Journal of Dynamical and Control Systems (2023) 29:1471–1508
https://doi.org/10.1007/s10883-023-09641-5

Risk-Sensitivity Vanishing Limit for Controlled Markov Processes

Yanan Dai[1] · **Jinwen Chen**[1]

Received: 31 August 2022 / Revised: 31 December 2022 / Accepted: 10 January 2023/
Published online: 16 March 2023

Abstract
In this paper, we prove that the optimal risk-sensitive reward for Markov decision processes with compact state space and action space converges to the optimal average reward as the risk-sensitive factor tends to 0. In doing so, a variational formula for the optimal risk-sensitive reward is derived. An extension of the Kreĭn-Rutman Theorem to certain nonlinear operators is involved. Based on these, partially observable Markov decision processes are also investigated. A portfolio optimization problem is presented as an example of an application of the approach, in which a duality-relation between the maximization of risk-sensitive reward and the maximization of upside chance for out-performance over the optimal average reward is established.

Keywords Markov decision processes · Partially observable Markov decision process · Risk-sensitive reward · Variational formula · Kreĭn-Rutman Theorem

Mathematics Subject Classification (2010) Primary 90C40 · 60F99 · Secondary 47J10 · 93E99

1 Introduction

In this paper, we study a risk-sensitive control problem for Markov decision processes (MDPs). The risk-sensitive control of MDPs has been widely investigated (see [10, 12, 16, 21, 22] and references cited therein). The basic target is to find the optimal solution to the following control problem

$$\lambda^{\pi}(x,\gamma) := \frac{1}{\gamma} \lim_{N\to\infty} \frac{1}{N} \log E_x^{\pi} \exp\left[\gamma \sum_{n=0}^{N-1} r(X_n, A_n)\right], \tag{1.1}$$

✉ Jinwen Chen
chenjw@mail.tsinghua.edu.cn

Yanan Dai
dyn18@mails.tsinghua.edu.cn

[1] Department of Mathematics, Tsinghua University, Beijing, China

where X_n is the state of the system at time n, x is the initial state, A_n is the decision made by the controller at time n, and π is the strategy for decision-making. The risk-sensitive factor γ represents the controller's risk preference. Regarding r as a reward, we concern the following maximization with $\gamma > 0$:

$$\lambda(x, \gamma) := \sup_{\pi} \lambda^{\pi}(x, \lambda).$$

It is well known that $\gamma = 0$ corresponds to the risk-neutral case in which the performance is evaluated according to the following typical long-run average reward:

$$v(x) := \sup_{\pi} \lim_{N\to\infty} \frac{1}{N} E^{\pi}\left[{}_x \sum_{n=0}^{N-1} r(X_n, A_n) \right].$$

Notice that for any $\gamma > 0$ if $E(e^{\gamma X})$ and $E(X)$ both exist, then

$$\lim_{\gamma\to 0} \frac{1}{\gamma} \log E(e^{\gamma X}) = E(X).$$

It is natural to consider the problem of whether the optimal risk-sensitive control converges to the optimal long-run average control as the risk-sensitive factor gets vanishing. It is the main purpose of this paper to prove that

$$\lim_{\gamma\to 0+} \lambda(x, \gamma) = v(x), \tag{1.2}$$

provided that both sides are well-defined (see the next section for an explicit description of this problem). This problem has been studied for minimizing risk-sensitive costs for MDPs; see the references cited above. A similar problem for optimal risk-sensitive portfolios has also been studied. Notice that in the framework of portfolio or other asset processes, maximizing rewards is a natural problem for consideration. This maximization problem is essentially different from the minimization problem, but both are of fundamental importance for applications. It is interesting that maximizing the risk-sensitive reward is dual to maximizing the upside chance or minimizing the downside risk under some conditions (see [18, 24] and [26]). These motivate us to study the asymptotics of optimal risk-sensitive rewards for MDPs. We shall show in this paper that for MDPs with compact state spaces and action spaces, under certain assumptions, the maximal risk-sensitive reward will converge to the maximal long-run average reward as the risk-sensitive factor gets down to 0.

We note that in the approach for deriving the asymptotics of minimal risk-sensitive cost, besides a few necessary continuity assumptions, some conditions on contraction and strong ergodicity for the transition probabilities were imposed, based on which span contraction of some properly defined operator can be verified which guarantee a solution to the corresponding Bellman equation. The strong ergodicity condition also makes it possible to apply some large deviation techniques (see (2.7) and (2.8) in Section 2 for the explicit conditions with more explanations given in Remark 2.1). In this paper, we shall use a quite different approach: inspired by Anantharam and Borkar [2], we use a nonlinear extension of the Krein-Rutman Theorem (see [23]) to find the eigenvalues of some properly defined operators on certain function spaces and characterize the optimal growth rate of the multiplicative reward with this eigenvalue. Using this characterization and a perturbation technique, we derive a variational formula for the optimal growth rate (Theorem 3.7) of the MDP without the ergodicity of transition probability. This variational formula is similar to the Donsker-Varadhan formula (see [14]), which is of independent significance. The vanishing risk-sensitivity limit of the maximal reward of MDPs follows as an application of this formula (Theorems 3.1 and 3.8), and its proof implies that for a risk-sensitive control

problem, the optimal policy can be taken to be a stationary one, even when the MDP is not communicating.

We also apply the approach to study the same problem for partially observable Markov decision processes (POMDPs) with compact state and action spaces. For POMDPs, see [4, 6, 8, 11, 17], and [19] and the references cited therein. A widely used approach for studying a POMDP is to transfer it into a completely observable MDP. However, the structure of the transferred MDP is usually much more complicated. Among the current results, such as those in the references mentioned above, few are on the risk-sensitivity vanishing limit. In [1], the limit of the minimal risk-sensitive cost as the risk-sensitive factor tends to 0 is derived for a class of POMDPs with a particular structure. This particular structure makes it possible to apply a large deviation approach. In [11], Di Masi and Stettner established the existence of the solution to the associated Bellman equation for cost-minimizing problems. However, they remarked that the limit as the risk-sensitive factor tends to 0 for general POMDPs had not been proven. Based on our investigation for MDPs, we prove that, as long as the solution to the associated Bellman equation exists, the maximal risk-sensitive reward converges to the maximal long-run average reward as the risk-sensitive factor tends to 0.

Finally, as an application of our approach, for portfolio optimization, we establish a duality-relation between maximizing the risk-sensitive reward and maximizing the chance for outperforming certain amounts of reward, with the range of the amounts being characterized by the optimal average reward (Theorem 5.2).

The paper is organized as follows. At the end of this section, we introduce some notations that will be frequently used in this paper. In Section 2, we define the decision model and derive some properties of the operator corresponding to the Bellman equation. The risk-neutral limit for MDPs is given in Section 3, in which the variational formula mentioned above is established. Section 4 is devoted to POMDPs. The portfolio optimization problem is investigated in Section 5.

Here are some notations and preliminaries. Given a separable and complete metric space (also called a *Polish* space) $(\mathcal{X}, \rho)$, let $\mathcal{M}(\mathcal{X})$ and $\mathcal{M}^+(\mathcal{X})$ denote the set of finite signed measures on $\mathcal{X}$ and the set of finite measures on $\mathcal{X}$, respectively. $\mathcal{P}(\mathcal{X})$ is the space of probability measures on $\mathcal{X}$, endowed with the weak topology. For $p, q \in \mathcal{P}(\mathcal{X})$, we use $p << q$ to denote that p is absolutely continuous with respect to q. As usual, $\delta_x(\cdot)$ denotes the Dirac measure on point $x \in \mathcal{X}$. When $\mathcal{X}$ is compact, $C(\mathcal{X})$, the real-valued continuous functions on $\mathcal{X}$, equipped with the supremum norm $\|\cdot\|$, is a Banach space. Let $C^+(\mathcal{X})$ denote the set of non-negative functions in $C(\mathcal{X})$. $C^+(\mathcal{X})$ is a *cone*, which means that for any $f, g \in C^+(\mathcal{X})$ and any $c > 0$, both $f + g$ and cf are in $C^+(\mathcal{X})$. $C^+(\mathcal{X})$ is convex, closed and satisfies that $C^+(\mathcal{X}) \cup (-C^+(\mathcal{X})) = \{0\}$ and that $interior(C^+(\mathcal{X})) \neq \emptyset$. We write $f \geq g, f > g, f >> g$ if $f - g \in C^+(\mathcal{X})$, $f - g \in C^+(\mathcal{X})\backslash\{0\}$, $f - g \in interior(C^+(\mathcal{X}))$, respectively. These facts form the basis for applying a nonlinear extension of the Krein-Rutman theorem (see Appendix) to the operator corresponding to the Bellman equation.

For two probability measures $p, q \in \mathcal{P}(\mathcal{X})$, the relative entropy of p with respect to q is defined by

$$D(p\|q) := \begin{cases} \int_{\mathcal{X}} \log\left(\frac{dp}{dq}(x)\right) p(dx), & p << q \\ \infty, & \text{otherwise} \end{cases}, \tag{1.3}$$

which plays an important role in the variational formula for the optimal reward.

Let $\mathrm{Lip}(\mathcal{X})$ denote the space of real-valued, bounded, and Lipschitz continuous functions on $\mathcal{X}$. Given $f \in \mathrm{Lip}(\mathcal{X})$, define its norm by

$$\|f\|_L := \max\left\{\sup_{x\in\mathcal{X}}|f(x)|,\ \sup_{\substack{x,y\in\mathcal{X}\\ x\neq y}}\frac{|f(x)-f(y)|}{\rho(x,y)}\right\}. \tag{1.4}$$

Then $(\mathrm{Lip}(\mathcal{X}),\ \|\cdot\|_L)$ is a Banach space when $\mathcal{X}$ is compact. Given $\mu \in \mathcal{M}(\mathcal{X})$, define the following Kantorovich-Rubinstein norm:

$$\|\mu\|_0 := \sup\left\{\int f d\mu,\ f \in \mathrm{Lip}(\mathcal{X}),\ \|f\|_L \le 1\right\}. \tag{1.5}$$

Then the weak topology on $\mathcal{P}(\mathcal{X})$ is generated by the Kantorovich-Rubinstein metric $d_0(\mu,\nu) := \|\mu-\nu\|_0$ (Theorem 8.3.2, pp. 193–194, in [7]). $\mathcal{P}(\mathcal{X})$ endowed with the weak topology is a Polish space since $\mathcal{X}$ is Polish. The space of Lipschitz functions and the Kantorovich-Rubinstein metric will be used in Section 4 to discuss the POMDPs.

Finally, as usual, $\mathbb{N}$ and $\mathbb{R}$ denote the sets of non-negative integers and real numbers, respectively.

2 Solution to the Bellman Equation

A discrete-time MDP can be represented as a four-tuple $M = \langle S, A, p(\cdot|\cdot,\cdot), r(\cdot,\cdot)\rangle$. S is the *state space*, A is the *action space*, and both are assumed to be compact metric spaces in the present paper. We assume for convenience that any action in A is admissible in any state. The transition kernel, which depends on actions, is denoted by $p(E|x,a)$, $E \subseteq S, x \in S, a \in A$. The last element in the tuple is the one-step reward function $r : S \times A \to \mathbb{R}$. To define a probability space and a stochastic process with the desired mechanism, let $\Omega = (S\times A)^\infty$ and $\mathcal{B}(\Omega)$ be the product Borel σ-field. Given a sample path $\omega = (x_1, a_1, x_2, a_2, ...) \in \Omega$, define $X_t := x_t, A_t := a_t, t \in \mathbb{N}$. At each time $t \in \mathbb{N}$, the system M occupies a state X_t, based on which the controller chooses an action A_t, and then the system moves to the next state according to the law $p(\cdot|X_t, A_t)$. A Markov *decision rule* at time t is a stochastic kernel $d_t \in \mathcal{P}(A|S)$, where $d_t(B|x)$ denotes the probability for taking action in $B \subseteq A$ when observing the current state $X_t = x$. A Markov *policy* π is a sequence of Markov decision rules. Let D_M denote the set of all the Markov decision rules. $\Pi_M = (D_M)^\infty$ is the set of all the Markov policies of M. Given an initial state $x \in S$ and a policy $\pi = (d_1, d_2, ...) \in \Pi_M$, we can define a unique probability measure P_x^π on $\mathcal{B}(\Omega)$ by the Ionescu-Tulcea theorem, such that for each $t \ge 0$,

$$\mathrm{P}_x^\pi(dx_1, da_1, ..., dx_{t-1}, da_{t-1}, dx_t) = \delta_x(dx_1)\left(\prod_{i=1}^{t-1} d_i(da_i|x_i)p(dx_{i+1}|x_i,a_i)\right).$$

The corresponding expectation operator is denoted by E_x^π. Since we view r as a reward, a typical criterion for evaluating the optimal policy is to maximize the average reward, i.e., we are interested in the following function

$$v(x) = \sup_{\pi\in\Pi_M}\liminf_{N\to\infty}\frac{1}{N}E_x^\pi\left[\sum_{t=1}^{N} r(X_t, A_t)\right], \tag{2.1}$$

which is risk-neutral. Informally, we notice that by the Taylor expansion, we see that for a small factor γ

$$\frac{1}{\gamma N}\log E_x^\pi \exp\left[\gamma\sum_{t=1}^N r(X_t, A_t)\right] = \frac{1}{N}E_x^\pi\left[\sum_{t=1}^N r(X_t, A_t)\right] + \frac{\gamma}{2N}\mathrm{Var}_x^\pi\left[\sum_{t=1}^N r(X_t, A_t)\right] + \frac{1}{N}\cdot o(\gamma^2). \tag{2.2}$$

Hence, we can use $\gamma \neq 0$ to evaluate the controller's risk preference. This leads to the following risk-sensitive criterion for maximization of reward (see [10]):

$$\lambda(x, \gamma) = \sup_{\pi\in\Pi_M}\frac{1}{\gamma}\liminf_{N\to\infty}\frac{1}{N}\log E_x^\pi \exp\left[\gamma\sum_{t=1}^N r(X_t, A_t)\right], \tag{2.3}$$

where $\gamma \neq 0$ is a constant evaluating the controller's risk preference. With "sup" replaced by "inf" in (2.3), we have the risk-sensitive criterion for minimization of cost. An interesting problem is the asymptotics of $\lambda(x, \gamma)$ as $\gamma \to 0$. Observe that if we define for $N \geq 1$

$$\lambda_N^\pi(x, \gamma) := \frac{1}{\gamma N}\log E_x^\pi \exp\left[\gamma\sum_{t=1}^N r(X_t, A_t)\right]. \tag{2.4}$$

Then (2.2) implies that

$$\lambda_N^\pi(x, 0) := \lim_{\gamma\to 0}\lambda_N^\pi(x, \gamma) = \frac{1}{N}E_x^\pi\left[\sum_{t=1}^N r(X_t, A_t)\right], \tag{2.5}$$

which motivates us to derive

$$\lambda(x, 0) := \lim_{\gamma\to 0}\lambda(x, \gamma) = v(x). \tag{2.6}$$

This is the main concern of this paper.

Remark 2.1 Replacing the sup in (2.3) with an inf to define $\tilde{\lambda}(x, \gamma)$, the $\gamma \to 0$ limit has already been established in [10]. They proved the existence of a solution to the Bellman equation with, in addition to some necessary continuity assumptions, the following two requirements:

$$p(E|x, a) - p(E|x', a') < \delta \text{ for some } \delta \in (0, 1) \text{ and for any } x, x' \in S,\ a, a' \in A; \tag{2.7}$$

and there exist an $\eta \in \mathcal{P}(S)$ and a continuous density $q(x, a, y)$ such that $p(E|x, a) = \int_E q(x, a, y)\eta(dy)$ for $E \in \mathcal{B}(S)$ and

$$q(x, a, y) > 0, \quad \sup_{x,x'\in S}\sup_{a\in A}\sup_{y\in S}\frac{q(x, a, y)}{q(x', a, y)} = K < \infty. \tag{2.8}$$

These conditions guarantee that the operator defining the corresponding Bellman equation is *span contractive*, and hence a solution exists. (2.8) also implies strong ergodicity for the family of transition probabilities defining the MDP. A consequence is the applicability of large deviation techniques for ergodic Markov processes. Instead of such conditions, we will use the following assumption (B1) on communication among the family of transition probabilities to guarantee that the eigenvector of the Bellman equation is strictly positive, based on which the variational formula holds. Then we will remove (B1) by a perturbation technique, which means that the limit can hold for completely observable MDPs without any communication requirements on the transition probability.

(B1) For any $x_1, x \in S$ and any open neighborhood U containing x, there exist an $N > 0$ and $a_1, .., a_N \in A$ such that

$$\int \mathbf{1}_U(x_{N+1}) p(dx_{N+1}|x_N, a_N)...p(dx_2|x_1, a_1) > 0,$$

Remark 2.2 When the model is finite (i.e., both S and A are finite), (B1) is the classical communicating condition.

Now let

$$v := \sup_{x \in S} v(x)$$

and

$$\lambda(\gamma) := \sup_{x \in S} \lambda(x, \gamma)$$

for $\gamma > 0$. The main objective of this paper is to show that

$$\lim_{\gamma \to 0} \lambda(\gamma) = v. \tag{2.9}$$

In order to do this, we make the following assumptions.

(A1) $r(x, a)$ is continuous in (x, a).
(A2) $(x, a) \mapsto \int_S p(dy|x, a) f(y)$ is continuous in (x, a) when $f \in C(S)$.
(A3) The family of functions

$$\left\{ x \mapsto \int_S f(y) p(dy|x, a), \ \ f \in C(S), \|f\| \leq 1, a \in A \right\}$$

is equicontinuous.

Moreover, if (B1) holds, we will see that, independent of the choice of the initial state $x \in S$, the value of $\lambda(x, \gamma)$ depends only on γ and

$$\lim_{\gamma \to 0} \lambda(x, \gamma) = \lim_{\gamma \to 0} \lambda(\gamma) = v. \tag{2.10}$$

Remark 2.3 A concrete case in which (A3) is satisfied is that

$$p(dy|x, a) = q(y|x, a)\Lambda(dy)$$

with $\Lambda \in \mathcal{P}(S)$ and $\{q(y|\cdot, a), y \in S, a \in A\}$ equicontinuous. The equicontinuity assumption (A3) is only used to prove the compactness of the operator related to the Bellman equation. In particular, for every finite MDP, (A1), (A2), and (A3) automatically hold. Combined with (B1), this compactness yields the existence of a positive eigenvalue and the associated positive eigenvector. The continuity assumptions of the result regarding the Bellman equation in [10] are the same as (A1) and (A2). But the $\gamma \to 0$ limit established in [10] requires that there exist an $\eta \in \mathcal{P}(S)$ and a density $q(x, a, y) > 0$ such that $p(E|x, a) = \int_E q(x, a, y)\eta(dy)$ for $E \in \mathcal{B}(S)$ and $(x, a, y) \to q(x, a, y)$ is continuous, which is more strict than (A3) when state space S and action space A are compact.

