Relaxed Equilibria for Time-Inconsistent Markov Decision Processes

Erhan Bayraktar* Yu-Jui Huang† Zhenhua Wang‡ Zhou Zhou§

Abstract

This paper considers an infinite-horizon Markov decision process (MDP) that allows for general non-exponential discount functions, in both discrete and continuous time. Due to the inherent time inconsistency, we look for a randomized equilibrium policy (i.e., relaxed equilibrium) in an intra-personal game between an agent's current and future selves. When we modify the MDP by entropy regularization, a relaxed equilibrium is shown to exist by a nontrivial entropy estimate. As the degree of regularization diminishes, the entropy-regularized MDPs approximate the original MDP, which gives the general existence of a relaxed equilibrium in the limit by weak convergence arguments. As opposed to prior studies that consider only deterministic policies, our existence of an equilibrium does not require any convexity (or concavity) of the controlled transition probabilities and reward function. Interestingly, this benefit of considering randomized policies is unique to the time-inconsistent case.

Keywords: time inconsistency, non-exponential discounting, Markov decision processes, relaxed controls, entropy regularization

1 Introduction

For Markov decision processes (MDPs) on an infinite horizon, discounting is a key feature that allows the expected total reward to take a finite value. A widespread assumption in the literature is exponential discounting, i.e., the discount rate is constant over time. There is, however, substantial evidence against exponential discounting. In behavioral economics (see e.g., Thaler [27], Loewenstein and Thaler [24], Laibson [23]), it is well documented that the discount rate is empirically time-varying and many non-exponential functions have been proposed to model empirical discounting; see e.g., Huang and Zhou [19, Remark 3.1].

For the past fifteen years or so, non-exponential discounting has been seriously considered and approached in stochastic control, and the involved mathematical challenge is now well understood. In a nutshell, non-exponential discounting causes *time inconsistency*: an optimal control policy an agent derives today will not be optimal from the eyes of the same agent tomorrow. As a dynamically optimal policy over the entire time horizon no longer exists, Strotz [26] suggests that one instead look for an *equilibrium policy* in an intra-personal game between one's current and future selves. A standard equilibrium (i.e., an equilibrium policy as a *deterministic* map on the state space) has

*Department of Mathematics, University of Michigan, Ann Arbor, MI 48109-1043, USA, email: erhan@umich.edu. Partially supported by National Science Foundation (DMS-2106556) and by the Susan M. Smith Professorship.

†Department of Applied Mathematics, University of Colorado, Boulder, CO 80309-0526, USA, email: yujui.huang@colorado.edu. Partially supported by National Science Foundation (DMS-2109002).

‡Department of Mathematics, University of Michigan, Ann Arbor, MI 48109-1043, USA, email: zhenhuaw@umich.edu.

§School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia, email: zhou.zhou@sydney.edu.au.

been studied under non-exponential discounting in diffusion models (e.g., Ekeland and Lazrak [12], Ekeland and Pirvu [14], Ekeland et al. [13], Yong [29], Björk et al. [9], Huang and Nguyen-Huu [17], Huang and Zhou [20], Bayraktar et al. [7, 8]), as well as in models of controlled Markov chains (e.g., Balbus et al. [3], Chatterjee and Eyigungor [10], Balbus et al. [4], Balbus et al. [2], Huang and Zhou [21], Balbus et al. [5], Bayraktar and Han [6]). Many of the studies establish the existence of a standard equilibrium for a general state space (e.g., a Borel space), but require very specific forms of discounting (e.g., quasi-hyperbolic or hyperbolic), the reward function (e.g., monotone and supermodular), and the controlled transition probabilities (e.g., decomposable into supermodular functions); see the assumptions in [3, 10, 4, 2, 5].[1] These conditions need not hold even in fairly simple examples, such that a standard equilibrium may fail to exist; see e.g., [21, Remark 10].

This motivates us to consider *relaxed equilibria* (i.e., *randomized* equilibrium policies), to possibly extend the existence of equilibria to cases where discounting, reward functions, and controlled transition probabilities are all general. To this end, it is natural to consider MDPs, which by definition considers randomized control policies (i.e., *relaxed controls*).

In the context of reinforcement learning, Alexander and Brown [1], Fedus et al. [15], and Schultheis et al. [25] design algorithms to compute optimal relaxed controls for MDPs under non-exponential discounting, but fail to realize that such policies are unsustainable due to time inconsistency. To the best of our knowledge, only the recent work Jaśkiewicz and Nowak [22] recognizes the issue of time inconsistency and applies Strotz's game-theoretic approach to MDPs: they prove that a relaxed equilibrium exists for discrete-time MDPs for a general state space, reward function (bounded and continuous), and transition probabilities (transition densities exist and are continuous), but require the discount function to be quasi-hyperbolic. In fact, their arguments rely crucially on the specific form of a quasi-hyperbolic discount function (which is stylized and tractable) and do not allow for other kinds of non-exponential discounting.

In this paper, we accommodate *general* discount functions and strive to establish the existence of a relaxed equilibrium for the resulting time-inconsistent MDPs in both discrete and continuous time. As we can no longer rely on the form of a discount function, our approach differs largely from that in Jaśkiewicz and Nowak [22].

Instead of working with the original MDP (i.e., (2.1) below) directly, we consider an entropy-regularized version, where the entropy of a randomized control policy (i.e., a *relaxed control*) is added to the functional to be maximized; see (2.4) below. Taking advantage of the form of the entropy term, we characterize relaxed equilibria for the entropy-regularized MDP (which we call *regular* relaxed equilibria) as fixed points of an operator, which takes a tractable Gibbs-measure form; see Proposition 2.1 and Corollary 2.1. By showing that the fixed-point operator continuously maps a compact domain to itself, we conclude from Brouwer's fixed-point theorem that a regular relaxed equilibrium exists; see Theorem 2.1. Note that for the operator to map a compact domain to itself, the growth of the entropy term needs to be contained appropriately, which is a known mathematical challenge. To achieve this, we assume that the reward function and controlled transition probabilities are Lipschitz in the action variable (Assumption 2.2) and the action space U fulfills a uniform cone condition (Assumption 2.3). Then, following an argument recently proposed in Huang et al. [18, Section 4.3], we get logarithmic growth of the entropy term uniformly in the state variable (Lemma 2.1), which facilitates the proof of Theorem 2.1.

For the original MDP (2.1), relaxed equilibria can also be characterized as fixed points of an operator (Proposition 2.2). However, the operator is an abstract set-valued map, which is, compared with the concrete single-valued operator under entropy regularization, much less tractable and much

[1] [6] considers a finite-horizon discrete-time model, for which a recursive backward induction provides an equilibrium. This method does not work for an infinite horizon and therefore cannot be applied to our study.

less promising for numerical implementation; see Remark 2.6. Thus, we approximate the original MDP by a sequence of entropy-regularized MDPs, with the degree of regularization (measured by $\lambda > 0$ in (2.4)) diminishing to zero. This yields a sequence of regular relaxed equilibria, one for each entropy-regularized MDP. Intriguingly, the value functions corresponding to this sequence are uniformly bounded: as shown in Lemma 2.2, the logarithmic growth of entropy (obtained in Lemma 2.1 for each fixed $\lambda > 0$) can be made uniform across all $\lambda > 0$ small enough, thereby giving a uniform bound for the value functions. Relying on this, a delicate probabilistic argument shows that the sequence of regular relaxed equilibria for entropy-regularized MDPs (i.e., fixed points of tractable Gibbs-form operators) converges weakly to a fixed point of the set-valued operator for the original MDP. That is, a relaxed equilibrium for the original MDP exists; see Theorem 2.2.

All the above, established in discrete time, can be extended to continuous time. For entropy-regularized continuous-time MDP, regular relaxed equilibria can again be characterized as fixed points of an operator of the Gibbs-measure form; see Proposition 3.1 and Corollary 3.1. The existence of a regular relaxed equilibrium is accordingly established in Theorem 3.1. Finally, we approximate a continuous-time MDP by a sequence of its entropy-regularized versions, with diminishing degree of regularization, and find that the sequence of regular relaxed equilibria (one for each entropy-regularized MDPs) converges to a relaxed equilibrium for the original MDP; see Theorem 3.2. Note that our continuous-time results are reconciled with those in discrete time: as shown in Theorem 5.1, when the time step tends to zero, (regular) relaxed equilibria in discrete time converge to one in continuous time.

Our mathematical setup generalizes that in Huang and Zhou [21], where controlled Markov chains (in both discrete and continuous time) are considered under non-exponential discounting. Specifically, one chooses transition probabilities directly (i.e., an action is simply a vector of transition probabilities) in Huang and Zhou [21], while we allow for general actions that may affect transition probabilities in a much more subtle way; see Remark 4.1 for details. In addition, Huang and Zhou [21] focuses on deterministic policies (i.e., standard controls), while we include randomized ones (i.e., relaxed controls). A key result in Huang and Zhou [21] states that a standard equilibrium exists, under a suitable condition on the transition probabilities and reward function; moreover, if such a condition fails, a standard equilibrium may not exist in general. This crucially explains the need of relaxed controls: unlike standard equilibria, a relaxed equilibrium generally exists, as proved in Theorems 2.2 and 3.2. Hence, in the face of general controlled transition probabilities and reward function, relaxed equilibria should certainly be considered; see Remark 4.2 and the discussion below it.

