}, \tau(\mathcal{X}, \mathcal{Y}))$ to $(\mathcal{Z}, \tau(\mathcal{Z}, \mathcal{V}))$.

A positive cone Co in $\mathcal{X}$ and its dual cone Co^* in $\mathcal{Y}$: We fix the following positive cone in $\mathcal{X}$, namely $Co = \{(\nu, b_1, \ldots, b_N) : \nu(dx \times da) \geq 0, b_n \geq 0, n = 1, \ldots, N\}$. Evidently, its dual cone is given by $Co^* = \{(f, e_1, \ldots, e_N) : f(x, a) \geq 0, e_n \geq 0, n = 1, \ldots, N\}$.

Two fixed points, namely $B \in \mathcal{Z}$ and $C \in \mathcal{Y}$: We take $B := (\gamma, d_1, \ldots, d_N)$ and $C := (c_0, 0, \ldots, 0)$.

Now we may rewrite Problem (2.7) as

$$\begin{aligned} &\langle X, C\rangle \to \min_{X \in \mathcal{X}} \\ s.t.& \\ &U \circ X = B; X \in Co. \end{aligned}$$

Its DLP, by the materials presented in [50, Chap.12], is

$$\begin{aligned} &\langle B, V\rangle \to \max_{V \in \mathcal{V}} \\ s.t.& \\ &C - U^* \circ V \in Co^*, \end{aligned}$$

which can be explicitly written out as follows:

$$
\begin{aligned}
& \int v'(x)\gamma(dx) + \sum_{n=1}^{N} d_n j'_n \to \max_{(v',j'_1,\dots,j'_N)\in\mathcal{V}} \\
s.t.\quad & \\
& c_0(x,a) - v'(x) + \int_{\mathbf{S}} v'(y)Q(dy|x,a) - \sum_{n=1}^{N} j'_n c_n(x,a) \geq 0; \\
& -j'_n \geq 0,\ \ n = 1,2,\dots N.
\end{aligned}
$$

After the change of variables through $j_n := -j'_n$ and $v(x) := v'(x) - \sum_{n=1}^{N} d_n j_n$, the above DLP takes the following more familiar form:

$$
\begin{aligned}
& \int_{\mathbf{S}} \gamma(dx)v(x) \to \max_{(v,j_1,\dots,j_N)} \qquad (2.13) \\
s.t.\quad & \\
& c_0(x,a) + \sum_{n=1}^{N} j_n\left(c_n(x,a) - d_n\right) - v(x) + \int_{\mathbf{S}} v(y)Q(dy|x,a) \geq 0; \\
& j_n \geq 0,\ \ n = 1,\dots,N; \\
& v \in \mathbf{B}_w(\mathbf{S}).
\end{aligned}
$$

Below we denote by $\inf(PLP)$ and $\sup(DLP)$ the values of the PLP and DLP, respectively.

Assumption 2.5 *(a)* $\forall\ x \in \mathbf{S}$, $A(x)$ *is compact.*
(b) $\forall\ x \in \mathbf{S}, n = 0,1,\dots,N$, $c_n(x,a)$ *is lower semicontinuous in* $a \in A(x)$.
(c) $\forall\ x \in \mathbf{S}$, $\int_{\mathbf{S}} u(y)Q(dy|x,a)$ *is continuous in* $A(x)$, *where* $u(\cdot)$ *is any bounded measurable function on* $\mathbf{S}$.
(d) $\forall\ x \in \mathbf{S}$, $\int_{\mathbf{S}} w(y)Q(dy|x,a)$ *is continuous in* $A(x)$.

Assumption 2.5 validates [50, Thm.9.6.10] about the *dynamic programming approach* for unconstrained problems, which we do not intend to investigate in detail in this chapter. In any case, the statement of [50, Thm.9.6.10] is quoted together with its direct consequences (see Lemma 2.1 below) and needed in the proof of Theorem 2.3 below.

Theorem 2.3 *Suppose Assumption 2.1, Assumption 2.2(a,b) and Assumption 2.5 are satisfied. Then the following assertions hold.*
(a)

$$-\infty < \sup(DLP(2.13)) \leq \inf(PLP(2.7)) < \infty.$$

Moreover, if X is feasible for PLP (2.7), V is feasible for DLP (2.13), and $\langle X, U^* \circ V\rangle = 0$, then X is optimal for PLP (2.7) and V is optimal for DLP (2.13).
(b) If in addition Assumption 2.4 is satisfied, and $c_0(x,a)$ is continuous on $\mathbf{K}$, then

$$-\infty < \sup(DLP(2.13)) = \inf(PLP(2.7)) < \infty.$$

The following lemma facilitates the proof of Theorem 2.3, and we only sketch its proof below.

Lemma 2.1 *For an absorbing MDP model, suppose Assumption 2.1 and Assumption 2.5 are satisfied. Let $\overline{c}(x,a) := c_0(x,a) + \sum_{n=1}^{N} j_n(c_n(x,a) - d_n)$, where $j_n \geq 0$ and $d_n \in \mathbb{R}$ are arbitrarily fixed. Then the following assertions hold.*
(a) There is a unique solution in $\boldsymbol{B}_w(\mathbf{S})$ to the following optimality equation

$$v^*(x) = \inf_{a\in A(x)} \left\{ \overline{c}(x,a) + \int_{\mathbf{S}} v^*(y) Q(dy|x,a) \right\}, \ x \in \mathbf{S}.$$

Moreover, this solution $v^*(x)$ is the optimal value of the unconstrained problem

$$E_x^{\pi}\left[\sum_{t=0}^{\infty} \overline{c}(x_t, a_t)\right] \to \min_{\pi\in U^H}.$$

(b) Let $v^*(\cdot)$ be the solution coming from part (a). Then $\int_{\mathbf{S}} \gamma(dx) v^*(x)$ is the optimal value of the unconstrained problem

$$E_\gamma^{\pi}\left[\sum_{t=0}^{\infty} \overline{c}(x_t, a_t)\right] \to \min_{\pi\in U^H}.$$

(c) Let $v^*(\cdot)$ be the solution coming from part (a). Then it solves the

following linear program

$$\begin{aligned} & \int_{\mathbf{S}} \gamma(dx)v(x) \to \max_v \qquad (2.14)\\ s.t. & \\ & \overline{c}(x,a) - v(x) + \int_{\mathbf{S}} v(y)Q(dy|x,a) \geq 0, \forall\ (x,a) \in \mathbf{K};\\ & v \in \boldsymbol{B}_w(\mathbf{S}). \end{aligned}$$

Sketched proof. (a) This part follows from [50, Thm.9.6.10].

(b) By Proposition 2.1, it is sufficient to be restricted to the class of stationary policies. Therefore, in this proof we always assume $\varphi \in U^S$. The proof now follows from the simple observation that for any measurable function u on S such that $u(\Delta_S) = 0$ and $\sup_{x\in\mathbf{S}} \frac{|u(x)|}{w(x)} < \infty$, we have $\forall\ m = 1, 2, \ldots,$

$$\begin{aligned} E_\gamma^\varphi \left[u(x_m)\right] = & \int_{\mathbf{S}} \gamma(dx)u(x)\\ + E_\gamma^\varphi & \left[\sum_{t=0}^{m-1} \left\{ \int_{\mathbf{S}} u(y) \int_A Q(dy|x_t,a)\varphi(da|x_t) - u(x_t) \right\} \right]. \qquad (2.15) \end{aligned}$$

Indeed, one only needs add to the both sides of (2.15) $E_\gamma^\varphi \left[\sum_{t=0}^{m} \overline{c}(x_t,a_t)\right]$, and pass to the limit $m \to \infty$, and use the facts that $v^* \in \mathbf{B}_w(\mathbf{S})$ (see part (a)) and $E_\gamma^\varphi \left[u(x_m)\right] \to 0$ as $m \to \infty$. Incidentally, equation (2.15) is just a simple version of Dynkin's formula.

(c) We start this proof with the observation that v^* is obviously feasible for Problem (2.14). It follows from (2.15) that for any $\varphi \in U^S$, if $u \in \mathbf{B}_w(\mathbf{S})$ and

$$u(x) \leq \int_A \varphi(da|x)\overline{c}(x,a) + \int_{\mathbf{S}} \int_A \varphi(da|x)Q(dy|x,a)u(y), \forall\ x \in \mathbf{S},$$

then $u(x) \leq E_x^\varphi \left[\sum_{t=0}^{\infty} \overline{c}(x_t,a_t)\right]$. From this we infer that for any feasible solution v to Problem (2.14), it holds that $v(x) \leq E_x^\varphi \left[\sum_{t=0}^{\infty} \overline{c}(x_t,a_t)\right]$, where $\varphi \in U^S$ can be arbitrarily fixed. Now let us arbitrarily fix such a feasible solution v, and suppose $\int_{\mathbf{S}} \gamma(dx)v(x) > \int_{\mathbf{S}} \gamma(dx)v^*(x)$. Then there exists some $\hat{x} \in \mathbf{S}$ and constant $\epsilon > 0$ such that $v^*(\hat{x}) < v(\hat{x}) - \epsilon$. Hence, $v^*(\hat{x}) < E_{\hat{x}}^\varphi \left[\sum_{t=0}^{\infty} \overline{c}(x_t,a_t)\right] - \epsilon$, where $\varphi \in U^S$ can be any stationary policy. However, this is against part (a) of this theorem, as desired. □

Proof of Theorem 2.3.(a) This part follows from Lemma 2.1 and [47, Thm.6.2.4].

(b) According to Assumption 2.4, Theorem 2.1(a), Remark 2.6(b), Proposition 2.1(d), the fact of $c_0(x,a)$ being continuous on $\mathbf{K}$ and [76, Ex.1", p.45], we have that [76, Thm.17(a)] is valid. As a result,

$$\inf(PLP(2.7)) = \sup_{j_n \geq 0, n=1,2,\ldots,N} \inf_{\nu \in \mathcal{D}} \left\{ \int_{\mathbf{K}} \overline{c}(x,a)\nu(dx \times da) \right\}, \quad (2.16)$$

where $\overline{c}(x,a) = c_0(x,a) + \sum_{n=1}^{N} j_n(c_n(x,a) - d_n)$ as defined in the statement of Lemma 2.1 above. Having fixed $j_n \geq 0, n = 1,2,\ldots,N$, the problem of

$$\int_{\mathbf{K}} \overline{c}(x,a)\nu(dx \times da) \to \min_{\nu \in \mathcal{D}} \quad (2.17)$$

takes the same form as PLP (2.7): it is just an unconstrained case. Its DLP takes the form of DLP (2.13):

$$\begin{aligned} &\int_{\mathbf{S}} \gamma(dx)v(x) \to \max_{v \in \mathbf{B}_w(\mathbf{S})} \quad (2.18)\\ s.t. \quad &\overline{c}(x,a) - v(x) + \int_{\mathbf{S}} v(y)Q(dy|x,a) \geq 0. \end{aligned}$$

By Lemma 2.1 we have

$$\begin{aligned} \inf(PLP(2.17)) &= \int_{\mathbf{S}} \gamma(dx)v^*(x) \\ = \quad \sup(DLP(2.18)) &= \sup_{v \in \mathbf{B}_w(\mathbf{S})} \int_{\mathbf{S}} \gamma(dx)v(x) \end{aligned}$$

(s.t. the constraints in (2.14) with fixed $j_n \geq 0$), where $v^*(\cdot)$ is the optimality function coming from Lemma 2.1(a). Now on the one hand, taking $\sup_{j_1 \geq 0, j_2 \geq 0, \ldots, j_N \geq 0}$ to the both sides of the above equalities yields

$$\sup_{j_1 \geq 0, j_2 \geq 0, \ldots, j_N \geq 0} \inf(PLP(2.17)) = \sup(DLP(2.13)).$$

On the other hand, we have

$$\inf(PLP(2.7)) = \sup_{j_1 \geq 0, j_2 \geq 0, \ldots, j_N \geq 0} \inf(PLP(2.17))$$

because of (2.16). Thus,

$$\sup(DLP(2.13)) = \inf(PLP(2.7)).$$

The fact that both the quantities are finite is easily seen. □

Part (a) of this theorem says that the weak duality and the complementary slackness hold, while part (b) shows the absence of the duality gap.

2.6 A discounted MDP model with a state-action-dependent discount factor

The constrained discounted MDP model under our consideration is the following 8-tuple

$$\{\mathbf{S}, \mathbf{A}, A(x), \hat{Q}(dy|x,a), \hat{c}_0(x,a), (\hat{c}_n(x,a), d_n)_{n=1,\ldots,N}, \gamma(dy), \beta(x,a)\},$$

where the state and action spaces $\mathbf{S}, \mathbf{A}$ are arbitrary non-empty Borel spaces, and $0 \leq \beta(x,a) < 1$ is a measurable function defined on $\mathbf{K}$, representing the state-action-dependent discount factor, and all the other primitives are understood similarly to previously.

Having fixed a policy π and constructed the probability measure $\hat{P}^\pi_\gamma$ in the canonical way, where $\hat{P}^\pi_\gamma$ signifies the underlying transition probability to be $\hat{Q}(dy|x,a)$, the discounted problem (assumed to be well defined for this moment) reads

$$V_0(\pi,\gamma) := \hat{E}^\pi_\gamma\left[\sum_{t=0}^{\infty}\prod_{m=0}^{t-1}\beta(x_m,a_m)\hat{c}_0(x_t,a_t)\right] \to \min_{\pi\in U^H} \tag{2.19}$$

s.t.

$$V_n(\pi,\gamma) := \hat{E}^\pi_\gamma\left[\sum_{t=0}^{\infty}\prod_{m=0}^{t-1}\beta(x_m,a_m)\hat{c}_n(x_t,a_t)\right] \le d_n, n=1,2,\dots,N.$$

As before, it is our standing assumption that $U^{Feasible} \neq \emptyset$, and U^H denotes the set of history-dependent policies.

As pointed out in [2, 50, 91], a standard discounted MDP model can be reformulated as an undiscounted one with total cost criteria

$$\{S, A, A(x), Q(dy|x,a), c_0(x,a), (c_n(x,a), d_n)_{n=1,\dots,N}, \gamma(dy)\}.$$

This also holds when the discount factor is state-action-dependent. Indeed, one may introduce an isolated point Δ_S and enlarge the state space from $\mathbf{S}$ to $S := \mathbf{S}\bigcup\{\Delta_S\}$, and correspondingly enlarge the action space from $\mathbf{A}$ to $A := \mathbf{A}\bigcup\{\Delta_A\}$ where Δ_A is the newly introduced isolated point corresponding to Δ_S, i.e., $A(\Delta_S) := \{\Delta_A\}$ and $A(x) \subseteq \mathbf{A}, \forall\ x \neq \Delta_S$. The transition probabilities $Q(dy|x,a)$ and the cost functions $c_n(x,a), n = 0,1,\dots,N$ of the transformed undiscounted MDP model with total cost criteria are given by

$$\begin{aligned}
&Q(dy\setminus\{\Delta_S\}|x,a) := \beta(x,a)\hat{Q}(dy|x,a),\\
&Q(\{\Delta_S\}|x,a) = 1-\beta(x,a), c_n(x,a) := \hat{c}_n(x,a), \forall\ (x,a)\in\mathbf{K};\\
&Q(\{\Delta_S\}|\Delta_S,\Delta_A) := 1, c_n(\Delta_S,\Delta_A) := 0.
\end{aligned}$$

The equivalence between the discounted and the transformed undiscounted models should be clear from the canonical construction and the form of performance functionals. Otherwise, a formal proof could be obtained by applying the same reasoning as in the proofs of [91, Lem.2,Thm.3] with some obvious notational complications.

It is clear that the policies for the original discounted model and those for the transformed model essentially characterize each other, see [91], so that without leading to confusion we use the same notation for both.

In order to apply the results derived in the previous section to analyze the underlying non-standard discounted MDP model, we need some conditions to guarantee the transformed model to be absorbing.

Assumption 2.6 *There exists a constant* $0 \leq \hat{\beta} < 1$ *satisfying* $\sup_{(x,a)\in \mathbf{K}} \beta(x,a) < \hat{\beta}$ *and a measurable function* w *on* S : $w(x) \geq 1$ *on* $\mathbf{S}$ *and* $w(\Delta_S) = 0$ *such that the following holds.*
(a) $\int_{\mathbf{S}} w(y)\hat{Q}(dy|x,a) \leq \frac{1}{\hat{\beta}} w(x)$, $\forall\ x \in \mathbf{S}$.
(b) $\sup_{x\in\mathbf{S}} \frac{\sup_{a\in A(x)} |\hat{c}_n(x,a)|}{w(x)} < \infty$, $\forall\ n = 0, 1, \ldots, N$.
(c) $\int_{\boldsymbol{S}} w(x)\gamma(dx) < \infty$.

It is a simple observation that under Assumption 2.6, the inequality presented in Remark 2.1(d) is satisfied with $b = 0 = l(x), \forall\ x \in S$, $\beta = \frac{\sup_{(x,a)\in\mathbf{K}} \beta(x,a)}{\hat{\beta}}$. Consequently, the transformed model is absorbing. Moreover, under Assumption 2.6 Problem (2.19) is well defined.

Assumption 2.7 *(a) The function* $\underline{\beta}(\cdot)$ *on* S *defined by* $\underline{\beta}(y) := \sup_{a\in A(y)} \beta(y,a)$ *on* $\mathbf{S}$ *and* $\underline{\beta}(\Delta_S) := 0$ *is measurable.*
(b) There exists a constant $0 \leq \delta < 1$ *such that* $\int_{\mathbf{S}} \underline{\beta}(y)\hat{Q}(dy|x,a) \leq \delta, \forall\ (x,a) \in \mathbf{K}$.
(c) The functions $\hat{c}_n(x,a), n = 0, 1, \ldots, N$ *are all bounded on* $\mathbf{K}$.

Assumption 2.7(a) is satisfied if $A(x)$ is compact and $\beta(x,a)$ is upper semicontinuous in $a \in A(x)\ \forall\ x \in \mathbf{S}$ because of [50, Prop.D.5(c)], or if $\beta(x,a)$ is a-independent, in which case by slightly abusing the notation we use $\beta(x)$ for $\beta(x,a)$.

Under Assumption 2.7, the transformed model is absorbing. Indeed, the inequality in Remark 2.1(d) is satisfied with $w(x) \equiv 1$ on $\mathbf{S}$, $\beta = 0$, $b = 1$, $\hat{l} = \delta$ and $l(x) = \underline{\beta}(x)$, because of

$$\int_{\mathbf{S}} 1 \cdot Q(dy|x,a) \leq \sup_{a\in A(x)} \beta(x,a) = \underline{\beta}(x),$$

$$\begin{aligned}&\int_{\mathbf{S}} \underline{\beta}(y)Q(dy|x,a) = \int_{\mathbf{S}} \sup_{a\in A(y)} \beta(y,a)Q(dy|x,a)\\&\leq \beta(x,a)\delta \leq \delta \sup_{a\in A(x)} \beta(x,a) = \delta\underline{\beta}(x),\end{aligned}$$

$$\begin{aligned}E_x^\pi\left[\underline{\beta}(x_t)\right] &= E_x^\pi\left[E_x^\pi\left[\underline{\beta}(x_{t+1})|x_0,a_0,\ldots,x_t,a_t\right]\right]\\&= E_x^\pi\left[\int_{\mathbf{S}} \underline{\beta}(y)Q(dy|x_t,a_t)\right]\\&\leq \delta E_x^\pi\left[\underline{\beta}(x_t)\right]\end{aligned}$$

and a simple inductive argument.

Remark 2.7 *On the one hand, the discounted MDP model with a state-action-dependent discount factor appears more general than the standard one (with a constant discount factor) prevailingly considered in the literature such as [2, 50, 72, 75]. On the other hand, it is easily seen that the requirement for the discount factor $\beta(x,a)$ to be separated from one, as in Assumption 2.6, is essentially the same as requiring the underlying dynamic programming operator to be contracting, which validates the reasoning in [50, Chap.8]. Hence, a discounted MDP model with such a state-action-dependent discount factor is essentially the same as a discounted model with a constant discount factor. Another variant of the discounted MDP model, which allows a randomized discount factor, is considered in [42]. However, it is treated in the same way as for the standard discounted model after using the trick of enlarging the state space. The constrained discounted MDP model in Borel spaces with unbounded cost functions and a state-action-dependent discount factor not separated from one seems to be still underdeveloped (but see [85]).*

Chapter 3

More constrained absorbing MDP with total cost criterion

3.1 Introduction

This chapter is organized as follows: Section 3.2 is about a more constrained MDP model, its reformulation as an unconstrained one and the relationship between two policies in each model; Section 3.3 is devoted to the dynamic programming setting , the optimality results and its proof.

3.2 Problem formulation and assumptions

A more constrained MDP model $\mathcal{M}$ in Borel spaces with total cost criterion is a 7-tuple

$$\{S, A, A(x), Q(dy|x,a), c_0(x,a), (c_n(x,a), d_n)_{n=1,\dots,N}, \gamma(dy)\},$$

All the components are understood identically to those in Chapter 2. Again, we are concerned with a MDP model which is absorbing in the sense of Definition 2.1.

Under Assumption 2.1 and 2.2, the more constrained problem of our interest is defined as follows,

$$W_0(\pi) := E_\gamma^\pi \left[\sum_{t=0}^{\infty} c_0(x_t, a_t) \right] \to \min_\pi \tag{3.1}$$

$s.t.$

$$W_n(\pi) := \sup_{T \geq 1} \; E_\gamma^\pi \left[\sum_{t=0}^{T-1} c_n(x_t, a_t) \right] \leq d_n, n = 1, 2, \ldots, N.$$

In order to facilitate our analysis, $\mathcal{M}$ is reformulated into an equivalent deterministic model $\widetilde{\mathcal{M}}$, denoted by

$$\{\tilde{S}, \tilde{A}, (\tilde{A}(P), P \in \mathcal{P}_w(S)), (\mathcal{H}(\tilde{a}), \mathcal{V}(\vec{W}, \tilde{a})), C(P, \vec{W}, \tilde{a})\}$$

where,

- $\tilde{S} := \mathcal{P}_w(S) \times \mathbb{R}^N$, recall that $S := \mathbf{S} \bigcup \{\Delta_S\}$ and Δ_S is an isolated point;

- $\tilde{A} := \mathcal{P}_w(\mathbb{K})$, recall that $A := \mathbf{A} \bigcup \{\Delta_A\}$, Δ_A is an isolated point, $\mathbb{K} := \{(x, a) \in S \times A : x \in S, a \in A(x)\}$;

- $\tilde{A}(P) := \{\tilde{a} \in \tilde{A} \mid \tilde{a}(dx \times A) = P(dx)\}$, where $P \in \mathcal{P}_w(S)$;

- $P_{t+1}(\Gamma_\mathbf{S}) = \mathcal{H}(\tilde{a}_t)(\Gamma_\mathbf{S}) := \int_{\mathbf{S} \times \mathbf{A}} \tilde{a}_t(dy \times da) Q(\Gamma_\mathbf{S} | y, a) \quad \forall \, \Gamma_\mathbf{S} \in \mathcal{B}(\mathbf{S})$
  $P_{t+1}(\Delta_S) = \mathcal{H}(\tilde{a}_t)(\Delta_S) := \int_{\mathbf{S} \times \mathbf{A}} \tilde{a}_t(dy \times da) Q(\Delta_S | y, a) + P_t(\Delta_S)$

- $W_{t+1}^n = \mathcal{V}(W_t^n, \tilde{a}_t) := W_t^n + \int_{\mathbf{S} \times \mathbf{A}} c_n(x, a) \tilde{a}_t(dx \times da) \quad n = 1, 2, \ldots, N.$

- $C(P_t, \vec{W}_t, \tilde{a}_t) := \begin{cases} \int_{\mathbf{S} \times \mathbf{A}} c_0(x, a) \tilde{a}_t(dx \times da) & \text{when} \quad \vec{W}_t \leq \vec{d} \\ +\infty & \text{otherwise} \end{cases}$

The new state is a probability measure on the original state space S along with an N-dimensional real-valued vector, and the new action is a probability measure on the graph $\mathbb{K}$. Note that these measures are finite in the w-norm, where $w(\cdot)$ comes from Definition 2.1(d). As mentioned in Remark 2.1(b), the absorbing set Δ_S can be compressively treated as a singleton $\{\Delta_S\}$, so is Δ_A as $\{\Delta_A\}$. The admissible action space is a measurable subset of $\tilde{A}$ for any $P \in \mathcal{P}_w(S)$, for each of which the projection coincides with P. Because of $w(\Delta_S) = 0$, $\mathcal{P}_w(\mathbf{S}) \subseteq \mathcal{P}_w(S)$, and $\mathcal{P}_w(\mathbf{K}) \subseteq \mathcal{P}_w(\mathbb{K})$. By the above definitions, the new policy is a sequence of newly defined admissible actions, i.e., $\tilde{\pi} = (\tilde{a}_t)_{t=0,1,2,\ldots}$. Note that only deterministic (Markov) policies are considered. Main consequences of the reformulation are embodied in two respects. Firstly, $\widetilde{\mathcal{M}}$ is essentially a deterministic model in the sense that the trajectory of the process is determined whenever the initial state (a probability measure) and a policy $\tilde{\pi}$ is given. In particular, we have transition functions $\mathcal{H}$ and $\mathcal{V}$ based on the transition kernel $Q(dy|x, a)$, which in turn corresponds to each part of $\tilde{S}$ respectively. Denote by $\vec{W}_t := (W_t^1, W_t^2, \ldots, W_t^N)$ and $\vec{d} := (d_1, d_2, \ldots, d_N)$ two real-valued vectors, and $\vec{W}_t \leq \vec{d}$ is understood in the sense of componentwise. Secondly, $\widetilde{\mathcal{M}}$ is an unconstrained model, where the source of constraints in the model $\mathcal{M}$ is incorporated into the structure of the new cost function $C(\cdot, \cdot)$ incorporates. $C(\cdot, \cdot)$ is defined similarly to the penalty function (see [65]), and keeps track of accumulated costs, which enables us to determine whether each of the constraints is satisfied up to every time step. To be specific, $C(\cdot, \cdot)$ penalizes the violation of any constraint with $+\infty$ cost immediately after its occurrence, otherwise computes the expected value of cost with respect to c_0 as usual when all the constraints are satisfied. The new initial state is a probability measure $P_0 \in \mathcal{P}_w(S)$ concentrated on $\mathbf{S}$ along with an N-dimensional vector $\vec{W}_0$. Accordingly, the weight function is modified as $\tilde{w}(P, \vec{W}) := \int_{\mathbf{S}} w(x)P(dx)$ for every $P \in \mathcal{P}_w(S)$.

Before we proceed with the statement of our main result, it is necessary to reveal and reassure some topological properties of new state and action space. In fact, $\tilde{S}$ and $\tilde{A}$ are Borel spaces. Firstly, a Borel space adjoining an isolated point is Borel. Secondly, the space of probability measures on a Borel space is Borel (see [12, Cor.7.25.1]). Thirdly, the space of probability measures on an arbitrary Borel space is homeomorphic to the space of probability measures on the same space with finite w-norm (see the paragraph prior to Remark 1.2 and [73] for more detail). Next, N-dimensional Euclidean space is Polish, and is Borel as well. Finally, the product space of two Borel spaces is Borel.

Within the context of the reformulated model $\widetilde{\mathcal{M}}$, Problem (3.1) can be rewritten in an integrated and neat form as follows,

$$J(\tilde{\pi},(P_0,\vec{W}_0)) := \sum_{t=0}^{\infty} C(P_t,\vec{W}_t,\tilde{a}_t) \to \min_{\tilde{\pi}} \tag{3.2}$$

Note that Problem (3.2) is well defined in view of [8, Thm.B.1.1]. Specifically, for each $P_0 \in \mathcal{P}_w(S)$, we have

$$\begin{aligned}
\sup_{\tilde{\pi}} \left[\sum_{t=0}^{\infty} C^-(P_t,\vec{W}_t,\tilde{a}_t)\right] &= \sup_{\pi\in U^H} E^{\pi}_{P_0}\left[\sum_{t=0}^{\infty}\left(\int_{\mathbf{S}\times\mathbf{A}} c_0(x,a)\tilde{a}_t(dx\times da)\right)^-\right] \\
&\leq \sup_{\pi\in U^H}\int_S E^{\pi}_y\left[\sum_{t=0}^{\infty}\hat{c}w(x_t)\right]P_0(dy) \\
&= \hat{c}k\int_S w(y)P_0(dy) = \hat{c}k\tilde{w}(P_0,\vec{W}_0) < \infty \qquad (3.3)
\end{aligned}$$

by noticing Definition 2.1(d).

In view of (3.2) we can see that the performance functional produces the value of $+\infty$ only for infeasible policies. For feasible ones, it is an easy observation that there exists a one-to-one correspondence between deterministic (Markov) policies $\tilde{\pi} = (\tilde{a}_t)_{t=0,1,2,\ldots}$ and Markov policies $\pi^M = (\pi^M_t)_{t=0,1,2,\ldots}$ for the model $\mathcal{M}$:

$$\tilde{\pi} \leftrightarrow \pi^M : \; \tilde{a}_t(dx\times da) = \pi^M_t(da|x)P_t(dx) \tag{3.4}$$

By the well-known Derman-Strauch Lemma (see Lemma 1.1, or [29] for the original version), the class of randomized Markov policies is sufficient for optimality problems with total cost criterion whenever the initial distribution is fixed. The connection between Problem (3.1) and Problem (3.2) is illustrated as follows: if there exists a deterministic optimal policy $\tilde{\pi} = (\pi_t)_{t=0,1,2,\ldots}$ to Problem (3.2), then the optimal value of Problem (3.1) is equal to $J(\tilde{\pi},(\gamma,\vec{0}))$, where $\vec{0} = (0,0,\ldots,0)$ is the N-dimensional null vector, and the optimal Markov policy to Problem (3.1) can be uniquely determined by the relation in (3.4). Here and below, we shall concentrate on Problem (3.2) and $\widetilde{\mathcal{M}}$ instead of (3.1). The main advantage is that this reformulation makes it possible not only to establish the optimality equation, but to characterize the optimal Markov policies explicitly which will be revealed in the next section.

3.3 Main statements and proofs

We define the operator in the context of $\widetilde{\mathcal{M}}$ as

$$Lu(P, \vec{W}) := C(P, \vec{W}, \tilde{a}) + u(\mathcal{H}(\tilde{a}), \mathcal{V}(\vec{W}, \tilde{a})), \tag{3.5}$$

and the corresponding optimality operator as

$$Tu(P, \vec{W}) := \inf_{\tilde{a} \in \tilde{A}(P)} \{C(P, \vec{W}, \tilde{a}) + u(\mathcal{H}(\tilde{a}), \mathcal{V}(\vec{W}, \tilde{a}))\}. \tag{3.6}$$

Let

$$J_n(\tilde{\pi}, (P_0, \vec{W}_0)) := \sum_{t=0}^{n} C(P_t, \vec{W}_t, \tilde{a}_t)$$

and

$$J_n((P_0, \vec{W}_0)) := \inf_{\tilde{\pi}} J_n(\tilde{\pi}, (P_0, \vec{W}_0))$$

be $(n+1)$-stage value function with respect to $\tilde{\pi}$, and $(n+1)$-stage optimal value function respectively.