The Bellman equation mentioned above is

$$\rho f(x) = \sup_{\nu\in\mathcal{P}(A)} \int_{S\times A} e^{\gamma r(x,a)} f(y) p(dy|x,a)\nu(da), \tag{2.11}$$

and the corresponding operator $L^{(\gamma)}$ on $C(S)$ is defined by

$$L^{(\gamma)} f(x) := \sup_{\nu\in\mathcal{P}(A)} \int_{S\times A} e^{\gamma r(x,a)} f(y) p(dy|x,a)\nu(da). \tag{2.12}$$

Since

$$\sup_{\nu\in\mathcal{P}(A)} \int_A \left(\int_S e^{\gamma r(x,a)} f(y) p(dy|x,a) \right) \nu(da) \le \sup_{a'\in A} \int_S e^{\gamma r(x,a')} f(y) p(dy|x,a')$$

and

$$\sup_{\nu\in\mathcal{P}(A)} \int_{S\times A} e^{\gamma r(x,a)} f(y) p(dy|x,a)\nu(da) \ge \sup_{a'\in A} \int_{S\times A} e^{\gamma r(x,a')} f(y) p(dy|x,a')\delta_{a'}(da),$$

we see that

$$L^{(\gamma)} f(x) = \sup_{a\in A} \int_S e^{\gamma r(x,a)} f(y) p(dy|x,a). \tag{2.13}$$

Assumptions (A1), (A2) and the compactness of $S \times A$ imply that when $f \in C(S)$, $L^{(\gamma)} f$ also belongs to $C(S)$. Combining these with (A3), we can prove that $L^{(\gamma)}$ is a compact operator, which is crucial to the existence of a positive eigenvalue, as we claimed before.

Proposition 2.1 *Assume (A1), (A2), and (A3). Then $L^{(\gamma)}$ is a compact operator mapping $C(S)$ into itself.*

Proof Notice that $r(\cdot,\cdot)$ is bounded under assumption (A1) and the compactness of $S \times A$. For convenience, we let r_M and r_m be the supremum and infimum of r, respectively. For any function f with $\|f\| \le K$, we have $\sup_{x\in S} \left|L^{(\gamma)} f(x)\right| \le K e^{\gamma r_M}$. Thus, to apply the Arzelà-Ascoli theorem, we need to verify that the family $\{L^{(\gamma)} f, f \in C(S), \|f\| \le K\}$ is equicontinuous.

To this end, let ρ denote the metric on S. According to (A3), for any $\varepsilon > 0$, there exists $\delta_1 > 0$ such that

$$\sup_{a\in A} \sup_{\substack{g\in C(S)\\ \|g\|\le 1}} \left| \int_S g(y) p(dy|x_1,a) - \int_S g(y) p(dy|x_2,a) \right| \le \varepsilon$$

for any x_1, x_2 with $\rho(x_1, x_2) \le \delta_1$. By the uniform continuity of $e^{\gamma r(\cdot,\cdot)}$, there exists $\delta_2 > 0$ such that

$$\sup_{a\in A} \left| e^{\gamma r(x_1,a)} - e^{\gamma r(x_2,a)} \right| \le \varepsilon$$

whenever $\rho(x_1, x_2) \leq \delta_2$. Consequently when $\|f\| \leq K$, for x_1, x_2 with $\rho(x_1, x_2) \leq \min\{\delta_1, \delta_2\}$, we have

$$\begin{aligned}\left|L^{(\gamma)}f(x_1) - L^{(\gamma)}f(x_2)\right| &\leq \sup_{a\in A}\left|\int_S e^{\gamma r(x_1,a)}f(y)p(dy|x_1,a) - \int_S e^{\gamma r(x_2,a)}f(y)p(dy|x_2,a)\right| \\ &\leq \sup_{a\in A}\left|e^{\gamma r(x_1,a)} - e^{\gamma r(x_2,a)}\right|\int_S |f(y)|\, p(dy|x_1,a) \\ &\quad + \sup_{a\in A} e^{\gamma r(x_2,a)}\left|\int_S f(y)p(dy|x_1,a) - \int_S f(y)p(dy|x_2,a)\right| \\ &\leq \left(K + e^{r_M}\right)\varepsilon.\end{aligned}$$

□

$L^{(\gamma)}$ has the following properties, which will be used to apply the non-linear Kreĭn-Rutman Theorem to prove the existence of the solution to (2.13).

(P1) Assume (A1). Then

$$(L^{(\gamma)})^N f(x) = \sup_{\pi\in\Pi_M} E_x^{\pi}\left[\exp\left(\sum_{t=1}^{N}\gamma r(X_t, A_t)\right)\cdot f(X_{N+1})\right], N \geq 1.$$

This property can be proven by induction using the Markov feature of $\pi \in \Pi_M$ and the fact that $e^{\gamma r(X_i,A_i)} \leq e^{\gamma r_M}$ (see Lemma 2.1 and its proof in [2]).

(P2) (Positive 1-homogeneity) $c(L^{(\gamma)}f) = L^{(\gamma)}(cf)$ for $c \geq 0$ and $f \in C(S)$.

(P3) (Order-preserving) If $f \geq g$, then $L^{(\gamma)}f \geq L^{(\gamma)}g$.

The following theorem shows that the spectral radius of $L^{(\gamma)}$ is an eigenvalue. For an operator $T : C(S) \to C(S)$, define

$$\|T\|^+ := \sup_{\substack{g\in C^+(S)\\ \|g\|\leq 1}} \{\|Tg\|\}.$$

It is not hard to check that

$$\|(L^{(\gamma)})^{m+n}\|^+ \leq \|(L^{(\gamma)})^m\|^+ \|(L^{(\gamma)})^n\|^+,$$

which implies that the limit

$$\rho(L^{(\gamma)}) := \lim_{n\to\infty}\left(\|(L^{(\gamma)})^n\|^+\right)^{\frac{1}{n}}$$

exists.

Theorem 2.2 *Assume (A1), (A2), and (A3). Then $\rho(L^{(\gamma)}) > 0$ and there exists an $f_\gamma \in C^+(S)$ depending on γ with $f_\gamma \neq 0$ such that*

$$\rho(L^{(\gamma)})f_\gamma = L^{(\gamma)}f_\gamma. \tag{2.14}$$

If in addition (B1) is satisfied, then $f_\gamma >> 0$ and $\gamma\lambda(x,\gamma) = \log\rho(L^{(\gamma)})$ is independent of $x \in S$.

Proof From (P1), we see that $\left\|(L^{(\gamma)})^n \mathbf{1}\right\| \geq e^{n\gamma r_m}$, which implies that $\rho(L^{(\gamma)}) \geq e^{\gamma r_m} > 0$. Since $L^{(\gamma)}$ is compact, positive 1-homogeneous and order-preserving, by Theorem A.1 in the Appendix, there exists an $f_\gamma \in C^+(S)$ satisfying (2.14). Moreover, from (A1) and (A2), we know that $\int_S e^{\gamma r(x,a)} f(y) p(dy|x,a)$ is continuous in a, which means that the supremum in (2.13) can be achieved. Hence, there exists a Markov decision rule d^* such that

$$L^{(\gamma)} f_\gamma(x) = \int_S e^{\gamma r(x,d^*(x))} f_\gamma(y) p(dy|x, d^*(x)).$$

Let $\pi^* = (d^*)^\infty$, then we have for $N \in \mathbb{N}$

$$\left[\rho(L^{(\gamma)})\right]^N f_\gamma(x) = E_x^{\pi^*}\left[e^{\gamma \sum_{t=1}^N r(X_t, A_t)} f_\gamma(X_{N+1})\right]. \tag{2.15}$$

Similarly, we have for any Markov policy π

$$\left[\rho(L^{(\gamma)})\right]^N f_\gamma(x) \geq E_x^{\pi}\left[e^{\gamma \sum_{t=1}^N r(X_t, A_t)} f_\gamma(X_{N+1})\right]. \tag{2.16}$$

Now, assume (B1). Since $f_\gamma \in C^+(S)$ and $f_\gamma \neq 0$, there exist $x_0 \in S$, $c_0 > 0$ and an open neighborhood U_0 containing x_0 such that $f_\gamma|_{U_0} > c_0 > 0$. It follows from (B1) that for any $x_1 \in S$, there exists $a_1, .., a_M$ such that

$$\begin{aligned} f_\gamma(x_1) &= \frac{1}{\left[\rho(L^{(\gamma)})\right]^M}\left(L^{(\gamma)}\right)^M f_\gamma(x_1) \\ &\geq \frac{1}{\left[\rho(L^{(\gamma)})\right]^M} \int \mathbf{1}_{U_0}(x_{M+1}) p(dx_{M+1}|x_M, a_M)...p(dx_2|x_1, a_1) \exp\left(\gamma \sum_{i=1}^M r(x_i, a_i)\right) f_\gamma(x_{M+1}) \\ &\geq \frac{c_0 \cdot e^{M\gamma r_m}}{\left[\rho(L^{(\gamma)})\right]^M} \cdot \int \mathbf{1}_{U_0}(x_{M+1}) p(dx_{M+1}|x_M, a_M)...p(dx_2|x_1, a_1) > 0. \end{aligned}$$

Thus, $f_\gamma >> 0$. Since S is compact, there are constants $0 < k_\gamma < K_\gamma < \infty$ such that $k_\gamma \leq f_\gamma \leq K_\gamma$. From (2.15) and (2.16), we see that for any $x \in S$

$$\frac{k_\gamma}{K_\gamma}\left(E_x^{\pi^*}\left[e^{\gamma \sum_{t=1}^N r(X_t, A_t)}\right]\right) \leq \left[\rho(L^{(\gamma)})\right]^N \leq \frac{K_\gamma}{k_\gamma}\left(E_x^{\pi^*}\left[e^{\gamma \sum_{t=1}^N r(X_t, A_t)}\right]\right)$$

and for any Markov policy π

$$\frac{k_\gamma}{K_\gamma}\left(E_x^{\pi}\left[e^{\gamma \sum_{t=1}^N r(X_t, A_t)}\right]\right) \leq \left[\rho(L^{(\gamma)})\right]^N.$$

Taking the logarithm and letting $N \to \infty$, we see that the limit

$$\log \rho(L^{(\gamma)}) = \lim_{N\to\infty} \frac{1}{N} \log E_x^{\pi^*}\left[e^{\gamma \sum_{t=1}^N r(X_t, A_t)}\right] \tag{2.17}$$

exists and

$$\log \rho(L^{(\gamma)}) = \gamma \lambda(x, \gamma)$$

for any $x \in S$. □

Remark 2.4 Assume (A1), (A2), (A3), and (B1). From the proof of Theorem 2.2 we can see that $\rho(L^{(\gamma)})$ is the unique positive eigenvalue of $L^{(\gamma)}$ restricted to $interior(C^+(S))$.

3 Risk-Sensitive Asymptotics of MDP

In this section, we shall apply the following variational formula for $\lambda(\gamma)$ to prove (2.9).

$$\lambda(\gamma) = \sup_{\beta\in\mathcal{I}} \left\{ \int_{S\times A} \left[r(x,a) - \frac{1}{\gamma} D(\beta_2(\cdot|x,a)\|p(\cdot|x,a)) \right] \beta'(dx,da) \right\}, \tag{3.1}$$

where $\mathcal{I}$ is defined by

$$\mathcal{I} := \{\beta \in \mathcal{P}(S\times A\times S) : \beta(S,A,dx) = \beta(dx,A,S)\}, \tag{3.2}$$

and for $\beta \in \mathcal{P}(S\times A\times S)$, the notations β_0, β_1, β_2, and β' are defined by

$$\beta(dx,da,dy) = \beta_0(dx)\beta_1(da|x)\beta_2(dy|x,a) = \beta'(dx,da)\beta_2(dy|x,a). \tag{3.3}$$

Obviously, $\mathcal{I}$ is nonempty and closed in $\mathcal{P}(S\times A\times S)$. Notice that $\mathcal{P}(S\times A\times S)$ is compact since $S\times A\times S$ is compact. Hence, $\mathcal{I}$ is compact, too. For $\beta\in\mathcal{P}(S\times A\times S)$, β_0 is the first 1-dimensional marginal of β, β' is the first 2-dimensional marginal of β, and β_1 and β_2 are the two successive conditional distributions of β. With these notations, $\mathcal{I}$ is seen to be the set of probability measures β on $S\times A\times S$ satisfying that β_0 is invariant under $\int_A \beta_1(da|x)\beta_2(dy|x,a)$.
The validity of (3.1) will be verified in Theorems 3.4 and 3.7. At present, we will apply (3.1) to get the limit of $\lambda(\gamma)$ as $\gamma\to 0$.

Theorem 3.1 *Assume (A1) and (A2). If (3.1) holds, then*

$$\lim_{\gamma\to 0} \lambda(\gamma) = v. \tag{3.4}$$

To prove the theorem, we need the following

Lemma 3.2 *If there exists a $\beta\in\mathcal{I}$ satisfying that $\beta_2 = p$, then*

$$\int_{S\times A} r(x,a)\beta'(dx,da) \le v,$$

where $\mathcal{I}$ is defined in (3.2).

Proof Since $\beta(S,A,dx) = \beta(dx,A,S)$, taking β_0 as the initial distribution and using the policy $\pi_\beta = (\beta_1(da|x))^\infty$, we see that

$$\begin{aligned} E_{\beta_0}^{\pi_\beta}[r(X_2,A_2)] &= \int_{S\times A\times S\times A} r(x_2,a_2)\beta_1(da_2|x_2)p(dx_2|x_1,a_1)\beta_1(da_1|x_1)\beta_0(dx_1) \\ &= \int_{S\times A\times S} \left(\int_A r(x_2,a_2)\beta_1(da_2|x_2) \right) \beta(dx_1,da_1,dx_2) \\ &= \int_{S\times A\times S} \left(\int_A r(x_1,a_2)\beta_1(da_2|x_1) \right) \beta(dx_1,da_1,dx_2) \\ &= \int_{S\times A} r(x_1,a_2)\beta_1(da_2|dx_1)\beta_0(dx_1) = E_{\beta_0}^{\pi_\beta}[r(X_1,A_1)]. \end{aligned}$$

The third equality is due to the coincidence of the first and the third marginal of β ensured by (3.2). By induction, we have

$$\begin{aligned} E_{\beta_0}^{\pi_\beta}\left[\sum_{i=1}^{N} r(X_i, A_i)\right] &= N \cdot E_{\beta_0}^{\pi_\beta}\left[r(X_1, A_1)\right] \\ &= N \int_{S\times A} r(x_1, a_1)\beta_1(da_1|x_1)\beta_0(dx_1) \\ &= N \int_{S\times A} r(x, a)\beta'(dx, da). \end{aligned}$$

Thus,

$$v \geq \liminf_{N\to\infty} E_{\beta_0}^{\pi_\beta}\left[\frac{1}{N}\sum_{i=1}^{N} r(X_i, A_i)\right] = \int_{S\times A} r(x, a)\beta'(dx, da).$$

□

Now we are ready to prove Theorem 3.1.

Proof of Theorem 3.1 By Hölder's inequality, for $\gamma \geq \gamma' > 0$, we have

$$\frac{1}{\gamma}\frac{1}{N}\log E_x^{\pi}\exp\left[\sum_{t=1}^{N}\gamma r(X_t, A_t)\right] \geq \frac{1}{\gamma'}\frac{1}{N}\log E_x^{\pi}\exp\left[\sum_{t=1}^{N}\gamma' r(X_t, A_t)\right]$$

and

$$\frac{1}{\gamma}\frac{1}{N}\log E_x^{\pi}\exp\left[\sum_{t=1}^{N}\gamma r(X_t, A_t)\right] \geq \frac{1}{N}E_x^{\pi}\left[\sum_{t=1}^{N} r(X_t, A_t)\right].$$

Therefore, $\lambda(\gamma)$ is non-decreasing in γ and $\lim_{\gamma\to 0}\lambda(\gamma) \geq v$. To prove (3.4), it suffices to verify that $\lim_{\gamma\to 0}\lambda(\gamma) \leq v$. To this end, we notice that it follows from (3.1) that for any $\varepsilon > 0$ and $\gamma > 0$, there exists $\beta_\gamma^\varepsilon \in \mathcal{I}$ such that

$$\lambda(\gamma) - \varepsilon \leq \int_{S\times A}\left[r(x, a) - \frac{1}{\gamma}D((\beta_\gamma^\varepsilon)_2(\cdot|x, a)\|p(\cdot|x, a))\right](\beta_\gamma^\varepsilon)'(dx, da). \tag{3.5}$$

Recall that $\mathcal{I} \subseteq \mathcal{P}(S \times A \times S)$ is compact, we can find a sequence $\{\gamma_n\}_{n\in\mathbb{N}}$ decreasing to 0 and a $\beta^\varepsilon \in \mathcal{I}$ such that

$$\lim_{\gamma\to 0}\lambda(\gamma) = \lim_{n\to\infty}\lambda(\gamma_n) \quad \text{and} \quad \lim_{n\to\infty}\beta_{\gamma_n}^\varepsilon = \beta^\varepsilon \text{ weakly.}$$

Therefore, from (A1), we know that

$$\lim_{n\to\infty}\lambda(\gamma_n) - \varepsilon \leq \lim_{n\to\infty}\int_{S\times A} r(x, a)\left(\beta_{\gamma_n}^\varepsilon\right)'(dx, da) = \int_{S\times A} r(x, a)\left(\beta^\varepsilon\right)'(dx, da), \tag{3.6}$$

which is finite. Now we claim that $(\beta^\varepsilon)_2 = p$. Indeed, we have

$$\begin{aligned} &\int_{S\times A} D\left(\left(\beta_{\gamma_n}^\varepsilon\right)_2(\cdot|x, a)\middle\|p(\cdot|x, a)\right)\left(\beta_{\gamma_n}^\varepsilon\right)'(dx, da) \\ &\quad = D\left(\beta_{\gamma_n}^\varepsilon(dx, da, dy)\middle\|\left(\beta_{\gamma_n}^\varepsilon\right)'(dx, da)p(dy|x, a)\right). \end{aligned}$$

It follows from the (joint) lower semicontinuity of $D(\cdot\|\cdot)$ and (A2) that

$$-\liminf_{n\to\infty} D\left(\beta^{\varepsilon}_{\gamma_n}(dx,da,dy)\middle\|\left(\beta^{\varepsilon}_{\gamma_n}\right)'(dx,da)p(dy|x,a)\right)\leq$$
$$-D\left(\beta^{\varepsilon}(dx,da,dy)\middle\|\left(\beta^{\varepsilon}\right)'(dx,da)p(dy|x,a)\right).$$

Thus if $(\beta^{\varepsilon})_2\neq p$, then

$$D\left(\beta^{\varepsilon}(dx,da,dy)\middle\|\left(\beta^{\varepsilon}\right)'(dx,da)p(dy|x,a)\right)>0.$$

Combining this with (3.5) and the fact that $\lambda(\gamma)\geq r_m$, we would have

$$r_m-\varepsilon\leq\lim_{n\to\infty}\lambda(\gamma_n)-\varepsilon\leq-\infty.$$

It is impossible. Thus, $\beta^{\varepsilon}\in\mathcal{I}$ and $(\beta^{\varepsilon})_2=p$. Recalling (3.6) and Lemma 3.2, we obtain that

$$\lim_{n\to\infty}\lambda(\gamma_n)-\varepsilon\leq\int_{S\times A}r(x,a)\left(\beta^{\varepsilon}\right)'(dx,da)\leq v.$$

(3.4) follows by letting $\varepsilon\to 0$. □

The remainder of this section is devoted to verifying (3.1) under certain conditions. This is carried out at first under assumptions including (B1), then with (B1) removed. Our assumption (A2) is slightly weaker than those in [2]. In [2], it is required that the family of functions

$$\left\{(x,a)\mapsto\int_S f(y)p(dy|x,a),\ \ f\in C(S),\|f\|\leq 1\right\}$$

is equicontinuous, while we assume that

$$\left\{x\mapsto\int_S f(y)p(dy|x,a),\ \ f\in C(S),\|f\|\leq 1,a\in A\right\}$$

is equicontinuous (Theorems 3.4 and 3.8). Moreover, it is worth mentioning that equicontinuity only plays a role in the existence of the positive eigenvalue and the eigenvector. Once $L^{(\gamma)}$ has a positive eigenvalue and a strictly positive eigenvector, only (A1) and (A2) are needed.