Interestingly, the need of relaxed controls is unique to the time-inconsistent case. In the time-consistent case of exponential discounting, even if one considers relaxed controls, the optimal value achieved by a relaxed control can always be achieved alternatively by a standard control—it is then necessary to consider only standard controls; see Remarks 4.4 and 4.6.

Let us stress that an entropy-regularized MDP, besides serving as a powerful approximation tool, has it own meaning and application. As introduced in Ziebart et al. [30], Fox et al. [16] and clearly explained in Wang et al. [28], an entropy-regularized MDP encodes the tradeoff between exploitation and exploration in reinforcement learning, with the exploration part represented by the added entropy term. This line of research has so far focused on time-consistent cases. The only exception we know of is Dai et al. [11], where a relaxed equilibrium is found for a mean-variance portfolio selection problem. Our results contribute to the burgeoning area of reinforcement learning under time inconsistency: the regular relaxed equilibrium we found represents a learning policy under time inconsistency induced by non-exponential discounting; see the discussion below Theorem 2.2 and Remark 2.4.

The paper is organized as follows. Section 2 introduces a time-inconsistent MDP in discrete time

that accommodates non-exponential discounting; an entropy-regularized version is also defined. We prove the existence of regular relaxed equilibria for entropy-regularized MDPs (Section 2.1) and obtain the existence of relaxed equilibria for the original MDP by an approximation argument (Section 2.2). Section 3.1 (resp. Section 3.2) extends results in Section 2.1 (resp. Section 2.2) to continuous time. Section 4 discusses when the use of relaxed controls is necessary. Section 5 shows that discrete-time (regular) relaxed equilibria converge to a (regular) relaxed equilibrium in continuous time, as time step tends to zero. The appendix collects (longer) proofs.

2 Relaxed Equilibria in Discrete Time

Set $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. Let $U \subset \mathbb{R}^\ell$, for a fixed $\ell \in \mathbb{N}$, be a compact action space with $\texttt{Leb}(U) > 0$. Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that supports a discrete-time Markov process $(X_t)_{t\in\mathbb{N}_0}$ that takes values in $\mathcal{S} = \{1, 2, ..., d\}$ for some $d \in \mathbb{N}$ and is controlled by a U-valued process α. Assume that the dynamics of X is time-homogeneous, i.e., for any $t \in \mathbb{N}_0$, $\mathbb{P}(X_{t+1} = j \mid X_t = i, \alpha_t = u) = \mathbb{P}(X_1 = j \mid X_0 = i, \alpha_0 = u)$ for all $i, j \in \mathcal{S}$ and $u \in U$. For any action $u \in U$, we will denote by p^u the associated transition matrix of X (i.e., $p^u_{ij} := \mathbb{P}(X_1 = j \mid X_0 = i, \alpha_0 = u)$ for all $i, j \in \mathcal{S}$) and write p^u_i for the i^{th} row of p^u for all $i \in \mathcal{S}$.

We call $\alpha : \mathcal{S} \to U$ a *standard feedback control* and denote by p^α the transition matrix induced by α. Specifically, the i^{th} row of p^α (denoted by p^α_i) is given by $p^\alpha_i = p^{\alpha(i)}_i$ for all $i \in \mathcal{S}$. Let $\mathcal{P}(U)$ (resp. $\mathcal{D}(U)$) denote the set of all probability measures (resp. density functions) on U. We call $\pi : \mathcal{S} \to \mathcal{P}(U)$ a *relaxed feedback control*. Note that a standard feedback control α can be viewed as a relaxed feedback control π by taking $\pi(i)$ to be the Dirac measure concentrated on $\alpha(i) \in U$ for all $i \in \mathcal{S}$. We denote by Π the set of all relaxed feedback controls.

Given $\pi \in \Pi$, the dynamics of $X = X^\pi$ is determined as follows. At any time $t \in \mathbb{N}_0$, given that $X_t = i \in \mathcal{S}$, we sample $u \in U$ according to the probability measure $\pi(i) \in \mathcal{P}(U)$. The realization of X_{t+1} is then governed by the transition probabilities $p^u_i = \{p^u_{ij}\}_{j\in\mathcal{S}}$.

Remark 2.1. *Given $\pi \in \Pi$, consider a Markov chain $\bar{X}^\pi$ whose transition matrix p^π is given by $p^\pi_{ij} := \int_U p^u_{ij}(\pi(i))(du)$ for all $i, j \in \mathcal{S}$. Note that $\bar{X}^\pi$ and X^π are different stochastic processes that share the same law. That is, the paths (i.e., realizations) taken by $\bar{X}^\pi$ and X^π are generally different, but the probability of which path will be taken is identical.*

Given a reward function $f : [0, \infty) \times \mathcal{S} \times U \to \mathbb{R}$, we define $f^\mu(t, i) := \int_U f(t, i, u)\mu(du)$ for all $\mu \in \mathcal{P}(U)$. For any $\pi \in \Pi$, the corresponding value function is given by

$$J^\pi(i) := \mathbb{E}_i\left[\sum_{k=0}^{\infty} f^{\pi(X^\pi_k)}(k, X^\pi_k)\right], \quad \forall i \in \mathcal{S}. \tag{2.1}$$

Remark 2.2. *The t variable in $f(t, i, u)$ does not represent "real calendar time" but "time difference," that is, the difference between the current time and the time of a future reward; see Huang and Zhou [21, Remark 1]. A typical example is*

$$f(t, i, u) = \delta(t)g(i, u), \tag{2.2}$$

where $g : \mathcal{S} \times U \to \mathbb{R}$ assigns a reward based on the current state i and the action u employed and $\delta : [0, \infty) \to [0, 1]$ is a discount function, assumed to be nonincreasing with $\delta(0) = 1$.

In this paper, we will investigate an entropy-regularized version of $J^\pi(i)$. To this end, let us first introduce the notion of a regular relaxed feedback control.

Definition 2.1. *A relaxed feedback control $\pi \in \Pi$ is regular if for each $i \in \mathcal{S}$, there exists $\rho_i \in \mathcal{D}(U)$ such that $(\pi(i))(A) = \int_A \rho_i(u)du$ for all Borel $A \subseteq U$ and its Shannon differential entropy satisfies $\mathcal{H}(\rho_i) > -\infty$, where*

$$\mathcal{H}(\rho) := -\int_U \ln(\rho(u))\rho(u)du, \quad \forall \rho \in \mathcal{D}(U). \tag{2.3}$$

We denote by Π_r the set of all regular feedback relaxed controls. Given $\pi \in \Pi_r$, we will often identify $\pi(i) \in \mathcal{P}(U)$ with its density $\rho_i \in \mathcal{D}(U)$ and write "$\pi \in \Pi_r$" and "$\{\rho_i\}_{i\in\mathcal{S}} \in \Pi_r$" interchangeably.

Now, for any $\lambda > 0$ and $\pi \in \Pi_r$, consider

$$J^\pi_\lambda(i) := \mathbb{E}_i\left[\sum_{k=0}^{\infty}\left(f^{\pi(X^\pi_k)}(k, X^\pi_k) + \lambda\delta(k)\mathcal{H}\big(\pi(X^\pi_k)\big)\right)\right], \quad \forall i \in \mathcal{S}, \tag{2.4}$$

where $\delta : [0,\infty) \to [0,1]$ is a discount function, assumed to be nonincreasing with $\delta(0) = 1$. To ensure that J^π in (2.1) and J^π_λ in (2.4) are well-defined, we impose the following.

Assumption 2.1. $M := \sum_{t=0}^{\infty}\left(\sup_{i\in\mathcal{S},u\in U}|f(t,i,u)| + \delta(t)\right) < \infty$.

By Assumption 2.1, J^π is clearly finitely valued for all $\pi \in \Pi$. On the other hand, under our assumption $0 < \texttt{Leb}(U) < \infty$, consider the uniform density $\nu \in \mathcal{D}(U)$ given by $\nu(u) := 1/\texttt{Leb}(U)$ for all $u \in U$. For any $\rho \in \mathcal{D}(U)$, we compute the Kullback-Leibler divergence

$$0 \le D_{KL}(\rho\|\nu) := \int_U \rho \ln\left(\frac{\rho}{\nu}\right)du = \int_U \rho\ln\rho du + \ln(\texttt{Leb}(U)), \tag{2.5}$$

which gives $\mathcal{H}(\rho) = -\int_U \ln(\rho(u))\rho(u)du \le \ln(\texttt{Leb}(U)) < \infty$ for all $\rho \in \mathcal{D}(U)$. This, along with Definition 2.1, indicates that for any $\pi \in \Pi_r$,

$$|\mathcal{H}(\pi(i))| < \infty \quad \text{for all } i \in \mathcal{S}. \tag{2.6}$$

By Assumption 2.1, (2.6), and $\mathcal{S}$ being a finite set, J^π_λ is finitely valued for all $\pi \in \Pi_r$.

An agent who aims to maximize $J^\pi(i)$ over all $\pi \in \Pi$ may run into the issue of time inconsistency. Specifically, given that $X_0 = i \in \mathcal{S}$, the time-0 agent's problem is $\sup_{\pi\in\Pi} J^\pi(i)$. At a later time $t > 0$ with $X_t = j \neq i$, the time-t agent's problem is $\sup_{\pi\in\Pi} J^\pi(j)$. As the two problems $\sup_{\pi\in\Pi} J^\pi(i)$ and $\sup_{\pi\in\Pi} J^\pi(j)$ need not share the same optimal control $\pi^* \in \Pi$, time-inconsistency may arise. Such inconsistency results from the "time difference" variable t in either the reward function f or the discount function δ; see Remark 2.2. In the typical setup (2.2), it is well-known that the optimization problems are time-consistent with $\delta(t) = e^{-\beta t}$ for some $\beta > 0$ (i.e., the case of exponential discounting) but time-inconsistent in general.