An extra set of assumptions is needed to fit the new model.

Assumption 3.1 *(a) The functions $c_n(x, a), n = 0, 1, 2, \ldots, N$ are all continuous on* $\mathbf{K}$*;*
(b) The state-action space $\mathbf{K}$ *is compact, or the state space* $\mathbf{S}$ *is denumerable;*
(c) $\int_S w(y) Q(dy|x, a)$ is continuous in $(x, a) \in \mathbf{K}$.

Theorem 3.1 *(a) Under Assumption 2.1, Assumption 2.2(a,b), Assumption 3.1,*

$$J_\infty(P, \vec{W}) := \inf_{\tilde{\pi}} J(\tilde{\pi}, (P, \vec{W})) = \lim_{n \to \infty} J_n((P, \vec{W}))$$

is lower semicontinuous, is the minimum solution to the optimality equation

$$u(P, \vec{W}) = Tu(P, \vec{W}) \tag{3.7}$$

out of the class of lower semicontinuous functions bounded by $\tilde{w}(P,\vec{W})$ *from the below. In addition, a measurable mapping* f^* *from* $\tilde{S}$ *to* $\tilde{A}$ *attaining the infimum in the right hand side of* (3.7) *corresponds to an optimal deterministic stationary policy.*
(b) $J_\infty(\gamma,\vec{0})$ *is the optimal value of Problem (3.1), and the above optimal policy* f^* *defines a randomized Markov policy for Problem (3.1).*

Proof. First of all, we make a series of preliminary observations, which as a whole form a version of compactness-continuity conditions.

Observation 1: $C(P,\vec{W},\tilde{a})$ is lower semicontinuous on $\tilde{S}\times\tilde{A}$.

Recall the definition of $C(P,\vec{W},\tilde{a})$,

$$C(P,\vec{W},\tilde{a}) := \begin{cases} \int_{\mathbf{S}\times\mathbf{A}} c_0(x,a)\tilde{a}(dx\times da) & \text{if} \qquad \vec{W}\le\vec{d} \\ \infty & \text{otherwise} \end{cases}$$

We set aside $\vec{W}$ for a while, and focus on the other two arguments. Let $(P_i,\tilde{a}_i)$ be an arbitrary sequence of pairs that converges to some point $(P,\tilde{a})$ in the w-weak topology where $P_i,P\in\mathcal{P}_w(S)$, and $\tilde{a}_i,\tilde{a}\in\mathcal{P}_w(\mathbb{K})$ for each $i=1,2,\ldots$. Note that P_i and P are projections of $\tilde{a}_i$ and $\tilde{a}$ respectively since all the actions are admissible. In view of Assumption 2.1(a) and 3.1(a), $c_0\in\mathbf{C}_w(\mathbf{K})$. That is, $C(P,\vec{W},\tilde{a})$ is continuous in $\mathcal{P}_w(S)\times\mathcal{P}_w(\mathbb{K})$ (see Remark 1.2). Indeed, $\vec{W}$ plays the same role of an indicator function, which maintains the value of $C(P,\vec{W},\tilde{a})$ calculated by the integration when $\vec{W}\le\vec{d}$. Thus, an easy observation gives that $\underline{\lim}_{i\to\infty}C(P_i,\vec{W}_i,\tilde{a}_i)\ge C(P,\vec{W},\tilde{a})$, for an arbitrary sequence $\vec{W}_i\to\vec{W}$ in the sense of componentwise.

Observation 2: For any continuous and $\tilde{w}$-bounded function u on $\tilde{S}$, $u'(P,\vec{W},\tilde{a}) := u(\mathcal{H}(\tilde{a}),\mathcal{V}(\vec{W},\tilde{a}))$ is continuous on $\tilde{S}\times\tilde{A}$ and also $\tilde{w}$-bounded.

It suffices to investigate the continuity of two transition functions

$\mathcal{H}(\tilde{a})$ and $\mathcal{V}(\vec{W}, \tilde{a})$ in $\tilde{a}$. Recall the definitions,

$$\begin{aligned}
\mathcal{H}(\tilde{a})(\Gamma_{\mathbf{S}}) &:= \int_{\mathbf{S}\times\mathbf{A}} \tilde{a}(dy \times da) Q(\Gamma_{\mathbf{S}}|y,a) \quad \mathbf{S} \in \mathcal{B}(\mathbf{S}) && (3.8)\\
\mathcal{H}(\tilde{a})(\Delta_S) &:= \int_{\mathbf{S}\times\mathbf{A}} \tilde{a}_t(dy \times da) Q(\Delta_S|y,a) + P_t(\Delta_S) && (3.9)
\end{aligned}$$

$$\mathcal{V}(W^n, \tilde{a}) := W^n + \int_{\mathbf{S}\times\mathbf{A}} c_n(x,a)\tilde{a}(dx \times da) \qquad (3.10)$$

For (3.8) and (3.9), let $(\tilde{a}_i)$ be a sequence that converges to $\tilde{a}$ in the w-weak topology where $\tilde{a}_i, \tilde{a} \in \tilde{A}(P)$ for each $i = 1, 2, \ldots$. We have

$$\begin{aligned}
\lim_{i\to\infty} \int_{\mathbf{S}} g(x) P_t(\tilde{a}_i)(dx) &= \lim_{i\to\infty} \int_{\mathbf{S}} g(x) \int_{\mathbf{S}\times\mathbf{A}} \tilde{a}_i(dy \times da) Q(dx|y,a)\\
&= \lim_{i\to\infty} \int_{\mathbf{S}\times\mathbf{A}} \int_{\mathbf{S}} g(x) Q(dx|y,a) \tilde{a}_i(dy \times da)\\
&= \lim_{i\to\infty} \int_{\mathbf{S}\times\mathbf{A}} g'(x,a) \tilde{a}_i(dy \times da)\\
&= \int_{\mathbf{S}\times\mathbf{A}} g'(x,a) \tilde{a}(dy \times da) && (3.11)
\end{aligned}$$

where $g(\cdot) \in \mathbf{C}_w(\mathbf{S})$. $g'(y,a) := \int_{\mathbf{S}} g(x) Q(dx|y,a)$ belongs to $\mathbf{C}_w(\mathbf{K})$ because of Assumption 2.2(b) and 3.1(c) and Lemma A.1. The proof for (3.10) follows in a similar manner by Assumption 2.1(a), Assumption 3.1(a) and the fact that the continuity of composition of two continuous functions preserves.

Note that $\mathcal{H}(\tilde{a}) \in \mathcal{P}_w(S)$ for each $P \in \mathcal{P}_w(S)$ by (3.11) with $g(\cdot)$ being replaced with $w(\cdot)$. So the $\tilde{w}$-boundedness of $u'(\cdot,\cdot)$ is justified.

Observation 3: The set-valued mapping $P \to \tilde{A}(P)$ is upper semicontinuous.

By upper semicontinuity of set-valued mapping $\tilde{A}(P)$, we refer to that for a sequence (P_i) that converges to P in the w-weak topology and arbitrarily chosen elements $\tilde{a}_i \in \tilde{A}(P_i)$, there exists a limit point $\tilde{a} \in \tilde{A}(P)$ for the sequence $(\tilde{a}_i)$; see Definition A.2(a). In view of Assumption 3.1(b), if $\mathbf{S}$ is denumerable, the continuity issue is automatically addressed. Otherwise, note that $\tilde{a}_i \in \tilde{A}(P_i) \subseteq \mathcal{P}_w(\mathbb{K})$. By Assumption 3.1(b) and [12, Prop.7.22], $\mathcal{P}(\mathbb{K})$ is compact in the usual weak topology, so is $\mathcal{P}_w(\mathbb{K})$ in the w-weak topology (see (1.5) and (1.6)). Therefore, $\bigcup_i \{\tilde{a}_i\}$ is relatively w-compact in $\mathcal{P}_w(\mathbb{K})$. That is, there exists a subsequence $(\tilde{a}_{i_k})$ that converges to some point $\tilde{a} \in \mathcal{P}_w(\mathbb{K})$ in the w-weak topology.

Let $g \in \mathbf{C}_w(\mathbf{S})$ be an arbitrary function. On the one hand, we have

$$\lim_{k\to\infty} \int_{\mathbf{S}\times\mathbf{A}} g(x)\tilde{a}_{i_k}(dx \times da) = \lim_{k\to\infty} \int_{\mathbf{S}} g(x)P_{i_k}(dx) = \int_{\mathbf{S}} g(x)P(dx)$$

which follows from $P_i \overset{w}{\to} P$ in w-weak topology. On the other hand,

$$\lim_{k\to\infty} \int_{\mathbf{S}\times\mathbf{A}} g(x)\tilde{a}_{i_k}(dx \times da) = \int_{\mathbf{S}\times\mathbf{A}} g(x)\tilde{a}(dx \times da) = \int_{\mathbf{S}} g(x)\tilde{a}(dx \times A)$$

as $\tilde{a}_{i_k} \overset{w}{\to} \tilde{a}$. Thus, $P(dx) = \tilde{a}(dx \times A)$, i.e., $\tilde{a} \in \tilde{A}(P)$.

Observation 4: $\tilde{A}(P)$ is w-weakly compact in $\mathcal{P}_w(\mathbb{K})$ for each $P \in \mathcal{P}_w(S)$.

It would be convenient to formulate an auxiliary model $\widehat{\mathcal{M}}$ which is viewed as a special case of $\mathcal{M}$.

- $\widehat{S} := S \times \{0, 1\}$;
- $\widehat{A} := A \bigcup \{\Delta_{\widehat{A}}\}$;
- $\widehat{A}((x, 0)) := A(x), \forall\ x \in S$; $\widehat{A}((x, 1)) := \Delta_{\widehat{A}}, \forall\ x \in S$;
- $\hat{Q}((dy, 1)|(x, 0), a) := Q(dy|x, a), \forall\ a \in A(x)$;
  $\hat{Q}((dy, 1)|(x, 1), \Delta_{\widehat{A}}) := \delta_x(dy)$;
- the initial distribution is $(P_0, 0)$, where $P_0(dx)$ is a probability measure on S.

For the sake of simplicity, we denote by $S_0 := S \times \{0\}$, $S_1 := S \times \{1\}$. S_0 and S_1 can be viewed and treated as two versions of the original state space S, which are understood in the same way below. From the above formulation, $S_1 = \{(x, 1) : x \in S\}$ is an absorbing set, thus further compressively treated as an absorbing point. Obviously, $\widehat{\mathcal{M}}$ satisfies Definition 2.1(a,b). Definition 2.1(c) is skipped as the cost is irrelevant to this model. For Definition 2.1(d), the weight function is modified as

$$\hat{w}(x,i) = \begin{cases} w(x) & \text{when} \quad x \in \mathbf{S}, i = 0 \\ 1 & \text{when} \quad x \in \Delta_S, i = 0 \\ 0 & \text{when} \quad i = 1 \end{cases} \tag{3.12}$$

Then the left hand side of Definition 2.1(d) is equal to 1. Note that the corresponding process automatically enters the cemetery S_1 after the first movement when the initial state is in S_0, and we are not concerned with the evolution afterwards. As a consequence, it suffices to consider the one-step policy taking the form $\pi = (\pi_0, \pi_1, \dots)$, where $\pi_0(da|x)$ is a stochastic kernel from S to A such that $\pi_0(A(x)|x) = 1$, and $\pi_1, \pi_2, \dots$ are arbitrarily selected. Accordingly, the policy $\hat{\pi} = (\hat{\pi}_0, \hat{\pi}_1, \dots)$ corresponding to $\widehat{\mathcal{M}}$ is defined as $\hat{\pi}_0(da|(x,0)) := \pi_0(da|x), \forall\ x \in S$, and $\hat{\pi}_n(da|(x,1)) := \delta_{\Delta_{\hat{A}}}((x,1)), \forall\ x \in S,\ n = 1, 2, \dots$. Note that under this formulation, only the first element of two polices defined above corresponds to each other. By canonical construction there exists a unique probability measure $\widehat{P}^{\hat{\pi}}_{\gamma}$ on the space of trajectories, $(\hat{x}_0, \hat{a}_0, \hat{x}_1, \hat{a}_1, \dots)$. One can write the explicit form of its occupation measure according to Definition 2.2,

$$\begin{aligned} \hat{\nu}^{\hat{\pi}}(\Gamma_{S_0} \times \Gamma_A) &= \sum_{t=0}^{\infty} \widehat{P}^{\hat{\pi}}_{P_0}(\hat{x}_t \in \Gamma_{S_0}, \hat{a}_t \in \Gamma_A) \\ &= \widehat{P}^{\hat{\pi}}_{P_0}(\hat{x}_0 \in \Gamma_{S_0}, \hat{a}_0 \in \Gamma_A) \\ &= \int_{\Gamma_{S_0}} \widehat{P}^{\pi}_{x}(\hat{a}_0 \in \Gamma_A) P_0(dx) \\ &= \int_{\Gamma_{S_0}} \int_{\Gamma_A} \pi_0(da|x) P_0(dx) \end{aligned}$$

where $\Gamma_{S_0} \times \Gamma_A \in \mathcal{B}(S_0 \times A)$, and $P_0 \in \mathcal{P}_w(S)$.

It is interesting to note that the occupation measure associated with the model $\widehat{\mathcal{M}}$ given the initial distribution $P_0(\cdot)$ coincides with the admissible action space in the model $\widetilde{\mathcal{M}}$ given the same measure $P_0(\cdot)$.

That is, $\widehat{D} = \tilde{A}(P_0)$, where $\widehat{D}$ denotes the space of occupation measures for the model $\widehat{\mathcal{M}}$. Therefore, Theorem 2.1(b) is applicable if all the conditions remains satisfied for the model $\widehat{\mathcal{M}}$ as well. Clearly, Assumption 2.1(b) and 2.2(a) hold directly from the definition of $\hat{w}(x,i)$; see (3.12). Assumption 2.1(c) and 2.2(b) are satisfied again by noting that the corresponding process is certain to enter the cemetery S_0 immediately after the first step, meaning that $\int_{S_0} u(y)Q(dy|(x,0),\hat{a}) \equiv 0$. Assumption 2.2(c) directly follows with the newly defined moment function $\hat{v}((\Delta_S,0),\Delta_A) := 0$ and $\hat{v}((x,0),a) := v(x,a), \forall x \in \mathbf{S}, a \in \mathbf{A}$.

With the above four observations in mind, we continue to verify remaining assumptions presented in Chapter 4. First of all, it is a simple observation that

$$R_{(P,\vec{W}),\tilde{a}}(v(\cdot)) := v(\mathcal{H}(\tilde{a}), \mathcal{V}(\vec{W},\tilde{a}))$$

is a coherent risk measure in accordance with Definition 4.1. Let $\mathbb{H}$ be the family of extended real-valued function $v(P,\vec{W})$ on $\tilde{S}$ such that

$$\sup_{(P,\vec{W})\in\tilde{S}} \frac{v^-(P,\vec{W})}{\tilde{w}(P,\vec{W})} < \infty.$$

Assumption 4.1 is automatically satisfied, since only deterministic Markov policies are under our consideration. Part (a) of Assumption 4.2 follows from (3.3), and part (b) is not needed by the assertion in Remark 4.2(b).

Note that the following relation holds for each initial state (respectively, predetermined probability measure) for the model $\widetilde{\mathcal{M}}$ (respectively, $\mathcal{M}$),

$$\begin{aligned} \sup_{\tilde{\pi}} \sum_{t=m+1}^{\infty} C^-(P_t,\vec{W}_t,\tilde{a}_t) &\leq \hat{c}\sup_{\tilde{\pi}} \sum_{t=m+1}^{\infty} \int_{\mathbf{S}} w(y)P_t(dy) \qquad (3.13)\\ &= \hat{c} \sup_{\pi^M\in U^M} \sum_{t=m+1}^{\infty} \int_{\mathbf{S}} w(y) P_{P_0}^{\pi^M}(x_t\in dy)\\ &= \hat{c} \sup_{\pi^M\in U^M} E_{P_0}^{\pi^M}\left[\sum_{t=m+1}^{\infty} w(x_t)\right] \end{aligned}$$

Thus, we define

$$\delta_m(P_0,\vec{W}_0) := \hat{c} \sup_{\pi^M\in U^M} E_{P_0}^{\pi^M}\left[\sum_{t=m+1}^{\infty} w(x_t)\right].$$

Observe that $\delta_m(\cdot)$ satisfies Assumption 4.3(a) because of Definition 2.1(d). Suppose $\pi^M = (\pi_0^M, \pi_1^M, \dots)$ is an arbitrary randomized Markov policy, define the corresponding one-stage shifted policy by $\pi^{M'} := (\pi_0, \pi_0^M, \pi_1^M, \ldots)$, where $\tilde{a}_0(dx\times da) = \pi_0(da|x)P_0(dx)$

$$\begin{aligned}
R_{(P_0,\vec{W}_0),\tilde{a}_0}(\delta_m(\cdot)) &= \delta_m(\mathcal{H}(\tilde{a}_0), \mathcal{V}(\vec{W}_0,\tilde{a}_0)) \\
&\le \sup_{\tilde{\pi}} \sum_{t=m+1}^{\infty} \int_{\mathbf{K}} c_0(y,a)\tilde{a}_t(dy\times da) \\
&\le \hat{c}\sup_{\tilde{\pi}} \sum_{t=m+1}^{\infty} \int_{\mathbf{S}} w(y)P_t(dy) \\
&= \hat{c} \sup_{\pi^M\in U^M} E_{\mathcal{H}(\tilde{a}_0)}^{\pi^M}\left[\sum_{t=m+1}^{\infty} w(x_t)\right] \\
&= \hat{c} \sup_{\pi^{M'}\in U^M} E_{P_0}^{\pi^{M'}}\left[\sum_{t=m+2}^{\infty} w(x_t)\right] \\
&= \delta_{m+1}(P_0,\vec{W}_0).
\end{aligned}$$

So, $C_m \equiv 1$ validates Assumption 4.3(b). Part (c) follows from (3.13) and the definition of $\delta_m(\cdot)$, whereas part (d) is satisfied by the fact $\delta_{-1}(P_0,\vec{W}_0) \le \hat{c}k\tilde{w}(P_0)$. The verification of Assumption 4.4 is trivial. Part (a) of Assumption 4.5 is a direct consequence of Observation 1-4 (see the discussion in [36]), whereas part (b) follows from Observation 2 and Definition 2.1.

All the statements in part (a) of Theorem 3.3, analogous to Theorem 4.1, should follow.

Denote by $\tilde{a}_t^* := f^*(P_t)$ for $t = 0,1,2,\ldots$ the induced optimal deterministic stationary policy for Problem (3.2). Obviously, one can disintegrate it in the same way as in (3.4),

$$\tilde{a}_t^*(dx\times da) = P_t(dx)\pi_t^{M^*}(da|x) \quad t = 0,1,2,\ldots$$

where $\pi^{M^*} = (\pi_t^{M^*})_{t=0,1,2,\ldots}$ corresponds to a randomized Markov policy for Problem (3.1). □

Remark 3.1 *(a) The value iteration and policy iteration algorithm are established in the statement of Theorem 3.1, see Remark 4.2 for details. This facilitates the determination of the optimal value, as well as approximating or obtaining an optimal stationary policy for Problem (3.1).*
(b) On the one hand, the establishment of an optimal randomized Markov policy for Problem (3.1) *does make sense due to the more constrained setup in the formulation of Problem* (3.1)*. On the other hand, the reformulation of $\mathcal{M}$ into $\widetilde{\mathcal{M}}$ makes it possible to deduce an optimal randomized Markov policy π^{M^*} for Problem (3.1), whenever an optimal deterministic stationary policy f^* for Problem (3.2) is obtained. As illustrated in [65], the latter fact is one of the advantages and aims of applying dynamic programming approach to constrained MDPs.*
(c) In addition, we observe that the converse statement of the last part in Theorem 3.1 remains valid. That is, if f^* is an optimal deterministic stationary policy to Problem 3.2, then

$$J_\infty(P,\vec{W}) = C(P,\vec{W},f^*(P)) + J_\infty(\mathcal{H}(f^*(P)),\mathcal{V}(\vec{W},f^*(P))).$$

Actually, a deterministic stationary policy satisfying the above equation is called a conserving policy. Briefly, a deterministic stationary policy is optimal to Problem 3.2 if and only if it is conserving. As a consequence, 3.7 is enough to characterize all the optimal policies to Problem 3.1.

Chapter 4

MDP with Iterated Coherent Risk Measures

4.1 Introduction

This section attempts to extend the dynamic programming for standard Markov decision processes (MDPs) with the expected total cost criterion to the case, where the (iterated) coherent risk measure of the cost is taken as the performance measure to be minimized.

This chapter is organized as follows. We introduce the notion of coherent risk measures, present the assumptions and state the optimal control problem in Section 4.2. Section 4.3 is about the optimality results together with its proof. Section 4.4 consists of the standard MDP and the MDP with iterated conditional value-at-risk as two illustrative examples.

4.2 Problem formulation and assumptions

In order to describe the concerned MDP model, we present some notations and definitions first.

As is seen in Chapter 1, the standard MDP model is made up of the five-tuple

$$\{S, A, (A(x), x \in S), Q(dy|x, a), c(x, a)\}.$$

We attempt to redefine one of its element, i.e., $Q(dy|x, a)$, to obtain our new model. In what follows, for any function $u(\cdot)$ and constant u, $u^+ := \max\{u, 0\}$ and $u^- := \max\{0, -u\}$. We also regard ∞ and $-\infty$ as constants.

Let $\mathcal{H}$ be a linear subspace of the space of all the extended real-valued measurable functions on S, and contains all the real constants.

Definition 4.1 *A mapping $R(\cdot)$ from $\mathcal{H}$ to $\overline{\mathbb{R}} := [-\infty, \infty]$ is called a sublinear expectation if the following conditions are satisfied;*

(a) $R(v(\cdot)) \leq R(u(\cdot))$ for each $v \leq u$ with $u, v \in \mathcal{H}$ (here and below the inequality $v \leq u$ is understood in the pointwise sense);

(b) $R(c) = c$ for each constant $c \in \mathbb{R}$;

(c) $R(v(\cdot) + u(\cdot)) \leq R(v(\cdot)) + R(u(\cdot))$ for each $u, v \in \mathcal{H}$; and

(d) $\lambda R(v(\cdot)) = R(\lambda v(\cdot))$ for each $v \in \mathcal{H}$ and $\lambda \in (0, \infty)$.

Throughout this chapter we put $R(\infty) := \infty$, $R(-\infty) := -\infty$, $\infty - \infty := \infty$ and $0 \cdot \pm\infty := 0$. Item (c) above is called the sublinearity of $R(\cdot)$, following from which we further have for each $u, v \in \mathcal{H}$,

$$R(u(\cdot) - v(\cdot)) \quad \geq \quad R(u(\cdot)) - R(v(\cdot)) \tag{4.1}$$

if $R(v(\cdot)) < \infty$.

It is easy to see that the sublinear expectation R is convex in the sense of $R(\lambda v(\cdot) + (1-\lambda)u(\cdot)) \leq \lambda R(v(\cdot)) + (1-\lambda)R(u(\cdot))$ for each $u, v \in \mathcal{H}$ and $\lambda \in [0, 1]$; and is translation invariant in the sense of $R(v(\cdot) + c) = R(v(\cdot)) + c$ for each $v \in \mathcal{H}$ and $c \in \mathbb{R}$. Thus, in consistency with [90] we call a sublinear expectation R a coherent risk measure throughout the rest of this chapter.

A coherent risk measure $R_{x,a}(\cdot)$ parameterized by $(x, a) \in \mathbb{K}$ is called a risk mapping (cf. [90]) if for each fixed $(x, a) \in \mathbb{K}$, $R_{x,a}(\cdot)$ is a coherent risk measure; and for each $v \in \mathcal{H}$, $R_{x,a}(v(\cdot))$ is measurable in $(x, a) \in \mathbb{K}$.

The MDP model under consideration is characterized by the following primitives $\{S, A, (A(x), x \in S), R_{x,a}(\cdot), c(x, a)\}$, for which we define the (Markov) policy as follows.

Definition 4.2 *A policy $\pi = (\pi_n)_{n=0,1,\ldots}$ is a sequence of stochastic kernels $\pi_n(da|x)$ on $\mathcal{B}(A)$ given $x \in S$ such that for each $x \in S$ and $n = 0, 1, 2, \ldots$, $\pi_n(A(x)|x) = 1$. A policy is called (randomized) stationary if the stochastic kernels π_n are independent of n. A stationary policy is called deterministic stationary if there is a measurable mapping f from S to A whose graph is contained in $\mathbb{K}$ such that, slightly but conventionally abusing the notation, the stochastic kernel π defining the policy can be written as $\pi(da|x) = \mathbf{1}_{\{f(x)\in da\}}$, where $\mathbf{1}_{\{\cdot\}}$ stands for the indicator function.*

Assumption 4.1 *There exists a subset $\mathbb{H} \subset \mathcal{H}$, which contains all the real constants and is closed under addition and nonnegative scalar multiplication (i.e., $\lambda v \in \mathbb{H}$ for each $v \in \mathbb{H}$ and $\lambda \in [0, \infty)$), such that the functions $\int_A c(\cdot, a)\pi(da|\cdot) \in \mathbb{H}$ and $\int_A R_{\cdot,a}(v(\cdot))\pi(da|\cdot) \in \mathbb{H}$ for each $v \in \mathbb{H}$ and stochastic kernel π from S to $\mathcal{B}(A)$ satisfying $\pi(A(x)|x) = 1$ for each $x \in S$. Furthermore, for each $u, v \in \mathbb{H}$, it holds that* $\min\{u, v\} \in \mathbb{H}$.

Throughout this chapter, all the assumptions, once introduced, are supposed to hold always without explicit indications. We define the operator

$$Lv(x, a) := c(x, a) + \beta R_{x,a}(v(\cdot))$$

for each $v \in \mathcal{H}$, where and in the sequel $\beta \in (0, 1]$ represents the discount factor. Note that $\beta = 1$ is also included so that the undiscounted problem is taken into consideration in the present setting. For notational convenience, let us denote

$$\begin{aligned} R_{x,\pi}(v(\cdot)) &:= \int_A R_{x,a}(v(\cdot))\pi(da|x), \\ c(x, \pi) &:= \int_A c(x, a)\pi(da|x), \text{ and} \\ L_\pi v(x) &:= c(x, \pi) + \beta R_{x,\pi}(v(\cdot)) \end{aligned}$$

for each stochastic kernel π from S to $\mathcal{B}(A)$ such that $\pi(A(x)|x) = 1$ for each $x \in S$.

For each fixed policy $\pi = (\pi_n)$ and $n \geq -1$, we define

$$\begin{aligned} J_n(x,\pi) &:= L_{\pi_0}L_{\pi_1}\dots L_{\pi_n}(0) \\ &= c(x,\pi_0) + \beta R_{x,\pi_0}(c(\cdot,\pi_1) + \beta R_{\cdot,\pi_1}(c(\cdot,\pi_2) \\ &\quad +\beta R_{\cdot,\pi_2}(\dots + \beta R_{\cdot,\pi_{n-1}}(c(\cdot,\pi_n))\dots), \end{aligned} \tag{4.2}$$

where if $n = -1$, $L_{\pi_0}L_{\pi_1}\dots L_{\pi_n}(0) := 0$.

We impose the following before we state the optimal control problem under consideration.

Assumption 4.2 *(a)* $\lim_{n\to\infty} J_n(\cdot,\pi) =: J_\infty(\cdot,\pi)$ *exists under each policy* π*;*
(b) For each $x \in S$*, if* $J_\infty(x,\pi) = \infty$ *for all* π*, then there exists some* n *such that* $\inf_\pi J_n(x,\pi) = \infty$.

Assumption 4.2(a) automatically holds if the cost function is nonnegative due to the monotone convergence, or more generally, under Assumption 4.3 imposed below by Lemma C.1. When $R_{x,a}(v(\cdot)) := \int_S v(y)Q(dy|x,a)$, the model described above becomes the standard MDP, for which, when $\beta = 1$, Assumption 4.2(a) is satisfied if e.g.,

$$\sup_\pi E_x^\pi \left[\sum_{t=0}^{\infty} \max\{-c(x_t,a_t),0\}\right] < \infty$$

for each $x \in S$, see Theorem A.3 of [56] or Remark 3.1(b) of [51]. Here E_x^π is taken with respect to the strategic measure P_x^π on the space of trajectories $(x_0,a_0,x_1,a_1,\dots)$ constructed in the canonical way, and x_t and a_t are the controlled and controlling processes [47, 50]. The previous inequality holds when e.g., the underlying model is absorbing and the negative part of the cost satisfies certain growth conditions, see Chapter 7 of [2] and Section 8 of [51].

Assumption 4.2(b) is not needed if the multifunction $A(\cdot)$ is compact-valued, see Remark 4.2 below.

Then the concerned optimal control problem reads

$$J_\infty(x,\pi) \to \min_\pi, \tag{4.3}$$

to which a policy π^* is called optimal if $J_\infty(x, \pi^*) = \inf_\pi J_\infty(x, \pi)$ for each $x \in S$.

Assumption 4.3 *For each $m \geq -1$, there exists a nonnegative real-valued upper semicontinuous function $\delta_m \in \mathbb{H}$ such that for each $x \in S$, (a) $\delta_m(x) \downarrow 0$ as $m \to \infty$; (b) $R_{x,a}(\delta_m(\cdot)) \leq C_m \delta_{m+1}(x)$ for some sequence of constants $C_m \in [0, \infty)$ satisfying $\sup_m \beta^m (\prod_{i=-1}^{m-2} C_i) < \infty$ and $C_m \delta_{m+1}(x) \to 0$ as $m \to \infty$ for each $x \in S$; (c)* [1]

$$\begin{aligned}\delta_m(x) \quad &\geq \quad \sup_{\pi=(\pi_n)} \beta^{m+1} R_{x,\pi_0}(R_{\cdot,\pi_1} \ldots R_{\cdot,\pi_m}(c^-(\cdot, \pi_{m+1}) + \beta R_{\cdot,\pi_{m+1}}(c^-(\cdot, \pi_{m+2}) \\ &\qquad + \cdots + \beta R_{\cdot,\pi_{n-1}}(c^-(\cdot, \pi_n) + \ldots) \ldots);\end{aligned}$$

and (d) for each convergent sequence $v_n \to v$ with $-\delta_{-1} \leq v_n \in \mathbb{H}$, it holds that $v \in \mathbb{H}$.