Proposition 3.3 *Assume (A1) and (A2). If there exist $\rho_\gamma>0$ and $f_\gamma\in C(S)$ such that $f_\gamma>>0$ and $L^{(\gamma)}f_\gamma=\rho_\gamma f_\gamma$, then (3.1) holds.*

Proof From the proof of Theorem 2.2, we know that $\log\rho_\gamma=\gamma\lambda(x,\gamma)$ for any $x\in S$. Thus, for any $\mu\in\mathcal{M}^+(S)$, we have

$$e^{\gamma\lambda(\gamma)}=\frac{\int L^{(\gamma)}f_\gamma d\mu}{\int f_\gamma d\mu}.$$

Therefore,

$$e^{\gamma\lambda(\gamma)}=\sup_{\mu\in\mathcal{M}^+(S)}\frac{\int L^{(\gamma)}f_\gamma d\mu}{\int f_\gamma d\mu}\geq\inf_{f>>0}\sup_{\mu\in\mathcal{M}^+(S)}\frac{\int L^{(\gamma)}f d\mu}{\int f d\mu}.\tag{3.7}$$

For any $f>>0$, we also have

$$\frac{L^{\gamma}f}{f}\leq\sup_{\mu\in\mathcal{M}^+(S)}\frac{\int L^{(\gamma)}f d\mu}{\int f d\mu},$$

which means that

$$L^{\gamma} f \leq \left(\sup_{\mu \in \mathcal{M}^{+}(S)} \frac{\int L^{(\gamma)} f d\mu}{\int f d\mu} \right) f.$$

Since under (A1) and (A2), properties (P2) and (P3) hold for $L^{(\gamma)}$, we can apply Theorem A.2 and A.3 in the Appendix to deduce that

$$e^{\gamma \lambda(\gamma)} = \rho(L^{(\gamma)}) \leq \left(\sup_{\mu \in \mathcal{M}^{+}(S)} \frac{\int L^{(\gamma)} f d\mu}{\int f d\mu} \right). \tag{3.8}$$

From (3.7) and (3.8), we have

$$\lambda(\gamma) = \frac{1}{\gamma} \log \inf_{f>>0} \sup_{\mu \in \mathcal{M}^{+}(S)} \frac{\int L^{(\gamma)} f d\mu}{\int f d\mu} = \frac{1}{\gamma} \log \inf_{f>>0} \sup_{\substack{\mu \in \mathcal{M}^{+}(S) \\ \int f d\mu = 1}} \int L^{(\gamma)} f d\mu.$$

Thus,

$$\begin{aligned}
\lambda(\gamma) &= \frac{1}{\gamma} \log \inf_{f>>0} \sup_{\substack{\mu \in \mathcal{M}^{+}(S) \\ \int f d\mu = 1}} \int_S \mu(dx) \sup_{a \in A} \int_S e^{\gamma r(x,a)} f(y) p(dy|x,a) \\
&= \frac{1}{\gamma} \log \inf_{f>>0} \sup_{\nu \in \mathcal{P}(S)} \int_S \nu(dx) \sup_{a \in A} \int_S e^{\gamma r(x,a) + \log f(y) - \log f(x)} p(dy|x,a) \\
&= \frac{1}{\gamma} \inf_{g \in C(S)} \sup_{x \in S} \sup_{a \in A} \log \int_S e^{\gamma r(x,a) + g(y) - g(x)} p(dy|x,a) \\
&= \frac{1}{\gamma} \inf_{g \in C(S)} \sup_{\eta \in \mathcal{P}(S \times A)} \log \int_{S \times A \times S} e^{\gamma r(x,a) + g(y) - g(x)} \eta(dx,da) p(dy|x,a).
\end{aligned}$$

Using the Gibbs variational formula (Proposition 1.4.2(a), pp. 33–34 in [15]), we see that

$$\lambda(\gamma) = \frac{1}{\gamma} \inf_{g \in C(S)} \sup_{\eta \in \mathcal{P}(S \times A)} \sup_{\beta \in \mathcal{P}(S \times A \times S)} \left\{ \int_{S \times A \times S} [\gamma r(x,a) + g(y) - g(x)] \beta(dx,da,dy) - D(\beta(dx,da,dy) \| \eta(dx,da) p(dy|x,a)) \right\}.$$

Since $D(\mu \| \nu)$ is jointly convex and lower semicontinuous in (μ, ν) (Lemma 1.4.3, pp. 36–38 in [15]) and $\mathcal{P}(S \times A)$, $\mathcal{P}(S \times A \times S)$ are both compact, the minimax theorem (Theorem 4.2 in [25]) can be applied to get

$$\lambda(\gamma) = \frac{1}{\gamma} \sup_{\beta \in \mathcal{P}(S \times A \times S)} \sup_{\eta \in \mathcal{P}(S \times A)} \inf_{g \in C(S)} \left\{ \int_{S \times A \times S} [\gamma r(x,a) + g(y) - g(x)] \beta(dx,da,dy) - D(\beta(dx,da,dy) \| \eta(dx,da) p(dy|x,a)) \right\}.$$

Furthermore, by the chain rule for relative entropy (Theorem D.13, pp. 357–359 in [9]), we have that

$$\lambda(\gamma) = \frac{1}{\gamma} \sup_{\beta \in \mathcal{P}(S \times A \times S)} \sup_{\eta \in \mathcal{P}(S \times A)} \inf_{g \in C(S)} \left\{ \int_{S \times A \times S} [\gamma r(x,a) + g(y) - g(x)] \beta(dx,da,dy) - D(\beta'(dx,da) \| \eta(dx,da)) - \int_{S \times A} D(\beta_2(dy|x,a) \| p(dy|x,a)) \beta'(dx,da) \right\}.$$

Since $D(\mu\|\nu) \geq 0$ and $D(\mu\|\nu) = 0$ iff $\mu = \nu$ (Lemma 1.4.1 , pp. 33, in [15]), the supremum over $\eta \in \mathcal{P}(S \times A)$ is attained at $\eta = \beta'$. Moreover, notice that when $\beta \in \mathcal{I}$, for any $g \in C(S)$,

$$\int_{S\times A\times S} [g(y) - g(x)]\beta(dx, da, dy) = 0,$$

and for $\beta \notin \mathcal{I}$,

$$\inf_{g\in C(S)} \int_{S\times A\times S} [g(y) - g(x)]\beta(dx, da, dy) = -\infty,$$

we obtain that

$$\begin{aligned}\lambda(\gamma) &= \sup_{\beta\in\mathcal{I}} \left\{ \int_{S\times A\times S} r(x,a)\beta(dx,da,dy) - \frac{1}{\gamma}\int_{S\times A} D(\beta_2(dy|x,a)\|p(dy|x,a))\beta'(dx,da)\right\} \\ &= \sup_{\beta\in\mathcal{I}} \left\{ \int_{S\times A} \left[r(x,a) - \frac{1}{\gamma} D(\beta_2(dy|x,a)\|p(dy|x,a))\right]\beta'(dx,da)\right\}.\end{aligned}$$

□

Combining Theorem 2.2 and Proposition 3.3, we obtain the following theorem immediately.

Theorem 3.4 *Assume (A1), (A2), (A3), and (B1). Then* (3.1) *holds.*

To remove (B1), we use a perturbation argument. For each $\epsilon > 0$, define a new MDP M_ϵ with the transition law and one-step reward given by

$$p_\epsilon^{(\gamma)}(dy|x,a) := \frac{\epsilon\Gamma(dy) + e^{\gamma r(x,a)} p(dy|x,a)}{\epsilon + e^{\gamma r(x,a)}} \text{ and } r_\epsilon^{(\gamma)}(x,a) := \log\left(\epsilon + e^{\gamma r(x,a)}\right) \tag{3.9}$$

respectively, where $\Gamma \in \mathcal{P}(S)$ with full support. It is not hard to check that M_ϵ satisfies (A1), (A2), (A3), and (B1). Using $E^\pi_{\epsilon,\gamma,x}$ to denote the corresponding expectation operator with initial state x and policy π, we define

$$L_\epsilon^{(\gamma)} f(x) := \sup_{\pi\in\Pi_M} E^\pi_{\epsilon,\gamma,x}\left[e^{r_\epsilon(X_1,A_1)} f(X_2)\right] = L^{(\gamma)} f(x) + \epsilon \int \Gamma(dy) f(y) \tag{3.10}$$

and

$$\lambda_\epsilon(x,\gamma) := \sup_{\pi\in\Pi_M} \frac{1}{\gamma} \liminf_{N\to\infty} \frac{1}{N} \log E^\pi_{\epsilon,\gamma,x} \exp\left[\sum_{t=1}^N r_\epsilon^{(\gamma)}(X_t, A_t)\right]. \tag{3.11}$$

By Theorem 2.2, $\lambda_\epsilon(x,\gamma)$ depends only on γ, and the limit inferior is actually a limit. Hence, we write it as $\lambda_\epsilon(\gamma)$. Without (B1), we will prove the variational formula by exploring properties of $\lambda_\epsilon(\gamma)$ and then letting $\varepsilon \to 0$.

Lemma 3.5 *Assume (A1) and (A2). Then $\lambda_\epsilon(\gamma)$ is non-decreasing in ϵ and $\lim_{\epsilon\to 0} \lambda_\epsilon(\gamma) \geq \lambda(\gamma)$.*

Proof From property (P1), we have

$$\lambda_\epsilon(\gamma) = \lim_{N\to\infty} \frac{1}{\gamma} \sup_{\pi\in\Pi_M} \frac{1}{N} \log E^\pi_{\epsilon,\gamma,x} \exp\left[\sum_{t=1}^N r_\epsilon^{(\gamma)}(X_t, A_t)\right] = \lim_{N\to\infty} \frac{1}{\gamma N} \log (L_\epsilon^{(\gamma)})^N \mathbf{1}(x) \tag{3.12}$$

for any $x \in S$. Thus, for any $\epsilon_1 > \epsilon_2 > 0$, by (3.10), we obtain that

$$\lambda_{\epsilon_1}(\gamma) \geq \lambda_{\epsilon_2}(\gamma) \geq \sup_{x\in S} \liminf_{N\to\infty} \frac{1}{\gamma N} \log (L^{(\gamma)})^N \mathbf{1}(\theta)$$
$$\geq \sup_{x\in S} \sup_{\pi\in\Pi_M} \liminf_{N\to\infty} \frac{1}{\gamma N} \log E_\theta^\pi \exp\left[\gamma \sum_{t=1}^N r(X_t, A_t)\right] = \lambda(\gamma).$$

□

In order to write the variational formula in a form that is more convenient for using in the following arguments, we define for given $\epsilon > 0$, $\gamma > 0$ and $\beta \in \mathcal{I}$

$$\phi(\beta, \gamma, \epsilon) := \int_{S\times A} \left[\tfrac{1}{\gamma} r_\epsilon^{(\gamma)}(x, a) - \tfrac{1}{\gamma} D\left(\beta_2(\cdot|x, a) \| p_\epsilon^{(\gamma)}(\cdot|x, a)\right)\right] \beta'(dx, da)$$

and

$$\phi(\beta, \gamma, 0) := \int_{S\times A} \left[r(x, a) - \tfrac{1}{\gamma} D\left(\beta_2(\cdot|x, a) \| p(\cdot|x, a)\right)\right] \beta'(dx, da).$$

To prove $\lambda(\gamma) \geq \lim_{\epsilon\to 0} \sup_{\beta\in\mathcal{I}} \phi(\beta, \gamma, \epsilon)$, we will show that

$$\lambda(\gamma) \geq \sup_{x\in S} \lambda_{SM}(x, \gamma) \geq \sup_{\beta\in\mathcal{I}} \phi(\beta, \gamma, 0) \geq \lim_{\epsilon\to 0} \sup_{\beta\in\mathcal{I}} \phi(\beta, \gamma, \epsilon),$$

where $\lambda_{SM}(x, \gamma)$ is defined by

$$\lambda_{SM}(x, \gamma) := \sup_{d\in D_M} \liminf_{N\to\infty} \frac{1}{\gamma N} \log E_x^{d^\infty} \exp\left[\gamma \sum_{t=1}^N r(X_t, A_t)\right],$$

with d^∞ denoting the *stationary* Markov policy whose decision rules at each time are the same $d \in D_M$.

Lemma 3.6 *Assume (A1) and (A2). Then* $\sup_{x\in S} \lambda_{SM}(x, \gamma) \geq \sup_{\beta\in\mathcal{I}} \phi(\beta, \gamma, 0)$.

Proof We need to prove that

$$\sup_{x\in S} \lambda_{SM}(x, \gamma) \geq \phi(\beta, \gamma, 0)$$

for each $\beta \in \mathcal{I}$. If $\phi(\beta, \gamma, 0) = -\infty$, the inequality holds trivially. Otherwise, $\beta \in \mathcal{I}$ with $\phi(\beta, \gamma, 0) > -\infty$ implies that $\beta_2(\cdot|x, a) << p(\cdot|x, a)$ $\beta'-$a.s.. Choosing the stationary Markov policy $\pi_\beta = (\beta_1(da|\theta))^\infty$ and the initial distribution $\beta_0 \in \mathcal{P}(S)$, we see that

$$\sup_{x\in S} \lambda_{SM}(x, \gamma) \geq \liminf_{N\to\infty} \frac{1}{\gamma N} \log E_{\beta_0}^{\pi_\beta} \exp\left[\sum_{t=1}^N \gamma r(X_t, A_t)\right].$$

Define ${}^{\beta}E_{\beta_0}^{\pi_\beta}$ as the expectation operator with respect to the probability measure determined by the initial distribution β_0, the transition law $\beta_2(dy|x,a)$ for $\{X_t\}_{t\in\mathbb{N}}$, and the policy π_β. Using the change of measure technique and Jensen's inequality, we obtain that

$$\begin{aligned}\log E_{\beta_0}^{\pi_\beta}\exp\left[\sum_{t=1}^{N}\gamma r(X_t,A_t)\right] &= \log {}^{\beta}E_{\beta_0}^{\pi_\beta}\exp\left\{\sum_{t=1}^{N}\left[\gamma r(X_t,A_t)-\log\left(\frac{d\beta_2(\cdot|X_t,A_t)}{dp(\cdot|X_t,A_t)}(X_{t+1})\right)\right]\right\}\\ &\geq {}^{\beta}E_{\beta_0}^{\pi_\beta}\sum_{t=1}^{N}\left[\gamma r(X_t,A_t)-\log\left(\frac{d\beta_2(\cdot|X_t,A_t)}{dp_0(\cdot|X_t,A_t)}(X_{t+1})\right)\right].\end{aligned}$$

Since $\beta\in\mathcal{I}$, the same argument as in proving Lemma 3.2 shows that

$$\begin{aligned}&{}^{\beta}E_{\beta_0}^{\pi_\beta}\sum_{t=1}^{N}\left[\gamma r(X_t,A_t)-\log\left(\frac{d\beta_2(\cdot|X_t,A_t)}{dp_0(\cdot|X_t,A_t)}(X_{t+1})\right)\right]\\ &= N\cdot{}^{\beta}E_{\beta_0}^{\pi_\beta}\left[\gamma r(X_1,A_1)-\log\left(\frac{d\beta_2(\cdot|X_1,A_1)}{dp(\cdot|X_1,A_1)}(X_2)\right)\right]\end{aligned}$$

Consequently,

$$\sup_{x\in S}\lambda_{SM}(x,\gamma)\geq\frac{1}{\beta}E_{\beta_0}^{\pi_\beta}\left[\gamma r(X_1,A_1)-\log\left(\frac{d\beta_2(\cdot|X_1,A_1)}{dp(\cdot|X_1,A_1)}(X_2)\right)\right]=\phi(\beta,\gamma,0).$$

□

Combining Theorem 3.4 and Lemmas 3.5 and 3.6, we obtain the following

Theorem 3.7 *Assume (A1), (A2), and (A3). Then* (3.1) *holds.*

Proof From Lemmas 3.5 and 3.6, we see that

$$\lim_{\epsilon\to0}\lambda_\epsilon(\gamma)\geq\lambda(\gamma)=\sup_{x\in S}\lambda(x,\gamma)\geq\sup_{x\in S}\lambda_{SM}(x,\gamma)\geq\sup_{\beta\in\mathcal{I}}\phi(\beta,\gamma,0).\tag{3.13}$$

Hence, (3.1) will follow once we prove that $\sup_{\beta\in\mathcal{I}}\phi(\beta,\gamma,0)\geq\lim_{\epsilon\to0}\lambda_\epsilon(\gamma)$. Since M_ϵ satisfies (A1), (A2), and (B1), by Theorems 2.2 and 3.4, we have

$$\lambda_\epsilon(\gamma)=\sup_{\beta\in\mathcal{I}}\phi(\beta,\gamma,\epsilon).\tag{3.14}$$