As proposed in Strotz [26], a sensible reaction to time inconsistency is to take future selves' disobedience into account and choose the best present action in response to that. Assuming that all future selves reason in the same way, the agent searches for a (subgame perfect) equilibrium strategy from which no future self has an incentive to deviate. To formulate such an equilibrium strategy, we introduce, for any $\pi, \pi' \in \Pi$, the concatenation of π' and π at time 1, denoted by $\pi' \otimes_1 \pi \in \Pi$. Using $\pi' \otimes_1 \pi \in \Pi$ means that the evolution of X is governed first by π' at time 0 and then by π from time 1 onward. That is, at time 0, given that $X_0 = i \in \mathcal{S}$, we sample $u \in U$ according to the measure $\pi'(i) \in \mathcal{P}(U)$ and the realization of X_1 is then governed by the transition probabilities $p^u_i = \{p^u_{ij}\}_{j\in\mathcal{S}}$. At any time $t \ge 1$, given that $X_t = j \in \mathcal{S}$, we sample $\bar{u} \in U$ according to the measure $\pi(j) \in \mathcal{P}(U)$ and the realization of X_{t+1} is then governed by $p^{\bar{u}}_j = \{p^{\bar{u}}_{jk}\}_{k\in\mathcal{S}}$.

Definition 2.2. *Given $\lambda > 0$, we say $\pi \in \Pi_r$ is a regular relaxed equilibrium (for (2.4)) if for any $i \in \mathcal{S}$,*

$$J_\lambda^{\pi' \otimes_1 \pi}(i) - J_\lambda^\pi(i) \leq 0, \quad \forall \pi' \in \Pi_r. \tag{2.7}$$

Similarly, we say $\pi \in \Pi$ is a relaxed equilibrium (for (2.1)) if for any $i \in \mathcal{S}$,

$$J^{\pi' \otimes_1 \pi}(i) - J^\pi(i) \leq 0, \quad \forall \pi' \in \Pi. \tag{2.8}$$

2.1 Existence of Regular Relaxed Equilibria for Entropy-Regularized MDP (2.4)

Let us first focus on the existence of regular relaxed equilibria (for (2.4)). Fix $\lambda > 0$. For any $\pi \in \Pi_r$, let us introduce the auxiliary value function

$$V_\lambda^\pi(i) := \mathbb{E}_i\left[\sum_{k=0}^\infty \Big(f^{\pi(X_k^\pi)}(1+k, X_k^\pi) + \lambda\delta(1+k)\mathcal{H}\big(\pi(X_k^\pi)\big)\Big)\right], \quad \forall i \in \mathcal{S}. \tag{2.9}$$

For convenience, we will commonly write V_λ^π as a vector in $\mathbb{R}^d$, i.e., $V_\lambda^\pi = (V_\lambda^\pi(1), V_\lambda^\pi(2), ..., V_\lambda^\pi(d))$. To understand the meaning of V_λ^π, we need to take a closer look at $J_\lambda^\pi(i)$ in (2.4). Given that $X_0 = i \in \mathcal{S}$, notice that the first term in the summation of (2.4) is no longer random and can be computed immediately. The remaining terms in the summation are still random and their expectation depends on the realization of X_1^π. Specifically,

$$\mathbb{E}_i\left[\sum_{k=1}^\infty \Big(f^{\pi(X_k^\pi)}(k, X_k^\pi) + \lambda\delta(k)\mathcal{H}\big(\pi(X_k^\pi)\big)\Big) \,\middle|\, X_1^\pi = j\right] = V_\lambda^\pi(j), \quad \forall j \in \mathcal{S}, \tag{2.10}$$

where the equality follows from the Markov property of X^π. That is, $V_\lambda^\pi(j)$ is the expectation of future rewards plus entropy conditioned on $X_1^\pi = j$.

Now, for the problem (2.4), given $X_0 = i \in \mathcal{S}$ and that all future selves will follow a relaxed control $\pi \in \Pi_r$, the agent at time 0 intends to find her best strategy (denoted by $\pi' \in \Pi_r$) in response to that. Observe from (2.4) and (2.10) that

$$\begin{aligned} J_\lambda^{\pi' \otimes_1 \pi}(i) &= f^{\pi'(i)}(0, i) + \lambda\mathcal{H}(\pi'(i)) + \mathbb{E}_i[V_\lambda^\pi(X_1^{\pi'})] \\ &= \int_U \Big(f(0, i, u) - \lambda \ln((\pi'(i))(u)) + p_i^u \cdot V_\lambda^\pi\Big)(\pi'(i))(u)du. \end{aligned}$$

Hence, the best strategy $\pi' \in \Pi_r$ for the time-0 agent should satisfy

$$\pi'(i) \in \operatorname*{arg\,max}_{\rho \in \mathcal{D}(U)} \int_U \Big(f(0, i, u) - \lambda \ln(\rho(u)) + p_i^u \cdot V_\lambda^\pi\Big)\rho(u)du, \quad \forall i \in \mathcal{S}. \tag{2.11}$$

Note that the set of maximizers on the right-hand side is a singleton and its unique element takes the explicit form

$$\rho_i^*(u) := \frac{e^{\frac{1}{\lambda}\left(f(0,i,u) + p_i^u \cdot V_\lambda^\pi\right)}}{\int_U e^{\frac{1}{\lambda}\left(f(0,i,u) + p_i^u \cdot V_\lambda^\pi\right)}du}, \quad u \in U.$$

As a result, by introducing a functional $\Gamma_\lambda : \mathbb{R}^d \times \mathcal{S} \to \mathcal{D}(U)$ defined by

$$u \mapsto \Gamma_\lambda(y, i)(u) := \frac{e^{\frac{1}{\lambda}\left(f(0,i,u) + p_i^u \cdot y\right)}}{\int_U e^{\frac{1}{\lambda}\left(f(0,i,v) + p_i^v \cdot y\right)}dv} \in \mathcal{D}(U), \quad \forall (y, i) \in \mathbb{R}^d \times \mathcal{S}, \tag{2.12}$$

we can re-write (2.11) as

$$\pi'(i) = \Gamma_\lambda(V_\lambda^\pi, i)(\cdot) \in \mathcal{D}(U), \quad \forall i \in \mathcal{S}. \tag{2.13}$$

We have argued so far that $\pi' \in \Pi_r$ in (2.13) is the time-0 agent's best response to her future selves using $\pi \in \Pi_r$. If it happens that the best response $\pi' \in \Pi_r$ coincides with future selves' strategy $\pi \in \Pi_r$ (i.e., $\Gamma_\lambda(V_\lambda^\pi, \cdot) = \pi$), then $\pi \in \Pi_r$ should be viewed as an equilibrium among the current and future selves, as it is a strategy that will be upheld over time. This motivates us to define an operator $\Phi_\lambda : \Pi_r \to \Pi_r$ by

$$\Phi_\lambda(\pi) := \Gamma_\lambda(V_\lambda^\pi, \cdot), \tag{2.14}$$

and we conjecture that a fixed point of Φ_λ is a regular relaxed equilibrium for (2.4). In addition, as $\Gamma_\lambda(V_\lambda^\pi, \cdot) = \pi$ consequently yields $V_\lambda^{\Gamma_\lambda(V_\lambda^\pi, \cdot)} = V_\lambda^\pi$, we also define an operator $\Psi_\lambda : \mathbb{R}^d \to \mathbb{R}^d$ by

$$\Psi_\lambda(y) := V_\lambda^{\Gamma_\lambda(y, \cdot)},$$

and conjecture that a fixed point of Ψ_λ must equal V_λ^π for some regular relaxed equilibrium $\pi \in \Pi_r$.

The next two results show that our conjectures are correct.

Proposition 2.1. *Let Assumption 2.1 hold. For any $\lambda > 0$,*

$$\pi \in \Pi_r \text{ is a regular relaxed equilibrium} \iff \Phi_\lambda(\pi) = \pi.$$

Proof. Given $\pi, \pi' \in \Pi_r$, since $J_\lambda^{\pi' \otimes_1 \pi}(i) = f^{\pi'(i)}(0, i) + \lambda \mathcal{H}(\pi'(i)) + \mathbb{E}_i[V_\lambda^\pi(X_1^{\pi'})]$ for all $i \in \mathcal{S}$, a direct calculation shows

$$\begin{aligned} J_\lambda^{\pi' \otimes_1 \pi}(i) - J_\lambda^\pi(i) = f^{\pi'(i)}(0, i) + \lambda \mathcal{H}(\pi'(i)) + \int_U (p_i^u \cdot V_\lambda^\pi)(\pi'(i))(u) du \\ - \left(f^{\pi(i)}(0, i) + \lambda \mathcal{H}(\pi(i)) + \int_U (p_i^u \cdot V_\lambda^\pi)(\pi(i))(u) du \right), \quad \forall i \in \mathcal{S}. \end{aligned}$$

It follows that (2.7) holds (i.e., π is a regular relaxed equilibrium) if and only if $\pi(i) \in \mathcal{D}(U)$ fulfills

$$\pi(i) \in \arg\max_{\rho \in \mathcal{D}(U)} \int_U \Big(f(0, i, u) - \lambda \ln(\rho(u)) + p_i^u \cdot V_\lambda^\pi \Big) \rho(u) du, \quad \forall i \in \mathcal{S}. \tag{2.15}$$

By the same arguments in (2.11)-(2.13), we can express (2.15) equivalently as $\pi(i) = \Gamma_\lambda(V_\lambda^\pi, i)$ for all $i \in \mathcal{S}$, which amounts to $\pi = \Phi_\lambda(\pi)$. □

Corollary 2.1. *Let Assumption 2.1 hold and $\lambda > 0$. For any $y \in \mathbb{R}^d$,*

$$y = V_\lambda^\pi \text{ for some regular relaxed equilibrium } \pi \in \Pi_r \iff \Psi_\lambda(y) = y.$$

In particular, $\Psi_\lambda(y) = y$ implies that $\Gamma_\lambda(y, \cdot) \in \Pi_r$ is a regular relaxed equilibrium.