Here and below the convergence of a sequence of functions is understood in the pointwise sense.

It now follows from (4.2) and (4.1) that for each $-1 \leq m \leq n$

$$\begin{aligned}J_n(x, \pi) \quad &\geq \quad c(x, \pi_0) + \beta R_{x,\pi_0}(c(\cdot, \pi_1) + \cdots + \beta R_{\cdot,\pi_{m-1}}(c(\cdot, \pi_m) \\ &\qquad -\beta R_{\cdot,\pi_m}(c^-(\cdot, \pi_{m+1}) + \beta R_{\cdot,\pi_{m+1}}(c^-(\cdot, \pi_{m+2}) + \ldots \\ &\qquad +\beta R_{\cdot,\pi_{n-1}}(c^-(\cdot, \pi_n)) \ldots) \\ &\geq \quad L_{\pi_0} L_{\pi_1} \ldots L_{\pi_m}(0) \\ &\qquad -\beta^{m+1} R_{x,\pi_0}(R_{\cdot,\pi_1} \ldots R_{\cdot,\pi_m}(c^-(\cdot, \pi_{m+1}) \\ &\qquad +\beta R_{\cdot,\pi_{m+1}}(c^-(\cdot, \pi_{m+2}) + \cdots + \beta R_{\cdot,\pi_{n-1}}(c^-(\cdot, \pi_n)) \ldots) \\ &= \quad J_m(x, \pi) - \beta^{m+1} R_{x,\pi_0}(R_{\cdot,\pi_1} \ldots R_{\cdot,\pi_m}(c^-(\cdot, \pi_{m+1}) \\ &\qquad +\beta R_{\cdot,\pi_{m+1}}(c^-(\cdot, \pi_{m+2}) + \ldots \\ &\qquad +\beta R_{\cdot,\pi_{n-1}}(c^-(\cdot, \pi_n)) \ldots),\end{aligned} \tag{4.4}$$

[1] If $m = -1$ in the next inequality, the term on the right hand side reads $\sup_{\pi=(\pi_n)}\{c^-(x, \pi_0) + \beta R_{x,\pi_0}(c^-(\cdot, \pi_1) + \cdots + \beta R_{\cdot,\pi_{n-1}}(c^-(\cdot, \pi_n) + \ldots) \ldots)\}$.

where if $m = n$,

$$\begin{aligned}&R_{x,\pi_0}(R_{\cdot,\pi_1}\dots R_{\cdot,\pi_m}(c^-(\cdot,\pi_{m+1})\\&+\beta R_{\cdot,\pi_{m+1}}(c^-(\cdot,\pi_{m+2})+\dots+\beta R_{\cdot,\pi_{n-1}}(c^-(\cdot,\pi_n))\dots):=0.\end{aligned}$$

Assumption 4.4 *For each sequence $-\delta_{-1} \leq v_n \in \mathbb{H}$ such that $v_n \uparrow v \in \mathbb{H}$ as $n \uparrow \infty$, $R_{x,a}(v_n(\cdot)) \uparrow R_{x,a}(v(\cdot))$ for each $(x,a) \in \mathbb{K}$.*

We now impose the last assumption for the optimality result below.

Assumption 4.5 *(a) The cost function c is $\mathbb{K}$-inf-compact, i.e., c is lower semicontinuous on $\mathbb{K}$, and for each $S \ni x_n \to x \in S$ and $a_n \in A(x_n)$ such that $c(x_n, a_n)$ is bounded from the above with respect to n, the sequence (a_n) has a limit point $a \in A(x)$.*
(b) For each lower semicontinuous function $v \in \mathbb{H}$, $R_{x,a}(v(\cdot))$ is lower semicontinuous in $(x,a) \in \mathbb{K}$, and there exists some lower semicontinuous function on S, say $u_v(x) > -\infty$ such that $u_v(x) \leq R_{x,a}(v(\cdot))$ for each $x \in S$ and $a \in A(x)$.

The concept of $\mathbb{K}$-inf-compactness comes from [37]. Assumption 4.5 is an extension of the (weak) continuity-compactness condition widely imposed to obtain the optimality results for the standard MDPs. However, as compared with the literature, we point out that with the help of the recent extension of the Berge's theorem in [37] quoted as Lemma A.3 in the appendix, the standard requirement for the multi-function $A(\cdot)$ to be compact-valued has been removed, and in fact, the condition on the upper semicontinuity of $A(\cdot)$ has also been relaxed given that additional properties are possessed by the cost function. We also refer the interested reader to [66] for another extension of a related Berge's theorem, which the authors apply to showing the continuity of the value function and the deterministic stationary optimal policy of a standard discounted MDP.

4.3 Main statements and proofs

Let us define the operator T by

$$Tv(x) := \inf_{a \in A(x)} \{c(x,a) + \beta R_{x,a}(v(\cdot))\}$$

for each $v \in \mathbb{H}$ and $x \in S$, and consider the function J_n given by that for each $x \in S$,

$$\begin{aligned} J_{-1}(x) &:= 0, \text{ and} \\ J_n(x) &:= \inf_{\pi} J_n(x, \pi) \end{aligned}$$

for each $n = 0, 1, 2, \ldots$.

Lemma 4.1 *For each lower semicontinuous $v \in \mathbb{H}$, $Tv(\cdot) \in \mathbb{H}$ is lower semicontinuous on S, and there exists a measurable mapping f from S to A such that $f(x) \in A(x)$, and $Tv(x) = L_f v(x)$ for each $x \in S$.*

Proof. Let some lower semicontinuous function $v \in \mathbb{H}$ be fixed. Then $c(x,a) + \beta R_{x,a}(v(\cdot))$ is lower semicontinuous in $(x,a) \in \mathbb{K}$ by Assumption 4.5. Now consider an arbitrarily fixed sequence $S \ni x_n \to x \in S$, and $a_n \in A(x_n)$ such that $(c(x_n, a_n) + \beta R_{x_n, a_n}(v(\cdot)))$ is bounded from the above. Under Assumption 4.5(b), we have

$$\inf_{(y,z) \in \{(x_n, a_n), n=1,2,\ldots\}} \beta R_{y,z}(v(\cdot)) \geq \inf_{y \in \{x_n, n=1,2,\ldots\} \bigcup \{x\}} u_v(y) > -\infty,$$

where the last inequality is due to the lower semicontinuity of the function u_v and the compactness of the set $\{x_n, n = 1, 2, \ldots\} \bigcup \{x\}$. It follows that the sequence $(c(x_n, a_n))$ is bounded from the above. This and the $\mathbb{K}$-inf-compactness of the function c asserts that the sequence (a_n) admits a limit point $a \in A(x)$. Hence, the function defined by $c(x,a) + \beta R_{x,a}(v(\cdot))$ is $\mathbb{K}$-inf-compact. One can refer to Lemma A.3 for the statement. □

Remark 4.1 *In the proof of Lemma 4.1, we have established the following useful fact; for each lower semicontinuous $v \in \mathbb{H}$, $Lv(x,a)$ is $\mathbb{K}$-inf-compact (under Assumption 4.5).*

Lemma 4.2 *For each $n \geq -1$, $J_n \in \mathbb{H}$ is lower semicontinuous on S, satisfying $J_{n+1}(x) = TJ_n(x)$ for each $x \in S$. Furthermore, there exists some measurable mapping f_{n+1} from S to A such that $f_{n+1}(x) \in A(x)$ and $J_{n+1}(x) = L_{f_{n+1}}J_n(x) = J_{n+1}(x, \pi^*)$ with $\pi^* = (f_{n+1}, f_n, \ldots, f_0)$ for each $x \in S$.*

Proof. Since $J_{-1}(x, \pi) := 0$ for each π, and the function c is $\mathbb{K}$-inf-compact, we see, according to Lemma A.3, the statement holds for the case of $n = -1$. Suppose the statement is true for $n \leq k$. Let us consider the case of $n = k + 1$. By the inductive supposition, J_k is lower semicontinuous and belongs to $\mathbb{H}$, so that by Lemma 4.1, J_{k+1} is lower semicontinuous on S and belongs to $\mathbb{H}$, and there exists a measurable mapping f_{k+2} such that $TJ_{k+1}(x) = L_{f_{k+2}}J_{k+1}(x)$. This, under Assumption 4.1, implies that TJ_{k+1} belongs to $\mathbb{H}$ and is lower semicontinuous on S by Lemma 4.1. Consider, for the finite horizon problem, the policy $\pi^* = (f_{k+2}, \ldots, f_0)$ given by the measurable selectors taking the corresponding infimums. To complete the inductive argument, it remains to show that $J_{k+2}(x) = TJ_{k+1}(x) = J_{k+2}(x, \pi^*)$. Indeed, for any policy $\pi = (\pi_n)$, by definition, $J_{k+2}(x, \pi) = L_{\pi_0}L_{\pi_1}\ldots L_{\pi_{k+2}}0 \geq L_{\pi_0}L_{f_{k+1}}\ldots L_{f_0}0 = L_{\pi_0}J_{k+1}(x) \geq TJ_{k+1}(x) = L_{f_{k+2}}J_{k+1}(x) = J_{k+2}(x, \pi^*)$. Since the policy π is arbitrarily fixed, we conclude $J_{k+2}(x) = TJ_{k+1}(x) = J_{k+2}(x, \pi^*)$. The proof is completed by induction. □

Lemma 4.3 $\lim_{n\to\infty} J_n(x) =: J(x)$ *exists, belongs to $\mathbb{H}$, and is lower semicontinuous. In addition,* $\lim_{n\to\infty} LJ_n(x, a) = LJ(x, a)$ *exists and is $\mathbb{K}$-inf-compact in $(x, a) \in \mathbb{K}$.*

Proof. The first part of the statement follows from Assumption 4.3, Lemma 4.2 and Lemma C.1, recalling (4.4), which leads to $J_n \geq J_m - \delta_m$ if $m \leq n$. We prove the second part by adopting the reasoning in the proof of Lemma 7.1.5 of [8]. Define $g_m(x) := \inf_{n\geq m} J_n(x)$ for each $x \in S$. Note that $g_m \in \mathbb{H}$ for each m. According to (4.4), Assumption 4.3 and the monotonicity of $R_{x,a}(\cdot)$, we obtain $g_m(x) \leq J_m(x) \leq g_m(x) + \delta_m(x)$ and $Lg_m(x, a) \leq LJ_m(x, a) \leq Lg_m(x, a) + C_m\delta_{m+1}(x)$ for each $x \in S$ and $a \in A(x)$. Thus, $J(x) = \lim_{m\to\infty} J_m(x) = \lim_{m\to\infty} g_m(x)$ and

$\lim_{m\to\infty} LJ_m(x,a) = \lim_{m\to\infty} Lg_m(x,a)$ exist due to the fact that g_m and L are nondecreasing. On the other hand, under Assumptions 4.3 and 4.4, we have $\lim_{m\to\infty} Lg_m(x,a) = L(\lim_{m\to\infty} g_m)(x,a) = LJ(x,a)$. Thus, we see $LJ_m(x,a) \to LJ(x,a)$. That LJ is $\mathbb{K}$-inf-compact is a result of Remark 4.1 and the first part of this statement. □

Now, we are in position to state our main optimality results in the infinite horizon case.

Theorem 4.1 *$J_\infty(x) := \inf_\pi J_\infty(x,\pi) = \lim_{n\to\infty} J_n(x)$ belongs to $\mathbb{H}$, is lower semicontinuous on S, and is the minimal solution out of the class of lower semicontinuous $v \in \mathbb{H}$ such that $v \geq -\delta_{-1}$ to the optimality equation*

$$J(x) = \inf_{a\in A(x)} \{c(x,a) + \beta R_{x,a}(J(\cdot))\} \quad \forall\ x \in S.$$

Moreover, any (and there exists at least one) measurable mapping f^* attaining the infimum in the optimality equation defines a deterministic stationary optimal policy to problem (4.3),

Proof. Let us define for each $n = -1, 0, 1, 2, \ldots, \infty$, and $x \in S$

$$A_n^*(x) := \{a^* \in A(x) : LJ_n(x,a^*) = TJ_n(x)\}, \tag{4.5}$$

and consider the topological upper limit of the sequence $(A_n^*(x))$ defined as

$$LsA_n^*(x) := \{a \in A(x) : \exists a_n \in A_n^*(x), n = 1, 2, \cdots : a_n \to a\}. \tag{4.6}$$

Let us arbitrarily fix some $x \in S$. By Lemma 4.2, $A_n^*(x) \neq \emptyset$ for each $n = -1, 0, \ldots$. So let us arbitrarily take $a_n \in A_n^*(x)$ for each n. Note that, for each $-1 \leq m \leq n$,

$$LJ_m(x,a_n) \leq TJ_n(x) + C_m\delta_{m+1}(x), \tag{4.7}$$

where we recall that δ_{m+1} is nonnegative and real-valued under Assumption 4.3. If $J_{n+1}(x) = TJ_n(x) = \infty$ for some n, then by (4.4) and Assumption 4.3, $J_n(x) = \infty$ for each big enough n. This and Lemma A.3 imply $A_n^*(x) = A(x)$ for all big enough n. Consequently, $LsA_n^*(x) = A(x) \neq \emptyset$ as desired. Suppose now that $J_k(x) < \infty$ for all k. Then Assumption 4.2(b) implies that $J(x) < \infty$ since $J(x) \leq J_\infty(x)$, which follows from that $J(x) = \lim_{n\to\infty} J_n(x) \leq \lim_{n\to\infty} J_n(x,\pi) = J_\infty(x,\pi)$ for each π. Moreover, it follows from (4.7) that $c(x, a_n) = LJ_{-1}(x, a_n) \leq \sup\{J_0(x), J_1(x), J_2(x), \dots, J(x)\} + C_{-1}\delta_0(x) < \infty$, where the last inequality is further by the compactness of the set $\{J_0(x), J_1(x), \dots, J(x)\}$ (recalling Lemma 4.3). Hence, the sequence $(c(x, a_n))$ is bounded from the above. Since c is $\mathbb{K}$-inf-compact, it follows from Lemma A.3 that the sequence (a_n) admits a limit point, and thus $LsA_n^*(x) \neq \emptyset$ as required.

Having established that $LsA_n^*(x) \neq \emptyset$ and keeping in mind (4.7), the proof of Theorem A.1.5 of [8] can be repeated in a word-by-word manner (from the second half of the sixth line in that proof on up to its end) to show that $LsA_n^*(x) \subset A^*(x) := \{a^* \in A(x) : LJ(x, a^*) = TJ(x)\}$, and

$$\lim_{n\to\infty} \inf_{a\in A(x)} LJ_n(x,a) = \inf_{a\in A(x)} \lim_{n\to\infty} LJ_n(x,a),$$

i.e.,

$$J(x) = TJ(x)$$

by Lemma 4.3. By Remark 4.1, there is some measurable mapping f^* from S to A such that $f^*(x) \in A(x)$ for each $x \in S$, satisfying $TJ(x) = L_{f^*}J(x) = L_{f^*}^m J(x)$ for each $x \in S$ and $m \geq 1$. By (4.1) and Assumption 4.3, we see

$$\begin{aligned} J(x) &= L_{f^*}^m J(x) \geq L_{f^*}^m (0 - \delta_{-1})(x) \geq L_{f^*}^m 0 - \beta^m (\prod_{i=-1}^{m-2} C_i)\delta_{m-1}(x) \\ &= J_{m-1}(x, f) - \beta^m (\prod_{i=-1}^{m-2} C_i)\delta_{m-1}(x). \qquad (4.8) \end{aligned}$$

Passing to the limit as $m \to \infty$ on the both sides of the previous inequality leads to $J(x) \geq J_\infty(x, f^*) \geq J_\infty(x)$ under Assumptions 4.2 and 4.3. On the other hand, it always holds that $J(x) \leq J_\infty(x)$ as explained earlier in this proof. Hence, $J(x) = J_\infty(x) = J_\infty(x, f^*)$, and the deterministic stationary policy f^* is optimal.

For the minimality of J_∞ as a solution to the optimality equation, let us consider any solution to the optimality equation $v \in \mathbb{H}$, which is lower semicontinuous and satisfies $v \geq -\delta_{-1}$. Then by Remark 4.1, there exists a measurable mapping f_v from S to A such that $Tv = L_{f_v} v(x)$. The reasoning of (4.8) can be repeated with f being replaced with f_v, J being replaced with v for $v(x) \geq J_\infty(x)$ for each $x \in S$. □

Remark 4.2 (Value iteration and policy iteration algorithms) *(a) According to Lemma 4.2, $J_{n+1}(x) = TJ_n(x)$. Thus, the first line in Theorem 4.1 gives the value iteration algorithm for problem (4.3).*
(b) If the multifunction $A(\cdot)$ is compact-valued, the statement of Theorem 4.1 remains true without requiring Assumption 4.2(b). Indeed, in this case, $LsA_n^*(x)$ defined by (4.6), as a subset of $A(x)$ is automatically nonempty. In addition, for each n, let the measurable mapping f_n^* be such that $f_n^*(x) \in A_n^*(x)$ for each $x \in S$, where $A_n^*(x)$ is defined by (4.5). Then $\bigcup_n \{f_n^*(x)\} \subset A(x)$ is relatively compact. Thus, by Lemma 4 of [84], there is a measurable mapping f^* satisfying $f^*(x) \in LsA_n^*(x) \subset A_\infty^*(x)$ for each $x \in S$, where $LsA_n^*(x)$ and $A_\infty^*(x)$ are as in (4.5) and (4.6). In fact, inspecting the proof of Theorem 4.1 reveals that this mapping f^*, regarded as a deterministic stationary policy, is indeed optimal, which leads to a policy iteration algorithm for Problem (4.3).

4.4 Illustrative examples

We illustrate the obtained results with two examples, where for simplicity we let $A(x)$ be compact. Note that the MDP models in both examples may not be covered by [90] since the cost functions in both examples can be arbitrarily unbounded from the above and can take the value of $+\infty$.

Example 4.1 (Standard MDP)

Let $\mathcal{H}$ be the space of all the extended real-valued measurable functions on S. Suppose $R_{x,a}(v(\cdot)) = \int_S v(y)Q(dy|x,a)$ for each $(x,a) \in \mathbb{K}$, where

$Q(dy|x,a)$ is a stochastic kernel from $\mathbb{K}$ to $\mathcal{B}(S)$. It is straightforward to verify that $R_{x,a}$ satisfies the conditions in Definition 4.1, and Problem (4.3) is reduced to the total cost criterion for the standard MDP. Suppose there exists some continuous (real-valued) function $w(x) \geq 1$ on S such that the $\mathbb{K}$-inf-compact cost function satisfies $c^-(x,a) \leq Cw(x)$ for each $x \in S, a \in A(x)$, and $\int_S w(y)Q(dy|x,a) \leq \alpha w(x)$ for each $x \in S, a \in A(x)$ with α satisfying $\alpha\beta < 1$. Furthermore, for each bounded continuous function f on S, $\int_S f(y)Q(dy|x,a)$, is continuous on $\mathbb{K}$, and so is $\int_S w(y)Q(dy|x,a)$. Then one can take $\mathbb{H}$ as the space of extended real-valued measurable functions u on S satisfying $u^- \leq C_u w$ for some real constant C_u, $\delta_m(x) = \frac{C(\alpha\beta)^{m+1}}{1-\alpha\beta} w(x)$ and $C_m = \frac{1}{\beta}$ for each $m = -1, 0, 1, \ldots$, and $u_v(x) = -\sup_{y \in S} \frac{v^-(y)}{w(y)} \alpha w(x)$ for each $v \in \mathbb{H}$. Then all the optimality results obtained in this chapter apply.

By the way, the condition of $\int_S w(y)Q(dy|x,a) \leq \alpha w(x)$ coincides with Assumption 8.3.2(b) of [50], and examples of MDPs satisfying this condition can be found in Section 8.6 in [50]. In general, this condition can be satisfied by both transient and recurrent (cf. Theorem 7.3.10 [50]) Markov chains (induced by stationary policies). On the other hand, it cannot be satisfied if $\alpha < 1$, unless $Q(dy|x,a)$ is substochastic [50]. Assume this is the case. Then one can complement it by adjoining an isolated cemetery point x_∞ to the state space. Under each stationary policy, this results in a (time-homogeneous) Markov chain in the state space $S \bigcup \{x_\infty\}$. Any subset $\Gamma \in \mathcal{B}(S)$ such that $\sup_{x \in \Gamma} w(x) < \infty$ is transient for this Markov chain, see Remark 8.7 of [51].

Example 4.2 (MDP with iterated conditional value-at-risk)

For the sake of completeness, we include the definitions of conditional value-at-risk and value-at-risk of a random loss or cost based on the tutorial [83]; see also [10, 71, 77, 78].

In line with [83], let X be a random variable, representing the loss with the distribution function given by $F_X(x) = P(X \leq x)$. Let $z \in (0,1)$ be the fixed confidence level. The value-at-risk of X is defined by the left continuous inverse of F_X

$$VaR_z(X) := \min\{x : F_X(x) \geq z\}.$$

It is known that $VaR_z(X)$ is not convex in X (and thus not a coherent risk measure), which makes it mathematically less tractable. In comparison, the conditional value-at-risk $CVaR_z(X)$, introduced in [77], is a coherent risk measure, see also [71, 78], and a single stage problem of optimizing a portfolio of financial instruments is solved with the objective of minimizing the conditional value-at-risk in [77]. In greater detail, the conditional value-at-risk of X is defined as the value of the problem [71]

$$CVaR_z(X) := \inf_{u\in\mathbb{R}} \left\{ u + \frac{1}{1-z} E\left[\max\{X-u, 0\}\right] \right\}.$$

Under some conditions on the distribution function F_X, it holds that

$$CVaR_z(X) = E\left[X | X \geq VaR_z(X)\right],$$

see [71, 78] for more properties of $CVaR_z(X)$. Intuitively, $CVaR_z(X) \leq L$ ensures that the average of $(1-z)\%$ highest losses does not exceed L so that the conditional value-at-risk measures the outcomes that hurt the most, see p.283 of [83], where the comparisons between the concepts of value-at-risk and conditional value-at-risk from the practitioner's viewpoint can be also found.

Let $\mathcal{H}$ be the space of all the extended real-valued measurable functions. Suppose $Q(dy|x,a)$ is a given stochastic kernel from $\mathbb{K}$ to $\mathcal{B}(S)$. Then the conditional value-at-risk at level $z \in (0,1)$ for each $v \in \mathcal{H}$ is defined as

$$\begin{aligned} CVaR_{x,a}(v) &:= R_{x,a}(v(\cdot)) \\ &:= \inf_{u\in\mathbb{R}} \left\{ u + \frac{1}{1-z} \int_S \max\{v(y)-u, 0\} Q(dy|x,a) \right\} \qquad (4.9) \end{aligned}$$

The reasoning in the proof of parts (i, ii, iv) of Proposition 2 in [71] can be easily adjusted to show that $CVaR_{x,a}$ satisfies the conditions in Definition 4.1. Indeed, it directly follows from the definitions of $CVaR_{x,a}$ that parts (a,b,d) of Definition 4.1 are satisfied. For completeness, we verify part (c) of Definition 4.1 as follows: for each $v_1, v_2 \in \mathcal{H}$,

$$
\begin{aligned}
& CVaR_{x,a}(v_1+v_2) \\
= \; & \inf_{u\in\mathbb{R}} \left\{ u + \frac{1}{1-z}\int_S \max\{v_1(y)+v_2(y)-u,0\}Q(dy|x,a) \right\} \\
= \; & \inf_{u_1,u_2\in\mathbb{R}} \left\{ u_1+u_2 + \frac{1}{1-z}\int_S \max\{v_1(y)+v_2(y)-u_1-u_2,0\}Q(dy|x,a) \right\} \\
\leq \; & u_1+u_2+\frac{1}{1-z}\int_S \max\{v_1(y)+v_2(y)-u_1-u_2,0\}Q(dy|x,a) \\
= \; & u_1+u_2+\frac{1}{1-z}\int_S \max\{(v_1(y)-u_1)+(v_2(y)-u_2),0\}Q(dy|x,a) \\
\leq \; & u_1+u_2+\frac{1}{1-z}\left\{ \int_S \max\{v_1(y)-u_1,0\}Q(dy|x,a) \right. \\
& \left. + \int_S \max\{v_2(y)-u_2,0\}Q(dy|x,a) \right\},
\end{aligned}
$$

where $u_1, u_2 \in \mathbb{R}$ are arbitrary. Taking the infimum with respect to $u_1, u_2 \in \mathbb{R}$ on the both sides of the above and using Lemma 3.2 of [56] show that part (c) of Definition 4.1 is satisfied.

Suppose the cost function c is nonnegative and $\mathbb{K}$-inf-compact. Then one can let $\mathbb{H}$ be the space of all extended real-valued nonnegative measurable functions bounded from below by 0, $\delta_m(x)=0$ for each $x\in S$, and $C_m=1$ for each $m=-1,0,1,\dots$, $u_v(x)=0$ for each $x\in S$ and $v\in\mathbb{H}$. Furthermore, we assume that the transition probability $Q(dy|x,a)$ is continuous in the sense that for each $\Gamma_S\in\mathcal{B}(S)$, $Q(\Gamma_S|x,a)$ is continuous in $(x,a)\in\mathbb{K}$.

For this example, all the aforementioned optimality results in this chapter are applicable. Now we give the detailed verifications for the less transparent Assumption 4.4 and Assumption 4.5(b) as follows.

For Assumption 4.4, note that for each $v \in \mathbb{H}$, one can write (4.9) as

$$\begin{aligned} CVaR_{x,a}(v) &= \inf_{u\in[0,\infty]} \left\{ u + \frac{1}{1-z}\int_S \max\{v(y)-u,0\}Q(dy|x,a) \right\} \\ &:= \inf_{u\in[0,\infty]} f(u,x,a), \end{aligned}$$

where we put $f(\infty,x,a) := \infty$. Keeping in mind the convention of $\infty - \infty := \infty$ and using the Fatou's lemma [81], one can see that for each $x \in S,\ a \in A(x)$ and $v \in \mathbb{H}$, $f(u,x,a)$ is lower semicontinuous in u with u from the compactified set $[0,\infty]$. Thus, for each $v_n \in \mathbb{H}$ that increases (pointwise) to $v \in \mathbb{H}$, one can refer to Proposition 10.1 of [85] for that

$$\begin{aligned} & \lim_{n\to\infty} CVaR_{x,a}(v_n) \\ = & \inf_{u\in[0,\infty]} \lim_{n\to\infty} \left\{ u + \frac{1}{1-z}\int_S \max\{v_n(y)-u,0\}Q(dy|x,a) \right\} \\ = & \inf_{u\in[0,\infty)} \lim_{n\to\infty} \left\{ u + \frac{1}{1-z}\int_S \max\{v_n(y)-u,0\}Q(dy|x,a) \right\} \\ = & \inf_{u\in[0,\infty)} \left\{ u + \frac{1}{1-z}\int_S \max\{v(y)-u,0\}Q(dy|x,a) \right\} \\ = & CVaR_{x,a}(v), \end{aligned}$$

where the second to the last equality is by the monotone convergence theorem. Thus, Assumption 4.4 is satisfied.

For Assumption 4.5(b), by using the generalized Fatou's lemma (see [81] and Theorem 4.1 of [38]), which is valid because of the continuity of the transition probability $Q(dy|x,a)$, we see that $f(u,x,a)$ is lower semicontinuous on $[0,\infty]\times\mathbb{K}$ for each lower semicontinuous $v \in \mathbb{H}$, which by Lemma A.3 leads to the lower semicontinuity of $CVaR_{x,a}(v)$ on $\mathbb{K}$ (the conditions of Lemma A.3 are satisfied because of the compactness of the constant multifunction $u \to [0,\infty]$). Thus, Assumption 4.5(b) is also verified.

Remark 4.3 *We point out that (4.2) and (4.9) define the iterated conditional value-at-risk, which was introduced in [45] under the name of iterated conditional tail expectation, where its application to an equity-linked insurance contract with maturity and death benefit guarantees is* *demonstrated, see also Section 3.3 of [70] for further arguments for the use of the iterated conditional value-at-risk as the performance measure. We underline that Example 4.2, which allows the cost to be arbitrarily unbounded from above, cannot be covered by [90], which imposes the condition that the growth of the cost must be bounded by some weight function of Lyapunov type, and thus, in particular, does not cover the* $+\infty$*-valued utility (or say cost) function with wide economic applications as considered in Example 2 of [62] and Example 4 of [63]. In this connection, Example 4.2 illustrates the economic applications of the optimality results of the present chapter that are not covered by [90].*

Chapter 5

Optimality of mixing policies for Constrained MDP with average criterion

5.1 Introduction

This chapter is organized as follows. Section 5.2 is about the description of the MDP model and problem formulation. We present both the maximization and minimization result for the unconstrained model in Section 5.3; We introduce the notion of stable policies and stable measures, show the compactness and closedness of the space of performance vectors, prove the sufficiency of stable policies for the constrained problem; Section 5.5 is about the characterization of extreme points in the space of performance vectors within the class of deterministic stationary policies; Section 5.6 is devoted to the optimality results and the existence of a mixing optimal policy to the constrained problem.

5.2 Problem formulation and assumptions

The constrained MDP is characterized similarly to that in Chapter 2, so we simplify the notations and present only basic components

$$\{S, A, (A(x), x \in S), Q(dy|x, a), c_0(x, a), (c_i(x, a), d_i)_{i=1,\ldots,M}, \gamma(dy)\}.$$

In this chapter, we are interested in another popular criterion, with the objective of minimizing long-run expected average cost. In order to have our problem well defined, the following assumption is introduced at the first place,

Assumption 5.1 *There exist a continuous function $w(\cdot) \geq 1$, a bounded measurable (possible constant) function $b(\cdot) \geq 0$, and nonnegative constants $\hat{c}$ and β, with $\beta < 1$, such that for every $x \in S$:*
(a) $\sup_{a\in A(x)} |c_i(x, a)| \leq \hat{c}w(x), \forall x \in S, i = 0, 1, \ldots, M$;
(b) $\int_S w^2(y)Q(dy|x, a)$ *is continuous in* $a \in A(x), \forall x \in S$;
(c) $\sup_{a\in A(x)} \int_S w^2(y)Q(dy|x, a) \leq \beta w^2(x) + b(x), \forall x \in S$;
(d) $\int_S w^2(x)\gamma(dx) < +\infty$, *for the initial distribution* $\gamma(\cdot)$.

Assumption 5.1 is of conventional Lyapunov type, where w^2 is viewed as the weight function in Assumption 5.1(b,c,d) and w is in use in Assumption 5.1(a). Here and below, $\mathbf{B}_{w^2}(S)$ denotes the Banach Space of all measurable functions on S bounded in the w^2-norm, and $\mathbf{C}_{w^2}(S)$ denotes the subspace of $\mathbf{B}_{w^2}(S)$ consisting of all the continuous functions. Accordingly, denote by $\mathcal{M}_{w^2}(S)$ the subspace of finite measures in $\mathcal{M}(S)$ satisfying

$$\int_S w^2(y)M(dy) < \infty, \ \forall M \in \mathcal{M}(S).$$

All the above notations follows the definition presented in Chapter 1.