Therefore, given $\xi>0$, for every $\epsilon>0$, there exists $\beta_\epsilon^\xi\in\mathcal{I}$ such that

$$\lambda_\epsilon(\gamma)<\sup_{\beta\in\mathcal{I}}\phi(\beta,\gamma,\epsilon)+\xi.$$

Since $\mathcal{I}$ is compact, there exists a sequence $\{\epsilon_n\}_{n\in\mathbb{N}}$ decreasing to 0 such that the weak limit $\lim_{n\to\infty}\beta_{\epsilon_n}^\xi=:\beta^\xi$ exists and $\in\mathcal{I}$. By (A1) and Dini's Theorem $r_\epsilon^{(\gamma)}(\cdot,\cdot)$ converges to $\gamma r(\cdot,\cdot)$ uniformly. Thus, we obtain that

$$\lim_{n\to\infty}\int_{S\times A}\frac{1}{\gamma}r_{\epsilon_n}^{(\gamma)}(x,a)\left(\beta_{\epsilon_n}^\xi\right)'(dx,da)=\int_{S\times A}r(x,a)\left(\beta^\xi\right)'(dx,da).$$

Recalling the definition of β_2 for $\beta \in \mathcal{I}$, by the lower semicontinuity of $D(\cdot\|\cdot)$, we see that

$$\begin{aligned}
\liminf_{n\to\infty} & \int_{S\times A} D\left(\left(\beta^{\xi}_{\epsilon_n}\right)_2(dy|x,a)\middle\| p^{(\gamma)}_{\epsilon}(dy|x,a)\right)\left(\beta^{\xi}_{\epsilon_n}\right)'(dx,da) \\
&= \liminf_{n\to\infty} D\left(\beta^{\xi}_{\epsilon_n}(dx,da,dy)\middle\| \left(\beta^{\xi}_{\epsilon_n}\right)'(dx,da)p^{(\gamma)}_{\epsilon}(dy|x,a)\right) \\
&\geq D\left(\beta^{\xi}(dx,da,dy)\middle\| \left(\beta^{\xi}\right)'(dx,da)p(dy|x,a)\right) \\
&= \int_{S\times A} D\left(\left(\beta^{\xi}\right)_2(dy|x,a)\middle\| p(dy|x,a)\right)\left(\beta^{\xi}\right)'(dx,da).
\end{aligned}$$

It then follows that

$$\begin{aligned}
\limsup_{n\to\infty}\phi(\beta^{\xi}_{\epsilon_n},\gamma,\epsilon_n) &\leq \int_{S\times A}\left[r(x,a)-\frac{1}{\gamma}D\left(\left(\beta^{\xi}\right)_2(\cdot|x,a)\|p(\cdot|x,a)\right)\right]\left(\beta^{\xi}\right)'(dx,da) \\
&= \phi(\beta^{\xi},\gamma,0)
\end{aligned}$$

Thus,

$$\sup_{\beta\in\mathcal{I}}\phi(\beta,\gamma,0)\geq\phi(\beta^{\xi},\gamma,0)\geq\limsup_{n\to\infty}\phi(\beta^{\xi}_{\epsilon_n},\gamma,\epsilon_n)\geq\lim_{n\to\infty}\lambda_{\epsilon_n}(\gamma)+\xi.$$

Letting $\xi\to 0$, we have $\sup_{\beta\in\mathcal{I}}\phi(\beta,\gamma,0)\geq\lim_{\epsilon\to 0}\lambda_{\epsilon}(\gamma)$. Now (3.1) follows. □

Remark 3.1 The proof shows that the inequalities in (3.13) are actually equalities, which indicates that the supremum over Markov policies in risk-sensitive MDP is tantamount to the supremum over stationary Markov policies, meaning that one should search for the optimal policy within the stationary policies even without the ergodicity of transition probability.

Combining Theorems 3.1, 3.4, and 3.7, we obtain the main result immediately.

Theorem 3.8 *Assume (A1), (A2), and (A3). Then*

$$\lim_{\gamma\to 0}\lambda(\gamma)=v.$$

In addition, if (B1) holds, then

$$\lim_{\gamma\to 0}\lambda(x,\gamma)=v.$$

for any $x\in S$.

Remark 3.2 Recalling the proof of Theorem 3.1, we see that under (A1), (A2), and (A3), there exists $\mu(dx,da)\in\mathcal{P}(S\times A)$ such that

$$\int_{S\times A}\mu(dx,da)p(dy|x,a)=\int_A\mu(dy,da).$$

.

A sufficient condition for the risk-neutral average optimal reward $v(x)$ to be independent of the initial state x is the *uniform ergodicity* (2.7) (see Section 5.5 in [20]). We rewrite it as

(B2) There exists $\delta<1$ such that

$$\sup_{U\in\mathcal{B}(S)}\sup_{x,x'\in S}\sup_{a,a'\in A}[P(U|x,a)-P(U|x',a')]\leq\delta. \tag{3.15}$$

We provide brief proof for this.

Theorem 3.9 *Assume (A1), (A2), and (B2), then $v(x)$ is independent of the initial state x.*

Proof Define an operator T on $C(S)$ by

$$Tf(x) := \sup_{a\in A}\left[r(x,a)+\int_S p(dy|x,a)f(y)\right].$$

It is not hard to check that under (A1) and (A2), T maps $C(S)$ into itself. Let

$$\|f\|_{\text{sp}} := \sup_{x\in S} f(x) - \inf_{x'\in S} f(x')$$

be the *span norm* on $C(S)$. For $f_1,\ f_2 \in C(S),\ x_1, x_2 \in S$, and $\varepsilon > 0$, there exist $a_1, a_2 \in A$ such that

$$Tf_i(x_i) \le r(x_i,a_i) + \int_S p(dy|x_i,a_i)f_i(y) + \varepsilon,\ i = 1,2.$$

Therefore, we obtain that

$$\begin{aligned}
Tf_1(x_1) &- Tf_2(x_1) - [Tf_1(x_2) - Tf_2(x_2)] \\
&\le \left[r(x_1,a_1) + \int_S p(dy|x_1,a_1)f_1(y) + \varepsilon\right] - \left[r(x_1,a_1) + \int_S p(dy|x_1,a_1)f_2(y)\right] \\
&- \left[r(x_2,a_2) + \int_S p(dy|x_2,a_2)f_1(y)\right] + \left[r(x_2,a_2) + \int_S p(dy|x_2,a_2)f_2(y) + \varepsilon\right] \\
&= \int_S p(dy|x_1,a_1)\,[f_1(y) - f_2(y)] - \int_S p(dy|x_2,a_2)\,[f_1(y) - f_2(y)] + 2\varepsilon \\
&\le [p(E|x_1,a_1) - p(E|x_2,a_2)] \cdot \|f_1 - f_2\|_{\text{sp}} + 2\epsilon \le \delta_p \cdot \|f_1 - f_2\|_{\text{sp}} + 2\varepsilon,
\end{aligned}$$

where E in the second to the last inequality comes from the Hahn-Jordan decomposition of $p(\cdot|x_1,a_1) - p(\cdot|x_2,a_2)$. Letting $\varepsilon \to 0$, we see that T is a contraction mapping on $(C(S), \|\cdot\|_{\text{sp}})$. Thus, by the Banach Fixed-Point Theorem, there exist a unique (up to an additive constant) $f_0 \in C(S)$ such that $\|Tf_0 - f_0\|_{\text{sp}} = 0$, which means that $Tf_0(x) - f_0(x)$ is a constant v_0. It follows that for any $x \in S$,

$$\begin{aligned}
v_0 = \lim_{N\to\infty}\frac{1}{N}T^N f_0(x) &= \lim_{N\to\infty}\sup_{\pi\in\Pi_M}\frac{1}{N}E_x^\pi\left[\sum_{t=1}^N r(X_t,A_t)\right] \\
&\ge \sup_{\pi\in\Pi_M}\liminf_{N\to\infty}\frac{1}{N}E_x^\pi\left[\sum_{t=1}^N r(X_t,A_t)\right]. \qquad (3.16)
\end{aligned}$$

Since $r(x,a) + \int p(dy|x,a)f_0(y)$ is in $C(A)$ due to (A2), for each $x \in S$, there exists an $d_0(x) \in A$ such that

$$Tf_0(x) = r(x,d_0(x)) + \int p(dy|x,d_0(x))f_0(y).$$

Letting $\pi_0 = (d_0)^\infty$, we have

$$\begin{aligned}
\lim_{N\to\infty}\sup_{\pi\in\Pi_M}\frac{1}{N}E_x^\pi\left[\sum_{t=1}^N r(X_t,A_t)\right] &= \lim_{N\to\infty}\frac{1}{N}E_x^{\pi_0}\left[\sum_{t=1}^N r(X_t,A_t)\right] \\
&\le \sup_{\pi\in\Pi_M}\liminf_{N\to\infty}\frac{1}{N}E_x^\pi\left[\sum_{t=1}^N r(X_t,A_t)\right]. \qquad (3.17)
\end{aligned}$$

(3.16) and (3.17) imply that $v_0 = v(x)$ for any $x \in S$. □

Combining the above theorem with our main result (Theorem 3.8), we obtain the following

Corollary 3.10 *Assume (A1), (A2), (A3), (B1), and (B2). Then*

$$\lim_{\gamma \to 0} \lambda(x, \gamma) = v(x),$$

Furthermore, this limit is indeed independent of $x \in S$.

4 Risk-Sensitive Asymptotics of POMDP

This section applies the approach explored in the last section to the *partially observable Markov decision process* (POMDP). Francesca Albertini, Paolo Dai Pra, and Chiara Prior established such a limit in [1] for processes described by $X_{n+1} = f(X_n, A_n, W_n)$, $Y_n = h(X_n, V_n)$, where X_n, A_n, and Y_n denote the state, control, and observation, respectively, and W, V are i.i.d random variables. As for general POMDPs, Di Masi and Stettner proved the existence of the solution to the associated Bellman equation for cost-minimizing problems and stated that the limit as $\gamma \to 0$ had not been proven (see Remark 2 in [11]). However, the method in [11] can not be applied to reward-maximizing problems since it requires the operator induced from the Bellman equation to preserve the concavity. However, in this case, the operator is convexity-preserving. Nevertheless, we can prove that given the existence of a solution to the Bellman equation, (3.4) holds for the maximal reward of POMDPs.

A POMDP can be represented as a six-tuple $M_P = \langle S, A, O, p(\cdot|\cdot,\cdot), q(\cdot|\cdot), r(\cdot,\cdot)\rangle$. S is the space of real but unobserved states, A is the action space, and both are assumed to be compact metric spaces. The *observation space* O is a Polish space. Like those in MDPs, p is the transition kernel depending on actions, and $r : S \times A \to \mathbb{R}$ is the reward function. $q(\cdot|x)$ denotes the observation probability when the system is in state $x \in S$. As mentioned in the introduction, a widely used technique for analyzing a POMDP is to transfer it into a completely observable MDP. We will also adopt this technique which allows us to employ the analysis for MDPs in the last section to establish

$$\lim_{\gamma \to 0} \lambda_P(\gamma) = v_P,$$

where

$$\lambda_P(\gamma) := \sup_{\theta \in \mathcal{P}(S)} \sup_{\pi \in \Pi} \frac{1}{\gamma} \liminf_{N \to \infty} \frac{1}{N} \log E_\theta^\pi \left[e^{\sum_{t=1}^N \gamma r(X_t, A_t)} \right]$$

and

$$v_P := \sup_{\pi \in \Pi} \liminf_{N \to \infty} \frac{1}{N} E_\theta^\pi \left[\sum_{t=1}^N r(X_t, A_t) \right].$$

The exact definitions of the set Π of policies and the expectation operator E_θ^π will be given after introducing the assumptions needed in this section.

In order to use the measure transformation technique, we first assume that

(C0) There exists a $\Lambda \in \mathcal{P}(O)$ with full support such that for every $x \in S$, $q(\cdot|x) << \Lambda$.

The corresponding density function is also denoted by $q(\cdot|\cdot)$, i.e.,

$$q(dy|x) = q(y|x)\Lambda(dy), x \in S.$$

To apply Theorem 3.1 and Proposition 3.3, we make the following assumptions to guarantee that the reward and the transition probability of the transferred MDP satisfy (A1) and (A2).

(C1) $r(\cdot,\cdot) \in C(S \times A)$.

(C2) $q(y|\cdot) \in \mathrm{Lip}(S)$. There exist $q_m > 0$ and $q_M > 0$ such that $q(y|\cdot) \geq q_m$ and $\|q(y|\cdot)\|_L \leq q_M$ for every $y \in O$.

(C3) There exists $K_p > 0$ such that

$$\|p(\cdot|x,a) - p(\cdot|x',a')\|_{KR} \leq K_p\left[\rho_S(x,x') + \rho_A(a,a')\right]$$

for any $x, x' \in A$ and $a, a' \in S$, where $\|\cdot\|_{KR}$ denotes the Kantorovich-Rubinstein norm defined by (1.5) on $\mathcal{P}(S)$, ρ_S denotes the metric on S, and ρ_A denotes the metric on A.

To define a probability space and a stochastic process with the desired mechanism, let $\Omega_p = S \times (A \times S \times O)^\infty$ and $\mathcal{B}(\Omega_p)$ be the product Borel σ-field. Given a sample path $\omega = (x_1, a_1, x_2, y_2, a_3, ...) \in \Omega_P$, define $X_t := x_t$, $A_t := a_t$, $t \geq 1$, and $Y_t := y_t$, $t \geq 2$. At each time $t \in \mathbb{N}$, the system M_P occupies a state X_t, which is unobservable. When $t = 1$, we know the distribution of X_1 and then choose an action A_1. When $t \geq 2$, we can observe a signal Y_t generated by X_t and then choose an action A_t.

The optimal policy in POMDP is usually not a Markovian one due to the unavailability of real states when making decisions. Hence, we introduce the *observed-history-dependent policy*. Let $\mathbb{H}_t$ denote the set of *observed histories* up to time $t \in \mathbb{N}$. Then, $\mathbb{H}_1 = \mathcal{P}(S)$ (the set of all the initial state distributions) and $\mathbb{H}_{t+1} = \mathbb{H}_t \times A \times O$. An *observed-history-dependent decision rule* at time t is a stochastic kernel $d_t \in \mathcal{P}(A|\mathbb{H}_t)$, where $d_t(B|h_t)$ denote the probability for taking action in $B \subseteq A$ when observing $h_t = (\theta_1, a_1, y_2, a_2, y_3, ..., a_{t-1}, y_t) \in \mathbb{H}_t$. An observed-history-dependent policy π is a sequence of such decision rules at different times. Let D_t denote all the observed-history-dependent decision rules at time t, and $\Pi = \prod_{t=1}^{\infty} D_t$ denote all the observable-history-dependent policies. Given an initial distribution of states $\theta_1 \in \mathcal{P}(S)$ and a policy $\pi = (d_1, d_2, ...) \in \Pi$, a unique probability measure $\mathrm{P}^\pi_{\theta_0}$ and the corresponding expectation operator on $\mathcal{B}(\Omega_p)$ is defined by the Ionescu-Tulcea theorem, such that for each $t \geq 1$,

$$\begin{aligned}&\mathrm{P}^\pi_{\theta_1}(dx_1, da_1, dx_2, dy_2, ...da_{t-1}, dx_t, dy_t)\\ &= \theta_1(dx_1)\left(\prod_{i=1}^{t-1} d_i(da_i|h_i)\,p(dx_{i+1}|x_i,a_i)q(y_{i+1}|x_{i+1})\Lambda(dy_{i+1})\right).\end{aligned}$$

The risk-sensitive criterion introduced in Section 2 is to optimize

$$\lambda_P(\theta,\gamma) := \sup_{\pi\in\Pi}\frac{1}{\gamma}\liminf_{N\to\infty}\frac{1}{N}\log E^\pi_\theta\left[e^{\sum_{t=1}^N \gamma r(X_t,A_t)}\right], \qquad \theta \in \mathcal{P}(S), \gamma > 0. \tag{4.1}$$

while the typical optimal average reward is

$$v_P(\theta) := \sup_{\pi\in\Pi}\liminf_{N\to\infty}\frac{1}{N}E^\pi_\theta\left[\sum_{t=1}^N r(X_t,A_t)\right], \qquad \theta \in \mathcal{P}(S). \tag{4.2}$$

Let $\lambda_P(\gamma) := \sup_{\theta\in\mathcal{P}(S)}\lambda_P(\theta,\gamma)$ and $v_P := \sup_{\theta\in\mathcal{P}(S)} v_P(\theta)$. We intend to apply Theorem 3.1 and Proposition 3.3 to prove the risk-sensitive asymptotics

$$\lim_{\gamma\to 0}\lambda_P(\gamma) = v_P. \tag{4.3}$$

It has already been shown that optimal control of a POMDP M_P under the average reward criterion can be converted to the optimal control of a properly transferred MDP M_0 (see, e.g., Section 7.2.1 in [3], and Section 5.3, pp. 157–159 in [5]), where the new states are the conditional distributions of real states given the observed history. The transition law p_0 and one-step reward r_0 of M_0 are

$$\begin{aligned} p_0(d\theta'|\theta,a) &:= \int_O \Delta^{(0)}_{a,y,\theta}(d\theta')\left(\int_{S\times S} q(y|x')p(dx'|x,a)\theta(dx)\right)\Lambda(dy), \qquad (4.4)\\ r_0(\theta,a) &:= \int_S \theta(dx)r(x,a), \qquad \theta\in\mathcal{P}(S), a\in A, \end{aligned}$$

where $\Delta^{(0)}_{a,y,\theta}$ is a measure on $\mathcal{P}(S)$ defined by

$$\Delta^{(0)}_{a,y,\theta}(U) := \begin{cases} \delta\left\{\frac{T^*_{a,y,0}(\theta)}{T^*_{a,y,0}(\theta)(S)}\right\}(U), & T^*_{a,y,0}(\theta)(S)\neq 0\\ 0, & T^*_{a,y,0}(\theta)(S)=0 \end{cases}, \qquad U\subseteq\mathcal{P}(S), \qquad (4.5)$$

and $T^*_{a,y,0}$ is an operator on $\mathcal{M}^+(S)$ given by

$$T^*_{a,y,0}(\mu)(E) := \int_{S\times S}\mathbf{1}_E(x')q(y|x')p(dx'|x,a)\mu(dx), \quad \mu\in\mathcal{M}^+(S), E\in\mathcal{B}(S). \quad (4.6)$$