Proof. If $y = V_\lambda^\pi$ for a regular relaxed equilibrium $\pi \in \Pi_r$, then by Proposition 2.1, $\pi = \Phi_\lambda(\pi) = \Gamma_\lambda(V_\lambda^\pi, \cdot) = \Gamma_\lambda(y, \cdot)$, which implies $y = V_\lambda^\pi = V_\lambda^{\Gamma_\lambda(y, \cdot)} = \Psi_\lambda(y)$. Conversely, if $y = \Psi_\lambda(y) = V_\lambda^{\Gamma_\lambda(y, \cdot)}$, set $\pi := \Gamma_\lambda(y, \cdot) \in \Pi_r$. Then, we have $y = V_\lambda^\pi$ and thus $\pi = \Gamma_\lambda(y, \cdot) = \Gamma_\lambda(V_\lambda^\pi, \cdot) = \Phi_\lambda(\pi)$. By Proposition 2.1, this implies that π is a regular relaxed equilibrium. □

To properly control the entropy of the fixed-point operator Φ_λ in (2.14), we need the following two assumptions.

Assumption 2.2. *The maps $u \mapsto f(t,i,u)$ and $u \mapsto p_i^u$ are Lipschitz, uniformly in (t,i), i.e.,*

$$\Theta := \sup_{t\in\mathbb{N}_0, i\in\mathcal{S}} \sup_{u_1,u_2\in U, u_1\neq n_2} \left\{ \frac{|f(i,t,u_1)-f(i,t,u_2)|}{|u_1-u_2|} + \frac{|p_i^{u_1}-p_i^{u_2}|}{|u_1-u_2|} \right\} < \infty. \tag{2.16}$$

We will also assume that the action space $U \subset \mathbb{R}^\ell$ fulfills a *uniform cone condition*. To properly state the condition, for any $\iota \in [0,\pi/2]$, we note that

$$\Delta_\iota := \{u = (u_1, ..., u_\ell) \in \mathbb{R}^\ell : u_1^2 + ... + u_{\ell-1}^2 \leq \tan^2(\iota) u_\ell^2\}$$

is a cone with vertex, axis, and angle being $0 \in \mathbb{R}^{d-1}$, $u_1 = u_2 = ... = u_{\ell-1} = 0$, and ι, respectively. Now, given $u \in \mathbb{R}^\ell$, the region obtained by a rotation of $u + \Delta_\iota$ in $\mathbb{R}^\ell$ about u will be called *a cone with vertex u and angle ι*.

Assumption 2.3. *When $\ell > 1$, there exists $\vartheta > 0$ and $\iota \in (0,\pi/2]$ such that for any $u \in U$, there is a cone with vertex u and angle ι (denoted by $\mathrm{cone}(u,\iota)$) that satisfies $(\mathrm{cone}(u,\iota) \cap B_\vartheta(u)) \subseteq U$. When $\ell = 1$, there exists $\vartheta > 0$ such that for any $u \in U$, either $[u-\vartheta, u]$ or $[u, u+\vartheta]$ is contained in U.*

Remark 2.3. *Assumption 2.3 states that a cone with a fixed size (determined by slant height ϑ and angle α) can be attached to any $u \in U$ (i.e., taking u as its vertex) such that the cone is contained entirely in U. This readily covers all polyhedrons and ellipsoids in $\mathbb{R}^\ell$.*

We can now establish a key estimate of the entropy of the fixed-point operator Φ_λ in (2.14), whose proof is relegated to Appendix A.1.

Lemma 2.1. *Let Assumptions 2.2 and 2.3 hold. Then,*

$$\sup_{i\in\mathcal{S}} |\mathcal{H}(\Gamma_\lambda(y,i)| \leq \phi(|y|), \quad \forall y \in \mathbb{R}^d, \tag{2.17}$$

where $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is defined by $\phi(z) := \kappa + \ell \ln(1+z)$, with $\kappa > 0$ depending on only ℓ, λ, $\texttt{Leb}(U)$, ι, ϑ, and Θ. Moreover, for $\lambda > 0$ small enough, (2.17) can be improved to

$$\sup_{i\in\mathcal{S}} |\mathcal{H}(\Gamma_\lambda(y,i))| \leq \varphi(\lambda, y) := \kappa_1 + \kappa_2 |\ln\lambda| + \ell \ln(1+|y|), \quad \forall y \in \mathbb{R}^d, \tag{2.18}$$

where $\kappa_1, \kappa_2 > 0$ depend on ℓ, $\texttt{Leb}(U)$, ι, ϑ, and Θ, but not on λ.

Now, we are ready to present the existence of a regular relaxed equilibrium, whose proof is relegated to Appendix A.2.

Theorem 2.1. *Let Assumptions 2.1, 2.2, and 2.3 hold. For any $\lambda > 0$, there exists $y \in \mathbb{R}^d$ such that $\Psi_\lambda(y) = y$. Hence, $\Gamma_\lambda(y,\cdot) \in \Pi_r$ is a regular relaxed equilibrium for (2.4).*

Interestingly, as the degree of regularization tends to zero (i.e., $\lambda \downarrow 0$ in (2.4)), the next result shows that the values generated by the corresponding regular relaxed equilibria (whose existence is guaranteed by Theorem 2.1) are uniformly bounded, thanks to the entropy estimate in Lemma 2.1. Its proof is relegated to Appendix A.3.

Lemma 2.2. *Let Assumptions 2.1, 2.2, and 2.3 hold. Given $\{\lambda_n\}_{n\in\mathbb{N}}$ in $(0,1]$ with $\lambda_n \downarrow 0$, consider $\{y^n\}_{n\in\mathbb{N}}$ in $\mathbb{R}^d$ such that $y^n = \Psi_{\lambda_n}(y^n)$. Then, $\sup_{n\in\mathbb{N}} |y^n| < \infty$.*

2.2 Existence of Relaxed Equilibria for MDP (2.1)

Now, we move on to prove the existence of a relaxed equilibrium for the original MDP (2.1). Similarly to (2.9), for any $\pi \in \Pi$, we introduce the auxiliary value function

$$V^\pi(i) := \mathbb{E}_i\left[\sum_{k=0}^{\infty} f^{\pi(X_k^\pi)}\left(1+k, X_k^\pi\right)\right], \quad \forall i \in \mathcal{S}. \tag{2.19}$$

For convenience, we will commonly write V^π as a vector in $\mathbb{R}^d$, i.e., $V^\pi = (V^\pi(1), V^\pi(2), ..., V^\pi(d))$. Suppose that $u \mapsto f(t,i,u)$ and $u \mapsto p_i^u$ are continuous. As U is compact, for any $(y,i) \in \mathbb{R}^d \times \mathcal{S}$,

$$E(y,i) := \arg\max_{u \in U} \{f(0,i,u) + p_i^u \cdot y\} \subseteq U \tag{2.20}$$

is nonempty and closed. For any $y \in \mathbb{R}^d$, consider the collection

$$\Gamma(y) := \{\pi \in \Pi : \mathrm{supp}(\pi(i)) \subseteq E(y,i),\ \forall i \in \mathcal{S}\}, \tag{2.21}$$

where $\mathrm{supp}(\rho)$ denotes the support of $\rho \in \mathcal{D}(U)$. We can then define a set-valued operator $\Psi : \mathbb{R}^d \to 2^{\mathbb{R}^d}$ by

$$\Psi(y) := \{V^\pi : \pi \in \Gamma(y)\} \subseteq \mathbb{R}^d. \tag{2.22}$$

Moreover, we can also define a set-valued operator $\Phi : \Pi \to 2^\Pi$ by

$$\Phi(\pi) := \Gamma(V^\pi) \subseteq \Pi. \tag{2.23}$$

Let us provide the following characterizations of relaxed equilibrium for (2.1).