Assumption 5.1(a,c,d) ensures the following optimization problem of our interest to be well defined, in the sense that each of long-run expected average costs is finite:

$$V_0(\pi,\gamma) := \overline{\lim_{n\to\infty}} \frac{1}{n} E_\gamma^\pi [\sum_{t=0}^{n-1} c_0(x_t, a_t)] \longrightarrow \min_{\pi \in U^H} \tag{5.1}$$

s.t.

$$V_i(\pi,\gamma) := \overline{\lim_{n\to\infty}} \frac{1}{n} E_\gamma^\pi [\sum_{t=0}^{n-1} c_i(x_t, a_t)] \leq d_i \quad i = 1, 2, \ldots, M$$

We denote by $U^{feasible} := \{\pi \in U^H : V_i(\pi,\gamma) \leq d_i, i = 1, \ldots, M\}$ the set of feasible policies for Problem (5.1).

Remark 5.1 *It is a standing assumption in this chapter that* $U^{feasible} \neq \emptyset$ *in order to avoid the concerned problem being trivial.*

Recall the definition $\mathbb{K} := \{(x,a) \in S \times A : x \in S, a \in A(x)\}$, which is the graph of the set-valued mapping $A(\cdot)$ on S.

Assumption 5.2 *(a)* $c_i(x,a)$ *is continuous in* $(x,a) \in \mathbb{K}$, $\forall\, i = 0, 1, \ldots, M$;
(b) There is a nonnegative moment (or strictly unbounded function) $v(x,a)$ *such that* $|v(x,a)| \leq \hat{v} w(x)$ *for some constant* $\hat{v} > 0$ *in* $x \in S, a \in A(x)$;
(c) $A(x)$ *is compact-valued,* $\forall x \in S$;
(d) $\int_S u(y) Q(dy|x,a)$ *is continuous in* $a \in A(x)$ $, \forall\, x \in S, u \in \mathbf{B}(S)$;
(e) $\int_S u(y) Q(dy|x,a)$ *is continuous in* $(x,a) \in \mathbb{K}$, $\forall\, u \in \mathbf{C}(S)$, *where* $\mathbf{C}(S)$ *denotes the space of continuous bounded real-valued functions on* S.

Assumption 5.2(a,c,d,e) constitutes a version of continuity-compactness conditions, in various forms, commonly imposed in the study of MDPs. Conditions of this type ensure the measurability of optimal value functions as well as the existence of at least one optimal deterministic stationary policy (of course, a measurable selector). The latter objective is achieved by means of measurable selection theorem or Berge's minimum theorem; see Lemma A.2, or [59, Prop.3.3,p.83] and [11, Thm.2,p.116]. Indeed, the present version is a mixture of both weak (**W**) and strong (**S**) continuity-compactness conditions based on the definitions from Schäl [85]. Although only minimization problem is under consideration in the present chapter, we assume cost functions to be jointly continuous in both arguments. The reason for this setup will be revealed latter.

For notational convenience, we introduce t-step transition kernel $Q_f^t(\Gamma_S|x)$ for each $\Gamma_S \in \mathcal{B}(S)$ associated with a deterministic policy $f \in U^{DS}$ for future reference.

$$Q_f^t(\Gamma_S|x) := Q^t(\Gamma_S|x, f) = P_x^f(x_t \in \Gamma_S) \tag{5.2}$$

For $t = 0$, (5.2) is reduced to

$$Q_f^0(\Gamma_S|x) = \delta_x(\Gamma_S) = \mathbf{1}_{\{x\in\Gamma_S\}} \tag{5.3}$$

Observe that Assumption 5.1(c) implies

$$\int_S w^2(y)Q_f(dy|x) \leq \beta w^2(x) + b(x) \quad \forall\ f \in U^{DS},\ x \in S. \tag{5.4}$$

Thus, multiplying by $w^2(x)^{-1}$ one can see that the family of stochastic kernels $\{Q_f,\ f \in \mathbb{F}\}$ possess a uniformly bounded w^2-norm. Explicitly, the w^2-norm of Q_f, i.e.,

$$\|Q_f\|_{w^2} := \sup_{x\in S} w^2(x)^{-1} \int_S w^2(y)Q_f(dy|x) \tag{5.5}$$

satisfies

$$\|Q_f\|_{w^2} \leq \beta + \|b\|_{w^2}, \tag{5.6}$$

alternatively,

$$\|Q_f\|_{w^2} \leq \beta + \|b\|, \tag{5.7}$$

where $\|b\| := \sup_{x\in S} |b(x)|$ is the sup-norm of $b(\cdot)$ as in Assumption 5.1 $b(\cdot)$ is assumed to be a bounded function. Likewise, according to the material presented in Chapter 1.3.1, all the aforementioned properties remain satisfied when $Q_f(\Gamma_S|x)$ is replaced by $Q_\varphi(\Gamma_S|x)$ with the following definition

$$Q^t_\varphi(\Gamma_S|x) := Q^t(\Gamma_S|x,\varphi) = \int_A Q^t(\Gamma_S|x,a)\varphi(da|x), \tag{5.8}$$

where $\varphi \in U^S$ is an arbitrary randomized stationary policy.

Definition 5.1 *A probability measure $\mu(dx)$ on S is called an invariant probability measure (i.p.m.) for a Markov chain $Q(dy|x)$, if for each $\Gamma_S \in \mathcal{B}(S)$*

$$\mu(\Gamma_S) = \int_S Q(\Gamma_S|x)\mu(dx).$$

We introduce our next general assumption immediately followed by its consequence.

Assumption 5.3 *For any deterministic stationary policy $f \in U^{DS}$, the Markov Chain $Q_f(dy|x)$ is w^2-geometrically ergodic, i.e., there exists a unique i.p.m. μ_f on S such that $\|Q^t_f - \mu_f\|_{w^2} \leq R\rho^t$, $\forall t = 0, 1, \ldots$, where $R \geq 0$ and $0 < \rho < 1$ are constants independent of f.*

In particular, Assumption 5.3 implies that $Q_f(dy|x)$ is positive Harris recurrent, and μ_f is the corresponding unique *i.p.m.* which belongs to the space $\mathcal{M}_{w^2}(S)$. To see the latter fact, we have for each $t > 0$,

$$\begin{aligned}\int_S w^2(y)\mu_f(dy) &\leq \int_S w^2(y)|\mu_f(dy) - Q^t(dy|x)| + \int_S w^2(y)Q^t(dy|x) \\ &\leq R\rho^t w^2(x) + \beta^t w^2(x) + \frac{1-\beta^t}{1-\beta}\|b\|.\end{aligned}$$

Letting $t \to \infty$ leads to

$$\|\mu_f\|_{w^2} := \int_S w^2(y)\mu_f(dy) \leq \frac{\|b\|}{1-\beta} < \infty. \tag{5.9}$$

Note that the above inequality holds uniformly for all $f \in U^{DS}$. Here, we emphasize that in general (5.9) need not hold for all stable policies.

5.3 Optimality of unconstrained models

This section concerns the average optimality problems for unconstrained model associated with any $c_i(x,a)$ when $i = 0, 1, \ldots, M$. We show the existence of optimal deterministic stationary policies for both maximization and minimization problems by referring to [44, Thm.4.1]. This is essential for the derivation of further results in the remaining sections. We remind the readers that the subscript i of $c(x,a)$ is omitted in this section, as all the cost functions $c_i(x,a) \in \mathbf{C}_w(\mathbb{K})$, $i = 0, 1, \ldots, M$, can be equally treated in view of Assumption 5.2.

Formally, for each $x \in S$ the unconstrained problem under our consideration in this section is defined as

$$\underline{V}(\pi,x) := \overline{\lim_{n\to\infty}} \frac{1}{n} E_x^{\pi}\left[\sum_{t=0}^{n-1} c(x_t,a_t)\right] \to \min_{\pi\in U^H} \tag{5.10}$$

$$\overline{V}(\pi,x) = \underline{\lim}_{n\to\infty} \frac{1}{n} E_x^{\pi}\left[\sum_{t=0}^{n-1} c(x_t,a_t)\right] \to \max_{\pi\in U^H} \tag{5.11}$$

Lemma 5.1 *For the unconstrained problem associated with any cost function $c(x,a)$, under Assumption 5.1, 5.2 and 5.3, there exist deterministic stationary policies f^*, $g^* \in U^{DS}$ such that $\underline{V}(f^*,x) = \inf_\pi \underline{V}(\pi,x) =: \rho_*$, and $\overline{V}(g^*,x) = \sup_\pi \overline{V}(\pi,x) =: \rho^*$ for each $x \in S$ respectively. Take the minimization problem as an example, there exists a triplet (ρ_*, h_0, f^*) which satisfies the average-cost optimal inequality (ACOI), namely,*

$$\begin{aligned} \rho_* + h_0(x) &\geq \min_{a\in A(x)}\left\{c(x,a) + \int_S h_0(y)Q(dy|x,a)\right\} \qquad (5.12)\\ &= c(x,f^*) + \int_S h_0(y)Q(dy|x,f^*) \quad \forall\ x \in S. \end{aligned}$$

Moreover, $\rho_* = \inf_\pi \underline{V}(\pi,x)$ for all $x \in S$, and indeed any measurable selector $f \in \mathbb{F}$ realizing the minimum of (5.12) defines an optimal deterministic stationary policy for Problem (5.10).

Proof. We implement the *vanishing discount factor approach*, which is commonly employed in dealing with average optimality problems. Firstly, we show that for any discount factor $0 < \alpha < 1$ and initial state $x \in S$, the value function for the discounted problem

$$E_x^\pi \left[\sum_{t=0}^\infty \alpha^t c(x_t, a_t) \right] \to \min_{\pi \in U^H} \tag{5.13}$$

is indeed uniformly bounded over all policies.

For every policy $\pi \in U^H$ and initial state $x \in S$, we have

$$\begin{aligned} J_\alpha(\pi, x) \quad &:= E_x^\pi \left[\sum_{t=0}^\infty \alpha^t c(x_t, a_t) \right] \\ &\leq E_x^\pi \left[\sum_{t=0}^\infty \alpha^t \hat{c} w^2(x_t) \right] \\ &< \hat{c} \sum_{t=0}^\infty \alpha^t \left[1 + \frac{\|b\|}{1-\beta} \right] w^2(x) \\ &= \frac{\hat{b}}{1-\alpha} w^2(x) < +\infty. \end{aligned} \tag{5.14}$$

where $\hat{b} := \hat{c} \left[1 + \frac{\|b\|}{1-\beta} \right]$.

The second to the last line holds by the following inequality which is obtained by successive iteration of Assumption 5.1(c) and will be frequently referred to in the sequel,

$$E_x^\pi[w^2(x_t)] \leq \beta^t w^2(x) + \frac{1-\beta^t}{1-\beta} \|b\| < \left[1 + \frac{\|b\|}{1-\beta} \right] w^2(x) \tag{5.15}$$

Let $z \in S$ be arbitrarily fixed. For every $0 < \alpha < 1$ we consider the *relative difference*

$$h_\alpha(x) := J_\alpha^*(x) - J_\alpha^*(z)$$

where $J_\alpha^*(x) := \inf_{\pi \in U^H} J_\alpha(\pi, x)$ denotes the corresponding optimal value function for each $x \in S$.

It is well known that the class in search for an optimal policy for Problem (5.13) associated with Lyapunov-like conditions, e.g., Assumption 5.1, can be reduced to the subclass of deterministic stationary policies (see [50, Thm.8.3.6]). Therefore, we need only consider U^{DS} in the following discussion.

For each deterministic stationary policy $f \in U^{DS}$ and $x \in S$,

$$\begin{aligned}
& \left|E_x^f[c(x_t, f)] - E_z^f[c(x_t, f)]\right| \\
\leq\ & \left|E_x^f[c(x_t, f)] - \int_S c(y, f)\mu_f(dy)\right| + \left|E_z^f[c(x_t, f)] - \int_S c(y, f)\mu_f(dy)\right| \\
\leq\ & \hat{c}\left(\left|E_x^f[w(x_t)] - \int_S w(y)\mu_f(dy)\right| + \left|E_z^f[w(x_t)] - \int_S w(y)\mu_f(dy)\right|\right) \\
\leq\ & \hat{c}\left(\int_S w^2(y)\left|Q_f^t(dy|x) - \mu_f(dy)\right| + \int_S w^2(y)\left|Q_f^t(dy|z) - \mu_f(dy)\right|\right) \\
\leq\ & \hat{c}R\rho^t[1 + w^2(z)]w^2(x), \qquad (5.16)
\end{aligned}$$

thus,

$$\begin{aligned}
|J_\alpha(f^\infty, x) - J_\alpha(f^\infty, z)| &\leq \sum_{t=0}^{\infty}\left|E_x^f[c(x_t, a_t)] - E_z^f[c(x_t, a_t)]\right| \\
&\leq \hat{c}R(1-\rho)^{-1}[1 + w^2(z)]w^2(x).
\end{aligned}$$

Furthermore,

$$|h_\alpha(x)| \leq \sup_{U^{DS}} |J_\alpha(f^\infty, x) - J_\alpha(f^\infty, z)|.$$

As a result, for each $0 < \alpha < 1$, $|h_\alpha(x)| \leq h_0(x)$, where

$$\hat{c}R(1-\rho)^{-1}[1 + w^2(z)]w^2(x) =: h_0(x) \in \mathbf{B}_{w^2}(S).$$

Note that $h_0(x)$ is independent of the discount factor α.

In view of [44, Thm.4.1(b)], all the conditions are now verified, which validates the stated result for minimization problem as required.

As far as the corresponding maximization problem is concerned, we observe that it can be easily treated as a minimization problem with respect to $-c(x, a)$. The same reasoning is applicable and similar results should follow. In a nutshell, there exists at least one deterministic optimal policy for either the maximization or minimization average problem in the unconstrained case for any cost function $c(\cdot, \cdot)$. □

Remark 5.2 *Note that each of cost functions $c_i(x, a)$ belongs to $\mathbf{C}_w(\mathbb{K})$,* *which is in turn an element of* $\mathbf{C}_{w^2}(\mathbb{K})$ *by the fact that* $\mathbf{C}_w(\mathbb{K}) \subseteq \mathbf{C}_{w^2}(\mathbb{K})$ *due to* $w(\cdot) \geq 1$*. Indeed, the weight function in the context of this chapter is* w^2 *compared with that in [44] or other literature.*

5.4 Stable policies, stable measures and the space of performance vectors

We draw our attention back to the the constrained Problem (5.1). The main objective of this section is to show that, with the help of Assumption 5.2, the search for optimal policies for the concerned problem can be reduced to the so-called stable policies. We further introduce the space of stable measures and the corresponding space of performance vectors. In addition, we justify the compactness and convexity of the space of performance vectors.

Accordingly, we define the notion of stable measures for future reference,

Definition 5.2 *A probability measure* $\mu(dx \times da)$ *on* $\mathbb{K}$ *is called stable if*
(a)

$$\int_{\mathbb{K}} w(x)\mu(dx \times da) < \infty;$$

(b) for each measurable set $\Gamma_S \in \mathcal{B}(S)$*,*

$$\mu(\Gamma_S \times A) = \int_{\mathbb{K}} Q(\Gamma_S|x, a)\mu(dx \times da).$$

Given a probability measure $\mu(dx \times da)$ concentrated on $\mathbb{K}$, one can represent it in terms of its projection (or, marginal) $\mu(dx \times A)$ as the following form

$$\mu(dx \times da) = \mu(dx \times A)\varphi^{\mu}(da|x),$$

where φ^{μ} is a stochastic kernel being unique (in the sense of $\mu(dx \times A)$ almost everywhere) which induces a randomized stationary policy. With this fact in mind, Definition 5.2(b) can be viewed as

$$\mu(\Gamma_S\times A) = \int_{\mathbb{K}} Q(\Gamma_S|x,a)\mu(dx\times da)\varphi^\mu(da|x) = \int_S Q(\Gamma_S|x,\varphi^\mu)\mu(dx\times A),$$

meaning $\mu(dx\times A)$ is an *i.p.m.* for the Markov chain $Q_{\varphi^\mu}(dy|x)$ in consistency with Definition 5.1. We called such a randomized stationary policy a stable policy. The space of stable measures is denoted by $\mathcal{D}\subseteq\mathcal{P}_w(S)$. In view of Assumption 5.3, each deterministic stationary policy $f\in U^{DS}$ is a stable policy with μ_f being its unique *i.p.m.*.

Assumption 5.4 *For any stable measure $\mu(dx\times da)$ and the predetermined initial distribution $\gamma(dx)$, we have the following relation*

$$\lim_{n\to\infty}\frac{1}{n}E_\gamma^{\varphi^\mu}\left[\sum_{t=0}^{n-1}c_i(x_t,a_t)\right] = \int_S c_i(x,\varphi^\mu)\mu(dx\times A) = \int_{\mathbb{K}} c_i(x,a)\mu(dx\times da)$$

for every Markov chain $Q_{\varphi^\mu}(dy|x)$ with respect to which $\mu(dx\times A)$ is its i.p.m., where $\mu(dx\times da)=\varphi^\mu(da|x)\mu(dx\times A)$, and $i=0,1,\ldots,M$.

Assumption 5.4 is similar to the traditional unichained assumption, which asserts that the controlled process is positive Harris recurrent under each stationary policy. In the present setting, Assumption 5.4 is satisfied if the unichained assumption holds for the class of stable policies. In particular, in the finite-state-finite-action case, Assumption 5.4 is automatically justified. In the denumerable-state-compact-action case, the conditions under which the above representation holds under the unichained assumption are provided in [3] which reveals the intrinsic necessity of uniform integrability of expected occupation measures (see the definition below in (5.18)) or one-sided boundedness of cost functions. More discussions will be given in the next chapter in the denumerable case, which essentially extends the results obtained in [3].

Lemma 5.2 *Under Assumptions 5.1, 5.2 and 5.4, for each $\pi\in U^{feasible}$ for Problem (5.1), there exists a stable policy $\varphi\in U^{stable}$ that performs as well as π.*

Proof. Note that from Remark 5.1, we can arbitrarily select and fix some policy $\pi \in U^{feasible}$ with $V_0(\pi,\gamma) < \infty$. Indeed, $V_0(\pi,\gamma) < \infty$ for every $\pi \in U^H$. From [50, Lem.10.4.1] and Assumption 5.1(c), we have

$$\begin{aligned} & E_\gamma^\pi[c_0(x_t,a_t)] \le E_\gamma^\pi[\hat{c}w^2(x_t)] \\ \le \ & \hat{c}\int_S E_y^\pi[w^2(x_t)]\gamma(dy) \le \hat{b}\int_S w^2(y)\gamma(dy) < \infty, \end{aligned}$$

where $\hat{b}$ comes from (5.14). Then,

$$V_0(\pi,\gamma) = \overline{\lim_{n\to\infty}} \frac{1}{n} E_\gamma^\pi\left[\sum_{t=0}^{n-1} c_0(x_t,a_t)\right] \le \hat{b}\int_S w^2(y)\gamma(dy) < \infty \tag{5.17}$$

For notational ease, we denote by $\mu_{\gamma,n}^\pi$ the n-stage occupation measure associated with (π,γ), that is,

$$\mu_{\gamma,n}^\pi(\Gamma) := \frac{1}{n}\sum_{t=0}^{n-1} P_\gamma^\pi\left((x_t,a_t)\in\Gamma\right) \quad \forall\ \Gamma \in \mathcal{B}(S\times A). \tag{5.18}$$

Accordingly, Problem (5.1) is rewritten in the following form

$$\begin{aligned} & V_0(\pi,\gamma) := \overline{\lim_{n\to\infty}} \int_{\mathbb{K}} c_0(x,a)\mu_{\gamma,n}^\pi(dx\times da) \to \min_\pi \\ s.t. & \\ & V_i(\pi,\gamma) := \overline{\lim_{n\to\infty}} \int_{\mathbb{K}} c_i(x,a)\mu_{\gamma,n}^\pi(dx\times da) \le d_i \quad i=1,2,\ldots,M \end{aligned}$$

By replacing $c_0(x_t,a_t)$ with $v(x_t,a_t)w(x_t)$ in (5.17), for any given $\epsilon > 0$, there exists a positive integer $N(\epsilon)$ such that,

$$\sup_{n\ge N(\epsilon)} \int v(x,a)w(x)\mu_{\gamma,n}^\pi(dx\times da) \tag{5.19}$$

$$\le \ \frac{\hat{v}\|b\|}{1-\beta}\int_S w^2(y)\gamma(dy) + \epsilon < \infty. \tag{5.20}$$

For each $\mu_{\gamma,n}^\pi$, a new measure is defined by

$$\tilde{\mu}_{\gamma,n}^\pi(dx\times da) := \mu_{\gamma,n}^\pi(dx\times da)w(x) \tag{5.21}$$

Note that $\mu^{\pi}_{\gamma,n} \in \mathcal{P}_{w^2}(S \times A) \subseteq \mathcal{P}_w(S \times A)$ because

$$\begin{aligned}\int_{S\times A} w^2(x)\mu^{\pi}_{\gamma,n}(dx \times da) &= \frac{1}{n}E^{\pi}_{\gamma}\left[\sum_{t=0}^{n-1} w^2(x_t)\right] \\ &\leq \left(1+\frac{\|b\|}{1-\beta}\right)\int_S w^2(x)\gamma(dx)\end{aligned}$$

from which it is observed that $\tilde{\mu}^{\pi}_{\gamma,n} \in \mathcal{M}_w(S \times A) \subseteq \mathcal{M}(S \times A)$. By [47, Prop.E.8] and Prohorov's Theorem (see Theorem B.1), there exists a subsequence $(\tilde{\mu}^{\pi}_{\gamma,n_k})$ such that $\tilde{\mu}^{\pi}_{\gamma,n_k} \to \tilde{\mu}^*$ for some $\tilde{\mu}^* \in \mathcal{M}(S\times A)$. It follows from (5.21) and Remark 1.2 that there is a corresponding subsequence (μ^{π}_{γ,n_k}) such that $\mu^{\pi}_{\gamma,n_k} \overset{w}{\to} \mu^*$. Since c_0 is w-bounded and continuous on $\mathbb{K}$, By Corollary A.2(b) we have

$$\begin{aligned}\int c_0(x,a)\mu^*(dx \times da) &= \lim_{k\to\infty}\int c_0(x,a)\mu^{\pi}_{\gamma,n_k}(dx \times da) \\ &\leq \overline{\lim_{n\to\infty}}\int c_0(x,a)\mu^{\pi}_{\gamma,n}(dx \times da) \\ &= V_0(\pi,\gamma) < \infty. \qquad (5.22)\end{aligned}$$

Next we show that μ^* is indeed a stable measure. Obviously, Definition 5.2(a) is satisfied, so we focus on part (b) of it. As usual, $\mu^*(dx \times da) = \mu^*(dx \times A)\varphi^*(da|x)$. For any $u \in \mathbf{B}_w(S)$, define the function $Tu \in \mathbf{B}_w(\mathbb{K})$ as

$$Tu(x,a) := \int_S u(y)Q(dy|x,a) - u(x).$$

Similarly to (2.15), a version of Dynkin's formula in the discrete-time case with respect to the prefixed $\pi = (\pi_t)_{t=0,1,2,\dots}$ takes the following form

$$\begin{aligned}E^{\pi}_{\gamma}\left[u(x_n)\right] &= \int_S u(x)\gamma(dx) \\ &\quad +E^{\pi}_{\gamma}\left[\sum_{t=1}^{n}\left\{\int_{\mathbb{K}} u(y)Q(dy|x_{t-1},a)\pi_{t-1}(da|h_{t-1}) - u(x_{t-1})\right\}\right] \\ &= \int_S u(x)\gamma(dx) + E^{\pi}_{\gamma}\left[\sum_{t=0}^{n-1} Tu(x_t,a_t)\right] \qquad (5.23)\end{aligned}$$

Multiplying $\frac{1}{n}$ on both sides of (5.23), rearranging the terms yields

$$\begin{aligned}
& \frac{1}{n}\left\{E_\gamma^\pi\left[u(x_n)\right] - \int_S u(x)\gamma(dx)\right\} && (5.24)\\
= \ & \frac{1}{n}E_\gamma^\pi\left[\sum_{t=0}^{n-1} Tu(x_t, a_t)\right] \\
= \ & \int_{\mathbb{K}} Tu(x,a)\mu_{\gamma,n}^\pi(dx \times da) && (5.25)
\end{aligned}$$

Replacing n with n_k when we derive μ^*, and letting $k \to \infty$ in (5.24) leads to

$$\int_{\mathbb{K}} Tu(x,a)\mu^*(dx \times da) = 0,$$

by Assumption 5.1(d) and (5.15), which justifies part (b) of Definition 5.2.

According to Assumption 5.4, we have

$$\begin{aligned}
\lim_{n\to\infty}\int_{\mathbb{K}} c_0(x,a)\mu_{\gamma,n}^{\varphi^*}(dx \times da) &= \int_{\mathbb{K}} c_0(x,a)\mu^*(dx \times da)\\
&= \int_S c_0(x,\varphi^*)\mu^*(dx \times A)
\end{aligned}$$

which together with (5.22) shows that the performance of the stable policy φ^* is at least as good as that of the prefixed policy π.

Meanwhile, all the constraints remain to be satisfied with the newly obtained stable policy φ^* by similar reasoning in the treatment of c_0. Explicitly,

$$\begin{aligned}
\lim_{n\to\infty}\int_{\mathbb{K}} c_i(x,a)\mu_{\gamma,n}^{\varphi^*}(dx \times da) &= \int_{\mathbb{K}} c_i(x,a)\mu^*(dx \times da)\\
&= \lim_{k\to\infty}\int_{\mathbb{K}} c_i(x,a)\mu_{\gamma,n_k}^{\varphi^*}(dx \times da)\\
&\leq \overline{\lim_{n\to\infty}}\int_{\mathbb{K}} c_i(x,a)\mu_{\gamma,n}^{\pi}(dx \times da)\\
&= V_i(\pi,\gamma) \leq d_i
\end{aligned}$$

$\forall\ i = 1, 2, \ldots, M$, as required. □

The proof appears similar to that of [53, Lem.3.5], in which the convergence is in the usual weak topology, whereas we are dealing with

w-weak topology here. The main consequence of Lemma 5.2 is that the space of stable policies is sufficient to solve Problem (5.1). Here and below, we shall consider only the space of stable policies, denoted by U^{stable}, whose definition is in consistency with the paragraph following Definition 5.2.

Another consequence of Lemma 5.2 is that Problem (5.1) can be rewritten in the form of a well-defined linear program as follows,

$$\begin{aligned} & \int_{\mathbb{K}} c_0(x,a)\mu(dx \times da) \to \min_{\mu} \\ s.t. & \\ & \int_{\mathbb{K}} c_n(x,a)\mu(dx \times da) \leq d_n, n = 1, 2, \ldots, M, \\ & \mu \in \mathcal{D}. \end{aligned} \tag{5.26}$$

From now on, we shall focus on Problem (5.26) instead of Problem (5.1) under Assumption 5.1, 5.2 and 5.4.

The space of performance vectors is introduced as follows.

Definition 5.3 *The space of performance vectors corresponding to a specific class of policies U is a subset of $\mathbb{R}^{M+1}$, defined as*

$$V(U,\gamma) := \left\{ (V_0(\pi,\gamma), V_1(\pi,\gamma), \ldots, V_M(\pi,\gamma)) \,;\, \pi \in U \right\}.$$

Define the linear mapping $\underline{Z} : \mathcal{D} \to \mathbb{R}^{M+1}$ by

$$\begin{aligned} \underline{Z}(\mu) \;\; := \;\; & \Big(\int c_0(x,a)\mu(dx \times da), \int c_1(x,a)\mu(dx \times da), \\ & \ldots, \int c_M(x,a)\mu(dx \times da) \Big), \end{aligned} \tag{5.27}$$

where $\mu \in \mathcal{D}$.

It follows from Assumption 5.4 that the space of performance vectors on the class of stable policies can be viewed as the complete image of $\underline{Z}$

on the space of stable measures $\mathcal{D}$, which takes the following form,

$$\begin{aligned} V(U^{stable}, \gamma) &= \left\{ \Big(\int_{\mathbb{K}} c_0(x,a)\mu(dx \times da), \ldots, \right. \\ &\qquad \left. \int_{\mathbb{K}} c_M(x,a)\mu(dx \times da) \Big) \, ; \mu \in \mathcal{D} \right\}. \end{aligned}$$

$V(U^{stable}, \gamma)$ is the main object and also the key tool that we are going to investigate in the remainder of this chapter. Some of its topological properties are revealed and proved in the following lemma.

Lemma 5.3 *Under Assumption 5.1, 5.2 and 5.4, the space of performance vectors $V(U^{stable}, \gamma)$ is compact and convex.*

Proof. We firstly show that the space of stable measures $\mathcal{D}$ is compact in $\mathcal{P}_w(\mathbb{K})$. We investigate the following functional,

$$\begin{aligned} \int_{\mathbb{K}} v(x,a)w(x)\mu^{\varphi}(dx \times da) &= \int_{S} v(x,\varphi)w(x)\mu^{\varphi}(dx \times A) \\ &\leq \hat{v} \int w^2(x)\mu^{\varphi}(dx \times A) < \infty \end{aligned}$$

which is justified by (5.19) as

$$\overline{\lim_{n\to\infty}} \int_{\mathbb{K}} v(x,a)w(x)\mu^{\pi}_{\gamma,n}(dx \times da) < \infty$$

for every $\pi \in U^H$.

The set $\tilde{\mathcal{D}}$, as the image of $\mathcal{D}$ via (1.5), is tight by Assumption 5.2(b) and [47, Prop.E.8] again, and is thus precompact by Prohorov's Theorem. The remainder is to show the closedness of $\mathcal{D}$ in $\mathcal{P}_w(\mathbb{K})$. Let $(\mu_n) \in \mathcal{D}$ be a sequence of stable measures that converges to $\mu \in \mathcal{P}_w(\mathbb{K})$ in w-weak topology. First, It is trivial that the corresponding family of projections

$\mu_n(dx \times A) \overset{w}{\to} \mu(dx \times A)$. Let $g \in \mathbf{C}_w(S)$, we have

$$\begin{aligned}
\int_S g(y)\mu(dy \times A) &= \lim_{n\to\infty} \int_S g(y)\mu_n(dy \times A) \\
&= \lim_{n\to\infty} \int_S g(y) \int_{\mathbb{K}} Q(dy|x,a)\mu_n(dx \times da) \\
&= \lim_{n\to\infty} \int_{\mathbb{K}} \left(\int_S g(y)Q(dy|x,a) \right) \mu_n(dx \times da) \\
&= \int_{\mathbb{K}} \left(\int_S g(y)Q(dy|x,a) \right) \mu(dx \times da)
\end{aligned}$$

The second to the last equality follows from Assumption 5.1(b), 5.2(e) and Lemma A.1, which asserts that $\int_S g(y)Q(dy|x,a) \in \mathbf{C}_w(\mathbb{K})$. By [12, Prop.7.18] and the fact that w-weak topology is at least as strong as the usual weak topology. Therefore, $\mathcal{D}$ is w-weakly compact in $\mathcal{P}_w(\mathbb{K})$.