In the case of risk-sensitive control, the transformed MDP is slightly different from the typical form of average reward control (see [4]). We present the transformation procedure with our notations. Assuming (C0), (C1), (C2), and (C3), we derive the new state and the corresponding transition mechanism of the transferred MDP first. For $t\geq 1$, define two filters $\mathcal{F}_t$ and $\mathcal{G}_t$ by

$$\mathcal{F}_t=\sigma(A_1,Y_2,...,A_{t-1},Y_t), \qquad \mathcal{G}_t=\sigma(X_1,A_1,X_2,Y_2,...,A_{t-1},X_t,Y_t).$$

respectively. Let $H_t=(\theta_1,A_1,Y_2,...,A_{t-1},Y_t)$ denote the observed history up to time t. Since $\theta_1\in\mathcal{P}(S)$ is fixed, $\mathcal{F}_t=\sigma(H_t)$. Define another probability measure $\widetilde{P}^\pi_{\theta_1}$ on $\mathcal{B}(\Omega_P)$ by

$$\begin{aligned} &\widetilde{\mathrm{P}}^\pi_{\theta_1}(dx_1,da_1,dx_2,dy_2,...da_{t-1},dx_t,dy_t)\\ &\quad=\theta_1(dx_1)\left(\prod_{i=1}^{t-1}d_i(da_i|h_i)p(dx_{i+1}|x_i,a_i)\Lambda(dy_{i+1})\right), \end{aligned}$$

or equivalently,

$$\left.\frac{d\mathrm{P}^\pi_{\theta_1}}{d\widetilde{\mathrm{P}}^\pi_{\theta_1}}\right|_{\mathcal{G}_t}=\prod_{i=2}^{t}q(Y_i|X_i)=:R_t.$$

Since S, A, O are all Polish spaces, Ω_P is also Polish. Thus, the following regular conditional expectations on $(\Omega_P,\mathcal{B}(\Omega_P),\widetilde{P}^\pi_{\theta_1})$ for bounded Borel functions f on S

$$\widetilde{E}^\pi_{\theta_1}\left[f(X_t)e^{\gamma\sum_{i=1}^{t-1}r(X_i,A_i)}R_t\middle|H_t=h_t\right], t\geq 2$$

have regular versions. Therefore, we can define an $\mathcal{M}^+(S)$-valued process $\{\psi^{(\gamma)}_t\}$ by

$$\psi^{(\gamma)}_1:=\theta_1, \qquad \psi^{(\gamma)}_t(f):=\widetilde{E}^\pi_{\theta_1}\left[f(X_t)e^{\gamma\sum_{i=1}^{t-1}r(X_i,A_i)}R_t\middle|H_t\right], t\geq 2, \qquad (4.7)$$

where f is bounded and measurable on S. For $a \in A$, $y \in O$, define $T_{a,y,\gamma}$ as an operator on the space of bounded Borel functions on S by

$$T_{a,y,\gamma}(f)(x) := \int_S e^{\gamma r(x,a)} q(y|x') f(x') p(dx'|x,a), \quad x \in S$$

Notice that under $\widetilde{P}^{\pi}_{\theta_0}$, Y_{t+1} is independent of X_{s+1}, A_{s+1} and Y_s with $s \le t$, and X_{t+1} depends on $\sigma(G_t \cup \sigma(A_t, Y_{t+1}))$ only through X_t and A_t, we have

$$\begin{aligned}
&\widetilde{E}^{\pi}_{\theta_1}\left[f(X_{t+1}) e^{\gamma \sum_{i=1}^{t} r(X_i,A_i)} R_{t+1} \middle| H_{t+1} = h_{t+1}\right] \\
&= \widetilde{E}^{\pi}_{\theta_1}\left[\widetilde{E}^{\pi}_{\theta_1}\left[f(X_{t+1}) e^{\gamma r(X_t,A_t)} q(Y_{t+1}|X_{t+1}) \middle| \sigma(G_t \cup \sigma(A_t, Y_{t+1}))\right] e^{\gamma \sum_{i=1}^{t-1} r(X_i,A_i)} R_t \middle| H_{t+1} = h_{t+1}\right] \\
&= \widetilde{E}^{\pi}_{\theta_1}\left[\widetilde{E}^{\pi}_{\theta_1}\left[f(X_{t+1}) e^{\gamma r(X_t,A_t)} q(Y_{t+1}|X_{t+1}) \middle| \sigma(X_t, A_t, Y_{t+1})\right] e^{\gamma \sum_{i=1}^{t-1} r(X_i,A_i)} R_t \middle| H_t = h_t, A_t = a_t, Y_{t+1} = y_{t+1}\right] \\
&= \widetilde{E}^{\pi}_{\theta_1}\left[T_{A_t,Y_{t+1},\gamma}(f)(X_t) e^{\gamma \sum_{i=1}^{t-1} r(X_i,A_i)} R_t \middle| H_t = h_t, A_t = a_t, Y_{t+1} = y_{t+1}\right] \\
&= \widetilde{E}^{\pi}_{\theta_1}\left[T_{a_t,y_{t+1},\gamma}(f)(X_t) e^{\gamma \sum_{i=1}^{t-1} r(X_i,A_i)} R_t \middle| H_t = h_t\right].
\end{aligned}$$

Therefore, we have

$$\psi^{(\gamma)}_{t+1}(f) = \psi^{(\gamma)}_t\left(T_{A_t,Y_{t+1},\gamma}(f)\right) = T^*_{A_t,Y_{t+1},\gamma}(\psi^{(\gamma)}_t)(f), \quad \widetilde{P}^{\pi}_{\theta_1}\ a.s.,$$

where $T^*_{a,y,\gamma}$ is the adjoint operator of $T_{a,y,\gamma}$ defined on $\mathcal{M}^+(S)$ by

$$\begin{aligned}
T^*_{a,y,\gamma}(\mu)(E) &:= \int_S T_{a,y,\gamma}(\mathbf{1}_E) d\mu \\
&= \int_{S\times S} \mathbf{1}_E(x') e^{\gamma r(x,a)} q(y|x') p(dx'|x,a) \mu(dx), \quad \mu \in \mathcal{M}^+(S), E \in \mathcal{B}(S).
\end{aligned}$$

From (C1), we know that $\psi^{(\gamma)}_t(S)$ is finite and strictly positive. Hence, we can define a new state process $\{\theta^{(\gamma)}_t\}$ taking values in $\mathcal{P}(S)$ by

$$\theta^{(\gamma)}_t := \frac{\psi^{(\gamma)}_t}{\psi^{(\gamma)}_t(S)}. \tag{4.8}$$

We call $\{\theta^{(\gamma)}_t\}$ the *information state process* since it represents the cumulative-reward-weighted conditional distribution of the real state given the observed history. The information state $\theta^{(\gamma)}_1$ at time $t = 1$ is still θ_1. Notice that the operator $T^*_{a,y,\gamma}$ is positively 1-homogeneous. We have

$$\theta^{(\gamma)}_{t+1} = \frac{T^*_{A_t,Y_{t+1},\gamma}(\psi^{(\gamma)}_t)}{T^*_{A_t,Y_{t+1},\gamma}(\psi^{(\gamma)}_t)(S)} = \frac{T^*_{A_t,Y_{t+1},\gamma}(\theta^{(\gamma)}_t)}{T^*_{A_t,Y_{t+1},\gamma}(\theta^{(\gamma)}_t)(S)}, \tag{4.9}$$

which implies the transition mechanism of $\theta^{(\gamma)}_t$.
As for the new reward function, we define

$$G_\gamma(a, y, \theta) := \log\left[T^*_{a,y,\gamma}(\theta)(S)\right]. \tag{4.10}$$

Then we have

$$\begin{aligned}
\psi^{(\gamma)}_t(S) &= T^*_{A_{t-1},Y_t,\gamma}(\psi^{(\gamma)}_{t-1})(S) = T^*_{A_{t-1},Y_t,\gamma}(\theta^{(\gamma)}_{t-1})(S) \cdot \psi^{(\gamma)}_{t-1}(S) \\
&= e^{G_\gamma(A_{t-1},Y_t,\theta^{(\gamma)}_{t-1})} \cdot \psi^{(\gamma)}_{t-1}(S), \quad t \ge 2
\end{aligned} \tag{4.11}$$

It then follows that

$$\begin{aligned} E^{\pi}_{\theta_1}\left[e^{\gamma\sum_{i=1}^{t} r(X_i,A_i)}\right] &= \widetilde{E}^{\pi}_{\theta_1}\left[e^{\gamma\sum_{i=1}^{t} r(X_i,A_i)}R_t\right] = \widetilde{E}^{\pi}_{\theta_1}\left[\widetilde{E}^{\pi}_{\theta_1}\left[e^{\gamma\sum_{i=1}^{t} r(X_i,A_i)}R_t\middle|\mathcal{F}_t\right]\right] \\ &= \widetilde{E}^{\pi}_{\theta_1}\left[\psi^{(\gamma)}_{t+1}(S)\right] \\ &= \widetilde{E}^{\pi}_{\theta_1}\left[e^{\sum_{i=1}^{t} G_\gamma(A_i,Y_{i+1},\theta_i^{(\gamma)})}\cdot\psi^{(\gamma)}_1(S)\right] \\ &= \widetilde{E}^{\pi}_{\theta_1}\left[e^{\sum_{i=1}^{t} G_\gamma(A_i,Y_{i+1},\theta_i^{(\gamma)})}\cdot\theta^{(\gamma)}_1(S)\right] \\ &= \widetilde{E}^{\pi}_{\theta_1}\left[e^{\sum_{i=1}^{t} G_\gamma(A_i,Y_{i+1},\theta_i^{(\gamma)})}\right]. \end{aligned} \tag{4.12}$$

Hence, we can consider G_γ as the new reward. Now, we can transfer M_P to the following completely observable model M'_γ with state space $\mathcal{P}(S)$ and action space A:

1. The initial information state is θ_1.
2. At time t, given the current information state $\theta_t^{(\gamma)}$, we take action A_t according to a pre-specified policy. Then, the system generates Y_{t+1}, which is independent of $\theta_s^{(\gamma)}$, A_s, and Y_s with $s \le t$, and distributed according to the law Λ. The next information state $\theta_{t+1}^{(\gamma)}$ is determined by
$$\theta_{t+1}^{(\gamma)} = \frac{T^*_{A_t,Y_{t+1},\gamma}(\theta_t^{(\gamma)})}{T^*_{A_t,Y_{t+1},\gamma}(\theta_t^{(\gamma)})(S)}.$$
3. Once Y_{t+1} is generated, the next state $\theta_{t+1}^{(\gamma)}$ is then obtained according to (4.9), and simultaneously the system generates one-step reward $G_\gamma(A_t, Y_{t+1}, \theta_t^{(\gamma)})$.

Remark 4.1 The one-step reward G_γ in (4.10) depends not only on the state θ and the action A but also on an independent signal Y under $\widetilde{P}^{\pi}_{\theta_1}$, which is slightly different from the typical form. We make the following changes to have a reward and a transition probability in a standard form in which assumptions (A1) and (A2) can be verified.

Define the completely observable Markov decision model M_γ with transition law p_γ and one-step reward r_γ by

$$\begin{aligned} p_\gamma(d\theta'|\theta,a) &:= \frac{1}{\int_O T^*_{a,y,\gamma}(\theta)(S)\Lambda(dy)}\left(\int_O \Delta^{(\gamma)}_{a,y,\theta}(d\theta')T^*_{a,y,\gamma}(\theta)(S)\Lambda(dy)\right) \\ &= \frac{1}{\int_S e^{\gamma r(x,a)}\theta(dx)}\left[\int_O \Delta^{(\gamma)}_{a,y,\theta}(d\theta')\Lambda(dy)\int_{S\times S} e^{\gamma r(x,a)}q(y|x')p(dx'|x,a)\theta(dx)\right], \\ r_\gamma(\theta,a) &:= \log\left(\int_O T^*_{a,y,\gamma}(\theta)(S)\Lambda(dy)\right) = \log\left(\int_S e^{\gamma r(x,a)}\theta(dx)\right). \end{aligned} \tag{4.13}$$

where $\Delta^{(\gamma)}_{a,y,\theta}$ is a measure on $\mathcal{P}(S)$ defined by

$$\Delta^{(\gamma)}_{a,y,\theta}(U) := \begin{cases} \delta\left\{\frac{T^*_{a,y,\gamma}(\theta)}{T^*_{a,y,\gamma}(\theta)(S)}\right\}(U), & T^*_{a,y,\gamma}(\theta)(S)\neq 0 \\ 0, & T^*_{a,y,\gamma}(\theta)(S)=0 \end{cases}, \qquad U\subseteq\mathcal{P}(S). \tag{4.14}$$

Use $E^{\pi}_{\gamma,\theta}$ to denote the expectation operator with respect to the transition probability p_γ with initial state θ and policy π. Since M_γ is an MDP, we consider the Markov policies of

M_γ, which consists of decision rules by choosing actions only through the current information state θ. Such policies are also called *separated policies* of the original model M_P (see, e.g., [19]). We let D_S denote the set of all the Markov decision rules of M_γ and $\Pi_S = (D_S)^\infty$. Then for $\pi_S \in \Pi_S$, by direct calculation, we have

$$\widetilde{E}_\theta^{\pi_S}\left[e^{\sum_{t=1}^N G_\gamma(A_t,Y_{t+1},\theta_t^{(\gamma)})} f(\theta_{t+1}^{(\gamma)})\right] = E_{\gamma,\theta}^{\pi_S}\left[e^{\sum_{t=1}^N r_\gamma(\theta_t^{(\gamma)},A_t)} f(\theta_{t+1}^{(\gamma)})\right] \tag{4.15}$$

for any bounded Borel function f on $\mathcal{P}(S)$. Π_S is a subset of Π since θ_t is $\mathcal{F}_t$-adaptive. Hence, from (4.12) and (4.15), we have for $\pi_S \in \Pi_S$,

$$\begin{aligned}\liminf_{N\to\infty}\frac{1}{N}\log E_\theta^{\pi_S}\left[e^{\sum_{t=1}^N \gamma r(X_t,A_t)}\right] &= \liminf_{N\to\infty}\frac{1}{N}\log \widetilde{E}_{\theta_1}^{\pi_S}\left[e^{\sum_{i=1}^N G_\gamma(A_i,Y_{i+1},\theta_i^{(\gamma)})}\right]\\ &= \liminf_{N\to\infty}\frac{1}{N}\log E_{\gamma,\theta}^{\pi_S}\left[e^{\sum_{t=1}^N r_\gamma(\theta_t^{(\gamma)},A_t)}\right].\end{aligned}$$

We define λ_S as the optimal value of separated policies, which is

$$\begin{aligned}\lambda_S(\theta,\gamma) &:= \sup_{\pi_S\in\Pi_S}\frac{1}{\gamma}\liminf_{N\to\infty}\frac{1}{N}\log E_\theta^{\pi}\left[e^{\sum_{t=1}^N \gamma r(X_t,A_t)}\right]\\ &= \sup_{\pi_S\in\Pi_S}\frac{1}{\gamma}\liminf_{N\to\infty}\frac{1}{N}\log E_{\gamma,\theta}^{\pi_S}\left[e^{\sum_{t=1}^N r_\gamma(\theta_t^{(\gamma)},A_t)}\right].\end{aligned} \tag{4.16}$$

We will show that under (C0), (C1), (C2), and (C3), $\lambda_S = \lambda_P$ and (4.3) hold if there exists $K > 0$ such that for every $\gamma \in (0, K)$, the Bellman equation

$$\rho_\gamma f_\gamma(\theta) = \sup_{a\in A}\int_{\mathcal{P}(S)} f_\gamma(\theta')e^{r_\gamma(\theta,a)}p_\gamma(d\theta'|\theta,a) \tag{4.17}$$

has a solution $\rho_\gamma > 0$ and $f_\gamma \in C(\mathcal{P}(S))$ with $f_\gamma >> 0$.
We first verify that r_γ satisfies (A1) and p_γ satisfies (A2), which implies that the corresponding operator $L_P^{(\gamma)}$ on $C(\mathcal{P}(S))$, defined by

$$L_P^{(\gamma)}f(\theta) := \sup_{a\in A}\int_{\mathcal{P}(S)} f(\theta')e^{r_\gamma(\theta,a)}p_\gamma(d\theta'|\theta,a), \tag{4.18}$$

maps $C(\mathcal{P}(S))$ into itself.

Lemma 4.1 *Assume (C1). Then $r_\gamma(\theta,a)$ is continuous in (θ,a).*

Proof For $\theta_n \to \theta$ weakly and $a_n \to a$, we have

$$\begin{aligned}\left|e^{r_\gamma(\theta_n,a_n)} - e^{r_\gamma(\theta,a)}\right| \le{}& \int_S \left|e^{\gamma r(x,a_n)} - e^{\gamma r(x,a)}\right|\theta_n(dx)\\ &+ \left|\int_S e^{\gamma r(x,a)}\theta_n(dx) - \int_S e^{\gamma r(x,a)}\theta(dx)\right|\end{aligned}$$

The second term tends to 0 due to the weak convergence of θ_n while the first term tends to 0 because of the uniform continuity of $r(\cdot,\cdot)$ on the compact set $S \times A$. Hence, $e^{r_\gamma(\theta,a)}$ is continuous in (θ,a). Notice that $e^{r_\gamma} \ge e^{r_m} > 0$, we see that $r_\gamma(\theta,a)$ is continuous in (θ,a). □

Lemma 4.2 *Assume (C0), (C1), (C2), and (C3). Then $(\theta,a) \mapsto \int_{\mathcal{P}(S)} p(d\theta'|\theta,a)f(\theta)$ is continuous in (θ,a) for $f \in C(\mathcal{P}(S))$.*

Proof Recall that r_M and r_m are the supremum and infimum of r, respectively. Fix $f \in C(\mathcal{P}(S))$. From (4.13) and direct calculation, we see that

$$\int_{\mathcal{P}(S)} p(d\theta'|\theta,a) f(\theta') = e^{-r_\gamma(\theta,a)} \int_O T^*_{a,y,\gamma}(\theta)(S) f\left(\frac{T^*_{a,y,\gamma}(\theta)}{T^*_{a,y,\gamma}(\theta)(S)}\right) \Lambda(dy),$$

where $e^{-r_\gamma(\theta,a)}$ is continuous in (θ,a) by Lemma 4.1. It suffices to show that

$$\int_O T^*_{a,y,\gamma}(\theta)(S) f\left(\frac{T^*_{a,y,\gamma}(\theta)}{T^*_{a,y,\gamma}(\theta)(S)}\right) \Lambda(dy) \tag{4.19}$$

is continuous in (θ,a). Once $\left\{(\theta,a) \mapsto T^*_{a,y,\gamma}(\theta),\ y \in O\right\}$ is equicontinuous, then from the uniform continuity of $f \in C(\mathcal{P}(S))$ ($\mathcal{P}(S)$ is compact since S is compact) and the fact that

$$T^*_{a,y,\gamma}(\theta)(S) = \int_S \theta(dx) e^{\gamma r(x,a)} \int_S q(y|x') p(dx'|x,a_n) \geq q_m e^{\gamma r_m} > 0$$

uniformly, we can see that

$$\left\{(\theta,a) \mapsto T^*_{a,y,\gamma}(\theta)(S) f\left(\frac{T^*_{a,y,\gamma}(\theta)}{T^*_{a,y,\gamma}(\theta)(S)}\right),\ y \in O\right\}$$

is equicontinuous, which proves this lemma.