Proposition 2.2. *Let Assumption 2.1 hold and the maps $u \mapsto f(t,i,u)$ and $u \mapsto p_i^u$ be continuous. Then,*

$$\pi \in \Pi \text{ is a relaxed equilibrium} \iff \pi \in \Phi(\pi). \tag{2.24}$$

Moreover, for any $y \in \mathbb{R}^d$,

$$y = V^\pi \text{ for some relaxed equilibrium } \pi \in \Pi \iff y \in \Psi(y).$$

Proof. By the same argument in the proof of Proposition 2.1 (while ignoring the term $\lambda\mathcal{H}(\cdot)$ therein), we observe that $\pi \in \Pi$ is a relaxed equilibrium if and only if $\pi(i) \in \mathcal{P}(U)$ fulfills

$$\pi(i) \in \arg\max_{\mu \in \mathcal{P}(U)} \int_U \left(f(0,i,u) + p_i^u \cdot V^\pi\right) \mu(du), \quad \forall i \in \mathcal{S},$$

which is equivalent to $\mathrm{supp}(\pi(i)) \subseteq E(V^\pi, i)$ for all $i \in \mathcal{S}$, i.e., $\pi \in \Gamma(V^\pi) = \Phi(\pi)$.

For any $y \in \mathbb{R}^d$ such that $y \in \Psi(y)$, in view of (2.22), $y = V^\pi$ for some $\pi \in \Gamma(y)$. It follows that $\pi \in \Gamma(y) = \Gamma(V^\pi) = \Phi(\pi)$. By (2.24), this implies that π is a relaxed equilibrium. Conversely, suppose that $y = V^\pi$ for some relaxed equilibrium $\pi \in \Pi$. By (2.24), $\pi \in \Phi(\pi) = \Gamma(V^\pi) = \Gamma(y)$. With $y = V^\pi$ and $\pi \in \Gamma(y)$, we immediately conclude $y \in \Psi(y)$. □

With the aid of Lemma 2.2 and Proposition 2.2, we are ready to present the existence of a relaxed equilibrium for (2.1), by approximating (2.1) using a sequence of entropy-regularized MDPs. The detailed proof is relegated to Appendix A.4.

Theorem 2.2. *Let Assumptions 2.1, 2.2, and 2.3 hold. Then, a relaxed equilibrium for (2.1) exists.*

This paper mainly takes the entropy-regularized MDP (2.4) as a powerful approximation tool for the original MDP (2.1), as shown in the proof of Theorem 2.2 (see Appendix A.4). Yet, (2.4) does have its own meaning and application. In the context of reinforcement learning (RL), an agent does not know the model perfectly (e.g., the MDP's transition probabilities may not be precisely known). She then chooses her control actions for two different purposes—one is to enlarge her cumulative payoff based on her present knowledge of the model (i.e., "exploitation"); the other is to obtain more information about the model based on the observed MDP evolution (i.e., "exploration"). To enhance "exploration," the agent randomizes her control actions (i.e., chooses a relaxed control) to more efficiently infer the model (from the more diverse MDP evolution), and the amount of information gained can be measured by Shannon's entropy of the randomization. This corresponds to the second term in the expectation of (2.4). The chosen relaxed control also needs to serve the "exploitation" purpose, which corresponds to the first term in the expectation of (2.4).

Remark 2.4. *For typical RL without time inconsistency (e.g., $\delta(t) = e^{-\beta t}$ for some $\beta > 0$ in (2.4)), the agent's goal is to find a relaxed control that maximizes (2.4), thereby striking a balance between "exploitation" and "exploration" (see e.g., Ziebart et al. [30], Fox et al. [16], Wang et al. [28]). When δ is a general discount function, the agent also needs to tackle the issue of time inconsistency. That is, besides striking a balance between "exploitation" and "exploration," she also wants to maintain the balance among all disobedient future selves. The agent then aims for a relaxed control for (2.4) that can be upheld by all current and future selves, i.e., a regular relaxed equilibrium in Definition 2.2. Theorem 2.1 asserts that such a desired RL policy exists.*

Remark 2.5. *For $\lambda > 0$, let $\pi_\lambda \in \Pi_r$ be a regular relaxed equilibrium for (2.4). Given $i \in \mathcal{S}$, $\pi_\lambda(i) \in \mathcal{P}(U)$ admits a density function for all $\lambda > 0$ (as π_λ is regular; see Definition 2.1). However, as $\lambda \downarrow 0$, the weak limit $\pi^*(i)$ of $\{\pi_\lambda(i)\}_{n\in\mathbb{N}}$ may not admit a density function. This is why in Theorem 2.2, we get only a relaxed equilibrium, which is not necessarily regular, for (2.1).*

Remark 2.6. *While Theorem 2.2 is a general existence result, its proof does suggest how we can actually find a relaxed equilibrium: as the original MDP (2.1) can be approximated by a sequence of entropy-regularized ones (indexed by $\lambda > 0$), one can compute a relaxed equilibrium for each regularized problem and the limit (as $\lambda \to 0$) will be a relaxed equilibrium of the original problem.*

This method is numerically viable as it circumvents the set-valued fixed-point operator associated with (2.1), i.e., Φ in (2.23). Indeed, it is difficult numerically to implement a set-valued fixed-point iteration, such that finding a relaxed equilibrium for (2.1) directly is not easy at all. By contrast, as each regularized problem is associated with a single-valued fixed-point operator, i.e., Φ_λ in (2.14), a fixed-point iteration can be implemented in a straightforward way. Certainly, to make this method fully rigorous, it remains to show the theoretic convergence of the single-valued fixed-point iteration, which is a nontrivial problem in itself and will be left for future research.

3 Relaxed Equilibria in Continuous Time

In this section, we take up the same setup in the first two paragraphs of Section 2, except that the controlled process X is now a continuous-time Markov chain. Specifically, each action $u \in U$ is associated with a $d \times d$ rate matrix (or, generator) q^u; namely, for each fixed $i \in \mathcal{S}$, $q^u_{ij} \geq 0$ for all $j \neq i$ and $q^u_{ii} = -\sum_{j\neq i} q^u_{ij}$. In addition, each $\mu \in \mathcal{P}(U)$ is associated with a $d \times d$ *relaxed* rate matrix Q^μ, defined by $Q^\mu_{ij} := \int_U q^u_{ij}\mu(du)$ for all $i, j \in \mathcal{S}$. At any current state $i \in \mathcal{S}$, given a relaxed feedback control $\pi : \mathcal{S} \to \mathcal{P}(U)$, the dynamics of X is dictated by $Q^{\pi(i)}_i$, the i^{th} row of $Q^{\pi(i)}$. That

is, the time until the next jump to other states is exponentially distributed with parameter $-Q_{ii}^{\pi(i)}$ and the jump will take X to state $j \neq i$ with probability $-Q_{ij}^{\pi(i)}/Q_{ii}^{\pi(i)}$.

For any $\pi \in \Pi$, the corresponding value function is given by

$$\widetilde{J}^{\pi}(i) := \mathbb{E}_i\bigg[\int_0^{\infty} f^{\pi(X_s^{\pi})}\left(s, X_s^{\pi}\right) ds\bigg], \quad \forall i \in \mathcal{S}. \tag{3.1}$$

Similarly to (2.4), for any $\lambda > 0$ and $\pi \in \Pi_r$, we consider

$$\widetilde{J}_{\lambda}^{\pi}(i) := \mathbb{E}_i\bigg[\int_0^{\infty} \Big(f^{\pi(X_s)}(s, X_s^{\pi}) + \lambda\delta(s)\mathcal{H}\big(\pi(X_s)\big)\Big) ds\bigg], \quad \forall i \in \mathcal{S}, \tag{3.2}$$

To ensure that $\widetilde{J}^{\pi}$ and $\widetilde{J}_{\lambda}^{\pi}$ are well-defined, we impose the following.

Assumption 3.1. $\widetilde{M} := \int_0^{\infty} \big(\sup_{i\in\mathcal{S}, u\in U} |f(s,i,u)| + \delta(s)\big)\, ds < \infty$.

We will commonly write $\widetilde{J}^{\pi}$ and $\widetilde{J}_{\lambda}^{\pi}$ as vectors in $\mathbb{R}^d$, i.e., $\widetilde{J}^{\pi} = (\widetilde{J}^{\pi}(1), \widetilde{J}^{\pi}(2), ..., \widetilde{J}^{\pi}(d))$ and $\widetilde{J}_{\lambda}^{\pi} = (\widetilde{J}_{\lambda}^{\pi}(1), \widetilde{J}_{\lambda}^{\pi}(2), ..., \widetilde{J}_{\lambda}^{\pi}(d))$.

Similarly to Definition 2.2, to formulate an equilibrium strategy in continuous time, we introduce, for any $\pi, \pi' \in \Pi$, the concatenation of π' and π at time $\varepsilon > 0$, denoted by $\pi' \otimes_{\varepsilon} \pi \in \Pi$. Using this concatenated relaxed control means that the evolution of X is governed first by π' on the time interval $[0, \varepsilon)$ and then by π on $[\varepsilon, \infty)$.

Definition 3.1. *Given $\lambda > 0$, we say $\pi \in \Pi_r$ is a regular relaxed equilibrium (for (3.2)) if*

$$\limsup_{\varepsilon \downarrow 0} \frac{\widetilde{J}_{\lambda}^{\pi' \otimes_{\varepsilon} \pi}(i) - \widetilde{J}_{\lambda}^{\pi}(i)}{\varepsilon} \leq 0, \quad \forall \pi' \in \Pi_r \text{ and } i \in \mathcal{S}. \tag{3.3}$$

Similarly, we say $\pi \in \Pi$ is a relaxed equilibrium (for (3.1)) if it satisfies (3.3) with $\widetilde{J}_{\lambda}^{\pi' \otimes_{\varepsilon} \pi}$, $\widetilde{J}_{\lambda}^{\pi}$, and $\pi' \in \Pi_r$ therein replaced by $\widetilde{J}^{\pi' \otimes_{\varepsilon} \pi}$, $\widetilde{J}^{\pi}$, and $\pi' \in \Pi$, respectively.