It is obvious that the linear mapping $\underline{Z}$ coming from (5.27) is (w-weakly) continuous on $\mathcal{D}$ (see Assumption 5.1 and 5.2(a)), which in turn demonstrates the compactness of $V(U_{stable}, \gamma)$.

The convexity is clear from the definition of stable measures and the linearity of $\underline{Z}$. □

5.5 Extreme points of the space of performance vectors

We have already shown the compactness and convexity of the space of performance vectors $V(U^{stable}, \gamma)$, which further ensures the existence of its extreme points. The main objective of this section is their characterization. To this end, we introduce one last assumption, and derive the corresponding result which is important in its own right, and meanwhile critical to the existence of optimal mixing policies shown in the next section.

Definition 5.4 *We call a Markov chain $Q(dy|x)$ λ-irreducible if there exists a measure λ on $\mathcal{B}(S)$ such that, for $\Gamma_S \in \mathcal{B}(S)$, whenever $\lambda(\Gamma_S) > 0$, we have*

$$P(\tau_{\Gamma_S} < \infty | x_0 = x) = 1 \quad \forall\, x \in \Gamma_S,$$

where $\tau_{\Gamma_S} := \min\{n \geq 1 : x_n \in \Gamma_S\}$.

Assumption 5.5 *There exists a σ-finite measure λ on $\mathcal{B}(S)$ with respect to which Q_f is λ-irreducible for each deterministic stationary policy $f \in U^{DS}$. In addition, λ is non-trivial in the sense of $\lambda(S) > 0$.*

Assumption 5.5 is called the uniform λ-irreducibility condition, which is commonly imposed to establish the average cost optimality equation (ACOE). We quote the above result coming from [50, Thm.10.3.6] in the following lemma. Note that subscript i of $c_i(x, a)$ is again omitted in line with what is presented in Lemma 5.1.

Lemma 5.4 *Under Assumption 5.1, 5.2, 5.3 and 5.5, consider any cost function $c(x, a)$ and the minimization problem (5.10) as in Lemma 5.1, there exists a constant ρ_* and a measurable function $h_* \in \mathbf{B}_{w^2}(S)$ such that the following average cost optimality equation hold.*

$$\rho_* + h_*(x) = \min_{a \in A(x)} \left\{ c(x, a) + \int_S h_*(y) Q(dy|x, a) \right\} \quad \forall\ x \in S, \quad (5.28)$$

where $\rho_* = \inf_{\pi \in U^H} \underline{V}(\pi, x)$ is the optimal value of Problem (5.10). Moreover, there exists a measurable selector $f_* \in \mathbb{F}$ realizing the minimum of (5.28), which in turn induces an optimal deterministic stationary policy for Problem (5.10).

As pointed out in [50], $f_* \in U^{DS}$ from Lemma 5.4 is called a canonical policy, which is certain to be optimal. However, the converse need not hold; that is, an optimal policy does not necessarily corresponds to a measurable function such that (5.28) holds for each $x \in S$.

Theorem 5.1 *Under Assumption 5.1, 5.2, 5.3, 5.4 and 5.5, for each extreme point of $V(U^{stable}, \gamma)$ denoted by $\vec{u}^{ex} = (u_0^{ex}, u_1^{ex}, \ldots, u_M^{ex})$, there exists a corresponding deterministic stationary (stable) policy $f^{ex} \in U^{DS}$ generating $\vec{u}^{ex}$, i.e.,*

$$\vec{u}^{ex} = V(f^{ex}, \gamma) = (V_0(f^{ex}, \gamma), V_1(f^{ex}, \gamma), \ldots, V_M(f^{ex}, \gamma))$$

Proof. By Lemma 5.3, the existence of extreme points of $V(U^{stable},\gamma)$ is automatically justified. Suppose $\vec{u}^{ex}=(u_0^{ex},u_1^{ex},\dots,u_M^{ex})$ is an extreme point of of $V(U^{stable},\gamma)$. Note that $\vec{u}^{ex}$ belongs to the boundary of $V(U^{stable},\gamma)$; otherwise, there is an open ball centered at $\vec{u}^{ex}$ which is a subset of $V(U^{stable},\gamma)$, such that $\vec{u}^{ex}$ can be represented by the convex combination of arbitrary two of its interior points.

Below we prove the statement by inductive argument. To start with, an auxiliary model

$$\mathcal{M}_0=\{S,A,A_0(x),Q,(c_i)_{i=0,1,\dots,M},\gamma\}$$

is formulated, where $A_0(x)\equiv A(x)$ for each $x\in S$. Note that $\mathcal{M}_0$ is identical to $\mathcal{M}$ without constraints. According to [13, Prop.2.4.1], there exists a supporting hyperplane $H_1:=\sum_{m=0}^{M}r_m^1u_m=r^1$ of $V(U^{stable},\gamma)$ at $\vec{u}^{ex}$, such that either $\sum_{m=0}^{M}r_m^1u_m\le r^1$, or $\sum_{m=0}^{M}r_m^1u_m\ge r^1$ for each $\vec{u}=(u_0,u_1,\dots,u_M)\in V(U^{stable},\gamma)$. Here, $r^1,r_0^1,\dots,r_M^1$ are some real constants that at least one of them is nonzero. By linearity of integration, we formulate an equivalent maximization or minimization problem associated with the cost function $c_1^{\vec{r}}(x,a):=\sum_{m=0}^{M}r_m^1c_m(x,a)$. What we aim is to construct a sub-model so that the space of performance vectors of it coincides with the exposed set $V(U^{stable},\gamma)\cap H_1$; see more discussion about concepts in convex analysis in [93].

Without loss of generality, we consider the minimization problem, i.e.,

$$\int_{\mathbb{K}}c_1^{\vec{r}}(x,a)\mu(dx\times da)\to\min_{\mu\in\mathcal{D}}$$

The sub-model is constructed by refining admissible action spaces on the original state space. Under Assumption 5.5 and by Lemma 5.4, the ACOE is established as follows,

$$r^1+h_1^{\vec{r}}(x)=\min_{a\in A_0(x)}\left\{c_1^{\vec{r}}(x,a)+\int_S h_1^{\vec{r}}(y)Q(dy|x,a)\right\}\quad\forall\ x\in S\quad(5.29)$$

for which we denote by

$$R(x,a) := c_1^{\vec{r}}(x,a) + \int_S h_1^{\vec{r}}(y)Q(dy|x,a)$$

for future reference. For each $x \in S$, define

$$L(x) := \left\{ a \in A_0(x) : r^1 + h_1^{\vec{r}}(x) = c_1^{\vec{r}}(x,a) + \int_S h_1^{\vec{r}}(y)Q(dy|x,a) \right\}$$

The new admissible action space is defined by $A_1(x) := L(x)$ for each $x \in S$.

Finally, the promised sub-model

$$\mathcal{M}_1 := \{S, A, A_1(x), Q, (c_i)_{i=0,1,\ldots,M}, \gamma\}$$

is obtained, with respect to which we make the following six observations.

Observation 1: $A_1(x) \subseteq A_0(x)$ is nonempty and compact for each $x \in S$.

Indeed, $A_1(x)$ is closed because $R(x,\cdot)$ is a continuous function on $A_0(x)$ by Assumption 5.2(a,d). $A_1(\cdot)$ is non-empty due to Lemma A.2(a), and the compactness of $A_1(x)$ follows from the fact that $A_0(x)$ is compact.

Observation 2: The graph

$$\mathbb{K}^1 := \{(x,a); x \in S, a \in A_1(x)\} \subseteq \mathbb{K}$$

is a product measurable subset of $S \times A$ and contains the graph of a measurable selector from S to A.

Suppose $G \subseteq A$ is an arbitrary closed set, we consider the complete preimage

$$A_1^{-1}[G] := \{x \in S; A_1(x) \bigcap G \neq \emptyset\}.$$

Observe that

$$A_1^{-1}[G] = \left\{ x \in S; \inf_{a \in A(x) \bigcap G} R(x,a) = r^1 + h_1^{\vec{r}}(x) \right\} \tag{5.30}$$

The set-valued mapping $A_0(x)$ is measurable by the definition and Proposition A.2. Note that $A_0(x) \bigcap G =: \tilde{A}(x)$ is a measurable set-valued mapping as well (cf. [54, Prop.2.4, Thm.4.1]). Further by [54, Prop.2.2], the domain of $\tilde{A}(x)$

$$Dom(\tilde{A}) := \{x \in S; \tilde{A}(x) \neq \emptyset\}$$

is measurable. Therefore,

$$E := \left\{ x \in Dom(\tilde{A}); \inf_{a \in \tilde{A}(x)} R(x,a) = r^1 + h_1^{\vec{r}}(x) \right\}$$

is measurable since $\inf_{a \in \tilde{A}(x)} R(x,a)$ is a measurable function on $Dom(\tilde{A})$ by Lemma A.2(a). Suppose there is $x \in S \backslash Dom(\tilde{A})$. We have

$$\inf_{a \in \tilde{A}(x)} R(x,a) = \infty > r^1 + h_1^{\vec{r}}(x),$$

which contradicts with the fact that $h_1^{\vec{r}} \in \mathbf{B}_{w^2}(S)$. Thus, $A_1^{-1}[G] = E$ is measurable, which in turn validates the measurability of $A_1(x)$ together with its graph

$$\mathbb{K}^1 := \{(x,a); x \in S, a \in A_1(x)\}.$$

Moreover, Lemma A.2(a) ensures the existence of a measurable function g from S to A.

Note that the above two observations ensures $\mathcal{M}_1$ is legally formulated in the sense that all of its components satisfies all the definitions in Chapter 1.

Observation 3: Assumption 5.2(b) remains to be satisfied, i.e., $v(\cdot,\cdot)$ is a moment with respect to the graph $\mathbb{K}^1$.

Suppose $(\mathbb{K}_n)$ is a non-decreasing sequence of compact subsets of $\mathbb{K}$ that converges to it. Define $\mathbb{K}_n^1 := \mathbb{K}_n \cap \mathbb{K}^1$. Apparently, $(\mathbb{K}_n^1)$ are non-decreasing and converges to $\mathbb{K}^1$, since $\mathbb{K}^1 \subseteq \mathbb{K}$. Next,

$$\mathbb{K}^1 \backslash \mathbb{K}_n^1 = \mathbb{K}^1 \cap {\mathbb{K}_n^1}^C = \mathbb{K}^1 \cap (\mathbb{K}_n^C \cup {\mathbb{K}^1}^C) = \mathbb{K}^1 \cap \mathbb{K}_n^C \subseteq \mathbb{K} \backslash \mathbb{K}_n.$$

Thus,

$$\lim_{n\to\infty} \inf_{(x,a)\in \mathbb{K}^1 \backslash \mathbb{K}_n^1} v(x,a) \geq \lim_{n\to\infty} \inf_{(x,a)\in \mathbb{K}_n^C} v(x,a) = \infty,$$

which validates that $v(\cdot,\cdot)$ remains a moment with respect to $\mathbb{K}^1$.

The above observation ensures the sufficiency of stable policies for the original Problem (5.1) by the same reasoning presented in Lemma 5.2. Accordingly, the space of stable measures reduces to

$$\begin{aligned} \mathcal{D}_1 \;\; := \;\; & \{\mu \in \mathcal{D};\; \mu(dx \times da) = \varphi(da|x)\mu_\varphi(dx \times A), \\ & \varphi(A_1(x)|x) = 1, \forall\; x \in S\} \end{aligned}$$

and the space of performance vectors of the sub-model $\mathcal{M}_1$ is defined accordingly by

$$\begin{aligned} V_1(U^{stable}, \gamma) := \Bigg\{ \Bigg(& \int_{\mathbb{K}} c_0(x,a)\mu(dx \times da), \ldots, \\ & \int_{\mathbb{K}} c_M(x,a)\mu(dx \times da) \Bigg) ; \mu \in \mathcal{D}_1 \Bigg\}, \end{aligned} \tag{5.31}$$

both of which are convex and compact (with respect to $(\mathcal{M}_w(\mathbb{K}), \tau(\mathcal{M}_w(\mathbb{K})))$ and $(M+1)$-th Euclidean topology, respectively).

Observation 4: $V_1(U^{stable}, \gamma)$ coincides with the exposed subset $V(U^{stable}, \gamma) \bigcap H_1$.

We show this fact by the bilateral inclusion.

Suppose $\vec{u} = (u_0, u_1, \ldots, u_M) \in V_1(U^{stable}, \gamma)$, there is $\mu_1 \in \mathcal{D}_1$ such that $\vec{u} = \underline{Z}(\mu_1)$. To be explicit,

$$(u_0, \ldots, u_M) = \left(\int_{\mathbb{K}} c_0(x,a)\mu_1(dx \times da), \ldots, \int_{\mathbb{K}} c_M(x,a)\mu_1(dx \times da) \right)$$

Since $\mathcal{D}_1 \subseteq \mathcal{D}$, $\vec{u} \in V(U^{stable}, \gamma)$. By the definition of $A_1(x)$, we have

$$r^1 + h_1^{\vec{r}}(x) = c_1^{\vec{r}}(x,a) + \int_S h_1^{\vec{r}}(y) Q(dy|x,a) \quad \forall\ x \in S,\ a \in A_1(x).$$

Replacing $a \in A_1(x)$ with the stationary policy φ_1 leads to

$$r^1 + h_1^{\vec{r}}(x) = c_1^{\vec{r}}(x,\varphi_1) + \int_S h_1^{\vec{r}}(y) Q(dy|x,\varphi_1) \quad \forall\ x \in S, \tag{5.32}$$

where φ_1 comes from the disintegration

$$\mu_1(dx \times da) = \varphi_1(da|x) \cdot \mu_1(dx \times A).$$

Taking integration on both sides of (5.32) with respect to the $\mu_1(dx \times A)$ yields

$$\begin{aligned} & r^1 + \int_S h_1^{\vec{r}}(x) \mu_1(dx \times A) \\ = \;& \int_S \left\{ c_1^{\vec{r}}(x,\varphi_1) + \int_S h_1^{\vec{r}}(y) Q(dy|x,\varphi_1) \right\} \mu_1(dx \times A). \end{aligned} \tag{5.33}$$

Then,

$$\sum_{m=0}^{M} r_m^1 u_m = \int_{\mathbb{K}} \sum_{m=0}^{M} c_m(x,a) \mu_1(dx \times da) = \int_{\mathbb{K}} c_1^{\vec{r}}(x,a) \mu_1(dx \times da) = r^1$$

where the last equality holds because of (5.33) by Definition 5.1. Thus, $\vec{u} \in H_1$ as well.

Conversely, assume $\vec{u} = (u_0, u_1, \ldots, u_M) \in V(U^{stable}, \gamma) \bigcap H_1$. That is, there exists $\mu \in \mathcal{D}$, such that $\vec{u} = \underline{Z}(\mu)$ and $\int_{\mathbb{K}} c_1^{\vec{r}}(x,a) \mu(dx \times da) = r^1$. This implies that μ solves the following linear problem

$$\int_{\mathbb{K}} c_1^{\vec{r}}(x,a) \mu(dx \times da) \to \min_{\mu \in \mathcal{D}}$$

Again, $\mu(dx \times da) = \mu(dx \times A) \cdot \varphi^{\mu}(da|x)$. Recall the definition

$$R(x,a) := c_1^{\vec{r}}(x,a) + \int_S h_1^{\vec{r}}(y) Q(dy|x,a).$$

Integrating on both sides of the ACOE, i.e., (5.29), with respect to μ gives

$$\begin{aligned}
&\int_S \min_{A_0(x)} R(x,a)\mu(dx\times A) = r^1 + \int_S h_1^{\vec{r}}(y)\mu(dy\times A)\\
= \quad & \int_{\mathbb{K}} c_1^{\vec{r}}(x,a)\mu(dx\times da) + \int_S h_1^{\vec{r}}(y)\int_{\mathbb{K}} Q(dy|x,a)\mu(dx\times da)\\
= \quad & \int_S \left(\int_A c_1^{\vec{r}}(x,a)\varphi^{\mu}(da|x)\right)\mu(dx\times A)\\
& +\int_S \left(\int_A\int_S h_1^{\vec{r}}(y)Q(dy|x,a)\varphi^{\mu}(da|x)\right)\mu(dx\times A)\\
= \quad & \int_S R(x,\varphi^{\mu})\mu(dx\times A)
\end{aligned}$$

Thus, there exists a measurable set $S^{\vec{r}}_{\varphi^\mu}\subseteq S$ with $\mu(S^{\vec{r}}_{\varphi^\mu})=1$, such that $\varphi^\mu(L(x)|x)=1$ for each $x\in S^{\vec{r}}_{\varphi^\mu}$. As a consequence, we define a policy taking the following form

$$\hat{\varphi}(da|x) := \begin{cases}\varphi(da|x) & x\in S^{\vec{r}}_{\varphi^\mu}\\ \hat{f}(x) & x\in S\backslash S^{\vec{r}}_{\varphi^\mu}\end{cases}$$

where $\hat{f}(x)\in L(x)$ is an arbitrarily selected measurable selector; so the obtained $\hat{\varphi}^\mu(da|x)$ is an admissible policy for the sub-model $\mathcal{M}_1$. Further observe that $\mu(dx\times da)=\mu(dx\times A)\cdot\varphi^\mu(da|x)=\mu(dx\times A)\cdot\hat{\varphi}^\mu(da|x)$, which leads to $\mu\in\mathcal{D}_1$. Note that for each $\hat{\eta}\in V(U^{stable},\gamma)\bigcap H_1$, one can always find a corresponding measurable $S^{\vec{r}}_{\varphi^\eta}\subseteq S$ (of course depending on φ^η) and carry out the same procedure to obtain a policy concentrated on $A_1(x)$. To conclude, $V_1(U^{stable},\gamma)=V(U^{stable},\gamma)\bigcap H_1$.

Observation 5: Each element $\vec{u}=(u_0,u_1,\ldots,u_M)\in V_1(U^{stable},\gamma)$ can be uniquely determined by $(u_0,u_1,\ldots,u_{M-1})$.

Indeed, this is a direct consequence of Observation 4. For each $\vec{u}=(u_0,u_1,\ldots,u_M)\in V_1(U^{stable},\gamma)$, we have $\vec{u}=(u_0,u_1,\ldots,u_M)\in V(U^{stable},\gamma)\bigcap H_1\subseteq H_1$. Thus, one can always identify the location of $\vec{u}$

even if the value u_k is missing for some $k = 0, 1, \ldots, M$, since

$$u_k = \frac{r^1 - \sum_{m \neq k}^{M} r_m^1 u_m}{r_k^1}.$$

In practice, one may simply get rid of any coordinate representing $V_1(U^{stable}, \gamma)$. Without loss of generality, we choose to drop the coordinate with the largest index, namely $(M+1)$-th. As a consequence, the space of performance vectors corresponding to $\mathcal{M}_1$ can be represented as

$$\begin{aligned} V^1(U^{stable}, \gamma) \quad := \quad & \left\{ \left(\int c_0(x,a)\mu(dx \times da), \ldots, \right. \right. \\ & \left. \left. \int c_{M-1}(x,a)\mu(dx \times da) \right) ; \mu \in \mathcal{D}_1 \right\} \qquad (5.34) \end{aligned}$$

Compare (5.31) and (5.34) to see the implication of the above observation.

Observation 6: $\vec{u}_1^{ex} = (u_0^{ex}, u_1^{ex}, \ldots, u_{M-1}^{ex})$, the projection of u^* on the $(M+1)$-th coordinate, is an extreme point of $V^1(U^{stable}, \gamma)$

Observation 4 and [13, Prop.3.3.1] directly yields the stated result.

The remainder is to carry out the above procedure in the recursive way. For example, if we have the sub-model

$$\mathcal{M}_k = \{S, A, A_k(x), Q, (c_i)_{i=0,1,\ldots,M-k+1}, \gamma\}$$

and the corresponding the space of performance vectors

$$\begin{aligned} V^k(U^{stable}, \gamma) \quad := \quad & \left\{ \left(\int_{\mathbb{K}} c_0(x,a)\mu(dx \times da), \ldots, \right. \right. \\ & \left. \left. \int_{\mathbb{K}} c_{M-k}(x,a)\hat{\mu}(dx \times da) \right) ; \hat{\mu} \in \mathcal{D}_k \right\} \end{aligned}$$

One can draw a supporting hyperplane $H_{k+1} : \sum_{m=0}^{M-k} r_m^{k+1} u_m = r^{k+1}$ at $\vec{u}_k^{ex} = (u_0^{ex}, u_1^{ex}, \ldots, u_{M-k}^{ex})$. Note that H_{k+1} could be unique; see [13, p.183]. Likewise, we consider the minimization problem associated with $c_{k+1}^{\vec{r}}(x,a) := \sum_{m=0}^{M-k+1} r_m^{k+1} c_m(x,a)$ with the optimal value r^{k+1}. There is a measurable function $h^{\vec{r}}_{k+1} \in \mathbf{B}_{w^2}(S)$, which satisfies the corresponding ACOE

$$\begin{aligned} & r^{k+1} + h^{\vec{r}}_{k+1}(x) \\ = \; & \min_{A_k(x)} \left\{ c^{\vec{r}}_{k+1}(x,a) + \int_S h^{\vec{r}}_{k+1}(y) Q(dy|x,a) \right\} \quad \forall\; x \in S. \quad (5.35) \end{aligned}$$

Similarly, the refinement of admissible action space is defined by

$$\begin{aligned} L_{k+1}(x) \;\; := \;\; & \big\{ a \in A_k(x); r^{k+1} + h^{\vec{r}}_{k+1}(x) \\ & = c^{\vec{r}}_{k+1}(x,a) + \int_S h^{\vec{r}}_{k+1}(y) Q(dy|x,a) \Big\} \end{aligned}$$

and consequently, the new admissible action space is defined by $A_{k+1}(x) := L_{k+1}(x)$ for each $x \in S$.

We obtain the sub-model

$$\mathcal{M}_{k+1} := \{S, A, A_{k+1}(x), Q, (c_i)_{i=0,1,\ldots,M-k}\},$$

which is well defined by the same reasoning as previously. The space of stable measures takes the form

$$\begin{aligned} \mathcal{D}_{k+1} \;\; := \;\; & \{\mu \in \mathcal{D}_k;\; \mu(dx \times da) = \varphi(da|x)\mu_\varphi(dx), \\ & \text{and } \varphi(A_{k+1}(x)|x) = 1 \text{ for each } x \in S\} \end{aligned}$$

and the new performance space is defined accordingly,

$$\begin{aligned} V_{k+1}(U^{stable}, \gamma) \;\; := \;\; & \left\{ \left(\int_{\mathbb{K}} c_0(x,a)\mu(dx \times da), \ldots, \right. \right. \\ & \left. \left. \int_{\mathbb{K}} c_{M-k}(x,a)\mu(dx \times da) \right) ; \mu \in \mathcal{D}_{k+1} \right\} \end{aligned}$$

Again by the similar reasoning, we obtain that

$$V_{k+1}(U^{stable}, \gamma) = V^k(U^{stable}, \gamma) \bigcap H_{k+1}.$$

Then, we project $V_{k+1}(U^{stable}, \gamma)$ onto $(M-k+1)$-th coordinate, and obtain

$$\begin{aligned} V^{k+1}(U^{stable}, \gamma) \;:=\; & \Bigg\{ \bigg(\int_{\mathbb{K}} c_0(x,a)\mu(dx \times da), \dots, \\ & \int_{\mathbb{K}} c_{M-k-1}(x,a)\mu(dx \times da) \bigg) ; \mu \in \mathcal{D}_{k+1} \Bigg\} \end{aligned}$$

Note that $\vec{u}^{ex}_{k+1} = (u^{ex}_0, u^{ex}_1, \dots, u^{ex}_{M-k-1})$, the projection of u^{ex}_k onto $(M-k+1)$-th coordinate, is still the extreme point of $V^{k+1}(U^{stable}, \gamma)$.

With the above sequence of steps being conducted M times, we obtain the one-dimensional compact-convex performance space

$$V^M(U^{stable}, \gamma) = \left\{ \int_{\mathbb{K}} c_0(x,a)\mu(dx \times da); \mu \in D_M \right\}$$

which is indeed a bounded interval in the space of real numbers. Denote by v_{max} and v_{min} the end points respectively. With slight abuse of notations, either $\vec{u}^{ex}_M = u^{ex}_0 = v_{max}$ or $u^{ex}_0 = v_{min}$. By Lemma 5.1, there exists a corresponding deterministic stationary policy f^{ex} such that either v_{max} and v_{min} is attained, i.e., $u^{ex}_0 = V_0(f^{ex}, \gamma)$. □

Corollary 5.1 *Under the same set of assumptions imposed in Theorem 5.1, Let*

$$V(U^{DS}, \gamma) := \{ (V_0(\pi, \gamma), V_1(\pi, \gamma), \dots, V_M(\pi, \gamma)) ; \pi \in U^{DS} \}$$

be the space of performance vectors generated by the set of deterministic stationary policies. Then $conv(V(U^{DS}, \gamma)) = V(U^{stable}, \gamma)$, where $conv(V(U^{DS}, \gamma))$ denotes the convex hull of $V(U^{DS}, \gamma)$.

Proof. By Lemma 5.3, Theorem 5.1 and Krein-Milman's Theorem (see Theorem C.2), it is direct to observe that $V(U^{stable}, \gamma) \subseteq conv(V(U^{DS}, \gamma))$. The reverse inclusion follows from Carathéodory's theorem (see Theorem C.3) and the convexity of $V(U^{stable}, \gamma)$. □

5.6 Existence of optimal mixing policies

For the source of constraints $(d_m)_{m=1,\ldots,M}$ and for each $m = 0, 1, \ldots, M$, define the closed half-plane in $\mathbb{R}^{M+1}$ by

$$Q_m := \{\vec{u} = (u_0, u_1, \ldots, u_M) \in \mathbb{R}^k; u_m \leq d_m\},$$

and the hyperplane by

$$L_m := \{\vec{u} = (u_0, u_1, \ldots, u_M) \in \mathbb{R}^k; u_m = d_m\}.$$

Further denote $Proj_0(\vec{u}) := u_0$ for any $\vec{u} = (u_0, u_1, \ldots, u_M) \in \mathbb{R}^{M+1}$ as the projection on the first coordinate.

The next auxiliary lemma comes from the main theorem and (5.7) of [30].

Lemma 5.5 *Let $\Lambda \subseteq \{1, 2, \ldots, M+1\}$ be an index set (possibly empty) with the cardinal number being denoted by $card(\Lambda)$. For a given compact and convex set $\mathcal{H} \subseteq \mathbb{R}^{M+1}$, the following two results holds.*
(a) Any extreme point of the set $\mathcal{H} \bigcap \left(\bigcap_{m\in\Lambda} L_m\right)$ can be expressed by the convex combination of no more than $card(\Lambda) + 1$ extreme points of $\mathcal{H}$;
(b) Any extreme point of the set $\mathcal{H} \bigcap \left(\bigcap_{m\notin\Lambda} Q_m \backslash L_m\right)$ is an extreme point of $\mathcal{H}$, where we adopt the convention that an intersection over the empty index set is the universal set.

Theorem 5.2 *Under Assumption 5.1, 5.2 and 5.4, the following statements hold:*
(a) Problem (5.26) is solvable, and further there exists a stable policy φ^{opt} optimal to Problem (5.1);
(b) If in addition Assumption 5.3 and 5.5 are also satisfied, there exist constants $\lambda^*_m, m = 1, 2, \ldots, M+1$ and stable measures $\mu^*_m, m = 1, \ldots, M+1$ such that $\lambda^*_n \geq 0$, $\sum_{m=1}^{M+1} \lambda^*_m = 1$, and $\mu^*_m, m = 1, 2, \ldots, M+1$ are generated by deterministic stationary policies, say f^*_m, and the stable measure defined by $\mu^* := \sum_{m=1}^{M+1} \lambda^*_m \mu^*_m$ solves the Problem (5.26).

Proof. (a) The space of feasible performance vectors for Problem (5.1) is defined by

$$\begin{aligned} V^{feasible} &:= \{\vec{u} = (u_0, \ldots, u_M) \in V(U^{stable}, \gamma) : u_m \leq d_m, m = 1, \ldots, M\} \\ &= V(U^{stable}, \gamma) \bigcap \left(\bigcap_{m=1}^{M} Q_m \right) \end{aligned}$$

It can be observed that $V^{feasible}$ is nonempty, compact and convex. Thus, the space of the projection of $V^{feasible}$, defined by

$$Proj_0(V^{feasible}) := \{u_0 \in \mathbb{R} : \vec{u} = (u_0, u_1, \ldots, u_M) \in V^{feasible}\},$$

is nonempty, compact and convex as well, so is $V^{feasible} \bigcap L_0$, where

$$L_0 = \{\vec{u} = (u_0, u_1, \ldots, u_M) \in \mathbb{R}^{M+1}; u_0 = \inf_{\vec{u} \in V^{feasible}} Proj_0(\vec{u})\}.$$

Therefore, there exists an extreme point $\vec{u}^{ex} = (u_0^{ex}, u_1^{ex}, \ldots, u_M^{ex})$ of $V^{feasible}$ such that

$$u_0^{ex} = \inf_{\vec{u} \in V^{feasible}} Proj_0(\vec{u}).$$

Denote by μ^{opt} such that $\underline{Z}(\mu^{Opt}) = \vec{u}^{ex}$, and $\mu^{opt} = \varphi^{opt} \cdot \mu^{opt}$. Then φ^{opt} is an optimal policy for Problem (5.1).