Now we prove that $\left\{(\theta,a) \mapsto T^*_{a,y,\gamma}(\theta),\ y \in O\right\}$ is equicontinuous. Fix $\theta \in \mathcal{P}(S)$ and $a \in A$. For $\theta_n \to \theta$ weakly and $a_n \to a$, we have

$$\begin{aligned}\sup_{y\in O} \left\| T^*_{a_n,y,\gamma}(\theta_n) - T^*_{a,y,\gamma}(\theta) \right\|_{KR} \leq & \sup_{y\in O} \left\| T^*_{a_n,y,\gamma}(\theta_n) - T^*_{a,y,\gamma}(\theta_n) \right\|_{KR} \\ & + \sup_{y\in O} \left\| T^*_{a,y,\gamma}(\theta_n) - T^*_{a,y,\gamma}(\theta) \right\|_{KR}.\end{aligned}$$

The first term $\left\| T^*_{a_n,y,\gamma}(\theta_n) - T^*_{a,y,\gamma}(\theta_n) \right\|_{KR}$ is

$$\begin{aligned}&\sup_{\substack{g\in \mathrm{Lip}(S)\\ \|g\|_L\leq 1}} \int_S \theta_n(dx) \left(e^{\gamma r(x,a_n)} \int_S g(x') q(y|x') p(dx'|x,a_n) - e^{\gamma r(x,a)} \int_S g(x') q(y|x') p(dx'|x,a) \right) \\ &\leq \sup_{\substack{g\in \mathrm{Lip}(S)\\ \|g\|_L\leq 1}} \int_S \theta_n(dx) e^{\gamma r(x,a_n)} \left(\int_S g(x') q(y|x') p(dx'|x,a_n) - \int_S g(x') q(y|x') p(dx'|x,a) \right) \\ &\quad + \sup_{\substack{g\in \mathrm{Lip}(S)\\ \|g\|_L\leq 1}} \int_S \theta_n(dx) \left(e^{\gamma r(x,a_n)} - e^{\gamma r(x,a)} \right) \left(\int_S g(x') q(y|x') p(dx'|x,a) \right)\end{aligned}$$

Notice that $\|g(\cdot)\|_L \leq 1$ and $\|q(y|\cdot)\|_L \leq q_M$ imply $\|g(\cdot)q(y|\cdot)\|_L \leq 2q_M$. Thus, by (C1), (C2), and (C3), we have that

$$\begin{aligned}&\int_S \theta_n(dx) e^{\gamma r(x,a_n)} \left(\int_S g(x') q(y|x') p(dx'|x,a_n) - \int_S g(x') q(y|x') p(dx'|x,a) \right) \\ &\leq 2q_M \cdot \int_S \theta_n(dx) e^{\gamma r(x,a_n)} \|p(\cdot|x,a_n) - p(\cdot|x,a)\|_{KR} \leq 2q_M e^{\gamma r_M} K_p \cdot \rho_A(a_n,a)\end{aligned}$$

and

$$\int_S \theta_n(dx)\left(e^{\gamma r(x,a_n)}-e^{\gamma r(x,a)}\right)\left(\int_S g(x')q(y|x')p(dx'|x,a)\right)\le q_M\cdot\int_S\left|e^{\gamma r(x,a_n)}-e^{\gamma r(x,a)}\right|\theta_n(dx).$$

holds for every $y\in O$. Hence, from the uniform continuity of $r(\cdot,\cdot)$, we know that

$$\lim_{n\to\infty}\sup_{y\in O}\left\|T^*_{a_n,y,\gamma}(\theta_n)-T^*_{a,y,\gamma}(\theta_n)\right\|_{KR}=0. \tag{4.20}$$

The second term $\left\|T^*_{a,y,\gamma}(\theta_n)-T^*_{a,y,\gamma}(\theta)\right\|_{KR}$ is

$$\sup_{\substack{g\in\mathrm{Lip}(S)\\ \|g\|_L\le 1}}\int_S\theta_n(dx)e^{\gamma r(x,a)}\int_S g(x')q(y|x')p(dx'|x,a)-\int_S\theta(dx)e^{\gamma r(x,a)}\int_S g(x')q(y|x')p(dx'|x,a).$$

Define two finite measures $\mu^a_n(dx):=\theta_n(dx)e^{\gamma r(x,a)}$ and $\mu^a(dx):=\theta(dx)e^{\gamma r(x,a)}$. Then from the continuity of $r(\cdot,a)$, we know that $\mu^a_n\to\mu^a$ weakly, which means that $\lim_{n\to\infty}\|\mu^a_n-\mu^a\|_{KR}=0$. Thus, we only need to verify that as a function of x, if $\|g\|_L\le 1$, then the Lipschitz constant $\left\|\int_S q(y|x')p(dx'|x,a)g(x')\right\|_L$ is bounded by a constant which is independent of y.
First, it is obvious that $\left|\int_S q(y|x')p(dx'|x,a)g(x')\right|\le q_M$. As for the Lipschitz constant, we have for $x_1,x_2\in S$,

$$\begin{aligned}&\left|\int_S g(x')q(y|x')p(dx'|x_1,a)-\int_S g(x')q(y|x')p(dx'|x_2,a)\right|\\ &\le\|g(\cdot)q(y|\cdot)\|_L\cdot\|p(\cdot|x_1,a)-p(\cdot|x_2,a)\|_{KR}\le 2q_MK_p\cdot\rho(x_1,x_2).\end{aligned}$$

Consequently,

$$\lim_{n\to\infty}\sup_{y\in O}\left\|T^*_{a,y,\gamma}(\theta_n)-T^*_{a,y,\gamma}(\theta)\right\|_{KR}\le\lim_{n\to\infty}\sup_{y\in O}\max\{q_M,2q_MK_p\}\cdot\|\mu^a_n-\mu^a\|_{KR}=0. \tag{4.21}$$

(4.20) and (4.21) show that $\left\{(\theta,a)\mapsto T^*_{a,y,\gamma}(\theta),\ y\in O\right\}$ is equicontinuous and the lemma is proved. □

Since r_γ satisfies (A1) and p_γ satisfies (A2), we can apply Proposition 3.3 to obtain the variational formula for λ_S.

Theorem 4.3 *Assume (C0), (C1), (C2), and (C3). If there exist $\rho_\gamma>0$ and $f_\gamma\in C(\mathcal{P}(S))$ with $f_\gamma>>0$ satisfying $\rho_\gamma f_\gamma=L^{(\gamma)}_P f_\gamma$, then*

$$\lambda_S(\gamma)=\sup_{\beta\in\mathcal{I}_P}\left\{\int_{\mathcal{P}(S)\times A}\left[\frac{1}{\gamma}r_\gamma(\theta,a)-\frac{1}{\gamma}D(\beta_2(\cdot|\theta,a)\|p_\gamma(\cdot|\theta,a))\right]\beta'(d\theta,da)\right\} \tag{4.22}$$

holds, where

$$\mathcal{I}_P:=\{\beta\in\mathcal{P}(\mathcal{P}(S)\times A\times\mathcal{P}(S)):\beta(\mathcal{P}(S),A,d\theta)=\beta(d\theta,A,\mathcal{P}(S))\}, \tag{4.23}$$

and the notations β_2 and β' are defined by (3.3).

Proof Lemmas 4.1 and 4.2 imply that M_γ satisfies (A1) and (A2). Notice that $\mathcal{P}(S)$ is compact, by Proposition 3.3, we have

$$\gamma\lambda_S(\gamma) = \sup_{\beta\in\mathcal{I}_P}\left\{\int_{\mathcal{P}(S)\times A}\left[r_\gamma(\theta,a) - D(\beta_2(\cdot|\theta,a)\|p_\gamma(\cdot|\theta,a))\right]\beta'(d\theta,da)\right\}.$$

Hence, (4.22) holds. □

By Hölder's inequality, for $\gamma \geq \gamma' > 0$, we have $\lambda_P(\gamma) \geq \lambda_P(\gamma') \geq \upsilon_P$. Therefore, $\lambda_P(\gamma)$ is non-decreasing in γ and $\lim_{\gamma\to 0}\lambda_P(\gamma) \geq \upsilon_P$. To get the desired assertion, we need that $\lambda_S = \lambda_P$.

Theorem 4.4 *Assume (C0), (C1), (C2), and (C3). If there exist $\rho_\gamma > 0$ and $f_\gamma \in C(\mathcal{P}(S))$ with $f_\gamma >> 0$ satisfying that $\rho_\gamma f_\gamma = L_P^{(\gamma)} f_\gamma$, then $\lambda_S(\gamma) = \lambda_P(\gamma)$.*

Proof We first show that it is true in the finite-horizon case. Fix $N > 0$ and an observed-history-dependent policy $\pi = (d_1, ..., d_N, ...)$. By (4.11), we have that

$$\psi_{t+1}^{(\gamma)}(S) = e^{G_\gamma(A_t,Y_{t+1},\theta_t^{(\gamma)})}\cdot\psi_t^{(\gamma)}(S), \quad t \geq 1.$$

Hence

$$\begin{aligned}E_\theta^\pi\left[e^{\gamma\sum_{t=1}^N r(X_t,A_t)}\right] &= \tilde{E}_\theta^\pi\left[e^{\gamma\sum_{t=1}^N r(X_t,A_t)}R_{N+1}\right] = \tilde{E}_\theta^\pi\left[\psi_{N+1}^{(\gamma)}(S)\right]\\ &= \tilde{E}_\theta^\pi\left[\psi_N^{(\gamma)}(S)e^{G_\gamma(A_N,Y_{N+1},\theta_N^{(\gamma)})}\right].\end{aligned}$$

For every $\psi \in \mathcal{M}^+(S)$ and $\varepsilon > 0$, there exists $d_N^*(\psi) \in A$ such that

$$\begin{aligned}&\tilde{E}_\theta^\pi\left[\psi_N^{(\gamma)}(S)e^{G_\gamma(d_N^*(\psi),Y_{N+1},\theta_N^{(\gamma)})}\middle|\psi_N^{(\gamma)} = \psi\right]\\ &\geq \sup_{a\in A}\tilde{E}_\theta^\pi\left[\psi_N^{(\gamma)}(S)e^{G_\gamma(a,Y_{N+1},\theta_N^{(\gamma)})}\middle|\psi_N^{(\gamma)} = \psi\right] - \frac{\varepsilon}{N}e^{(N-1)(r_m-r_M)}.\end{aligned}$$

d_N^* actually depends only on $\theta = \frac{\psi}{\psi(S)}$. Given $\psi' \neq \psi$ with $\frac{\psi'}{\psi'(S)} = \theta = \frac{\psi}{\psi(S)}$, we can assume that

$$e^{(N-1)\gamma r_m} \leq \psi'(S) \quad \text{and} \quad \psi(S) \leq e^{(N-1)\gamma r_M}$$

since $e^{(N-1)\gamma r_m} \leq \psi_N^{(\gamma)}(S) \leq e^{(N-1)\gamma r_M}$. Thus, for any $a \in A$, we have

$$\begin{aligned}&\tilde{E}_\theta^\pi\left[\psi_N^{(\gamma)}(S)e^{G_\gamma(a,Y_{N+1},\theta_N^{(\gamma)})}\middle|\psi_N^{(\gamma)} = \psi'\right] = \psi'(S)\int_O e^{G_\gamma(a,y,\theta)}\Lambda(dy)\\ &= \frac{\psi'(S)}{\psi(S)}\cdot\psi(S)\int_O e^{G_\gamma(a,y,\theta)}\Lambda(dy) = \frac{\psi'(S)}{\psi(S)}\cdot\tilde{E}_\theta^\pi\left[\psi_N^{(\gamma)}(S)e^{G_\gamma(a,Y_{N+1},\theta_N^{(\gamma)})}\middle|\psi_N^{(\gamma)} = \psi\right].\end{aligned}$$

Hence,

$$\begin{aligned}&\tilde{E}_\theta^\pi\left[\psi_N^{(\gamma)}(S)e^{G_\gamma(d_N^*(\psi),Y_{N+1},\theta_N^{(\gamma)})}\middle|\psi_N^{(\gamma)} = \psi'\right]\\ &\geq \frac{\psi'(S)}{\psi(S)}\cdot\sup_{a\in A}\tilde{E}_\theta^\pi\left[\psi_N^{(\gamma)}(S)e^{G_\gamma(a,Y_{N+1},\theta_N^{(\gamma)})}\middle|\psi_N^{(\gamma)} = \psi\right] - \frac{\psi'(S)}{\psi(S)}\cdot\frac{\varepsilon}{N}e^{(N-1)(r_m-r_M)}\\ &= \sup_{a\in A}\tilde{E}_\theta^\pi\left[\psi_N^{(\gamma)}(S)e^{G_\gamma(a,Y_{N+1},\theta_N^{(\gamma)})}\middle|\psi_N^{(\gamma)} = \psi'\right] - \frac{\psi'(S)}{\psi(S)}\cdot\frac{\varepsilon}{N}e^{(N-1)(r_m-r_M)}\\ &\geq \sup_{a\in A}\tilde{E}_\theta^\pi\left[\psi_N^{(\gamma)}(S)e^{G_\gamma(a,Y_{N+1},\theta_N^{(\gamma)})}\middle|\psi_N^{(\gamma)} = \psi'\right] - \frac{\varepsilon}{N}.\end{aligned}$$

Thus, we see that for every $\psi \in \mathcal{M}^+(S)$, there exists $d_N^*(\theta) \in A$ depending only on $\theta = \frac{\psi}{\psi(S)}$ such that

$$\tilde{E}_\theta^\pi \left[\psi_N^{(\gamma)}(S) e^{G_\gamma\left(d_N^*\left(\theta_N^{(\gamma)}\right), Y_{N+1}, \theta_N^{(\gamma)}\right)} \middle| \psi_N^{(\gamma)} = \psi \right] \geq \sup_{a \in A} \tilde{E}_\theta^\pi \left[\psi_N^{(\gamma)}(S) e^{G_\gamma\left(a, Y_{N+1}, \theta_N^{(\gamma)}\right)} \middle| \psi_N^{(\gamma)} = \psi \right] - \frac{\varepsilon}{N}.$$

Now modify the policy π by simply replacing d_N with d_N^* to get a new policy $\pi_N^* = (d_1, ..., d_{N-1}, d_N^*, ...)$, we obtain that

$$\tilde{E}_\theta^{\pi_N^*} \left[\psi_N^{(\gamma)}(S) e^{G_\gamma\left(A_N, Y_{N+1}, \theta_N^{(\gamma)}\right)} \right] \geq \tilde{E}_\theta^\pi \left[\psi_N^{(\gamma)}(S) e^{G_\gamma\left(A_N, Y_{N+1}, \theta_N^{(\gamma)}\right)} \right] - \frac{\varepsilon}{N}.$$

Continue this procedure by successively replacing d_j with d_j^* for $j = N-1, \cdots, 1$, with each d_j^* depending only on the information state and satisfying that

$$\tilde{E}_\theta^{\pi_{n-1}^*} \left[\psi_{n-1}^{(\gamma)}(S) e^{\sum_{j=n-1}^N G_\gamma\left(A_j, Y_{j+1}, \theta_j^{(\gamma)}\right)} \right] \geq \tilde{E}_\theta^{\pi_n^*} \left[\psi_{n-1}^{(\gamma)}(S) e^{\sum_{j=n-1}^N G_\gamma\left(A_j, Y_{j+1}, \theta_j^{(\gamma)}\right)} \right] - \frac{\varepsilon}{N}.$$

where $\pi_n^* = (d_1, ..., d_{n-1}, d_n^*, ..., d_N^*, ...)$. In this way, with $\pi_1^* = (d_1^*, d_2^*, ...d_N^*, ...)$, we obtain that

$$\tilde{E}_\theta^{\pi_1^*} \left[e^{\sum_{t=1}^N G_\gamma\left(A_t, Y_{t+1}, \theta_t^{(\gamma)}\right)} \right] \geq \tilde{E}_\theta^\pi \left[e^{\sum_{t=1}^N G_\gamma\left(A_t, Y_{t+1}, \theta_t^{(\gamma)}\right)} \right] - \varepsilon.$$

Noticing that the decision rules after time N is irrelevant and recalling (4.12), we have proved that

$$\sup_{\pi \in \Pi_S} E_\theta^\pi \left[e^{\gamma \sum_{t=1}^N r(X_t, A_t)} \right] \geq \sup_{\pi \in \Pi} E_\theta^\pi \left[e^{\gamma \sum_{t=1}^N r(X_t, A_t)} \right] - \varepsilon.$$

for any $\varepsilon > 0$. Obviously,

$$\sup_{\pi \in \Pi_S} E_\theta^\pi \left[e^{\gamma \sum_{t=1}^N r(X_t, A_t)} \right] \leq \sup_{\pi \in \Pi} E_\theta^\pi \left[e^{\gamma \sum_{t=1}^N r(X_t, A_t)} \right].$$

Consequently,

$$\sup_{\pi \in \Pi_S} E_\theta^\pi \left[e^{\gamma \sum_{t=1}^N r(X_t, A_t)} \right] = \sup_{\pi \in \Pi} E_\theta^\pi \left[e^{\gamma \sum_{t=1}^N r(X_t, A_t)} \right].$$

Letting $N \to \infty$, we obtain that

$$\lambda_S(\gamma) \leq \lambda_P(\gamma) \leq \sup_{\theta \in \mathcal{P}(S)} \frac{1}{\gamma} \liminf_{N \to \infty} \sup_{\pi \in \Pi} \frac{1}{N} E_\theta^\pi \left[e^{\gamma \sum_{t=1}^N r(X_t, A_t)} \right] \tag{4.24}$$

$$= \sup_{\theta \in \mathcal{P}(S)} \frac{1}{\gamma} \liminf_{N \to \infty} \sup_{\pi \in \Pi_S} \frac{1}{N} E_\theta^\pi \left[e^{\gamma \sum_{t=1}^N r(X_t, A_t)} \right]. \tag{4.25}$$

Thus, it suffices to verify that

$$\lambda_S(\gamma) \geq \sup_{\theta \in \mathcal{P}(S)} \frac{1}{\gamma} \liminf_{N \to \infty} \sup_{\pi \in \Pi_S} \frac{1}{N} E_\theta^\pi \left[e^{\gamma \sum_{t=1}^N r(X_t, A_t)} \right]. \tag{4.26}$$

Recall that by assumption, we have $\rho_\gamma f_\gamma = L_P^{(\gamma)} f_\gamma$ with $\rho_\gamma > 0$ and $f_\gamma >> 0$. From Lemmas 4.1 and 4.2, we know that

$$\int_{\mathcal{P}(S)} f_\gamma(\theta') e^{r_\gamma(\theta, a)} p_\gamma(d\theta' | \theta, a)$$

is continuous in a. Due to the compactness of A, for every $\theta \in \mathcal{P}(S)$, there exists $d^*(\theta) \in A$ such that

$$\rho_\gamma f_\gamma(\theta) = L_P^{(\gamma)} f_\gamma(\theta) = \int_{\mathcal{P}(S)} f_\gamma(\theta') e^{r_\gamma(\theta, d^*(\theta))} p_\gamma(d\theta' | \theta, d^*(\theta)).$$

Let $\pi^* = (d^*)^\infty$. Similarly to the argument used to derive (2.17) in the proof of Theorem 2.2, we see that

$$\lim_{N\to\infty} \sup_{\pi\in\Pi_S} \frac{1}{N} E_\theta^\pi \left[e^{\gamma \sum_{t=1}^N r(X_t, A_t)} \right] = \log(\rho_\gamma) = \lim_{N\to\infty} \frac{1}{N} E_\theta^{\pi^*} \left[e^{\gamma \sum_{t=1}^N r(X_t, A_t)} \right] \le \gamma \lambda_S(\theta, \gamma).$$

Hence, the inequalities in (4.24) are equalities, which gives that $\lambda_S(\gamma) = \lambda_P(\gamma)$. □

With Theorems 4.3 and 4.4, using a similar argument as in the proof of Theorem 3.1, we can now extend the risk-sensitive asymptotics to POMDPs, which is the main result of this section.