Similarly to (2.19), for any $\pi \in \Pi$, we introduce the auxiliary value function

$$\widetilde{V}^{\pi}(t, i) := \mathbb{E}_i\bigg[\int_0^{\infty} f^{\pi(X_s)}(t+s, X_s) ds\bigg], \quad \forall (t, i) \in [0, \infty) \times \mathcal{S}. \tag{3.4}$$

In addition, similarly to (2.9), for any $\lambda > 0$ and $\pi \in \Pi_r$, we introduce the auxiliary value function

$$\widetilde{V}_{\lambda}^{\pi}(t, i) := \mathbb{E}_i\left[\int_0^{\infty} \Big(f^{\pi(X_s)}(t+s, X_s) + \lambda\delta(t+s)\mathcal{H}(\pi(X_s))\Big) ds\right], \quad \forall (t, i) \in [0, \infty) \times \mathcal{S}. \tag{3.5}$$

For convenience, we will commonly write $\widetilde{V}_{\lambda}^{\pi}(t)$ as a vector in $\mathbb{R}^d$, i.e.,

$$\widetilde{V}_{\lambda}^{\pi}(t) = (\widetilde{V}_{\lambda}^{\pi}(t, 1), \widetilde{V}_{\lambda}^{\pi}(t, 2), ..., \widetilde{V}_{\lambda}^{\pi}(t, d)) \in \mathbb{R}^d.$$

We will write $\widetilde{V}^{\pi}$ as a vector in $\mathbb{R}^d$ in the same manner.

3.1 Existence of Regular Relaxed Equilibria for Entropy-Regularized MDP (3.2)

Lemma 3.1. *Let Assumption 3.1 hold and $f(\cdot,i,u)$ and $\delta(\cdot)$ be continuous on $[0,\infty)$. Given $\lambda>0$, it holds for all $i\in\mathcal{S}$ and $\pi,\pi'\in\Pi_r$ that*

$$\widetilde{J}_\lambda^{\pi'\otimes_\varepsilon\pi}(i)=\widetilde{V}_\lambda^{\pi}(\varepsilon,i)+\Big(f^{\pi'(i)}(0,i)+\lambda\mathcal{H}(\pi'(i))+Q_i^{\pi'(i)}\cdot\widetilde{V}_\lambda^{\pi}(\varepsilon)\Big)\varepsilon+o(\varepsilon),\quad \text{as } \varepsilon\downarrow 0. \tag{3.6}$$

Similarly, it holds for all $i\in\mathcal{S}$ and $\pi,\pi'\in\Pi$ that

$$\widetilde{J}^{\pi'\otimes_\varepsilon\pi}(i)=\widetilde{V}^{\pi}(\varepsilon,i)+\Big(f^{\pi'(i)}(0,i)+Q_i^{\pi'(i)}\cdot\widetilde{V}^{\pi}(\varepsilon)\Big)\varepsilon+o(\varepsilon),\quad \text{as } \varepsilon\downarrow 0. \tag{3.7}$$

Proof. The result follows from a similar argument in Huang and Zhou [21, Lemma 1]. □

Based on Lemma 3.1, we can now generalize the fixed-point characterization in Proposition 2.1 to continuous time. Consider a functional $\widetilde{\Gamma}_\lambda:\mathbb{R}^d\times\mathcal{S}\to\mathcal{D}(U)$ defined by

$$u\mapsto\widetilde{\Gamma}_\lambda(y,i)(u):=\frac{e^{\frac{1}{\lambda}\left(f(0,i,u)+q_i^u\cdot y\right)}}{\int_U e^{\frac{1}{\lambda}\left(f(0,i,v)+q_i^v\cdot y\right)}dv}\in\mathcal{D}(U),\quad\forall(y,i)\in\mathbb{R}^d\times\mathcal{S}. \tag{3.8}$$

It follows that $\widetilde{\Gamma}_\lambda(y,\cdot)\in\Pi_r$ for all $y\in\mathbb{R}^d$. We can then define an operator $\widetilde{\Psi}_\lambda:\mathbb{R}^d\to\mathbb{R}^d$ by

$$\widetilde{\Psi}_\lambda(y):=\widetilde{J}_\lambda^{\widetilde{\Gamma}_\lambda(y,\cdot)}.$$

Moreover, for any $\pi\in\Pi_r$, as $\widetilde{\Gamma}_\lambda(\widetilde{J}_\lambda^\pi,\cdot)\in\Pi_r$, we can define an operator $\widetilde{\Phi}_\lambda:\Pi_r\to\Pi_r$ by

$$\widetilde{\Phi}_\lambda(\pi):=\widetilde{\Gamma}_\lambda\big(\widetilde{J}_\lambda^\pi,\cdot\big).$$

Proposition 3.1. *Let Assumption 3.1 hold and $f(\cdot,i,u)$ and $\delta(\cdot)$ be continuous on $[0,\infty)$. For any $\lambda>0$,*

$$\pi\in\Pi_r \text{ is a regular relaxed equilibrium}\iff\widetilde{\Phi}_\lambda(\pi)=\pi.$$

The proof of Proposition 3.1 is relegated to Appendix A.5.

Corollary 3.1. *Let Assumption 3.1 hold and suppose that $f(\cdot,i,u)$ and $\delta(\cdot)$ are continuous on $[0,\infty)$. Given $\lambda>0$, it holds for all $y\in\mathbb{R}^d$,*

$$y=\widetilde{J}_\lambda^\pi \text{ for some regular relaxed equilibrium } \pi\in\Pi_r\iff\widetilde{\Psi}_\lambda(y)=y.$$

In particular, $\widetilde{\Psi}_\lambda(y)=y$ implies that $\widetilde{\Gamma}_\lambda(y,\cdot)\in\Pi_r$ is a regular relaxed equilibrium for (3.2).

Proof. If $y=\widetilde{J}_\lambda^\pi$ for a regular relaxed equilibrium $\pi\in\Pi_r$, then by Proposition 3.1, $\pi=\widetilde{\Phi}_\lambda(\pi)=\widetilde{\Gamma}_\lambda(\widetilde{J}_\lambda^\pi,\cdot)=\widetilde{\Gamma}_\lambda(y,\cdot)$, which implies $y=\widetilde{J}_\lambda^\pi=\widetilde{J}_\lambda^{\widetilde{\Gamma}_\lambda(y,\cdot)}=\widetilde{\Psi}_\lambda(y)$. Conversely, if $y=\widetilde{\Psi}_\lambda(y)=\widetilde{J}_\lambda^{\widetilde{\Gamma}_\lambda(y,\cdot)}$, set $\pi:=\widetilde{\Gamma}_\lambda(y,\cdot)\in\Pi_r$. Then, we have $y=\widetilde{J}_\lambda^\pi$ and thus $\pi=\widetilde{\Gamma}_\lambda(y,\cdot)=\widetilde{\Gamma}_\lambda(\widetilde{J}_\lambda^\pi,\cdot)=\widetilde{\Phi}_\lambda(\pi)$. By Proposition 3.1, this implies that π is a regular relaxed equilibrium. □

By a similar argument in the proof of Theorem 2.1 and using Corollary 3.1, we can establish the existence of regular relaxed equilibria for (3.2).

Theorem 3.1. *Let Assumptions 3.1, 2.2 (with p_i^u therein replaced by q_i^u), and 2.3 hold. Suppose that $f(\cdot,i,u)$ and $\delta(\cdot)$ are continuous on $[0,\infty)$. For any $\lambda>0$, there exists $y\in\mathbb{R}^d$ such that $\widetilde{\Psi}_\lambda(y)=y$. Hence, $\widetilde{\Gamma}_\lambda(y,\cdot)\in\Pi_r$ is a regular relaxed equilibrium for (3.2).*

3.2 Existence of Relaxed Equilibrium for MDP (3.1)

Let us now move on to prove the existence of a relaxed equilibrium for (3.1). Suppose that $u \mapsto f(t,i,u)$ and $u \mapsto q_i^u$ are continuous. As U is compact, for any $(y,i) \in \mathbb{R}^d \times \mathcal{S}$,

$$\widetilde{E}(y,i) := \arg\max_{u \in U} \{f(0,i,u) + q_i^u \cdot y\} \subseteq U \tag{3.9}$$

is nonempty and closed. For any $y \in \mathbb{R}^d$, consider the collection

$$\widetilde{\Gamma}(y) := \{\pi \in \Pi : \operatorname{supp}(\pi(i)) \subseteq E(y,i),\ \forall i \in \mathcal{S}\}, \tag{3.10}$$

where $\operatorname{supp}(\rho)$ denotes the support of $\rho \in \mathcal{D}(U)$. We can then define a set-valued operator $\widetilde{\Psi} : \mathbb{R}^d \to 2^{\mathbb{R}^d}$ by

$$\widetilde{\Psi}(y) := \left\{\widetilde{J}^\pi : \pi \in \widetilde{\Gamma}(y)\right\} \subseteq \mathbb{R}^d. \tag{3.11}$$

Moreover, we can also define a set-valued operator $\widetilde{\Phi} : \Pi \to 2^\Pi$ by

$$\widetilde{\Phi}(\pi) := \widetilde{\Gamma}(\widetilde{J}^\pi) \subseteq \Pi.$$

We can then generalize the fixed-point characterizations in Proposition 2.2 to continuous time.