(b) Let $\Lambda := \{1 \leq m \leq M : u_m^{ex} = d_m\}$, we have

$$\begin{aligned} \vec{u}^{ex} \in V(U^{stable}, \gamma) \ \ \bigcap \ \ & \left(\bigcap_{m \in \Lambda} L_m \right) \bigcap \left(\bigcap_{m \notin \Lambda} Q_m \backslash L_m \right) \\ \subseteq \ \ & V(U^{stable}, \gamma) \bigcap \left(\bigcap_{m=1}^{M} Q_m \right) \end{aligned}$$

Since the $\vec{u}^{ex}$ is an extreme point of $V^{feasible}$, it is also an extreme point of

$$V(U^{stable}, \gamma) \bigcap \left(\bigcap_{m \in \Lambda} L_m \right) \bigcap \left(\bigcap_{m \notin \Lambda} Q_m \backslash L_m \right).$$

By Lemma 5.5(b), $\vec{u}^{ex}$ is an extreme point of $V(U^{stable}, \gamma) \bigcap \left(\bigcap_{m \in \Lambda} L_m \right)$, where the cardinal number of Λ can not exceed M. It further follows from

Lemma 5.5(a) that there exist extreme points $\vec{u}_m^{ex}$, $m = 1, 2, \ldots, M+1$, of $V(U^{stable}, \gamma)$ and nonnegative constants λ_m^*, $m = 1, 2, \ldots, M+1$, such that $\sum_{m=1}^{M+1} \lambda_m^* = 1$ and $\vec{u}^{ex} = \sum_{m=0}^{M+1} \lambda_m^* \vec{u}_m^{ex}$. Denote by $\mu_m^* \in \mathcal{D}$, $m = 1, 2, \ldots, M+1$, the preimage of $\vec{u}_m^{ex}$, $m = 1, 2, \ldots, M+1$, under the linear mapping $\underline{Z}$ coming from (5.27), and further define a measure $\mu^* := \sum_{m=1}^{M+1} \lambda_m^* \mu_m^*$. μ^* is an optimal solution to Problem (5.1) as $\underline{Z}(\mu^*) = \vec{u}^{ex}$, and $\mu^* \in \mathcal{D}$ because of the convexity of $\mathcal{D}$. Note that by applying Theorem 5.1, there exists f_m, $m = 1, 2, \ldots, M+1$ such that

$$\underline{Z}(\mu_m^*) = \vec{u}_m^{ex} = (V_0(f_m, \gamma), V_1(f_m, \gamma), \ldots, V_M(f_m, \gamma))$$

for $m = 1, 2, \ldots, M+1$.

Finally, φ^* corresponding to μ^* is the promised optimal mixing policy optimal for Problem (5.1). □

Chapter 6

Average optimality of an MDP with the general cost function

6.1 Introduction

This chapter is organized as follows. Section 6.2 is a description of the model and the average optimality problem. We introduce a set of assumptions about sufficiency to the concerned problem along with some discussions in the denumerable case. Section 6.4 is devoted to the main optimality result together with its proof. Section 6.5 involves an illustrative example.

6.2 Problem formulation and preliminaries

We solve this problem by using *vanishing discount approach* under the condition that there exists a class of sufficient policies inducing corresponding processes that possess some ergodic properties. In connection with formulations in related literature (cf. [36], [46], [47], [50], [86], [87], [88]), the cost defined in this chapter is allowed to be unbounded from the below while keeping its positive part to be arbitrarily unbounded. As a consequence, we deal with the positive and negative parts of the cost separately.

As usual, a standard MDP is characterized with the following primitives,

$$\{S, A, (A(x), x \in S), Q(dy|x, a), c(x, a)\}$$

Let $\mathbf{B}_w^-(S)$ denotes the space of measurable functions on S, each of which the negative part is bounded by $w(\cdot)$. A strong version of compactness-continuity conditions is imposed,

Assumption 6.1 *(a) $c(x, a)$ is lower semi-continuous in $a \in A(x)$, for each $x \in S$;*
(b) $A(x)$ is compact-valued for each $x \in S$;
(c) $\int_S u(y)Q(dy|x, a)$ is continuous in $a \in A(x)$ $\forall x \in S, u \in \mathbf{B}(S)$.

For an arbitrary policy $\pi \in U^H$, we define the value function and optimal value function for a discount (total) problem over n-stage horizon respectively by

$$\begin{aligned} J_\alpha^n(\pi, x) &:= E_x^\pi[\sum_{t=0}^{n-1} \alpha^t c(x_t, a_t)] \\ J_\alpha^{*n}(x) &:= \inf_\pi J_\alpha^n(\pi, x) \\ J^n(\pi, x) &:= V_1^n(\pi, x) = E_x^\pi[\sum_{t=0}^{n-1} c(x_t, a_t)] \end{aligned}$$

and those over the infinite horizon by

$$\begin{aligned} J_\alpha(\pi, x) &:= E_x^\pi[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t)] \\ J_\alpha^*(x) &:= \inf_\pi J_\alpha(\pi, x) \end{aligned}$$

The positive and negative parts of $c(x, a)$ are defined by

$$c^+(x, a) := \max\{0, c(x, a)\}, \quad c^-(x, a) := \max\{0, -c(x, a)\}$$

Accordingly, we have similar presentations of problems associated with positive and negative parts of the cost function,

$$\begin{aligned}
J_\alpha^{(+)}(\pi,x) &:= E_x^\pi[\sum_{t=0}^{\infty}\alpha^t c^+(x_t,a_t)] \\
J_\alpha^{*(+)}(x) &:= \inf_\pi J_\alpha^{(+)}(\pi,x) \\
J_\alpha^{(-)}(\pi,x) &:= E_x^\pi[\sum_{t=0}^{\infty}\alpha^t c^-(x_t,a_t)] \\
J_\alpha^{*(-)}(x) &:= \sup_\pi J_\alpha^{(-)}(\pi,x)
\end{aligned}$$

Note that, $c^-(x,a)$ is upper semicontinuous in $a \in A(x)$ for each $x \in S$, for which we investigate the maximization problem associated with it.

The average cost optimal problem of our interest in this chapter takes the following form,

$$V(\pi,x) := \overline{\lim_{n\to\infty}} \frac{1}{n} J^n(\pi,x) \longrightarrow \min_{\pi\in U^H} \tag{6.1}$$

To have our average and auxiliary discounted problems well defined, negative part of the cost function is assumed to be controlled by a weight function, which satisfies Lyapunov-like condition.

Assumption 6.2 *There exists a continuous weight function $w(\cdot) \geq 1$ on S, a bounded measurable function (possible constant) function $b(\cdot)$, and nonnegative constants $\hat{c}$ and β, with $\beta < 1$, such that,*
(a) $c^-(x,a) \leq \hat{c}w(x)$, for each $x \in S$ and $a \in A(x)$;
(b) $a \to \int_S w(y)Q(dy|x,a)$ is continuous in $a \in A(x)$, for each $x \in S$;
(c) $\sup_{A(x)} \int_S w(y)Q(dy|x,a) \leq \beta w(x) + b(x)$, for each $x \in S$.

It is worth mentioning that Assumption 6.2 is quite standard in the study of so-called *negative dynamic programming* problem with total cost criterion first examined by Strauch in [94]. Actually, Assumption 6.2(a,c) are sufficient for the following two general conditions with $0 < \alpha < 1$ to hold,

$$\delta(x) := \sup_\pi E_x^\pi[\sum_{t=0}^{\infty}\alpha^t c^-(x_t,a_t)] < \infty \tag{6.2}$$

$$\lim_{n\to\infty}\sup_{\pi} E_x^{\pi}[\sum_{t=n}^{\infty}\alpha^t c^-(x_t,a_t)] = 0 \tag{6.3}$$

both of which, along with Assumption 6.1, validate *Structure Assumption* in [8, p.199]. Therefore, standard optimality results, i.e., the establishment of optimality equation and existence of an optimal deterministic stationary policy, follow accordingly. We quote them for future reference in the following lemma.

Lemma 6.1 *Let* $0 < \alpha < 1$ *be an arbitrarily fixed discount factor. Under Assumption 6.1 and 6.2,* $J_\alpha^{*(-)}(x)$ *is the maximum solution out of the class* $u \in \mathbf{B}_w(S)$ *to the optimality equation*

$$u(x) = \sup_{a\in A(x)}\left\{c^-(x,a) + \alpha\int_S u(y)Q(dy|x,a)\right\} \quad \forall\ x \in S.$$

Moreover, there exists a measurable mapping f_α *that attains the supremum in the right hand side, i.e.,*

$$J_\alpha^{*(-)}(x) = c^-(x,f_\alpha) + \alpha\int_S J_\alpha^{*(-)}(y)Q(dy|x,f_\alpha) \quad \forall\ x \in S.$$

In addition, f_α *induces a deterministic stationary policy which is optimal (not necessarily unique) to the discounted problem*

$$E_x^{\pi}[\sum_{t=0}^{\infty}\alpha^t c^-(x_t,a_t)] \to \max_{\pi}$$

Proof. see [8, 85]. □

6.3 Sufficient policies and discussions in the denumerable case

At first, we formally define the notion of sufficiency with respect to our concerned average optimal problem.

Definition 6.1 *A family of policies* U^{suff} *is called sufficient for Problem (6.1), if and only if for any policy* $\pi \in U^H$*, there exists a corresponding* *policy $\hat{\pi} \in U^{suff}$ such that*

$$V(\hat{\pi}, x) \leq V(\pi, x) \quad \forall\ x \in S.$$

A set of new assumptions below is imposed, on the one hand, to avoid Problem (6.1) being trivial, and on the other hand, to assume some ergodic properties on the mixture of the dynamics and negative part of the cost.

Assumption 6.3 *(a)There is a sufficient class of policies U^{suff} for Problem (6.1) such that the following two assertions are satisfied:*

(i) There exist a policy $\hat{\pi} \in U^{suff}$ and a state $z \in S$ that satisfy

$$V(\hat{\pi}, z) < \infty.$$

(ii) For each policy $\hat{\pi} \in U^{suff}$ and each state $x \in S$,

$$\lim_{n\to\infty} \frac{1}{n} E_x^{\hat{\pi}}[\sum_{t=0}^{n-1} c^-(x_t, a_t)]$$

exists and is finite.

(b) For every deterministic stationary policy $f \in U^{DS}$, there exists a unique probability measure μ_f on S such that

$$\|Q_f^t - \mu_f\|_w \leq R\rho^t \quad \forall\ t = 0, 1, \cdots$$

where $R > 0$ and $0 < \rho < 1$ are constants independent of $f \in U^{DS}$.

Note that Assumption 6.3(a,i) is equivalent to the statement that there exists a policy $\pi \in U^H$ such that

$$V(\pi, z) < \infty,$$

which should be clear from the definition of a sufficient policy. In comparison, Assumption 6.3(a,ii) is more subtle, and seems ambiguous and restrictive at first sight. In practice, the restriction to a subclass of sufficient policies is necessary. Otherwise, there could be policies such that

Assumption 6.3(a,ii) fails to hold whenever the action is relevant to either negative part of the cost or the dynamics (the corresponding transition kernel); see Example 6.1. This sort of setting is too restrictive to make it meaningful to allow for negative part of the cost being unbounded. Note that sufficiency does not have to guarantee the existence of an optimal policy, but instead narrow the search for optimal policies to a smaller scale. As is seen in Chapter 5, the family of stable policies is an example of sufficient class of policies to Problem (5.1). The issue of sufficiency is discussed in greater detail in the present setting.

Example 6.1

Let $S = \{0, 1, 2, \ldots\}$ be a countable state space, and $A = \{0, 1\}$ be a finite action spaces, and $A(i) \equiv A$. Set $Q(i+1|i, 1) = 1$, and $Q(0|i, 0) = 1$, $\forall\ i \geq 1$, whereas $Q(1|0, 1) = 1$, $Q(0|0, 0) = 1$, $c(i, a) = Ci$, where C is a positive constant. Note that the selected a is only related to the dynamics Q.

We choose a policy $\pi = (\pi_n)$ as $\pi_{n_k} = 0$, where (n_k) is a subsequence defined as $n_k = \frac{k^2+3k}{2} - 1$, $k = 1, 2, \ldots$, and otherwise $\pi_t = 1$. The trajectory of the controlled process x_t under the specified policy π is $h_t = (0, 1, 0, 1, 2, 0, 1, 2, 3, 0, \ldots)$. We can see

$$\begin{aligned}
\varliminf_{n\to\infty} \frac{1}{n} E_0^{\pi}\left[\sum_{t=0}^{n-1} c(x_t, a_t)\right] &= \lim_{k\to\infty} \frac{1}{n_k+2} E_0^{\pi}\left[\sum_{t=0}^{n_k+1} c(x_t, a_t)\right] \\
&= \lim_{k\to\infty} \frac{2}{k^2+3k+2} \sum_{l=1}^{k} \frac{l(l+1)}{2} \\
&= \lim_{k\to\infty} \frac{2}{k^2+3k+2} \frac{k(k+1)(k+2)}{6} \\
&= \lim_{k\to\infty} \frac{k}{3} = \infty
\end{aligned}$$

which violates Assumption 6.2(a,ii). Indeed, the policy can be selected in a simpler way.

The state space S is assumed to be denumerable throughout the remainder of this section. A set of assumptions is introduced as follows,

Assumption 6.4 *(a) There is a moment (strictly unbounded function) $w'(\cdot)$ on S such that for every policy $\pi \in U^H$,*

$$\overline{\lim_{n\to\infty}} \int_{\mathbb{K}} w'(x)w(x)\mu^{\pi}_{\gamma,n}(dx \times da) < \infty, \tag{6.4}$$

where the sequence of occupation measures $\mu^{\pi}_{\gamma,n}(dx \times da)$ stems from (5.18);

(b) The state space S forms a single positive recurrent class for every stationary policy $\varphi \in U^S$;

(c) There exits a finite set $C \subset S$ such that

$$\sup_{a\in A(i)} \sum_{j\in S} w(j)Q(j|i,a) \leq \beta w(i) + b(i)\mathbf{1}_{\{i\in C\}} \tag{6.5}$$

Note that Assumption 6.4(c) is more restrictive than Assumption 6.2(c). The purpose of this modification is to provide sufficient conditions for Assumption 6.3(b), which will be revealed later.

Lemma 6.2 *(a) Under Assumption 6.2(c) and Assumption 6.4(b) for each stationary policy $\varphi \in U^S$,*

$$E^{\varphi}_i[\sum_{t=0}^{\tau^i-1} w(x_t)] < \infty \tag{6.6}$$

$$\sum_{i\in S} w(i)\mu^{\varphi}(i) < \infty \tag{6.7}$$

holds, where

$$\tau^i := \inf\{n \geq 1 : x_0 = i, x_\nu \neq i, 1 \leq \nu \leq n-1, x_n = i\}$$

(φ is omitted for simplicity) denotes the first return time to a prefixed state $i \in S$ when starting from the same state, and $\mu^{\varphi}(i)$ stands for the unique i.p.m. of the Markov chain $Q_{\varphi}(j|i)$;
(b) if, in addition, Assumption 6.2(c) is replaced with Assumption 6.4(c), Assumption 6.3(b) holds.

Proof. (a) Firstly, we arbitrarily select and fix a stationary policy φ. Let us put

$$p_{ij}^{(n)} = Q^n(j|i,\varphi) := P_i^\varphi(x_n = j)$$

and

$$_kp_{ij}^{(n)} := P_0^\varphi(x_n = j,\ x_\nu \neq k,\ 0 < \nu < n)$$

denote the n-step transition probability, and the n-step transition probability with a taboo state $k \in S$ for any $i, j \in S$ respectively (see [24, §9] for greater details regarding "taboo probabilities"). Assume that τ^i follows the distribution $f_{ii}^{(n)}$ (or equivalently, $_ip_{ii}^{(n)}$). Note that all the notations τ^i, $f_{ii}^{(n)}$, $p_{ij}^{(n)}$ and $_kp_{ij}^{(n)}$ correspond to the Markov chain $Q_\varphi(j|i)$, and the prefixed stationary policy φ is omitted for the sake of simplicity.

For (6.6), we fix an arbitrary state $0 \in S$, and denote by $(\tau_k^0)_{k=0,1,2,\ldots}$ the increasing infinite sequence ($\tau_0^0 := 0$ and $\tau_1^0 = \tau^0$) of all values of $n \geq 1$ for which $x_n = 0$, and by

$$\eta_k^0 := \tau_k^0 - \tau_{k-1}^0$$

the k-th return time to $0 \in S$. From [24, §13], it is observed that (η_k^0) is a sequence of independent random variables with common distribution $f_{00}^{(n)}$. Thus, it suffices to consider the first cycle letting $x_0 = 0$. Moreover, τ_1^0 is simplified as τ^0 and $\|b\| := \sup_{i \in S} b(i)$ is simplified as b in the sequel.

$$\begin{aligned}
E_0^\varphi\left[\sum_{t=0}^{\tau^0-1} w(x_t)\right] &= \sum_{n=1}^{\infty} E_0^\varphi\left[\sum_{t=0}^{n-1} w(x_t)|\tau^0 = n\right] f_{00}^{(n)} \\
&= \sum_{n=2}^{\infty}\left\{w(0) + E_0^\varphi\left[\sum_{t=1}^{n-1} w(x_t)|\tau^0 = n\right]\right\} f_{00}^{(n)} \\
&= w(0) + \sum_{n=2}^{\infty}\sum_{t=1}^{n-1}\sum_{j\neq 0} w(j)P_0^\varphi(x_t = j|\tau^0 = n)f_{00}^{(n)} \\
&= w(0) + \sum_{t=1}^{\infty}\sum_{n=t+1}^{\infty}\sum_{j\neq 0} w(j)P_0^\varphi(x_t = j, \tau^0 = n) \\
&= w(0) + \sum_{t=1}^{\infty}\sum_{n=t+1}^{\infty}\sum_{j\neq 0} w(j)_0p_{j0}^{(n-t)}{}_0p_{0j}^{(t)}
\end{aligned}$$

$$
\begin{aligned}
&= w(0) + \sum_{j\neq 0} w(j) \sum_{t=1}^{\infty} {}_0p_{0j}^{(t)} \sum_{n=1}^{\infty} {}_0p_{j0}^{(n)} \\
&= w(0) \sum_{t=1}^{\infty} {}_0p_{00}^{(t)} + \sum_{j\neq 0} w(j) \sum_{t=1}^{\infty} {}_0p_{0j}^{(t)} \\
&= \sum_{j\in S} w(j) {}_0p_{0j}^{*} \\
&= \sum_{t=1}^{\infty} \sum_{j\in S} w(j) {}_0p_{0j}^{(t)}
\end{aligned}
$$

where ${}_0p_{0j}^{*} := \sum_{t=1}^{\infty} {}_0p_{0j}^{(t)}$ is finite and ${}_0p_{j0}^{*} := \sum_{n=1}^{\infty} {}_0p_{j0}^{(n)} = 1$ by Assumption 6.4(b) and the Corollary to [24, Thm.9.6]. The fifth equality, noting that $t < n$, comes from the following relation where the Markov property is in use,

$$
\begin{aligned}
& P_0^{\varphi}(x_t = j, \tau^0 = n) \\
= \quad & P_0^{\varphi}(x_t = j, (x_\nu \neq 0, 0 < \nu < n), x_n = 0) \\
= \quad & P_0^{\varphi}(x_n = 0, (x_\nu \neq 0, t < \nu < n) | x_t = j, (x_\nu \neq 0, 0 < \nu < t)) \\
& \cdot P_0^{\varphi}(x_t = j, (x_\nu \neq 0, 0 < \nu < t)) \\
= \quad & P_j^{\varphi}(x_{n-t} = j, (x_\nu \neq 0, 0 < \nu < n-t)) \cdot P_0(x_t = j, (x_\nu \neq 0, 0 < \nu < t)) \\
= \quad & {}_0p_{j0}^{(n-t)} \cdot {}_0p_{0j}^{(t)}
\end{aligned}
$$

All the interchanges of order of summations are justified due to nonnegativity of all the terms by *Tonelli's Theorem* (cf. [1, Thm.11.28]). Note that the above representation is identical to what appears in [24, Thm.14.5], but instead without extra restrictions on the convergence of the concerned series, again owing to nonnegativity.

We extract the term $\sum_{j\in S} w(j){}_0p_{0j}^{(n)}$, which is denoted by A_n, and take a look at it,

$$
\begin{aligned}
\sum_{j\in S} w(j) {}_0p_{0j}^{(n)} &= \sum_{j\in S} w(j) \sum_{k\neq 0} p_{kj} \cdot {}_0p_{0k}^{(n-1)} \\
&= \sum_{k\neq 0} {}_0p_{0k}^{(n-1)} \sum_{j\in S} w(j) p_{kj}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{k\in S} {}_0p_{0k}^{(n-1)}(\beta w(k)+b) - {}_0p_{00}^{(n-1)}(\beta w(0)+b) \\
&= \beta \sum_{k\in S} w(k) {}_0p_{0k}^{(n-1)} + b \sum_{\nu=n-1}^{\infty} f_{00}^{(\nu)} - f_{00}^{(n-1)}(\beta w(0)+b)
\end{aligned}
$$

where $\sum_{k\in S} {}_0p_{0k}^{(n-1)} = \sum_{\nu=n-1}^{\infty} f_{00}^{(\nu)}$ by [24, Thm.9.6], which gives the following recursive formula,

$$
A_n \leq \beta A_{n-1} + b \sum_{\nu=n-1}^{\infty} f_{00}^{(\nu)} - f_{00}^{(n-1)}(\beta w(0)+b).
$$

Based on the above inequality together with the initial condition

$$
A_1 \leq \beta w(0) + b,
$$

the explicit form of the bound of A_n can be expressed as

$$
A_n \leq \sum_{t=0}^{n-1} \beta^t \left[b \sum_{\nu=n-t-1}^{\infty} f_{00}^{(\nu)} - b f_{00}^{(n-t-1)} - f_{00}^{(n-t)} w(0) \right] + f_{00}^{(n)} w(0) + \beta^n w(0).
$$

Then,

$$
\begin{aligned}
E_0^{\varphi}[\sum_{t=0}^{\tau^0-1} w(x_t)] &= \sum_{n=1}^{\infty} A_n \\
&\leq \sum_{n=1}^{\infty} f_{00}^{(n)} w(0) + \sum_{n=1}^{\infty} \beta^n w(0) \\
&\quad + \sum_{t=0}^{\infty} \sum_{n=t+1}^{\infty} \left[b\beta^t \sum_{\nu=n-t-1}^{\infty} f_{00}^{(\nu)} - \beta^t b f_{00}^{(n-t-1)} - \beta^t f_{00}^{(n-t)} w(0) \right] \\
&= w(0) + \frac{\beta}{1-\beta} w(0) \\
&\quad + \sum_{t=0}^{\infty} \left[b\beta^t \sum_{\nu=0}^{\infty} \sum_{n=t+1}^{\nu+t+1} f_{00}^{(\nu)} - b\beta^t - w(0)\beta^t \right] \\
&= w(0) + \frac{\beta}{1-\beta} w(0) \\
&\quad + \sum_{t=0}^{\infty} \left[b\beta^t \sum_{\nu=0}^{\infty} (\nu+1) f_{00}^{(\nu)} - b\beta^t - w(0)\beta^t \right]
\end{aligned}
$$

$$\begin{aligned} &= w(0) + \frac{\beta}{1-\beta} w(0) + \frac{bm_{00} - w(0)}{1-\beta} \\ &= \frac{b\ m_{00}}{1-\beta} < \infty \end{aligned}$$

where state $0 \in S$ can be replaced by any state $i \in S$, and

$$m_{ii} := \sum_{n=1}^{\infty} n f_{ii}^{(n)} < \infty$$

holds for every state $i \in S$ by Assumption 6.4(b). (6.6) is justified.

(6.7) follows from [24, Thm.9.6 & Thm.9.7] with a bit computation that

$$\begin{aligned} \sum_{j \in S} w(j) \mu^{\varphi}(j) &= \sum_{j \in S} w(j) \frac{{}_i p_{ij}^{*}}{m_{ii}} \\ &= \frac{b\ m_{ii}}{1-\beta} \frac{1}{m_{ii}} = \frac{b}{1-\beta} < \infty, \end{aligned}$$

which completes the proof of part (a).

(b) can be directly derived from [92, Thm.1] or [58, (2.3) & Prop.2.4] by noting that C is finite and $\beta < 1$. □

Now we are in position to proceed with our discussion of the remaining statements in Assumption 6.3, and incidentally, show that the whole family of stationary policies is a sufficient class for Problem (6.1).

Lemma 6.3 *Under Assumption 6.1, 6.2 and 6.4, the following assertions hold:*
(a) For each stationary policy $\varphi \in U^S$ and each state $x \in S$, the following representation holds,

$$\overline{\lim_{n \to \infty}} \int_{\mathbb{K}} c(y,a) \mu_{x,n}^{\varphi}(dy \times da) = \int_{\mathbb{K}} c(y,a) \mu_{x}^{\varphi}(dy \times da); \qquad (6.8)$$

(b) The whole family of stationary policies U^S forms the promised sufficient class for Problem (6.1);
(c) Assumption 6.3(a,ii) is valid.

Proof. (a) In view of the proof of Lemma 2.3 in [3], if we fix a stationary policy φ, Lemma 6.2 (specifically, the validity of (6.6) and (6.7)) is sufficient for (6.8) to hold.

(b) It is easy to deduce from Assumption 6.4(a) that the sequence of occupation measures $(\mu^{\pi}_{x,n})$ is tight for every $\pi \in U^H$, which yields the existence of a subsequence such that $\mu^{\pi}_{x,n_i} \xrightarrow{w} \mu^{\pi^*}_x$ (recall that "$\xrightarrow{w}$" stands for convergence in the w-weak topology).By Assumption 6.2(a) and Corollary A.2(b), we have

$$\begin{aligned} \int_S c(y,a)\mu^{\varphi}_x(dy \times da) &\leq \varliminf_{i\to\infty} \int_S c(y,a)\mu^{\pi}_{x,n_i}(dy \times da) \\ &\leq \varlimsup_{n\to\infty} \int_S c(y,a)\mu^{\pi}_{x,n}(dy \times da) < \infty \end{aligned}$$

as required. So part (b) follows further from part (a).

(c) This is an immediate consequence of Lemma 6.2(a) and [24, Thm.15.2 & Thm.15.3]. □

6.4 Main statements and proofs

We start with an important auxiliary result to implement *vanishing discount factor approach* in the sequel.

Lemma 6.4 *Under Assumption 6.2 and 6.3(a),* $-\infty < \varliminf_{\alpha\uparrow 1}(1-\alpha)J^*_\alpha(z) \leq \varlimsup_{\alpha\uparrow 1}(1-\alpha)J^*_\alpha(z) < \infty$.

Proof.

$$\begin{aligned} & \varlimsup_{\alpha\uparrow 1}(1-\alpha)J_\alpha(\hat{\pi}, z) \\ \leq\ & \varlimsup_{\alpha\uparrow 1}(1-\alpha)E^{\hat{\pi}}_z[\sum_{t=0}^{\infty}\alpha^t c^+(x_t,a_t)] - \varliminf_{\alpha\uparrow 1}(1-\alpha)E^{\hat{\pi}}_z[\sum_{t=0}^{\infty}\alpha^t c^-(x_t,a_t)] \\ \leq\ & \varlimsup_{n\to\infty}\frac{1}{n}E^{\hat{\pi}}_z[\sum_{t=0}^{n-1} c^+(x_t,a_t)] - \varliminf_{n\to\infty}\frac{1}{n}E^{\hat{\pi}}_z[\sum_{t=0}^{n-1} c^-(x_t,a_t)] \\ =\ & \varlimsup_{n\to\infty}\frac{1}{n}E^{\hat{\pi}}_z[\sum_{t=0}^{n-1} c^+(x_t,a_t)] - \lim_{n\to\infty}\frac{1}{n}E^{\hat{\pi}}_z[\sum_{t=0}^{n-1} c^-(x_t,a_t)] \end{aligned}$$

$$
\begin{aligned}
&= \overline{\lim_{n\to\infty}} \frac{1}{n} E_z^{\hat{\pi}}[\sum_{t=0}^{n-1} c(x_t, a_t)] \\
&= V(\hat{\pi}, z)
\end{aligned}
$$

The second inequality holds by virtue of Abelian (Taubarian) Theorem (see Theorem C.1), and the first equality comes from (ii) of Assumption 6.3(a). Taking infimum over all policies on the left hand side leads to

$$\overline{\lim_{\alpha\uparrow 1}}(1-\alpha)J_\alpha^*(z) \le V(\hat{\pi}, z) < \infty.$$

For the other part of the statement, we observe

$$(1-\alpha)J_\alpha(\pi, z) \ge -(1-\alpha)J_\alpha^{(-)}(\pi, z),$$

thus,

$$(1-\alpha)J_\alpha^*(z) \ge -(1-\alpha)\sup_\pi J_\alpha^{(-)}(\pi, z).$$

A direct implementation of Lemma 6.1 yields the existence of a deterministic stationary policy f_α such that,

$$\sup_\pi J_\alpha^{(-)}(\pi, z) = J_\alpha^{(-)}(f_\alpha, z).$$

Moreover, we have the following,

$$
\begin{aligned}
& J_\alpha^{(-)}(f_\alpha, z) \\
= \;& E_z^{f_\alpha}\left[\sum_{t=0}^{\infty} \alpha^t c^-(x_t, a_t)\right] \\
\le \;& \hat{c}\sum_{t=0}^{\infty} \alpha^t E_z^{f_\alpha}\left[w(x_t)\right] \\
\le \;& \hat{c}\sum_{t=0}^{\infty} \alpha^t \left[\left|\int_S w(y)Q_{f_\alpha}^t(dy|z) - \int_S w(y)\mu_{f_0}(dy)\right| + \int_S w(y)\mu_{f_\alpha}(dy)\right] \\
\le \;& \hat{c}\sum_{t=0}^{\infty} \alpha^t \left[R\rho^t w(z) + \int_S w(y)\mu_{f_\alpha}(dy)\right] = \hat{c}\left[\frac{Rw(z)}{1-\rho\alpha} + \frac{1}{1-\alpha}\int_S w(y)\mu_{f_\alpha}(dy)\right]
\end{aligned}
$$

which further follows from Assumption 6.3(b). Recall (5.9) in Chapter 5 we have,

$$\|\mu_{f_\alpha}\|_w := \int_S w(y)\mu_{f_\alpha}(dy) \leq \frac{\|b\|}{1-\beta} < \infty$$

Therefore,

$$\begin{aligned} & \overline{\lim_{\alpha\uparrow 1}}\,(1-\alpha)\sup_{\pi} J_\alpha^{(-)}(\pi, z) \\ \leq \quad & \overline{\lim_{\alpha\uparrow 1}}\,\hat{c}\left[\frac{1-\alpha}{1-\alpha\rho}Rw(z) + \frac{\|b\|}{1-\beta}\right] \\ = \quad & \frac{\hat{c}\|b\|}{1-\beta} < \infty \end{aligned}$$

To summarize,

$$-\hat{c}\left[\frac{\|b\|}{1-\beta}\right] \leq \underline{\lim}_{\alpha\uparrow 1}(1-\alpha)J_\alpha^*(z) \leq \overline{\lim_{\alpha\uparrow 1}}(1-\alpha)J_\alpha^*(z) \leq V(\hat{\pi}, z).$$

as desired. □

Generally, Lemma 6.4 asserts that $(1-\alpha)J_\alpha^*(z)$ is bounded as α increases to 1, which in turn ensures the existence of a subsequence that converges to some limit point. To simplify notations, we introduce the following,

$$h_\alpha(x) \quad := \quad J_\alpha^*(x) - J_\alpha^*(z) \tag{6.9}$$

$h_\alpha(x)$ is called *relative difference* or *differential discounted value function* (see [4] for details), which is introduced to facilitate the presentation of next assumption. Loosely speaking, the positive part of cost can take values more freely, so we use conditions of the type Condition (**B**) in [86], which together with Assumption 6.3 is claimed to be equivalent to Assumption 5.4.1 in [47].