Theorem 4.5 *Assume (C0), (C1), (C2), and (C3). If there exists $K > 0$ such that for every $\gamma \in (0, K)$, there exist $\rho_\gamma > 0$ and $f_\gamma \in C(\mathcal{P}(S))$ with $f_\gamma >> 0$ satisfying $\rho_\gamma f_\gamma = L_P^{(\gamma)} f_\gamma$, then*

$$\lim_{\gamma\to 0} \lambda_P(\gamma) = v_P. \tag{4.27}$$

Before proving Theorem 4.5, we present a lemma to show that p_γ converges to p_0 weakly and uniformly, where p_0 is defined in (4.4).

Lemma 4.6 *Assume (C0), (C1), (C2), and (C3). Then $p_\gamma(\cdot|\theta, a)$ weakly converges to $p_0(\cdot|\theta, a)$, uniformly in $\theta \in \mathcal{P}(S)$ and $a \in A$, i.e., if $f \in C(\mathcal{P}(S) \times A \times \mathcal{P}(S))$, then*

$$\lim_{\gamma\to 0} \sup_{\substack{\theta\in\mathcal{P}(S)\\ a\in A}} \left| \int_{\mathcal{P}(S)} f(\theta, a, \theta') p_\gamma(d\theta'|\theta, a) - \int_{\mathcal{P}(S)} f(\theta, a, \theta') p_0(d\theta'|\theta, a) \right| = 0. \tag{4.28}$$

Proof Fix an $f \in C(\mathcal{P}(S) \times A \times \mathcal{P}(S))$. Then

$$\begin{aligned} &\left| \int_{\mathcal{P}(S)} f(\theta, a, \theta') p_\gamma(d\theta'|\theta, a) - \int_{\mathcal{P}(S)} f(\theta, a, \theta') p_0(d\theta'|\theta, a) \right| \\ &= \left| e^{-r_\gamma(\theta,a)} \int_O T^*_{a,y,\gamma}(\theta)(S) f\left(\theta, a, \frac{T^*_{a,y,\gamma}(\theta)}{T^*_{a,y,\gamma}(\theta)(S)}\right) \Lambda(dy) - \int_O T^*_{a,y,0}(\theta)(S) f\left(\theta, a, \frac{T^*_{a,y,0}(\theta)}{T^*_{a,y,0}(\theta)(S)}\right) \Lambda(dy) \right|. \end{aligned}$$

Since

$$|e^{r_\gamma(\theta,a)} - 1| \le \int_S |e^{\gamma r(x,a)} - 1| \theta(dx) \le \max\{|e^{\gamma r_m} - 1|,\ |e^{\gamma r_M} - 1|\},$$

It follows that $e^{r_\gamma(\theta,a)}$ converges uniformly to 1 as $\gamma \to 1$. Then similarly as in the proof of Lemma 4.2, it suffices to verify that $T^*_{a,y,\gamma}(\theta)$ converges weakly to $T^*_{a,y,0}(\theta)$, uniformly in

$a \in A$, $y \in O$, and $\theta \in \mathcal{P}(S)$. In fact, recalling the definition of the Kantorovich-Rubinstein norm, we see that

$$\begin{aligned}\left\| T^*_{a,y,\gamma}(\theta) - T^*_{a,y,0}(\theta) \right\|_{KR} &= \sup_{\substack{g \in \mathrm{Lip}(S) \\ \|g\|_L \le 1}} \int_S \theta(dx) \left(e^{\gamma r(x,a)} - 1 \right) \int_S g(x') q(y|x') p(dx'|x,a) \\ &\le \sup_{\substack{g \in \mathrm{Lip}(S) \\ \|g\|_L \le 1}} \int_S \theta(dx) \left| e^{\gamma r(x,a)} - 1 \right| \int_S q_M p(dx'|x,a) \\ &\le \max\{|e^{\gamma r_m} - 1|, |e^{\gamma r_M} - 1|\} \cdot q_M.\end{aligned}$$

Consequently,

$$\lim_{\gamma \to 0} \sup_{\substack{\theta \in \mathcal{P}(S) \\ a \in A, y \in O}} \left\| T^*_{a,y,\gamma}(\theta) - T^*_{a,y,0}(\theta) \right\|_{KR} = 0$$

and thus (4.28) follows. □

Proof of Theorem 4.5 We already knew that $\lim_{\gamma \to 0} \lambda_P(\gamma) \ge v_P$. Hence, by Theorem 4.4, it suffices to verify that

$$\lim_{\gamma \to 0} \lambda_S(\gamma) \le v_P.$$

From Theorem 4.3, we know that for any $\varepsilon > 0$ and $\gamma > 0$, there exists $\beta^\varepsilon_\gamma \in \mathcal{I}$ such that

$$\lambda_S(\gamma) - \varepsilon \le \int_{\mathcal{P}(S) \times A} \left[\frac{1}{\gamma} r_\gamma(\theta, a) - \frac{1}{\gamma} D((\beta^\varepsilon_\gamma)_2(\cdot|\theta, a) \| p_\gamma(\cdot|\theta, a)) \right] (\beta^\varepsilon_\gamma)'(d\theta, da).$$

Recall that $\mathcal{I} \subseteq \mathcal{P}(\mathcal{P}(S) \times A \times \mathcal{P}(S))$ is compact. We can find a sequence $\{\gamma_n\}_{n \in \mathbb{N}}$ monotonically tending to 0 and a $\beta^\varepsilon \in \mathcal{I}_P$ such that

$$\lim_{\gamma \to 0} \lambda_P(\gamma) = \lim_{n \to \infty} \lambda_P(\gamma_n) \text{ and } \lim_{n \to \infty} \beta^\varepsilon_{\gamma_n} = \beta^\varepsilon$$

weakly. Since the relative entropy is non-negative, we obtain that

$$\lambda_S(\gamma_n) - \varepsilon \le \int_{\mathcal{P}(S) \times A} \frac{1}{\gamma_n} r_{\gamma_n}(\theta, a) (\beta^\varepsilon_{\gamma_n})'(d\theta, da).$$

Monotonicity follows from Hölder's inequality. Thus, we have as $\gamma_n \to 0$ that

$$\frac{1}{\gamma_n} r_{\gamma_n}(\theta, a) = \frac{1}{\gamma_n} \log \left(\int_S \theta(dx) e^{\gamma_n r(x,a)} \right) \downarrow \int_S \theta(dx) r(x, a) = r_0(\theta, a),$$

where r_0 is defined in (4.4). Hence, by Dini's theorem, $\gamma_n^{-1} r_{\gamma_n}$ converge to r_0 uniformly. From the weak convergence of $\beta^\varepsilon_{\gamma_n}$, we then obtain that

$$\lim_{n \to \infty} \int_{\mathcal{P}(S) \times A} \frac{1}{\gamma_n} r_{\gamma_n}(\theta, a) (\beta^\varepsilon_{\gamma_n})'(d\theta, da) = \int_{\mathcal{P}(S) \times A} r_0(\theta, a) (\beta^\varepsilon)'(d\theta, da).$$

Now, we claim that $(\beta^\varepsilon)_2 = p_0$, where p_0 is defined in (4.4). Notice that

$$\begin{aligned}\lim_{n\to\infty}\lambda_S(\gamma_n)-\varepsilon \le & \int_{\mathcal{P}(S)\times A} r_0(\theta,a)(\beta^\varepsilon)'(d\theta,da)\\ & -\liminf_{n\to\infty}\frac{1}{\gamma_n}\int_{\mathcal{P}(S)\times A}\left(\beta^\varepsilon_{\gamma_n}\right)'(d\theta,da)D\left(\left(\beta^\varepsilon_{\gamma_n}\right)_2(\cdot|\theta,a)\|p_{\gamma_n}(\cdot|\theta,a)\right)\\ = & \int_{\mathcal{P}(S)\times A} r_0(\theta,a)(\beta^\varepsilon)'(d\theta,da)\\ & -\liminf_{n\to\infty}\frac{1}{\gamma_n}D\left(\beta^\varepsilon_{\gamma_n}(d\theta,da,d\theta')\middle\|\left(\beta^\varepsilon_{\gamma_n}\right)'(d\theta,da)p_{\gamma_n}(d\theta'|\theta,a)\right)\\ \le & \, r_M-\liminf_{n\to\infty}\frac{1}{\gamma_n}D\left(\beta^\varepsilon_{\gamma_n}(d\theta,da,d\theta')\middle\|\left(\beta^\varepsilon_{\gamma_n}\right)'(d\theta,da)p_{\gamma_n}(d\theta'|\theta,a)\right).\end{aligned}$$

From the lower semicontinuity of $D(\cdot\|\cdot)$ and Lemma 4.6, we deduce that

$$\begin{aligned}& -\liminf_{n\to\infty}D\left(\beta^\varepsilon_{\gamma_n}(d\theta,da,d\theta')\middle\|\left(\beta^\varepsilon_{\gamma_n}\right)'(d\theta,da)p_{\gamma_n}(d\theta'|\theta,a)\right)\le\\ & -D\left(\beta^\varepsilon(d\theta,da,d\theta')\middle\|\left(\beta^\varepsilon\right)'(d\theta,da)p_0(d\theta'|\theta,a)\right).\end{aligned}$$

Thus, if $(\beta^\varepsilon)_2 \neq p_0$, then

$$D\left(\beta^\varepsilon(d\theta,da,d\theta')\middle\|\left(\beta^\varepsilon\right)'(d\theta,da)p_0(d\theta'|\theta,a)\right)>0,$$

and we would have

$$-r_m-\varepsilon\le\lambda(0)-\varepsilon\le\lim_{n\to\infty}\lambda(\gamma_n)-\varepsilon=-\infty.$$

It is impossible, so $\beta^\varepsilon\in\mathcal{I}$ and $(\beta^\varepsilon)_2=p_0$. Now we can employ the same argument as used in the proof of Lemma 3.2 to derive that

$$\lim_{n\to\infty}\lambda_S(\gamma_n)-\varepsilon\le\int_{\mathcal{P}(S)\times A}r_0(\theta,a)(\beta^\varepsilon)'(d\theta,da)\le v_P.$$

Then (4.27) follows by letting $\varepsilon\to 0$. □

Remark 4.2 From the proof of Theorem 4.5, we can see that the existence of a solution to the risk-sensitive Bellman equation guarantees the existence of the invariant probability measure for p_0.

We end this section with a simple example.

Example 4.7 Consider a finite POMDP with $S=\{x_1,x_2\}$, $A=\{a_1,a_2\}$, $O=\{y_1,y_2\}$. The transition probability p, observation probability q, and reward r are described by

$$\begin{bmatrix}p(x_1|x_1,a_1) & p(x_2|x_1,a_1)\\ p(x_1|x_2,a_1) & p(x_2|x_2,a_1)\end{bmatrix}=\begin{bmatrix}1/2 & 1/2\\ 1/2 & 1/2\end{bmatrix},\ \begin{bmatrix}p(x_1|x_1,a_2) & p(x_2|x_1,a_2)\\ p(x_1|x_2,a_2) & p(x_2|x_2,a_2)\end{bmatrix}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix},$$

$$\begin{bmatrix}q(y_1|x_1) & q(y_2|x_1)\\ q(y_1|x_2) & q(y_2|x_2)\end{bmatrix}=\begin{bmatrix}1/2 & 1/2\\ 1/2 & 1/2\end{bmatrix},\ \begin{bmatrix}r(x_1,a_1) & r(x_1,a_2)\\ r(x_2,a_1) & r(x_2,a_2)\end{bmatrix}=\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}.$$

The state space of the transferred MDP is $\mathcal{P}(S)$, which is isomorphic to $[0,1]$. We use $(1-t,t)$ to denote the probability distribution in $\mathcal{P}(S)$, where $0\le t\le 1$ represents the probability assigned on x_2.

On the one hand, to apply Theorem 4.5, by straightforward calculations, we have for $f \in C([0,1])$,

$$L_P^{(\gamma)} f(\theta) = \max\left\{ f\left(\frac{1}{2}\right), (1-t+te^{\gamma}) f\left(\frac{te^{\gamma}}{1-t+te^{\gamma}}\right)\right\}.$$

Let

$$\rho_\gamma = e^{\gamma}, \quad f_\gamma(t) = \begin{cases} 1, & t \in [0, \frac{1}{2}e^{-\gamma}) \\ 2e^{\gamma}t, & t \in [\frac{1}{2}e^{-\gamma}, 1] \end{cases}.$$

We can verify that $L_P^{(\gamma)} f_\gamma = \rho_\gamma f_\gamma$. Then all the assumptions in Theorem 4.5 are fulfilled. Thus,

$$v_P = \lim_{\gamma \to 0+} \lambda_P(\gamma) = \lim_{\gamma \to 0+} \frac{1}{\gamma} \log \rho_\gamma = 1.$$

On the other hand, given an initial distribution $t \in [0,1]$, we can see that the optimal average reward is $v_P(t) = t$. Hence, $v_P = \sup_{t\in[0,1]} v_P(t) = 1$, which coincides with Theorem 4.5. Furthermore, this example illustrates that there are circumstances in which the optimal risk-sensitive reward is independent of the initial distribution while the optimal average reward is not.

5 A Portfolio Optimization Example

In this section, as an example of applications of the approach developed in the previous sections, we consider a problem for portfolio optimization. Given a market with m securities and k price affecting factors. Let $V(n)$ denote the portfolio's value at time n. We assume that the portfolio dynamics are determined by

$$\frac{V(n+1)}{V(n)} = e^{F(X(n),H(n),W(n))}, \tag{5.1}$$

where $X(n) = (X^1(n), ..., X^k(n))$ denotes the factor process, which is a Markov chain with transition kernel $P(dx'|x)$, $H(n) = (H^1(n), ..., H^m(n))$ represents the portfolio strategy, i.e., the proportions of capital invested in the m securities at time n, $\{W(n), n \geq 1\}$ is the i.i.d. random noise which is independent of the factor process and has a common law η. F is a Borel measurable function. $X(n)$ and $H(n)$ take values in some compact subsets $S \subset \mathbb{R}^k$ and $A \subset \mathbb{R}^m$ respectively, while the noise $W(n)$ takes values in a Polish space Z.

This model was extensively studied in [26] for the dual relationship between maximizing the probability of outperforming over a given benchmark and optimizing the long-term risk-sensitive reward. In this section, we will demonstrate that our approach guarantees the convergence of the optimal risk-sensitive reward to the optimal risk-neutral reward as the risk-sensitive factor tends to 0. As a consequence, we show that the optimal risk-neutral reward can be taken as a benchmark appearing in the duality mentioned above, complementing the studies of [26].

Given an initial state $X(1) = x$, we use P_x to denote the corresponding probability measure on $\mathcal{B}((S \times Z)^\infty)$ and E_x the expectation under P_x. Let $\mathcal{A}$ denote the set of all Markov portfolio strategies. Given a risk factor $\gamma > 0$, the risk-sensitive optimal value is

$$\lambda(x,\gamma) = \sup_{H\in\mathcal{A}} \frac{1}{\gamma} \liminf_{N\to\infty} \frac{1}{N} \log E_x \exp\left[\gamma \sum_{t=1}^{N} F(X(n), H(n), W(n))\right]. \tag{5.2}$$

In what follows, we state the two assumptions on F and P for fitting (A1), (A2), and (A3).

(H1) For each $w \in Z$, $F(\cdot,\cdot,w) \in C(S \times A)$, and there is an $\eta-$integrable random variable $g(w)$, such that

$$e^{|F(x,h,w)|} \le g(w) \ \ \forall x \in S,\ h \in A \text{ and } w \in Z; \tag{5.3}$$

(H2) The family of functions $\left\{x \mapsto \int f(x')P(dx'|x),\ f \in C(S), \|f\| \le 1\right\}$ is equicontinuous.

Remark 5.1 (1) If (H1) holds, then $|F(x,h,w)| \le g(w)$ and hence $\hat{F}(x,h) := \int F(x,h,\cdot)d\eta$ is bounded continuous in (x,h). Let F_m and F_M be the infimum and supremum of $\hat{F}$. Then from Jensen's inequality, it follows that for $\gamma > 0$

$$\log \int e^{\gamma F(x,h,\cdot)} d\eta \ge \gamma F_m. \tag{5.4}$$

(2) A particular case in which (H2) is true is that $P(dx'|x) = Q(x'|x)\Lambda(dx')$ with $\{Q(x'|\cdot), x' \in S\}$ equicontinuous and $\Lambda \in \mathcal{P}(S)$.