Proposition 3.2. *Let Assumption 3.1 hold and the maps $u \mapsto f(t,i,u)$ and $u \mapsto q_i^u$ be continuous. Then,*

$$\pi \in \Pi \text{ is a relaxed equilibrium} \iff \pi \in \widetilde{\Phi}(\pi). \tag{3.12}$$

Moreover, for any $y \in \mathbb{R}^d$,

$$y = \widetilde{J}^\pi \text{ for some relaxed equilibrium } \pi \in \Pi \iff y \in \widetilde{\Psi}(y).$$

Proof. By the same argument in the proof of Proposition 3.1 (except that now we use (3.7) instead of (3.6)), we observe that $\pi \in \Pi$ is a relaxed equilibrium if and only if $\pi(i) \in \mathcal{P}(U)$ fulfills

$$\pi(i) \in \arg\max_{\mu \in \mathcal{P}(U)} \int_U \left(f(0,i,u) + q_i^u \cdot \widetilde{J}^\pi\right) \mu(du), \quad \forall i \in \mathcal{S},$$

which is equivalent to $\operatorname{supp}(\pi(i)) \subseteq \widetilde{E}(\widetilde{J}^\pi, i)$ for all $i \in \mathcal{S}$, i.e., $\pi \in \widetilde{\Gamma}(\widetilde{J}^\pi) = \widetilde{\Phi}(\pi)$.

For any $y \in \mathbb{R}^d$ such that $y \in \widetilde{\Psi}(y)$, in view of (3.11), $y = \widetilde{J}^\pi$ for some $\pi \in \widetilde{\Gamma}(y)$. It follows that $\pi \in \widetilde{\Gamma}(y) = \widetilde{\Gamma}(\widetilde{J}^\pi) = \widetilde{\Phi}(\pi)$. By (2.24), this implies that π is a relaxed equilibrium. Conversely, suppose that $y = \widetilde{J}^\pi$ for some relaxed equilibrium $\pi \in \Pi$. By (2.24), $\pi \in \widetilde{\Phi}(\pi) = \widetilde{\Gamma}(\widetilde{J}^\pi) = \widetilde{\Gamma}(y)$. With $y = \widetilde{J}^\pi$ and $\pi \in \widetilde{\Gamma}(y)$, we immediately conclude $y \in \widetilde{\Psi}(y)$. □

Given an arbitrary sequence $\lambda_n \to 0+$ with $v^n = \widetilde{\Psi}_{\lambda_n}(v^n)$ for all $n \in \mathbb{N}$, i.e., $\widetilde{\Gamma}_{\lambda_n}(v)$ is a relaxed equilibrium under λ_n, the next theorem follows from similar arguments in Theorem 2.2.

Theorem 3.2. *Let Assumptions 3.1, 2.2 (with p_i^u therein replaced by q_i^u), and 2.3 hold. Then, a relaxed equilibrium for (3.1) exists.*

4 Discussion: When Do We Need Relaxed Equilibria?

4.1 The Discrete-time Case

Let us consider the discrete-time setup in Section 2. We first draw a detailed comparison between Theorem 2.2 and Huang and Zhou [21, Theorem 4].

Remark 4.1. *Huang and Zhou [21] consider the special case where $d = \ell$ and*

$$p_i^u := u_i \quad \forall i \in \mathcal{S}, \quad \text{with } u_i \in U \subseteq \{y \in \mathbb{R}^d : y \geq 0,\ y_1 + y_2 + ... + y_d = 1\}. \tag{4.1}$$

This essentially states that one can directly "decide" (rather than just "influence") the transition probabilities. By contrast, Theorem 2.2 only requires the map $u \mapsto p_i^u$ to be Lipschitz (Assumption 2.2), without specifying any specific form of it. That is, Theorem 2.2 allows for more general (and potentially more realistic) dependence of transition probabilities on the control action $u \in U$.

Remark 4.2. *By [21, Theorem 4], a standard equilibrium for (2.1) exists, provided that (i) (4.1) holds, (ii) U is convex, and (iii) $f(0, i, \cdot)$ is concave for all $i \in \mathcal{S}$. If one of the conditions fails, it is unclear whether standard equilibria exist.*

In particular, when (iii) fails, [21, Remark 10] shows that standard equilibria do not exist in a concrete example, where $\mathcal{S} = \{1, 2\}$, $\delta(\cdot)$ is a quasi-hyperbolic discount function, $f(t, i, u)$ is of the form (2.2) with $g(i, u)$ bounded and continuous (but non-concave) in u, and $p_i^u := u_i$ $\forall i \in \{1, 2\}$ with $u_i \in U = \{y \in \mathbb{R}^2 : y \geq 0, y_1 + y_2 = 1\}$. This simple setup allows us to verify Assumptions 2.1, 2.2, and 2.3 immediately. Hence, although there is no standard equilibrium, a relaxed equilibrium exists in this example by Theorem 2.2.

Remarks 4.1 and 4.2 show the importance of Theorem 2.2: unlike standard equilibria, a relaxed equilibrium for (2.1) exists much more generally, without the need of conditions (i), (ii), and (iii) in Remark 4.2. Hence, in the face of general controlled transition probabilities (beyond the special form (4.1)) and non-concave reward functions, relaxed equilibria should certainly be considered.

After learning that relaxed equilibria are needed in cases where standard equilibria may not exist, let us now investigate the other side of the picture: for cases where both standard and relaxed equilibria are known to exist, does the consideration of relaxed equilibria provide any added value? As shown in the next result, it is *not* necessary to consider relaxed equilibria in such a case, as any value achieved by a relaxed equilibrium can be recovered by a suitable standard equilibrium.

Proposition 4.1. *Assume (4.1) and that U therein is convex. Suppose that $f(t, i, u)$ takes the form (2.2), with $g(i, \cdot)$ therein concave for all $i \in \mathcal{S}$. Also, let Assumptions 2.1, 2.2, and 2.3 hold. Then, for any relaxed equilibrium $\pi \in \Pi$ for (2.1), $\alpha^\pi : \mathcal{S} \to \mathbb{R}^d$ defined by $\alpha^\pi(i) := \int_U u(\pi(i))(du)$, $\forall i \in \mathcal{S}$, is a standard equilibrium such that $J^{\alpha^\pi}(\cdot) = J^\pi(\cdot)$.*

Proof. Note that the convexity of U ensures that $\alpha^\pi(i) \in U$ for all $i \in \mathcal{S}$. As $\pi \in \Pi$ is a relaxed equilibrium for (2.1), by (2.24),

$$\text{supp}(\pi(i)) \subseteq E(V^\pi, i) = \arg\max_{u \in U} \{g(i, u) + u \cdot V^\pi\}, \quad \forall i \in \mathcal{S}, \tag{4.2}$$

where the equality follows from (2.20) and (4.1). Given $i \in \mathcal{S}$, if $E(V^\pi, i)$ is a singleton, (4.2) implies that $\pi(i) \in \mathcal{P}(U)$ concentrates on the unique element in $E(V^\pi, i)$, which coincides with α^π by definition. Hence, $f^{\pi(i)}(t, i) = f^{\alpha^\pi}(t, i)$ trivially holds for all $t \in \mathbb{N}_0$. If $E(V^\pi, i)$ is not a singleton, in view of its form in (4.2), $u \mapsto g(i, u)$ must be linear on $E(V^\pi, i)$. It follows that

$$f^{\pi(i)}(t, i) = \int_U \delta(t) g(i, u)(\pi(i))(du) = \delta(t) g(i, \alpha^\pi) = f^{\alpha^\pi}(t, i), \quad \forall t \in \mathbb{N}_0.$$

With $f^{\pi(i)}(t,i) = f^{\alpha^\pi}(t,i)$ for all $t \in \mathbb{N}_0$ and $i \in \mathcal{S}$, we conclude from (2.1) that $J^{\alpha^\pi}(\cdot) = J^\pi(\cdot)$. □

It is interesting to note that Jaśkiewicz and Nowak [22] also establish results in the same spirit as Proposition 4.1, under a specific form of discounting.

Remark 4.3. *Under quasi-hyperbolic discounting, [22, Theorem 3.4] asserts the existence of a relaxed equilibrium π that is almost a standard one: for each state i, the support of the probability $\pi(i)$ contains at most two points in the action space. Moreover, [22, Theorem 3.5] shows that additional "atomless" assumptions can further reduce the support of $\pi(i)$ to contain only one point, i.e., the relaxed equilibrium reduces to a standard one. As opposed to Proposition 4.1, these results hold for fairly general reward functions (without the need of concavity) and transition probabilities (without the need of linear dependence on u).*

Such generality, however, hinges crucially on quasi-hyperbolic discounting. In a nutshell, by using the structure of quasi-hyperbolic discounting, the original time-inconsistent problem can be expressed as a functional of an auxiliary time-consistent problem (under exponential discounting); see [22, (2.3)-(2.4)]. A detailed analysis is then performed on the time-consistent problem, which contributes to proving [22, Theorems 3.4 and 3.5]; see [22, Section 7]. This method, quite specific to quasi-hyperbolic discounting, cannot be easily extended to the case of a general discount function.