Assumption 6.5 *There exists a constant* $0 < \alpha_0 < 1$ *such that*

$$\sup_{\alpha_0 \leq \alpha < 1}\left(J_\alpha^{*(+)}(x) - \inf_{x\in S} J_\alpha^{*(+)}(x)\right) < \infty \quad x \in S$$

Combined with Assumptions 6.2 and 6.3(b), the boundedness of $h_\alpha(x)$ viewed as a function of α for each fixed $x \in S$ is established and elaborated in the following corollary with a sketched proof.

Corollary 6.1 *There exist two nonnegative real-valued functions $u_1(x)$ (measurable) and $u_2(x)$, such that $-u_1(x) \leq h_\alpha(x) \leq u_2(x)$ for all $x \in S$ and $\alpha_0 \leq \alpha < 1$. In addition, $u_1(x) \in \mathbf{B}_w(S)$.*

Proof. For the positive part, it can be seen from [47, Thm.5.4.6] that there is a constant $N \geq 0$ and a nonnegative real-valued (not necessarily measurable) function $d(\cdot)$ on S such that $-N \leq h_\alpha^+(x) \leq d(x)$ for every $x \in S$ and $\alpha \in [\alpha_0, 1)$, where α_0 stems from Assumption 6.5. The result in regard to its negative part follows directly from Lemma 6.1 and [50, Lem.10.4.2] under Assumption 6.3(b). That is,

$$|h_\alpha^-(x)| \leq \frac{\hat{c}R}{(1-\rho)}[1 + w(z)]w(x).$$

To conclude,

$$u_1(x) := N + \frac{\hat{c}R}{(1-\rho)}[1 + w(z)]w(x)$$

and

$$u_2(x) := d(x) + \frac{\hat{c}R}{(1-\rho)}[1 + w(z)]w(x)$$

are the two promised bounding functions, where, in particular, $u_1 \in \mathbf{B}_w(S)$. □

Corollary 6.1 further implies that $J_\alpha^*(x)$ is finite-valued for $\alpha_0 \leq \alpha < 1$. Refer to Lemma 6.4 and (6.9) to see this fact. Select and fix an increasing sequence of discount factors (α_n) that converges to 1 such that

$$\rho^* := \lim_{n\to\infty}(1 - \alpha_n)J_{\alpha_n}^*(z) = \overline{\lim_{\alpha\uparrow 1}}(1 - \alpha)J_\alpha^*(z) \tag{6.10}$$

where ρ^* is claimed to be a real constant by referring to Lemma 6.4. In view of Corollary 6.1, we are able to extend (6.10) to each state $x \in S$.

Corollary 6.2 *From Lemma 6.4 and Corollary 6.1, for any state $x \in S$ and prefixed sequence (α_n), we have $\lim_{n\to\infty}(1-\alpha_n)J^*_{\alpha_n}(x) = \rho^*$.*

Proof.

$$\begin{aligned}
& |(1-\alpha_n)J^*_{\alpha_n}(x) - \rho^*| \\
\leq \quad & (1-\alpha_n)|h_{\alpha_n}(x)| + |(1-\alpha_n)J^*_{\alpha_n}(z) - \rho^*| \\
\leq \quad & (1-\alpha_n)\max\{u_1(x), u_2(x)\} + |(1-\alpha_n)J^*_{\alpha_n}(z) - \rho^*| \\
\to \quad & 0 \quad as \quad n \to \infty
\end{aligned}$$

as required. □

Furthermore, we obtain

$$\rho^* = \lim_{n\to\infty}(1-\alpha_n)J^*_{\alpha_n}(x) \leq \overline{\lim_{\alpha\uparrow 1}}(1-\alpha)J^*_{\alpha}(x) \leq V(\hat{\pi}, x),$$

where $\hat{\pi} \in U^{suff}$. Taking infimum over U^H on the right hand side along with the sufficiency of U^{suff} for Problem (6.1) yields

$$\rho^* \leq \inf_{\pi\in U^{suff}} V(\pi, x) = \inf_{\pi\in U^H} V(\pi, x) \quad \forall\ x \in S. \tag{6.11}$$

In doing so, the optimality equation for the discounted problem can be expressed in terms of $J^*_{\alpha_n}(z)$ and $h_{\alpha_n}(x)$, that is,

$$\begin{aligned}
& (1-\alpha_n)J^*_{\alpha_n}(z) + h_{\alpha_n}(x) \\
= \quad & \min_{a\in A(x)}\left\{c(x,a) + \alpha_n\int_S h_{\alpha_n}(y)Q(dy|x,a)\right\} \quad \forall\ x \in S. \qquad (6.12)
\end{aligned}$$

Theorem 6.1 *Under Assumptions 6.1, 6.2, 6.3 and 6.5, there exists a constant ρ^* and a real-valued measurable function $h^*(x) \in \mathbf{B}^-_w(S)$ such that the average cost optimal inequality (ACOI) is satisfied, i.e.,*

$$\rho^* + h^*(x) \geq \min_{a\in A(x)}\left\{c(x,a) + \int_S h^*(y)Q(dy|x,a)\right\} \quad \forall\ x \in S$$

Moreover, there is a measurable selector f^* that attains the minimum in the right hand side of the ACOI, which is turn induces an optimal deterministic stationary policy f^* to Problem (6.1).

Proof. We define

$$h^*(x) := \varliminf_{n\to\infty} h_{\alpha_n}(x) = \lim_{n\to\infty}\inf_{k\geq n} h_{\alpha_k}(x) =: \lim_{n\to\infty} g_{\alpha_n}(x). \tag{6.13}$$

Note that, $g_{\alpha_n}(x) \leq h_{\alpha_n}(x)$ by definition and (g_{α_n}) form a nondecreasing sequence of real-valued measurable functions that converges to $h^*(x)$ in the pointwise sense. From Corollary 6.1, $-u_1(x) \leq h^*(x)$ (also, $g_{\alpha_n}(x)) \leq u_2(x),\ \forall x \in S,\ 0 < \alpha < 1$. Pass to the lower limit $n \to \infty$ on both sides of (6.12), we obtain,

$$\begin{aligned}
\rho^* + h^*(x) &= \varliminf_{n\to\infty} \min_{a\in A(x)} \left\{c(x,a) + \alpha_n \int_S h_{\alpha_n}(y)Q(dy|x,a)\right\} \\
&\geq \varliminf_{n\to\infty} \min_{a\in A(x)} \left\{c(x,a) + \alpha_n \int_S g_{\alpha_n}(y)Q(dy|x,a)\right\} \\
&= \lim_{n\to\infty} \min_{a\in A(x)} \left\{c(x,a) + \alpha_n \int_S g_{\alpha_n}(y)Q(dy|x,a))\right\} \\
&= \lim_{i\to\infty} \left\{c(x,a_{n_i}(x)) + \alpha_{n_i} \int_S g_{\alpha_{n_i}}(y)Q(dy|x,a_{n_i}(x))\right\} \\
&\geq c(x,a^*(x)) + \int_S h^*(y)Q(dy|x,a^*(x)) \\
&\geq \min_{a\in A(x)} \left\{c(x,a) + \int_S h^*(y)Q(dy|x,a)\right\}
\end{aligned}$$

The replacement of the lower limit with the ordinary limit in the third line is due to the monotonicity of $g_{\alpha_n}(\cdot)$. The existence of a subsequence $(a_{n_i}(x))$ that converges to $a^*(x)$ in the fourth line follows from Assumption 6.1(b). At last, the second to the last inequality follows by applying Assumption 6.1(a,c), Corollary 6.1 and the *extended Fatou's Lemma* (see [50, Lem.8.3.7(b)]).

By [47, App.D.6], there exists a measurable selector f^* attaining the infimum, which reads

$$\rho^* + h^*(x) \geq c(x,f^*) + \int_S h^*(y)Q(dy|x,f^*)$$

Iterating itself leads to

$$n\rho^* + h^*(x) \geq J^n(f^*,x) + \int_S h^*(y)Q^n(dy|x,f^*) \tag{6.14}$$

Indeed, by Assumption 6.2 and Corollary 6.1 we have

$$\begin{aligned}
& \varlimsup_{n\to\infty} \frac{1}{n} \int_S h^*(y) Q^n(dy|x, f^*) \\
\leq \quad & \varlimsup_{n\to\infty} \frac{1}{n} \int_S u_1(y) Q^n(dy|x, f^*) \\
\leq \quad & k \varlimsup_{n\to\infty} \frac{1}{n} \int_S w(y) Q^n(dy|x, f^*) = 0.
\end{aligned}$$

for some nonnegative constant k. Dividing by n and then passing to the upper limit $n \to \infty$ on both sides of (6.14) leads to $\rho^* \geq V(f^*, x) \geq \inf_\pi V(\pi, x)$, which together with (6.11) yields $V(f^*, x) = \inf_\pi V(\pi, x) = \rho^* < \infty$. That is, ρ^* is the optimal value of Problem (6.1) and f^* is the promised optimal deterministic stationary policy. □

6.5 An illustrative example

As is shown Section 6.3, a denumerable model is illustrated as an example. $S = \mathbb{N}_0 = 0, 1, 2, \ldots$ and $A = [0, +\infty)$ are state and action spaces. The admissible action space is defined as $A(i) = [0, i]$, $\forall\, i = 1, 2, \ldots$ and $A(0) = [a_0, b_0]$, where $0 < a_0 < b_0 < 1$, which is compact-valued for every $i \in S$. The evolution of the process is only controllable at state $0 \in S$, which is characterized as follows

$$\begin{aligned}
& Q(0|0, a) = a, \quad Q(i|0, a) = (1 - a)p_i \\
& Q(0|i, a) = 1 \quad \forall\, i = 1, 2, \ldots
\end{aligned}$$

where (p_i) is the prefixed probability mass function, and $a \in [a_0, b_0]$. The positive part of cost is free of control, whereas the negative part is controlled by a weight function.

Formally, we put

$$\begin{aligned}
& c^+(i) = k_i \\
& c^-(i, a) = m_i + B(a) \\
& W(i) = Cr^i \quad C > 0, r > 1
\end{aligned}$$

where (k_i) is a nondecreasing and sequence of non-negative real constants, m_i is an arbitrary sequence of nonnegative constants, and $B(a)$ is a nonnegative continuous function in $a \in A$, such that

$$\sup_{a \in A(i)} c^-(i,a) \leq \sqrt{W(i)} \tag{6.15}$$

A specific assumption is introduced for this example.

Assumption 6.6 *Both of the following two series,* $\sum_{i \neq 0} p_i r^i$ *and* $\sum_{i \neq 0} p_i k_i$*, converge.*

Within the present setting under the above assumption, negative part of the cost is polynomially bounded and the positive part, which is certain to depend on the explicit form of k_i, is allowed to be arbitrarily unbounded. Note that the generalization promised in Chapter 6 is achieved and justified by this setting. Next let us show that $W(i)$ satisfies Assumption 6.2(d),

$$\begin{aligned}
&\sum_{i \neq 0} W(i) Q(i|0,a) = aC + (1-a)C \sum_{i \neq 0} p_i r^i \leq \beta C + b \\
&W(0) = C \leq \beta C r^i + b \quad i = 1, 2, \ldots
\end{aligned}$$

By inspecting Assumption 6.6, it is not difficult to select $0 < \beta < 1$ and $b > 0$ so that both of these two inequalities are satisfied. In addition, both $w(\cdot) := \sqrt{W(\cdot)}$ and $w'(\cdot) := w(\cdot)$, in accordance with the definitions in Assumption 6.2 and 6.4, satisfy Assumption 6.2(c), with the same constants β and b as well as $r > 1$. Therefore, Assumption 6.4(a) is satisfied. Assumption 6.1 and the remainder of Assumption 6.2 are justified in a straightforward manner. Assumption 6.5 follows from the following reasoning. The model of positive part of the cost functions induces a *negative programming problem* with the discounted expected total cost criterion, the optimality equation is eligible for the determination of the optimal value function, i.e.,

$$V_0 = k_0 + \alpha a^* V_0 + \alpha(1-a^*)\sum_{i=1}^{\infty} p_i V_i$$
$$V_i = k_i + \alpha V_0 \quad i = 1, 2, \ldots$$

where $a^* = f^*(0)$ is the promised optimal deterministic stationary policy. We solve the above system of equations, and obtain

$$V_0 = \frac{k_0 + \alpha(1-a^*)\sum_{i=1}^{\infty} p_i k_i}{1 - \alpha a^* - \alpha^2(1-a^*)}$$

which is finite by Assumption 6.6. It is a direct observation that (V_i) is nondecreasing for $i \geq 1$ due to the monotonicity of (k_i). Therefore, V_i is finite for each state i and

$$\min_i V(i) = \min\{V_0, V_1\}.$$

Two possible cases are discussed respectively. If $V_0 < V_1$,

$$V_i - V_0 = (V_i - V_1) + (V_1 - V_0) = k_i - k_1 + (V_1 - V_0) \quad i = 2, 3, \ldots \tag{6.16}$$

otherwise,

$$V_i - V_1 = k_i - k_1 \quad i = 2, 3, \ldots.$$

Therefore, in either case we need only examine the boundedness of $V_1 - V_0$ as a function of the discount factor $\alpha \in (0, 1]$. Concretely,

$$\begin{aligned}
\overline{\lim_{\alpha\uparrow 1}}(V_1 - V_0) \quad &= \overline{\lim_{\alpha\uparrow 1}}(k_1 - (1-\alpha)V_0) \\
&= k_1 - \overline{\lim_{\alpha\uparrow 1}} \frac{(1-\alpha)k_0 + \alpha(1-\alpha)(1-a^*)\sum_{i=1}^{\infty} p_i k_i}{1 - \alpha a^* - \alpha^2(1-a^*)} \\
&= k_1 - \frac{1}{2-a^*}k_0 - \frac{1-a^*}{2-a^*}\sum_{i=1}^{\infty} p_i k_i > -\infty \; and < +\infty
\end{aligned}$$

where the last line follows from L'Hôpital's rule. Thus, Assumption 6.5 is justified.

As has been discussed in Chapter 6.3, Assumption 6.4(b,c) and Assumption 6.3(a,i) will be verified in the sequel. It is worth mentioning that for each $f \in U^{DS}$ (indeed, for each $\varphi \in U^S$), the corresponding Markov chain Q_f (respectively, Q_φ) is positive recurrent and aperiodic, or equivalently, ergodic. So, Assumption 6.4(b) is satisfied. For the sake of brevity, we shall consider only deterministic stationary policies U^{DS} below. As a consequence, there exists a unique *i.p.m.* $\mu_f(i)$ which is indeed a limiting distribution. This fact along with $W(\cdot)$ being a moment function satisfying (6.4) yields the sufficiency of stationary polices, or equivalently, Lemma 6.3(b). Let us evaluate the following expectation for any $f \in U^{DS}$ provided that $0 \in S$ is the initial state.

$$\begin{aligned} E_0^f[\sum_{t=1}^{\tau^0-1} c^-(x_t, f)] &\leq E_0^f[\sum_{t=0}^{\tau^0-1} W(x_t)] \\ &\leq (1-f(0))C\sum_{j=1}^{\infty} p_j r^j < \infty \end{aligned} \tag{6.17}$$

Incidentally, the same expectation with respect to the positive part of cost is computed,

$$E_0^f[\sum_{t=1}^{\tau^0-1} c^+(x_t)] \leq (1-f(0))\sum_{j=1}^{\infty} p_j k_j < \infty \tag{6.18}$$

again by Assumption 6.6. Thus, Assumption 6.3(a,ii) is justified be referring to [24, Thm.15.3]. In particular, one can derive the explicit form the *i.p.m.* for the Markov chain Q_f, where $f(0) = a$, taking the form

$$\begin{aligned} \mu_f(0) &= \frac{1}{2-a}; \\ \mu_f(i) &= \frac{1-a}{2-a} p_i \quad \forall i = 1, 2, \ldots \end{aligned}$$

where $a \in [a_0, b_0]$. In this case, Assumption 6.3(a,i) can be shown to be satisfied for every stationary policy by (6.17), (6.18) and again [24, Thm.15.3]. That is, we have the following,

$$\overline{\lim_{n\to\infty}} \sum_{j=1}^{\infty} c(j,f)\hat{\mu}_{i,f}^{n}(j) = \sum_{j=1}^{\infty} c(j,f)\mu_f(j) < \infty$$

Finally, Assumption 6.3(b) is verified by [50, Prop.10.2.5]. In particular,

$$\begin{aligned} &l_a(i) = 1 - a \\ &\nu(i) = p_i, \quad i = 0, 1, 2, \ldots \end{aligned}$$

Note that everything except p_0 comes from the primitives of our model. A simple computation yields,

$$\begin{cases} p_0 \leq \frac{a_0}{1-a_0} & \text{when} \quad a_0 \leq 1/2 \\ p_0 \leq 1 & \text{when} \quad a_0 > 1/2 \end{cases}$$

which justify all the definitions and conditions required in [50, Prop.10.2.5]. Therefore, the corresponding optimality result follows.

Chapter 7

An inventory-production system

7.1 Introduction

In this chapter we present an inventory-production system to illustrate main results in Chapter 2 and 5 (Theorem 2.2, 5.1 and 5.2). We follow the notations in Chapter 2 for consistency.

We consider a $\mathbf{S}-$valued controlled processes (x_t) of the form

$$x_{t+1} = F(x_t, a_t, z_t), \quad t = 0, 1, \ldots, \tag{7.1}$$

and always suppose the following assumption.

Assumption 7.1 *(a) The so-called disturbance sequence (z_t) are composed of independent and identically distributed (i.i.d.) random variables with values in a Borel Space Z, and (z_t) is independent of the initial distribution γ. The common distribution of z_t is denoted by G;*
(b) $F : \mathbf{K} \times Z \to \mathbf{S}$ is a given measurable function, where $\mathbf{K} \subset \mathbf{S} \times \mathbf{A}$ is the graph of admissible action space $A(\cdot)$ defined in Chapter 1.

Let $\pi \in U^H$ be an arbitrary control policy. By Assumption 7.1(a), the variables (x_t, a_t) and z_t are independent for each $t = 0, 1, \ldots$. Then the transition kernel Q is given by

$$
\begin{aligned}
Q(\Gamma_S|x,a) &= Prob(x_{t+1} \in \Gamma_S | x_t = x, a_t = a) \\
&= \int_Z \mathbf{1}_{\{F(x,a,z)\in\Gamma_S\}} G(dz) \qquad (7.2)
\end{aligned}
$$

for every $\Gamma_S \in \mathcal{B}(\mathbf{S}), (x,a) \in \mathbf{K}$, and $t = 0, 1, \ldots$. Moreover, for every measurable function $u \in \mathbf{B}(\mathbf{S})$, we have

$$
\begin{aligned}
u'(x,a) : &= \int_S u(dy) Q(dy|x,a) = E[u(x_{t+1})|x_t = x, a_t = a] \\
&= \int_Z u[F(x,a,z)] G(dz). \qquad (7.3)
\end{aligned}
$$

7.2 An inventory-production system

The state variable x_t, the control action a_t, and the disturbance z_t, for every $t = 0, 1, \ldots$, have the following practical meanings:

- x_t denotes the inventory level at the beginning of period t;
- a_t stands for the amount of products ordered (or produced immediately) at the beginning of period t;
- z_t represents the amount of sales during the period t.

Setting the conventional notation $c^+(x,a) := \max\{c(x,a), 0\}$, the inventory level is assumed to evolve in the following way

$$
x_{t+1} = (x_t + a_t - z_t)^+, \quad t = 0, 1, \ldots, \qquad (7.4)
$$

given the initial inventory level x_0. The state space is thereby $\mathbf{S} := [0, +\infty)$, while the production variable a_t takes value in a compact interval $\mathbf{A} := [0, \theta]$, for some given constant $\theta > 0$, irrespective of the present value, meaning that the admissible action space $A(x)$ satisfy $A(x) = \mathbf{A}$ for every $x \in \mathbf{S}$.

In addition, we suppose that the sales process (z_t) satisfies Assumption 7.1 with $Z := [0, +\infty)$, so that z_t is non-negative for each t, and possesses a common distribution G with the following properties:

- G has a continuous bounded density g, i.e., $G(dz) = g(z)dz$;
- G has finite mean value $\overline{z}$, i.e., $\overline{z} := E(z_0) = \int_0^\infty zG(dz) < \infty$.

To complete the description of the control model

$$\mathcal{M} := \{\mathbf{S}, \mathbf{A}, (A(x), x \in \mathbf{S}), Q, c_0, c_1, \gamma\},$$

where Q is given by (7.2) and (7.4), we shall consider a primary cost function c_0 that represents a *net cost* of the form

$$\textit{production cost} \ + \ \textit{maintenance (or holding) cost} \ - \ \textit{sales revenue}$$

given by

$$c_0(x, a) := p \cdot a + m \cdot (x + a) - s \cdot E[\min(x + a, z_0)] \tag{7.5}$$

and a secondary cost function c_1 that stands for a *pure cost* of the form

$$\textit{production cost} \ + \ \textit{maintenance cost}$$

given by

$$c_1(x, a) := p \cdot a + m \cdot (x + a) \tag{7.6}$$

where in both definitions, p, m, and s are positive constants. The unit production p and the unit maintenance cost m do not exceed the unit sale price, that is,

$$p, m \leq s. \tag{7.7}$$

7.2.1 The absorbing model

In this subsection, we present a discount model to illustrate Theorem 2.2 in Chapter 2. As is shown in [50, 91], there is a close relationship between a discount model and a transient model. Indeed, the former can be viewed and treated as a special case of the latter one. We state a simplified version of such a relation below, and one can get insight into more subtle discussion in the above references. We modify Assumption 2.1(c) so that the weight function $w(\cdot)$ meets Lyapunov-like condition which will be specified below.

Suppose we have a discounted model

$$\mathcal{M}_1 := \{\mathbf{S}, \mathbf{A}, (A(x), x \in \mathbf{S}), Q, c_0, c_1, \gamma\},$$

define the corresponding objective function as

$$J(\pi, x) := E_x^\pi \sum_{t=0}^{\infty} \alpha^t c(x_t, a_t)$$

We attempt to transform it into an absorbing model in the following manner. Firstly, we adjoin two isolated points, $\triangle_S$ and $\triangle_A$ to the original state and action space, $\mathbf{S}$ and $\mathbf{A}$, respectively. That is, we have two enlarged spaces that take the form

$$S := \mathbf{S} \cup \triangle_S, \;\; A := \mathbf{A} \cup \triangle_A$$

Accordingly, the admissible action space is modified as

$$\tilde{A}(x) := A(x) \quad \forall x \in \mathbf{S}, \quad \tilde{A}(\triangle_S) := \triangle_A$$

Consequently, the transition kernel is modified as

$$\begin{aligned} &\tilde{Q}(y|x, a) := \alpha Q(y|x, a) \quad \forall x \in \mathbf{S}, y \in \mathbf{S}; \\ &\tilde{Q}(\triangle_S|x, a) := 1 - \alpha \qquad\quad \forall x \in \mathbf{S}; \\ &\tilde{Q}(\triangle_S|\triangle_S, a) := 1, \end{aligned}$$

and the cost functions as

$$\tilde{c}_0 \text{ (and, } \tilde{c}_1)(\triangle_S, \triangle_A) := 0, \quad \tilde{c}_0 \text{ (and, } \tilde{c}_1)(x, a) := c(x, a) \quad \forall\, x \in \mathbf{S}.$$

In doing so, we obtain an enlarged model

$$\tilde{\mathcal{M}}_1 := \{S, A, (\tilde{A}(x), x \in S), \tilde{Q}, \tilde{c}_0, \tilde{c}_1, \gamma\}.$$

Let $\tilde{P}_x^{\tilde{\pi}}$ and $\tilde{E}_x^{\tilde{\pi}}$ be the induced probability measure and expectation operator for a given policy π in $\mathcal{M}$ and initial state $x \in \mathbf{S}$. Finally, let

$$\tilde{J}(\tilde{\pi}, x) := \tilde{E}_x^{\tilde{\pi}} \sum_{t=0}^{\infty} \tilde{c}(x_t, a_t). \tag{7.8}$$

This objective function is well defined under the condition that the following Lyapunov-like inequality (7.9) holds,

$$\sup_{a \in A(x)} \int_{\mathbf{S}} \tilde{w}(y) Q(dy|x, a) \leq \beta \tilde{w}(x) + b \quad \forall\, x \in \mathbf{S} \tag{7.9}$$

where $0 < \beta < 1/\alpha$, and b is a real constant, and $\tilde{w}(x) := w(x)$ for every $x \in \mathbf{S}$ and $\tilde{w}(\triangle_S) = 0$.

The thing is to show the equivalence between $\mathcal{M}_1$ and $\tilde{\mathcal{M}}_1$ in the sense that for each π with respect to $\mathcal{M}_1$ there is $\tilde{\pi}$ with respect to $\tilde{\mathcal{M}}_1$ such that

$$J(\pi, x) = \tilde{J}(\tilde{\pi}, x)$$

Clearly, the two policies characterize each other, which has been explained in Chapter 2. Consequently, $\tilde{\mathcal{M}}_1$ satisfies Definition 2.1.

With this observation in mind, we are ready to continue our discussion of the inventory-production problem by focusing on $\tilde{\mathcal{M}}_1$. Next, we verify Assumption 2.1, 2.2 and 2.4.

Firstly, we observe that

$$E[\min(x + a, z_0)] = (x + a)[1 - G(x + a)] + \int_0^{x+a} z G(dz), \tag{7.10}$$

which implies that both $c_0(x, a)$ and $c_1(x, a)$ are continuous in $(x, a) \in \mathbf{K}$

since $G(\cdot)$ has a continuous and bounded density function. Assumption 2.3 and 2.4(b) are validated. On the other hand,

$$\begin{aligned} u'(x,a) &= \int_0^\infty u[(x+a-z)^+]g(z)dz \\ &= u(0)[1-G(x+a)] + \int_0^{x+a} u(x+a-z)g(z)dz. \quad (7.11) \end{aligned}$$

Thus, an elementary change of variable in the latter integral yields

$$u'(x,a) = u(0)[1-G(x+a)] + \int_0^{x+a} u(z)g(x+a-z)dz \qquad (7.12)$$

and so we see that $u'(x,a)$ is continuous in $(x,a) \in \mathbf{K}$ for each measurable function $u \in \mathbf{B}(\mathbf{S})$. This implies that Assumption 2.2(b) is satisfied.

We move on to the determination of a proper weight function w. To this end, let us consider the moment generating function ψ of $\theta - z_0$,

$$\psi(r) := E[e^{r(\theta - z_0)}], \quad \forall\ r \geq 0. \qquad (7.13)$$

As $\psi(0) = 1$ and ψ is continuous, for each $\epsilon > 0$ there is a positive number $\hat{r}$ such that

$$\psi(\hat{r}) \leq 1 + \epsilon. \qquad (7.14)$$

Define

$$W(x) := k \cdot e^{\hat{r}(x+2\bar{z})}, \qquad x \in \mathbf{S}. \qquad (7.15)$$

Then, substituting u with W in (7.11) yields

$$W'(x,a) = W(0)[1-G(x+a)] + W(x)\int_0^{x+a} e^{\hat{r}(a-z)}G(dz), \qquad (7.16)$$

so that, since $1 - G(x+a) \leq 1$ and $\hat{r}(a-z) \leq \hat{r}(\theta - z)$ for all $a \in \mathbf{A}$, we obtain

$$W'(x,a) \leq W(0) + \psi(\hat{r})W(x) \leq \beta W(x) + b \quad \forall x \in \mathbf{S}, \qquad (7.17)$$

with

$$\beta := 1 + \epsilon \quad \textit{and} \quad b := W(0) \tag{7.18}$$

On the other hand, a direct computation using (7.7) and (7.10) shows that $\sup_{a \in \mathbf{A}} |c_0(x,a)| \le s(x + l_0 \bar{z})$, and $\sup_{a \in \mathbf{A}} |c_1(x,a)| \le s(x + l_1 \bar{z})$, for every $x \in \mathbf{S}$, where l_0 and l_1 are some constants. We can always find such constants as they depend merely on θ and $\bar{z}$. Define

$$w(x) = w'(x) = \sqrt{W(x)} = \sqrt{k} \cdot e^{\frac{\bar{r}}{2}(x + 2\hat{z})}, \tag{7.19}$$

therefore,

$$\sup_{a \in \mathbf{A}} |c_0(x,a)| \le k_0 \cdot w(x) \tag{7.20}$$

$$\sup_{a \in \mathbf{A}} |c_1(x,a)| \le k_1 \cdot w(x) \tag{7.21}$$

for some constants k_0 and k_1 sufficiently large. Thus, $w(\cdot)$ is continuous in $x \in \mathbf{S}$, and Assumption 2.1(a) and 2.2(a) are satisfied. The initial distribution $\gamma(\cdot)$ is defined to be a probability measure on $\mathbf{S}$ such that $\int W(y)\gamma(dy) < \infty$ which naturally validates Assumption 2.1(b). Consider a stationary policy $f_0(x) \equiv 0$, $\forall\ x \in S$, so that

$$\begin{aligned} J_1(f_0, \gamma) &= E_\gamma^{f_0}[\sum_{t=0}^{\infty} \alpha^t c_1(x_t, a_t)] \\ &\le \int_{\mathbf{S}} E_x^{f_0}[\sum_{t=0}^{\infty} k_1 \alpha^t w(x_t)]\gamma(dx) \\ &= \frac{k_1}{1-\alpha}\left[\int_{\mathbf{S}} w(x)\gamma(dx) + \frac{b}{1-\beta}\right] < \infty \end{aligned}$$

Assumption 2.4(a) is satisfied if $d_1 > V_1(f_0, \gamma)$. We refer the readers to Remark 2.1(b) for a sufficient condition for Assumption 2.2(c). In particular, (i) and (ii) of Remark 2.1(b) are trivial. For part (iii), it is the case that $W(x) = w(x)w'(x)$ has been verified by (7.17) whereas $w(x)$

itself is not. By the same reasoning in the treatment of $W(x)$ we obtain

$$\begin{aligned} w'(x,a) &= w(0)[1-G(x+a)] + w(x)\int_0^{x+a} w^{\frac{\hat{r}}{2}(a-z)}G(dz) \\ &\leq w(0) + w(x)\sqrt{\psi(\hat{r})} \\ &\leq w(0) + w(x)\psi(\hat{r}) \end{aligned}$$

The second line follows from Jensen's inequality and that $\sqrt{\cdot}$ is a concave function. Part (iii) immediately follows from the fact that $\mathcal{P}_w(\mathbf{S}) \subseteq \mathcal{P}_W(\mathbf{S})$ and the same argument when $W(\cdot)$ is replaced by $w(\cdot)$. Incidentally, Assumption 2.1(c) is shown to be satisfied. Note that ϵ is allowed to be chosen arbitrarily close to 0, which indicates that we are always able to find a β that corresponds to any discount factor α such that $\beta < 1/a$.