The one-step reward F in (5.1) depends not only on the state x and the action h but also on W, which is slightly different from the typical form. We make the following changes to get a reward in such a standard form. Define a new Markov decision model with the transition law $p^{(\gamma)}$ and the one-step reward $r^{(\gamma)}$ defined respectively by

$$p^{(\gamma)}(dx'|x,h) := \frac{1}{\int_Z e^{\gamma F(x,h,w)}\eta(dw)} \left(\int_Z P(dx'|x)e^{\gamma F(x,h,w)}\eta(dw)\right), \tag{5.5}$$

$$\text{and} r^{(\gamma)}(x,h) := \log\left(\int_Z e^{\gamma F(x,h,w)}\eta(dw)\right).$$

By a direct calculation, we see that $p^{(\gamma)}(dx'|x,h)$ is actually $P(dx'|x)$, and for any $N \ge 1$

$$E_x \exp\left[\sum_{n=1}^{N} r^{(\gamma)}(X(n),H(n))\right] = E_x \exp\left[\gamma \sum_{n=1}^{N} F(X(n),H(n),W(n))\right]. \tag{5.6}$$

Notice that the transition kernel of this MDP is still P, but the reward is $r^{(\gamma)}$ instead of $\gamma \cdot r$. Assumption (H1) implies that $r^{(\gamma)}(x,h)$ is continuous in (x,h). Thus, with an extra discussion about the convergence of $\frac{1}{\gamma}r^{(\gamma)}$ as $\gamma \to 0$, we can obtain the limit with the same argument as the one in the proof of Theorem 3.1. In particular, it is not hard to check that (H1) and (H2) imply that $r^{(\gamma)}$ and P satisfy (A1), (A2), and (A3). Therefore, setting the risk-sensitive coefficient in Theorem 3.7 to be one and then dividing both sides by γ, we have the following variational formula for $\lambda(\gamma) = \sup_{x\in S}\lambda(\gamma,x)$.

$$\lambda(\gamma) = \frac{1}{\gamma}\sup_{\beta\in\mathcal{I}}\left\{\int_{S\times A}\left[r^{(\gamma)}(x,h) - D(\beta_2(\cdot|x,h)\|P(\cdot|x))\right]\beta'(dx,dh)\right\}, \tag{5.7}$$

where $\mathcal{I}$ is defined in (3.2). Although Theorem 3.8 can not be directly applied due to the difference between (5.7) and (3.1), the risk-neutral limit $\lim_{\gamma\to 0}\lambda(\gamma)$ can still be derived by an argument similar to the one used in proving Theorem 3.1. To see this, we still use v to denote the average optimal return, i.e.,

$$v = \sup_{x\in S} v(x) = \sup_{x\in S}\sup_{H\in\mathcal{A}} \liminf_{N\to\infty} \frac{1}{N} E_x\left[\sum_{n=1}^{N} F(X(n),H(n),W(n))\right]. \tag{5.8}$$

Theorem 5.1 *Assume (H1) and (H2). Then*

$$\lim_{\gamma\to 0}\lambda(\gamma)=v. \tag{5.9}$$

Proof By Hölder's inequality, we see that $\lambda(\gamma)$ is nondecreasing in γ and $\liminf_{\gamma\to 0}\lambda(\gamma)\geq v$. We will apply (5.7) to prove that $\limsup_{\gamma\to 0}\lambda(\gamma)\leq v$. Similarly to the argument in the proof of Theorem 3.1, for any $\varepsilon>0$, we can find a sequence $\{\gamma_n\}_{n\in\mathbb{N}}$ decreasing to 0 with $\lim_{\gamma\to 0}\lambda(\gamma)=\lim_{n\to\infty}\lambda(\gamma_n)$ and $\beta^{\varepsilon}_{\gamma_n},\beta^{\varepsilon}\in\mathcal{I}$ with $\lim_{n\to\infty}\beta^{\varepsilon}_{\gamma_n}=\beta^{\varepsilon}$ weakly such that

$$\lambda(\gamma_n)-\varepsilon\leq\int_{S\times A}\left[\frac{1}{\gamma_n}r^{(\gamma_n)}(x,h)-\frac{1}{\gamma_n}D((\beta^{\varepsilon}_{\gamma_n})_2(\cdot|x,h)\|P(\cdot|x))\right](\beta^{\varepsilon}_{\gamma_n})'(dx,dh). \tag{5.10}$$

Therefore,

$$\lambda(\gamma_n)-\varepsilon\leq\int_{S\times A}\frac{1}{\gamma_n}r^{(\gamma_n)}(x,h)(\beta^{\varepsilon}_{\gamma_n})'(dx,dh).$$

Monotonicity follows from Hölder's inequality, and thus we have as $\gamma_n\to 0$ that

$$\frac{1}{\gamma_n}r^{(\gamma_n)}(x,h)=\frac{1}{\gamma_n}\log\left(\int_Z e^{\gamma_n F(x,h,w)}\eta(dw)\right)\downarrow\int_Z F(x,h,w)\eta(dw)=:r^{(0)}(x,h).$$

Therefore, it follows from Dini's theorem that $\frac{1}{\gamma_n}r^{(\gamma_n)}$ converge to $r^{(0)}$ uniformly. Combining this fact with the weak convergence of $\beta^{\varepsilon}_{\gamma_n}$, we obtain that

$$\lim_{n\to\infty}\int_{S\times A}\frac{1}{\gamma_n}r^{(\gamma_n)}(x,h)(\beta^{\varepsilon}_{\gamma_n})'(dx,dh)=\int_{S\times A}r^{(0)}(x,h)(\beta^{\varepsilon})'(dx,dh).$$

Now we claim that $(\beta^{\varepsilon})_2=P$. Indeed, from the joint semicontinuity of the relative entropy, we see that

$$\begin{aligned}
&-\liminf_{n\to\infty}\int_{S\times A}D\left(\left(\beta^{\varepsilon}_{\gamma_n}\right)_2(\cdot|x,h)\middle\|P(\cdot|x)\right)\left(\beta^{\varepsilon}_{\gamma_n}\right)'(dx,dh)\\
=&-\liminf_{n\to\infty}D\left(\beta^{\varepsilon}_{\gamma_n}(dx,dh,dx')\middle\|\left(\beta^{\varepsilon}_{\gamma_n}\right)'(dx,dh)P(dx'|x)\right)\\
\leq&-D\left(\beta^{\varepsilon}(dx,dh,dx')\middle\|\left(\beta^{\varepsilon}\right)'(dx,dh)P(dx'|x)\right).
\end{aligned}$$

Thus, if $(\beta^{\varepsilon})_2\neq P$, then

$$D\left(\beta^{\varepsilon}(dx,dh,dx')\middle\|\left(\beta^{\varepsilon}\right)'(dx,dh)P(dx'|x)\right)>0.$$

From assumption (H1), (5.4), (5.5), and (5.6), together with (5.2) and (5.10), we would have

$$F_m-\varepsilon\leq\lim_{n\to\infty}\lambda(\gamma_n)-\varepsilon\leq-\infty.$$

It is impossible, implying that it must be that $\beta^{\varepsilon}\in\mathcal{I}$ and $(\beta^{\varepsilon})_2=P$. Then it is routine to follow the same argument as that of Theorem 3.2 to check that $\int_{S\times A}r^{(0)}(x,h)(\beta^{\epsilon})'(dx,dh)\leq v$. Consequently, (5.9) follows by letting $\varepsilon\to 0$. □

As claimed in the introduction, it was shown in [18, 24], and [26] that the risk-sensitive portfolio optimization is a dual problem to the maximization of the outperformance

probability (upside chance) when assuming differentiability for the optimal value. To describe this more precisely, for $b \in \mathbb{R}$, $\gamma > 0$, $x \in S$, define $I_x(b)$ by

$$I_x(b) := \sup_{H \in \mathcal{A}} \liminf_{N \to \infty} \frac{1}{N} \log \mathrm{P}_x \left[\frac{1}{N} \sum_{n=1}^{N} F(X(n), H(n), W(n)) \geq b \right], \quad (5.11)$$

$$\text{and} \quad I(b) := \sup_{x \in S} I_x(b).$$

Then by Chebyshev's inequality,

$$\gamma \lambda(\gamma, x) - \gamma b \geq I_x(b), \text{ and } \gamma \lambda(\gamma) - \gamma b \geq I(b).$$

Thus,

$$- \sup_{\gamma \in [0, K)} \{\gamma b - \gamma \lambda(\gamma)\} \geq I(b) \quad (5.12)$$

for a pre-specified $K > 0$.

Let $\Lambda(\gamma) := \gamma \lambda(\gamma)$ for convenience. It has already been established in [26] that if $\Lambda(\gamma)$ is differentiable on $[0, K)$ and the limit

$$\lim_{N \to \infty} \frac{1}{N} \log E_x \exp \left[\gamma \sum_{n=1}^{N} F(X(n), H(n), W(n)) \right] \quad (5.13)$$

exists and does not depend on the initial state x, then the duality

$$- \sup_{\gamma \in [0, K)} \{\gamma b - \Lambda(\gamma)\} = I(b) \quad (5.14)$$

holds whenever $b \in \{\Lambda'^{+}(\gamma), \gamma \in [0, K)]\}$ or $b \leq \Lambda'^{+}(0)$, where $\Lambda'^{+}(\gamma)$ denote the righthand derivative of $\Lambda(\gamma)$ (see Theorem 2.7 in [26]). In the meantime, our result shows that under (H1) and (H2),

$$\Lambda'^{+}(0) = \lim_{\gamma \to 0+} \frac{\Lambda(\gamma) - \Lambda(0)}{\gamma - 0} = \lim_{\gamma \to 0} \lambda(\gamma) = v,$$

which reveals the connection between the outer-performance probability, the risk-neutral average return, and the risk-sensitive average growth rate.

In order to guarantee the differentiability, we add the following assumptions for P and $r^{(\gamma)}$ in accordance with Theorem 3.1 in [26], which also implies that the transition law satisfies (B2).

(H3) There exists $\delta_p < 1$ such that

$$\sup_{U \in \mathcal{B}(S)} \sup_{x, x' \in S} [P(U|x) - P(U|x')] \leq \delta_p. \quad (5.15)$$

(H4) There exists a $K_\gamma > 0$ such that the mapping $\gamma \to \sup_{h \in A} r^{(\gamma)}(x, h)$ is differentiable on $[0, K_\gamma)$ for any $x \in S$.

Remark 5.2 Let F_m and F_M be defined in Remark 5.1(1). Then (H1), (H2), (H3), and the condition $\gamma \leq -\frac{\log \delta_p}{F_M - F_m}$ guarantee that the limit inferior in (5.2) is actually a limit and $\lambda(\gamma, x)$ does not depend on x (see Theorem 1 in [13]). So the constant K in (5.12) can be determined.

Theorem 5.2 *Assume (H1), (H2), (H3), and (H4). Let* $K = \min\{-\frac{\log \delta_p}{F_M - F_m}, K_\gamma\}$. *Then* $v(x) = v$ *is a constant, and the duality* (5.14) *holds for every* $b \leq v$.

Proof Combining Theorem 3.1 in [26] and the above remark, we see that $\Lambda(\gamma)$ is differentiable on $[0, K)$. Thus, by Theorem 2.7 in [26],

$$-\sup_{\gamma\in[0,K)}\{\gamma b-\Lambda(\gamma)\}=I(b)$$

holds for every $b \le \Lambda'^{+}(0)$. Theorem 3.7 implies that

$$\Lambda'^{+}(0)=\lim_{\gamma\to 0}\frac{1}{\gamma}\Lambda(\gamma)=\lim_{\gamma\to 0}\sup_{x\in S}\lambda(x,\gamma)=\sup_{x\in S}v(x),$$

where $v(x)$ is indeed a constant due to (H3) and Lemma 3.9. These complete the proof. □

We end this section with an illustrative example.

Example 5.3 Let $S=\{-1,1\}$, $A=\{(h_1,h_2), h_i\ge 0,\ h_1+h_2=1\}$ and $\{W_n, n\ge 1\}$ i.i.d. with the standard Normal distribution $N(0,1)$. F is given by

$$F(x,h,w)=\alpha\cdot x\cdot(h_1-h_2)\cdot w^2,\ x\in S,\ h\in A,\ w\in W,$$

where $\alpha\in(0,1/2)$ is a constant. Then $e^{|F(x,h,w)|}\le g(w)=e^{\alpha w^2}$ with g being $\eta-$integrable. Thus (H1), (H2), and (H4) are fulfilled, and (5.9) holds. If the transition probabilities are chosen to satisfy that

$$\min_{x,x'\in S}P(x'|x)>0,$$

then (H3) is also satisfied; therefore, the assertions of Theorem 5.2 hold true.

Appendix A nonlinear extension of the Kreĭn-Rutman theorem

The following theorem is an extension of the Kreĭn-Rutman theorem for nonlinear operators on Banach space (see Proposition 3.1.5, Lemma 3.1.3, and Lemma 3.1.7 in [23]).
An *ordered Banach space* is a real Banach space $(X,\|\cdot\|)$ endowed with a *ordered cone* K, which means K is a closed convex cone with vertex at 0 such that $K\cap(-K)=\{0\}$. Assume $interior(K)\ne\emptyset$ and $\dim X\ge 2$. For $x_1,x_2\in X$, we write $x_1\ge x_2$ if $x_1-x_2\in K$, $x_1>x_2$ if $x_1-x_2\in K\backslash\{0\}$ and $x_1>>x_2$ if $x_1-x_2\in interior(K)$. For an operator $L:K\to K$, define

$$\sigma^{+}(L):=\{\rho>0: Lx=\rho x,\ \text{for some } x>0\}.$$
$$\|L\|^{+}:=\sup_{x\in K,\|x\|\le 1}\{\|Lx\|\}.$$

Since $\|L^{m+n}\|^{+}\le\|L^m\|^{+}\|L^n\|^{+}$, $m,n\ge 1$, we can define

$$\rho(L):=\lim_{n\to\infty}\left(\|L^n\|^{+}\right)^{\frac{1}{n}}.$$

The following properties for operator L on K will be required:

1. (Compactness) L is a compact operator, meaning that L maps any bounded subset into a relatively compact one.
2. (Positively 1-homogeneity) $c(Lx)=L(cx)$ for any $x\in K, c>0$.
3. (Order-preserving) $Lx_1\ge Lx_2$ for any $x_1\ge x_2,\ x_1,x_2\in K$.

Theorem A.1 *Let $L:K\to K$ be a compact, positively 1-homogeneous, and order-preserving operator. If $\rho(L)>0$, then $\rho(L)\in\sigma^{+}(L)$ and $\rho(L)=\max\sigma^{+}(L)$.*

Theorem A.2 *Let $L : K \to K$ be a positively 1-homogeneous and order-preserving operator. If there exits $\rho' > 0$ and $f >> 0$ such that $Lf = \rho' f$, then $\rho' = \rho(L)$.*

Theorem A.3 *Let $L : K \to K$ be a positively 1-homogeneous and order-preserving operator. If there exits $\rho' \geq 0$ and $f >> 0$ such that $Lf \leq \rho' f$, then $\rho' \geq \rho(L)$.*

Acknowledgements This paper is part of the first author's dissertation. The authors are grateful to the referee for the careful review of the manuscript and for the helpful comments and suggestions for improvement.

Funding This work is supported by the NSFC 11671226.

Data Availability Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

References

1. Albertini F, Dai Pra P, Prior C. Small parameter limit for ergodic, discrete-time, partially observed, risk-sensitive control problems. Math Control Signals Syst. 2001;14(1):1–28.
2. Anantharam V, Borkar VS. A variational formula for risk-sensitive reward. SIAM J Control Optim. 2017;55(2):961–988.
3. Arapostathis A, Borkar VS, Fernández-Gaucherand E, Ghosh MK, Marcus SI. Discrete-time controlled Markov processes with average cost criterion: a survey. SIAM J Control Optim. 1993;31(2):282–344.
4. Baras JS, James MR. Robust and risk-sensitive output feedback control for finite state machines and hidden Markov models (summary). J Math Syst Estimation Control. 1997;7:371–374.
5. Bäuerle N, Rieder U. Markov decision processes with applications to finance. Berlin: Springer; 2011.
6. Bäuerle N, Rieder U. Partially observable risk-sensitive Markov decision processes. Math Oper Res. 2017;42(4):1180–1196.
7. Bogachev VI. Measure Theory, vol II. Berlin: Springer; 2007.
8. Cavazos-Cadena R, Hernández-Hernández D. Successive approximations in partially observable controlled Markov chains with risk-sensitive average criterion. Stochast: an Int J Probab Stochast Process. 2005;77(6):537–568.
9. Dembo A, Zeitouni O. Large deviations techniques and applications. Stochastic modelling and applied probability, 38. Berlin: Springer; 2010.
10. Di Masi GB, Stettner L. Risk-sensitive control of discrete-time Markov processes with infinite horizon. SIAM J Control Optim. 1999;38(1):61–78.
11. Di Masi GB, Stettner L. Risk sensitive control of discrete time partially observed Markov processes with infinite horizon. Stochast: an Int J Probab Stochast Process. 1999;67(3-4):309–322.
12. Di Masi GB, Stettner L. Remarks on risk neutral and risk sensitive portfolio optimization. In From stochastic calculus to mathematical finance. pp 211-226. Berlin: Springer; 2006.
13. Di Masi GB. Infinite horizon risk sensitive control of discrete time Markov processes with small risk. Syst Control Lett. 2000;40(1):15–20.
14. Donsker MD, Varadhan SS. On a variational formula for the principal eigenvalue for operators with maximum principle. Proc Natl Acad Sci. 1975;72(3):780–783.
15. Dupuis P, Ellis RS. A weak convergence approach to the theory of large deviations. New York: Wiley; 1997.
16. Fleming WH, Hernández-Hernández D. Risk-sensitive control of finite state machines on an infinite horizon I. SIAM J Control Optim. 1997;35(5):1790–1810.
17. Fleming WH, Hernández-Hernández D. Risk-sensitive control of finite state machines on an infinite horizon II. SIAM J Control Optim. 1999;37(4):1048–1069.
18. Hata H, Nagai H, Sheu SJ. Asymptotics of the probability minimizing a "down-side" risk. Annals Appl Probab. 2010;20(1):52–89.
19. Hernández-Hernández D. Partially observed control problems with multiplicative cost. In: Stochastic analysis, control, optimization and applications. pp 41-55. Birkhäuser, Boston; 1999.
20. Hernández-Lerma O, Lasserre JB. Discrete-time Markov control processes: basic optimality criteria. Vol 30. Berlin: Springer; 2012.
21. Howard RA, Matheson JE. Risk-sensitive Markov decision processes. Manag Sci. 1972;18(7):356–369.

22. Jaśkiewicz A. Average optimality for risk-sensitive control with general state space. Ann Appl Probab. 2007;17(2):654–675.
23. Ogiwara T. Nonlinear Perron-Frobenius problem on an ordered Banach space. Japan J Math. 1995;21(1):43–103.
24. Pham H. A risk-sensitive control dual approach to a large deviations control problem. Syst Control Lett. 2003;49(4):295–309.
25. Sion M. On general minimax theorems. Pac J Math. 1958;8(1):171–176.
26. Stettner L. Duality and risk sensitive portfolio optimization. Contemp Math. 2004;351:333–348.