7.2.2 The average model

This subsection is devoted to illustrating Theorem 5.1 and Theorem 5.2. The model $\mathcal{M}$ with Lyapunov-like condition (7.9) is maintained, and in addition, $\bar{z} := E(z_0)$, $\theta < \bar{z}$. The extra assumption states that the expected demand $\bar{z}$ should exceed the maximum allowed production. This is slightly more restrictive than what is imposed in the previous example, where no relation between $\bar{z}$ and θ is specified. In contrast to the constant β in (7.9) which takes the value $1+\epsilon$ greater than 1, we will show that β can possibly be smaller than 1 within our new setting.

Let $\psi(r) := Ee^{r(\theta - z_0)}$, $r \geq 0$, be the moment generating function of $\theta - z_0$. Due to that $\psi(0) = 1$ and $\psi'(0) = E(\theta - z_0) = \theta - \bar{z} < 0$, there is a positive number r_* such that $\psi(r_*) < 1$. Therefore, the newly-defined weight function takes the form,

$$W(x) := e^{r_*(x+2\bar{z})}, \quad x \in \mathbf{S}, \tag{7.22}$$

we see that $w^2(\cdot) := W(\cdot)$ satisfies (7.16) and (7.17) with $\hat{r}$ being replaced by r_*. In particular, (7.17) becomes

$$W'(x,a) \leq \beta W(x) + b \quad x \in \mathbf{S},\ a \in \mathbf{A} \tag{7.23}$$

with

$$\beta := \psi(r_*) < 1, and \quad b := W(0) \tag{7.24}$$

The verification of the set of assumptions are quite similar to what has been done in the previous case. We pick out those who are different from, or additional to the previous and provide a simple discussion. For Assumption 5.2(b), the moment function takes the following form,

$$v(x,a) := \sqrt{W(x)}, \quad \forall x \in \mathbf{S},\ a \in \mathbf{A}.$$

Assumption 5.2(d) follows from the aforementioned argument which states that $\int_S u(y)Q(dy|x,a)$ is continuous in $(x,a) \in \mathbf{K}$ for every $u \in \mathbf{C}(\mathbf{S})$. This leads to the conclusion that both Assumption 5.2(d,e) are met. Assumption 5.3 and 5.4 are verified by the same argument as in [50, Exp 10.9.3] with $f \in U^{DS}$ being replaced with or extended to $\varphi \in U^S$. The resulting process is a homogeneous Markov chain by Proposition 1.1. Briefly, the process $Q_\varphi(dy|x)$ is positive Harris recurrent for every stationary policy φ. Consequently, the above two assumptions follow from individual ergodic theorem (see [52, Thm.2.3.4]). Assumption 5.5 follows exactly the same as in [50, Exp 10.9.3], which puts an end to our work. To conclude, all assumptions in Chapter 5 are satisfied so that Theorem 5.1 and 5.2 hold.

Chapter 8

Conclusion

In this chapter we briefly summarize the material presented in this dissertation.

Chapter 2 tackles the constrained absorbing MDP model in Borel spaces with possibly unbounded (from both the above and below) cost functions, and the constrained discounted MDP model with a state-action-dependent discount factor not necessarily separated from one. The latter problem is addressed as a specific example of the former, along with the observation of an equivalence between a discounted model and an absorbing one. Incidentally, we present some topological properties of occupation measures in a proper topology. In addition, duality results are derived by the *linear programming approach*.

In Chapter 3 we propose a similar problem, but in a more constrained context than that in Chapter 2. To the best of our knowledge, such a problem has not been studied before, partly because of the lack of related techniques. Nevertheless, it could be resolved by a variant of *dynamic programming approach* similarly to [65]. The original model is reformulated into an unconstrained one, and corresponding optimality results are derived with the same method obtained in Chapter 4 which deals with standard MDP models. In addition, the correspondence between policies in two models is shown. It should be emphasized that only optimal randomized Markov policies are justified, as distinct from the sufficiency of randomized stationary policies for the traditional constrained problems, which is deemed to be an open problem.

In Chapter 4 we are concerned with a risk-sensitive MDP, i.e., attempting to incorporate the notion of risk measures into the standard MDP model. The risk aggregating method applied in the present work is *iterated coherent risk measure*. Again, we allow our cost functions to be defined in a fairly relaxed way. To be specific, the positive part is arbitrarily unbounded, whereas the negative part is controlled by a weight function. In this chapter, optimality results for either the finite or infinite horizon are obtained using *dynamic programming approach* under an extension of Berge's Theorem.

Chapter 5 and 6 concern the average MDP problem with unbounded cost functions. In Chapter 5, we study a constrained MDP model formulated similarly to what is considered in Chapter 2. The main difference resides in that the process of our interest here exhibit some ergodic behaviours, as compared with an absorbing model in the former case. Under mild conditions, we establish the existence of optimal mixing policies and characterize the extreme points of the space of performance vectors, by considering a sufficient class of stable policies. In Chapter 6, we turn back to a unconstrained model allowing cost functions to be unbounded below, and letting its positive part be arbitrarily unbounded. A discussion of sufficiency in the denumerable case is provided. We establish the average cost optimality inequality (ACOI) and the optimal deterministic stationary policy as well. The extension of Chapter 6.3 to the general Borel state space remains an open problem, as the definitions of irreducibility and ergodicity are quite different from those in the denumerable case.

Appendix A

Semicontinuous functions, set-valued mappings and measurable selectors

Let S and A be two Borel spaces. A set-valued mapping (also known as a multifunction, or a correspondence) $A(\cdot)$ from S to A is a function such that $A(x)$ is a nonempty subset of A for each $x \in S$. The graph of the set-valued mapping $A(\cdot)$ is the subset of $S \times A$ defined as

$$\mathbb{K} := \{(x,a) \in S \times A : x \in S, a \in A(x)\}.$$

Definition A.1 *(a) A measurable function g on S is lower semicontinuous, if for each $x \in S$, and each sequence $S \ni x_n \to x \in S$,*

$$\varliminf_{n\to\infty} g(x_n) \geq g(x);$$

(b) A measurable function g on $\mathbb{K}$ is $\mathbb{K}$-inf-compact, if g is lower semicontinuous on $\mathbb{K}$, and for each $S \ni x_n \to x \in S$ and $a_n \in A(x_n)$ such that $c(x_n, a_n)$ is bounded from the above with respect to n, the sequence (a_n) admits a limit point $a \in A(x)$.

The following proposition is a standing one, and can be viewed as a way of characterizing lower semicontinuous functions, also known as Baire functions, and consequently, is often referred to as Baire's theorem.

Proposition A.1 *A measurable function g on S is lower semicontinuous and bounded below, if and only if there exists a sequence of continuous and bounded functions $g_n \in \mathbf{C}(S)$ such that $g_n \uparrow g$ in the pointwise sense.*

Proof. See [12, §7.5]. □

Corollary A.1 *Given a continuous weight function $w \geq 1$ on S, a measurable function g on S is lower semicontinuous and bounded from the below in the w-norm, if and only if there exists a sequence of continuous and w-bounded functions $g_n \in \mathbf{C}_w(S)$ such that $g_n \uparrow g$ in the pointwise sense.*

The following result is a natural consequence of Proposition A.1 and Corollary A.1, thus whose proof is omitted.

Corollary A.2 *(a) Suppose a sequence of finite measures $\mu_n \in \mathcal{M}(S)$ converges to some $\mu \in \mathcal{M}(S)$ in the usual weak topology, denoted by $\mu_n \to \mu$, and g is a lower semicontinuous function on S and bounded from the below, then*

$$\varliminf_{n\to\infty} \int_S g(x)\mu_n(dx) \geq \int_S g(x)\mu_n(dx);$$

(b) Given a continuous weight function $wge1$ on S, suppose a sequence of finite measures $\mu_n \in \mathcal{M}_w(S)$ converges to some $\mu \in \mathcal{M}_w(S)$ in the w-weak topology, denoted by $\mu_n \overset{w}{\to} \mu$, and g is a lower semicontinuous function on S and bounded from the below in the w-norm, then

$$\varliminf_{n\to\infty} \int_S g(x)\mu_n(dx) \geq \int_S g(x)\mu_n(dx).$$

Lemma A.1 *Let $w \geq 1$ be a continuous weight function on S, and $Q(dy|x,a)$ be a stochastic kernel on $\mathcal{B}(S) \times \mathbb{K}$, then the following two statements are equivalent:*
(a) $\int_S g(y)Q(dy|x,a)$ is lower semicontinuous in $(x,a) \in \mathbb{K}$ for each lower semicontinuous function g on S, which is bounded from the below in the w-norm.
(b) $\int_S w(y)Q(dy|x,a)$ is continuous in $(x,a) \in \mathbb{K}$, and $\int_S g(y)Q(dy|x,a)$ is continuous in $(x,a) \in \mathbb{K}$ and bounded for each $g \in \mathbf{C}(S)$.

Proof. See [8, Lem.2.4.7]. □

Definition A.2 *(a) A set-valued mapping $A(\cdot)$ from S to A is called measurable, if $A^{-1}(\Gamma_A)$ is a Borel subset of S for every open set $\Gamma_A \subseteq A$;*
(b) A set-valued mapping $A(\cdot)$ from S to A is called upper semicontinous, if for each $S \in x_n \to x \in S$ and $a_n \in A(x_n)$, the sequence (a_n) admits a limit point in $A(x)$.

Proposition A.2 *Let $A(\cdot)$ be a compact-valued set-valued mapping from S to A. The following two assertions are equivalent:*
(a) $A(\cdot)$ is measurable;
(b) $\mathbb{K}$, the graph of the set-valued mapping $A(\cdot)$, is a Borel subset of $S \times A$.

Proof. See [55] and [85]. □

Given a set-valued mapping $A(\cdot)$ from S to A, a measurable function $f : S \to A$ such that $f(x) \in A(x)$ for each $x \in S$ is called a measurable selector for the set-valued mapping $A(\cdot)$. Moreover, $g : \mathbb{K} \to \mathbb{R}$ is a given measurable function and

$$g^*(x) := \inf_{a \in A(x)} g(x, a), \quad \forall\, x \in S.$$

If $g(x, \cdot)$ attains its minimum at some point in $A(x)$, we use "min" instead of "inf".

The following lemma is a version of measurable selection theorem, which is vital to justify the existence of an optimal deterministic stationary policy.

Lemma A.2 *Suppose that the given set-valued mapping $A(\cdot)$ is measurable and compact-valued. The following two assertions hold.*
(a) If $g(x, \cdot)$ is lower semicontinuous on $A(\cdot)$ for each $x \in S$, then there exists a measurable selector f^* such that

$$g(x, f^*(x)) = g^*(x) = \min_{a \in A(x)} g(x, a) \quad \forall\, x \in S, \tag{A.1}$$

and g^* is measurable on S;

(b) If $A(\cdot)$ is upper semicontinuous and $g(\cdot,\cdot)$ is lower semicontinuous and bounded from the below on $\mathbb{K}$, then there exists f^* for which (A.1) *holds, and g^* is lower semicontinuous and bounded from the below on S.*

Proof. See [55] and [85]. □

Lemma A.3 *For any $\mathbb{K}$-inf-compact (extended real-valued) function g on $\mathbb{K}$, there is a measurable mapping f^* from S to A satisfying $f^*(x) \in A(x)$ and*

$$g(x, f^*(x)) = \inf_{a \in A(x)} g(x,a).$$

Moreover, $\inf_{a \in A(x)} g(x,a)$ is lower semicontinuous on S, and

$$A_g^*(x) := \{b \in A(x) : g(x,b) = \inf_{a \in A(x)} g(x,a)\}$$

is compact if $\inf_{a \in A(x)} g(x,a) < \infty$, and $A_g^*(x) = A(x)$ if $\inf_{a \in A(x)} g(x,a) = \infty$.

Proof. See [37]. □

Appendix B

Prohorov's theorem

The following well-known definitions come from [1, 47].

Definition B.1 *A nonnegative measurable function $v(x)$ on a Borel space S is called a moment (or strictly unbounded function) if there exists a nondecreasing sequence of compact subsets $S_n \uparrow S, n = 1, 2, \dots$ such that $\lim_{n\to\infty} \inf_{x\in S_n^C} v(x) = \infty$. Here we adopt the convention that the infimum taken over the empty set is ∞.*

Definition B.2 *(a) A family $\mathcal{G}$ of finite measures on a Borel space S is called tight if $\forall\ \epsilon > 0$, there exists a compact subset $S_\epsilon \subseteq S : \forall\ \mu \in \mathcal{G}, \mu(S_\epsilon^C) < \epsilon$.*
(b) A set S_0 is called relatively compact (or, precompact) in a Borel space S, where $S_0 \subseteq S$, if for any sequence (b_n), $b_n \in S_0$, $n = 1, 2, \dots,$ there exists a subsequence (b_{n_k}) such that b_{n_k} converges to some point $b^* \in S$ in a proper topology. If $b^* \in S_0$, S_0 is called compact in the prescribed topology.

The following result is celebrated Prohorov's theorem, although the original statement concerns the space of probability measures $\mathcal{P}(S)$.

Theorem B.1 *Let $\mathcal{G}$ be a family of finite measures on a Borel space S. The following assertion holds.*
(a) If $\mathcal{G}$ is tight, then it is relatively compact in $\mathcal{M}(S)$ in the usual weak topology;

(b) Suppose that S is separable and complete. If $\mathcal{G}$ is relatively compact in $\mathcal{M}(S)$ in the usual topology, then it is tight.

Proof. See [16, Thm.8.6.2]. □

Appendix C

Miscellaneous

The following result is the well-celebrated Tauberian (Abelian) theorem.

Theorem C.1 *Let* $(u_t)_{t=0,1,2,...}$ *is a sequence of nonnegative real numbers. Then*

$$\varliminf_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} u_t \leq \varliminf_{\alpha\uparrow 1}(1-\alpha)\sum_{t=0}^{n-1}\alpha^t u_t \leq \varlimsup_{\alpha\uparrow 1}(1-\alpha)\sum_{t=0}^{n-1}\alpha^t u_t \leq \varlimsup_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} u_t$$

Proof. See [89, ThmA.4.2]. □

The following two results are referred as Krein-Milman theorem and Carathéodory convexity theorem respectively in various literature.

Theorem C.2 *Let* S *be a compact and convex subset of the* n*-dimensional Euclidean space* $\mathbb{R}^n$*, where* n *is a natural number. Then,* S *is equal to the convex hull of its extreme points.*

Proof. See [13, Prop.3.3.1(c)], or [1, Thm.7.68]. □

Theorem C.3 *Let* S *be a nonempty subset of the* n*-dimensional Euclidean space* $\mathbb{R}^n$*, where* n *is a natural number. Then every vector in the convex hull of* S *can be represented by a convex combination of no more than* $n+1$ *vectors in* S*.*

Proof. See [13, Prop.1.3.1(b)], or [1, Thm.5.32]. □

The following lemma justifies the convergence of a sequence of weakly monotone functions.

Lemma C.1 *Suppose $v_n(\cdot) \geq v_m(\cdot) - \sigma_m(\cdot)$ for each $n \geq m$, where v_k are (extended real-valued) measurable function on a Borel space S, and the nonnegative real-valued functions σ_m on S satisfy $\sigma_m(x) \to 0$ as $m \to \infty$, then $\lim_{n\to\infty} v_n$ exists. If v_k are lower semicontinuous, and σ_k are upper semicontinuous, then $\lim_{n\to\infty} v_n$ is also lower semicontinuous.*

Proof. See the proof of Lemma A.1.4(c) of [8], which applies to the case of extended real-valued functions. See also Proposition 10.1 of [85]. □

Obviously, $\sigma_m \equiv 0$ satisfies the conditions needed in Lemma C.1, which reduces to the conventional monotone convergence theorem.

Bibliography

[1] Aliprantis, C. and Border, K. *Infinite-dimensional analysis.* Springer, New York, 2007.

[2] Altman, E. *Constrained Markov decision processes.* Chapman and Hall/CRC, Boca Raton, 1999.

[3] Altman, E. and Shwartz, A.: Markov decision problems and state-action frequencies. *SIAM J. Control. and Optim.*, **29(4)** (1991) 786-809.

[4] Arapostathis, A. et al.: Discrete-time controlled Markov processes with average cost criterion: a survey. *SIAM J. Control. and Optim.*, **31** (1993) 282-344.

[5] Artzner, P., Delbaen, F., Eber, J. & Heath, D.: Coherent measures of risk. *Mathe. Finance*, **9** (1999) 203-228.

[6] Bather, J.: Optimal decision procedures for finite Markov chains. Part I: Examples. *Adv. in Appl. Probab.* **5(2)** (1973) 328-339.

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

[8] Bäuerle, N. and Rieder, U. *Markov Decision Processes with Applications to Finance.* Springer-Verlag, Berlin, 2011.

[9] Bäuerle, N. and Rieder, U.: More risk-sensitive Markov decision processes. *Mathematical Operations Research*, **39** (2014) 105-120.

[10] Ben-Tal, A. & Teboulle, M.: An old-new concept of convex risk measures: the optimized certainty equivalent. *Math. Finance*, **17** (2007) 449-476.

[11] Berge, C. *Topological Spaces*, Macmillan, New York, 1963.

[12] Bertsekas, D. and Shreve, S. *Stochastic optimal control.* Academic Press, NY, 1978.

[13] Bertsekas, D., Nedic, A. and Ozdaglar, A. E. *Convex Analysis and Optimization.* Athena Scientific, Belmont, Massachusetts, 2003.

[14] Blackwell, D.: Discrete dynamic programming. *Ann. Math. Statist.*, **33(2)** (1962) 719-726.

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

[16] Bogachev, V. I. *Measure Theory (Volume II)*, Springer-Verlag, Berlin, 2007

[17] Borkar, V.: Control of Markov chains with long-run average cost criterion. *in Stochastic Defferential Systems, Stochastic Control Theory and Applications (W. Fleming and P. L. Lions, eds.), The IMA Volumes in Mathematics and Its Application*, **10**, Springer-Verlag, Berlin (1988) 57-77.

[18] Borkar, V.: A convex analytic approach to Markov decision processes *Probab. Th. Rel. Fields.*, **78** (1988) 583-602.

[19] Borkar, V.: Control of Markov chains with long-run average cost criterion: the dynamic programming equations. *SIAM. J. Control. Optimization* (1989) 642-657.

[20] Borkar, V: Ergodic control of Markov chains with constraints–the general case. *SIAM J. Control and Optimization* (1994) 176-186.

[21] Cavazos-Cadena, R.: A counterexample on the optimality equation in Markov decision chains with the average cost criterion. *Systems and Control Letters.* **16(5)** (1991) 387-392.

[22] Cavazos-Cadena, R. and Montes-de-Oca, R.: Optimal stationary policies in risk sensitive dynamic programs with finite state space and nonnegative rewards. *Applied Mathematics(Warsaw)*, **27** (2000) 153-165.

[23] Chu, S. and Zhang, Y.: Markov decision processes with iterated coherent risk measures. *International Journal of Control.* **87** (2014) 2286-2293.

[24] Chung, K. L. *Markov Chains with Stationary Transition Probabilities.* Springer-Verlag, Berlin, 1967.

[25] Costa, O. L. V. and Dufour F.: Average control of Markov decision processes with Feller transition probabilities and general action spaces. *J. Math. Anal. Appl.* **396** (2012) 58-69.

[26] Denardo, E., Feinberg, E. and Rothblum, U.: Splitting randomized stationary policies in Total-reward Markov decision processes. *Mathematics of Operations Research.* **37(1)** (2012) 129-153.

[27] Derman, C.: On sequential decisions and Markov chains. *Management Sci.* **9(1)** (1962) 16-24.

[28] Derman, C.: Denumerable state Markov decision processes. *Ann. Math.* **37** (1966) 1545-1553.

[29] Derman, C. and Strauch, R. E.: A note on memoryless rules for controlling sequential control processes. *Ann. Math. Statist.* **37** (1966) 276-278.

[30] Dubins, L. E. On Extreme Points of Convex Sets. *Journal of Mathe. Ana. and Appli.* **5** (1962) 237-244.

[31] Dynkin, E. and Yushkevich, A. *Controlled Markov processes.* Springer-Verlag, NY, 1979.

[32] Feinberg, E.: An ϵ-optimal control of a finite Markov chain. *Theor. Probability Appl.* **25(1)** (1980) 70-81.

[33] Feinberg, E. and Shwartz, A.: Constrained Markov decision models with weighted discounted rewards. *Mathematics of Operations Research.* **20(2)** (1995) 302-320.

[34] Feinberg, E. and Shwartz, A.: Constrained discounted dynamic programming. *Math. Oper. Res.* **21** (1996) 922-945.

[35] Feinberg, E. and Lewis, M. E.: Optimality inequalities for average cost Markov decision processes and the stochastic cash balance problem. *Math. Oper. Res.* **32(4)** (2007) 769-783.

[36] Feinberg, E., Kasyanov, P. and Zadoianchuk, N.: Average cost Markov decision processes with weakly continuous transition probabilities. *Math. Oper. Res.* **37(4)** (2012) 591-607.

[37] Feinberg, E., Kasyanov, P. and Zadoianchuk, N.: Berge's theorem for noncompact image sets. *J. Math. Anal. Appl.*, **397** (2013) 255-259.

[38] Feinberg, E., Kasyanov, P. and Zadoianchuk, N.: Fatou's lemma for weakly convergent probabilities. Preprint (2013b) available at arxiv:1206.4073v2.

[39] Frid, E.: On optimal strategies in control problems with constraints. *Theory. Probab. Appl.* **17** (1972) 188-192.

[40] Gemignani, M. *Elementary topology.* Dover, NY, 1990.

[41] González-Hernández and Hernández-Lerma, O.: Extreme points of sets of randomized strategies in constrained optimization and control problems. *SIAM J. Optim.* **15** (2005) 1085-1104.

[42] González-Hernández, J., López-Martińez, R. and Pérez-Hernández, J.: Markov control processes with randomized discounted cost. *Math. Meth. Oper. Res.* **65** (2007) 27-44.

[43] González-Hernández, J. and Villarreal, E.: Optimal policies for constrained average-cost Markov decision processes. *Top.* **19** (2011) 107-120.

[44] Guo, X. and Zhu, Q. Average Optimality for Markov Decision Processes in Borel Spaces: A New Condition and Approach. *Journal of Applied Probability.* **43** (2006) 318-334.

[45] Hardy, M. & Wirch, J.: The iterated CTE: a dynamic risk measure. *North Amer. Act. Journal.* **8** (2004) 62-75.

[46] Hernández-Lerma, O.: Average optimality in dynamic programming on Borel spaces-Unbounded costs and controls. *Systems and Control Lett.* **17(3)** (1991) 237-242.

[47] Hernández-Lerma, O. and Lasserre, J. *Discrete-time Markov control processes.* Springer-Verlag, NY, 1996.

[48] Hernández-Lerma, O. and González-Hernández, J.: Infinite linear programming and multichain Markov control processes in uncountable spaces. *SIAM J. Control. Optim.* **36(1)** (1998) 313-335.

[49] Hernández-Lerma, O. and Vega-Amaya, O.: Infinite-horizon Markov control processes with undiscounted cost criteria: from average to overtaking optimality. *Applicationes Mathematicae* **25** (1998) 153-178.

[50] Hernández-Lerma, O. and Lasserre, J. *Further topics on discrete-time Markov control processes.* Springer-Verlag, NY, 1999.

[51] Hernández-Lerma, O., Carrasco, G. & Pérez-Hernández, R.: Markov control processes with the expected total cost criterion: optimality, stability, and transient models. *Acta Appl. Math.*, **59** (1999) 229-269.

[52] Hernández-Lerma, O. and Lasserre, J. *Markov chains and invariant probabilities* Birkhäuser, Berlin, 2003.

[53] Hernández-Lerma, O., González-Hernández, J. and López-Martínez, R. Constrained Average Cost Markov Control Processes in Borel Spaces. *SIAM J. Control. Optim.* **42** (2003) 442-468.

[54] Himmelberg, C. J. Measurable relations. *Fundamenta Mathematicae* **87(1)** (1975) 53-72.

[55] Himmelberg, C. J., Parthasarathy, T., and Van Vleck, F. S. Optimal plans for dynamic programming problems. *Math. Oper. Res.*, **1** (1976) 390-394.

[56] Hinderer, K.: *Foundations of non-Stationary dynamic programming with discrete time parameter*, Berlin, Springer, 1970.

[57] Hordijk, A. *Dynamic Programming and Markov Potential Theory.* Math. Centre Tract, **51**, Mathematisch Centrum, Amsterdam, 1974.

[58] Hordijk, A. and Spieksma, F.: On ergodicity and recurrence properties of a Markov chain with an application to an open Jackson Network. *Advances in Applied Probability* **24** (1992) 343-376.

[59] Hu, S. and Papageorgiou, N.S. *Handbook of Multivalued Analysis. Volume I: Theory.* Kluwer, Dordrecht, 1997.

[60] James, H. and Collins, E.: An analysis of transient Markov decision processes. *J. Appl. Probab.* **43** (2006) 603-621.

[61] Jaśkiewicz, A.: A note on negative dynamic programming for risk-sensitive control. *Operations Research Letters.* **36** (2008) 531-534.

[62] Jaśkiewicz, A. and Nowak, A.: Discounted dynamic programming with unbounded returns: application to economic models. *J. Math. Anal. Appl.*, **378** (2011) 450-562.

[63] Jaśkiewicz, A. & Nowak, A.: Stochastic games with unbounded payoffs: applications to robust control in economics. *Dyn. Games Appl.* **1** (2011) 253-279.

[64] Lang, S. *Linear Algebra.* Springer, USA, 2000.

[65] Mao, X. and Piunovskiy, A.: Constrained Markovian decision processes: the dynamic programming approach. *Stoch. Anal. Appl.* **18** (2000) 755-776.

[66] Montes-de-Oca, R. and Lemus-Rodríguez, E.: An unbounded Berge's minimum theorem with applications to discounted Markov decision processes. *Kybernetika (Prague)*, **48** (2012) 268–286.

[67] Meyer, P. *Probability and Potentials.* Blaisdell Publishing Company, Waltham, 1966.

[68] Meyn, S.P. and Tweedie, R.L. *Markov Chains and Stochastic Stability* Springer-Verlag, London, 1993.

[69] Osogami, T. *Overcoming limitations of expected utility with iterated risk measures* (Tech. Rep. No. RT0921). (2010) Tokyo, Japan: IBM Research-Tokyo.

[70] Osogami, T.: Iterated risk measures for risk-sensitive Markov decision processes with discounted cost. In the proceedings of *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, Barcelona, Spain, July 2011 (pp. 567–574).

[71] Pflug, C.: Some remarks on the value-at-risk and the conditional value-at-risk In S. Urysaev (Ed.), *Probabilistic constrained optimization: Methodology and applications*(pp. 272-281). (2000) New York, NY: Springer.

[72] Piunovskiy, A. *Optimal control of random sequences in problems with constraints.* Kluwer, Dordrecht, 1997.

[73] Piunovskiy, A. and Zhang, Y.: Discounted continuous-time Markov decision processes with unbounded rates: the convex analytic approach. *SIAM J. Control. Optim.* **49(5)** (2011) 2032-2061.

[74] Prokhorov, Yu.: Convergence of random processes and limit theorems in probability theory. *Theory. Probab. Appl.* **1** (1956) 157-214.

[75] Puterman, M. *Markov decision processes: discrete stochastic dynamic programming.* Wiley, NY, 1994.

[76] Rockafellar, R. *Conjugate duality and optimization.* SIAM, Philadelphia, 1974.

[77] Rockafellar, R.T. & Uryasev, S.: Optimization of conditional value-at-risk, *J. Risk*, **2** (2000) 493-517.

[78] Rockafellar, R.T. & Urysaev, S.: Conditional value-at-risk for general loss distributions. *J. Bank. Financ.*, **26** (2002) 1443-1471.

[79] Ross, S. M., On the nonexistence of ϵ-optimal randomized stationary policies in average cost Markov decision models. *Ann. Math. Statist.* **42(5)** (1971) 1767-1768.

[80] Ross, S. M., *Introduction to stochastic dynamic programming.* Academic Press, NY, 1983.

[81] Royden, H. *Real Analysis.* New Jersey, Prentice-Hall, 1988.

[82] Ruszczyński, A. Risk-averse dynamic programming for Markov decision processes. *Math. Program.(B)*, **125** (2010) 235-261.

[83] Sarykalin, S., Serraino, G. & Urysaev, S.: Value-at-Risk vs. conditional Value-at-Risk in risk management and optimization. *Tutorials in Operations Research*, (2008) 270-294, doi 10.1287/educ.1080.0052.

[84] Schäl, M.: A selection theorem for optimization problems. *Arch. Math.* **25** (1974) 219-224.

[85] Schäl, M.: Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. *Z. Wahrscheinlichkeitstheorie verw. Gebiete.* **32** (1975) 179-196

[86] Schäl, M.: Average optimality in dynamic programming with general state space. *Mathematics of Operations Research* **18(1)** (1993) 163-172.

[87] Sennott, L.: A new condition for the existence of optimal stationary policies in average cost Markov decision processes. *Oper. Res. Lett.* **5(1)** (1986) 17-23.

[88] Sennott, L.: Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs. *Oper. Res.* **37** (1989) 626-633.

[89] Sennott, L. *Stochastic dynamic programming and the control of queueing systems*, NY, John Wiley and Sons, 1999.

[90] Shen, Y., Stannat, W. & Obermayer, K. (2013). Risk-sensitive Markov control processes. *SIAM J. Control Optim.*, **51** (2013) 3652-3672.

[91] Shwartz, A.: Death and discounting. *IEEE Trans. Automat. Contr.* **46** (2001) 644-647.

[92] Spieksma, F.: Ave, lyapunov functions! *Mathematical Methods in Operational Research.* Special issue in honour of Arie Hordijk. (2005) http://www.math.vu.nl/ koole/articles/mmor05/

[93] Stoer, J. & Witzgall, C.: *Convexity and Optimization in Finite Dimensions I*, Berlin, Springer-Verlag, 1970.

[94] Strauch, R: Negative dynamic programming. *Ann. Math. Stat.* **37** (1966) 871-890.

[95] Varadarajan, V.: Weak convergence of measures on separable metric spaces. *Sankhyā* **19** (1958) 15-22.

[96] Vega-Amaya, O.: The average cost optimality equation: A fixed point approach. *Bol. Soc. Math. Mexicana.* **9** (2003) 185-195.