Finally, we would like to point out that the need of relaxed controls is unique to the case of time inconsistency. When the problem $\sup_{\pi\in\Pi} J^\pi(i)$ is time-consistent (e.g., under exponential discounting), one can consider without loss of generality only standard controls.

Remark 4.4. *Let $f(t,i,u)$ takes the form (2.2) with $\delta(t) = e^{-\beta t}$ for some $\beta > 0$. As there is no time inconsistency under exponential discounting, the optimal value $J^*(i) := \sup_{\pi\in\Pi} J^\pi(i)$ fulfills the Bellman equation*

$$-J^*(i) + \sup_{\mu\in\mathcal{P}(U)} \int_U \left(g(i,u) + e^{-\beta} p_i^u \cdot J^*\right) \mu(du) = 0, \quad \forall i \in \mathcal{S}. \tag{4.3}$$

Suppose that an optimal relaxed control $\pi^* \in \Pi$ exists, i.e., $J^{\pi^*} = J^*$. Our goal is to show that J^* can be achieved by a standard control—such that there is no need to consider relaxed controls.

Given $i \in \mathcal{S}$, let $u \mapsto p_i^u$ and $u \mapsto g(i,u)$ be continuous such that $\arg\max_{u\in U}\{g(i,u) + e^{-\beta} p_i^u \cdot J^*\}$ is nonempty. For any standard control α with $\alpha(i) \in \arg\max_{u\in U}\{g(i,u) + e^{-\beta} p_i^u \cdot J^*\}$ for all $i \in \mathcal{S}$,

$$J^{\alpha\otimes_1\pi^*}(i) = g(i,\alpha(i)) + e^{-\beta} p_i^{\alpha(i)} \cdot J^* = J^*(i), \quad \forall i \in \mathcal{S},$$

where the second equality follows from (4.3) and $\alpha(i) \in \arg\max_{u\in U}\{g(i,u) + e^{-\beta} p_i^u \cdot J^*\}$. Similarly,

$$J^{\alpha\otimes_2\pi^*}(i) = g(i,\alpha(i)) + e^{-\beta} p_i^{\alpha(i)} \cdot J^{\alpha\otimes_1\pi^*} = g(i,\alpha(i)) + e^{-\beta} p_i^{\alpha(i)} \cdot J^{\pi^*} = J^*(i), \quad \forall i \in \mathcal{S}.$$

Iterating this argument yields $J^{\alpha\otimes_m\pi^*} = J^*$ for all $m \in \mathbb{N}$. It follows that

$$J^*(i) = \lim_{m\to\infty} J^{\alpha\otimes_m\pi^*}(i) = \lim_{m\to\infty} \left(J^\alpha(i) + e^{-\beta m}\mathbb{E}\big[p_{X_{m-1}}^{\alpha(X_{m-1})} \cdot (J^* - J^\alpha)\big]\right) = J^\alpha(i), \quad \forall i \in \mathcal{S}.$$

That is, the optimal value is achieved by the standard feedback control α.

4.2 The Continuous-time Case

Let us consider the continuous-time setup in Section 3. As we will see, many observations we made in Section 4.1 still hold in continuous time. To begin with, we draw a detailed comparison between Theorem 3.2 and Huang and Zhou [21, Theorem 3].

Remark 4.5. *In [21], one considers the special case where $\ell = d-1$ and*

$$q_i^u := \left(u_1, ..., u_{i-1}, -\sum_{j\neq i} u_j, u_{i+1}, ..., u_d\right) \quad \forall i \in \mathcal{S}, \quad \text{with } u \in U \subseteq \{y \in \mathbb{R}^{d-1} : y \geq 0\}. \tag{4.4}$$

This essentially states that one can directly "decide" (rather than just "influence") the transition rates. By contrast, Theorem 3.2 only requires the map $u \mapsto q_i^u$ to be Lipschitz (Assumption 2.2, with p_i^u therein replaced by q_i^u), without specifying any specific form of it. That is, Theorem 3.2 allows for more general (and potentially more realistic) dependence of transition rates on the control action $u \in U$.

By [21, Theorem 3], a standard equilibrium for (3.1) exists, provided that (i) (4.4) holds, (ii) U is convex, and (iii) $f(0, i, \cdot)$ is concave for all $i \in \mathcal{S}$. If one of the conditions fails, it is unclear whether standard equilibria exist. In particular, when (iii) fails, the next example has no standard equilibrium.

Example 4.1. *Let $\mathcal{S} = \{1, 2\}$ and $U = [0, 1]$. For any $u \in U$, we take the rate matrix q^u to be $q_1^u = (-u, u)$ and $q_2^u = (u, -u)$. Consider $\delta(t) := (e^{-t} + e^{-2t})/2$ for $t \geq 0$ and take $f(t, i, u) = \delta(t)g_i(u)$, $i = 1, 2$, where $g_1(u) := -\frac{7}{8}\sqrt{u}$ and $g_2(u) := \frac{19}{9} - \sqrt{1-u}$ are strictly convex on U. Thanks to [21, Theorem 1], $\alpha = (\alpha(1), \alpha(2)) = (a^*, b^*)$ is a standard equilibrium if and only if*

$$\begin{aligned}
&\arg\max_{a\in[0,1]} \left\{g_1(a) - a(V^{(a^*,b^*)}(0,1) - V^{(a^*,b^*)}(0,2))\right\} = a^*,\\
&\arg\max_{b\in[0,1]} \left\{g_2(b) + b(V^{(a^*,b^*)}(0,1) - V^{(a^*,b^*)}(0,2))\right\} = b^*.
\end{aligned}$$

The strict convexity of g_1 and g_2 then implies that there are only four possibilities for a standard equilibrium $\alpha = (a^*, b^*)$, i.e., $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$. By calculations similar to [21, (37)-(39)], we have

$$\begin{cases}
V^{(0,0)}(0,1) - V^{(0,0)}(0,2) &= -\frac{3}{4}\cdot\frac{10}{9} = -\frac{5}{6};\\
V^{(1,0)}(0,1) - V^{(1,0)}(0,2) &= -\frac{5}{12}\cdot(\frac{15}{8} + \frac{1}{9}) \in (-\frac{7}{8}, 0);\\
V^{(0,1)}(0,1) - V^{(0,1)}(0,2) &= -\frac{5}{12}\cdot(2 + \frac{1}{9}) < -\frac{7}{8};\\
V^{(1,1)}(0,1) - V^{(1,1)}(0,2) &= -\frac{7}{24}\cdot(\frac{23}{8} + \frac{1}{9}) \in (-\frac{7}{8}, 0),
\end{cases}$$

which imply

$$\begin{cases}
\arg\max_{b\in[0,1]}\{g_2(b) + b(V^{(0,0)}(0,1) - V^{(0,0)}(0,2))\} = 1 \neq 0;\\
\arg\max_{a\in[0,1]}\{g_1(a) - a(V^{(1,0)}(0,1) - V^{(1,0)}(0,2))\} = 0 \neq 1;\\
\arg\max_{a\in[0,1]}\{g_1(a) - a(V^{(0,1)}(0,1) - V^{(0,1)}(0,2))\} = 1 \neq 0;\\
\arg\max_{a\in[0,1]}\{g_1(a) - a(V^{(1,1)}(0,1) - V^{(1,1)}(0,2))\} = 0 \neq 1.
\end{cases}$$

Therefore, we conclude that there exists no standard equilibrium. Despite this, since Assumptions 3.1, 2.2 (with p_i^u therein replaced by q_i^u), and 2.3 can be immediately verified in the present setting, Theorem 3.2 asserts that a relaxed equilibrium exists.

Remark 4.5 and Example 4.1 show the importance of Theorem 3.2: unlike standard equilibria, a relaxed equilibrium for (2.1) exists much more generally, without the need of conditions (i), (ii), and (iii) mentioned above Example 4.1. Hence, in the face of general controlled transition rates (beyond the special form (4.4)) and non-concave reward functions, relaxed equilibria should certainly be considered.

Let us now investigate the other side of the picture: for cases where both standard and relaxed equilibria are known to exist, does the consideration of relaxed equilibria provide any added value? As shown in the next result, it is *not* necessary to consider relaxed equilibria in such a case, as any value achieved by a relaxed equilibrium can be recovered by a suitable standard equilibrium.

Proposition 4.2. *Assume (4.4) and that U therein is convex. Suppose that $f(t,i,u)$ takes the form (2.2), with $g(i,\cdot)$ therein concave for all $i \in \mathcal{S}$. Also, let Assumptions 2.1, 2.2 (with p_i^u therein replaced by q_i^u), and 2.3 hold. Then, for any relaxed equilibrium $\pi \in \Pi$ for (2.1), $\alpha^\pi : \mathcal{S} \to \mathbb{R}^d$ defined by $\alpha^\pi(i) := \int_U u(\pi(i))(du)$, $\forall i \in \mathcal{S}$, is a standard equilibrium such that $J^{\alpha^\pi}(\cdot) = J^\pi(\cdot)$.*

The proof of Proposition 4.2 is similar to that of Proposition 4.1, with p_i^u and $V^\pi(\cdot)$ therein replaced by q_i^u and $\widetilde{V}^\pi(0,\cdot)$, respectively.

Finally, we would like to point out that, in line with the discrete-time setting, the need of relaxed controls is unique to the case of time inconsistency. When the problem $\sup_{\pi\in\Pi} \widetilde{J}^\pi(i)$ is time-consistent (e.g., under exponential discounting), one only needs to consider standard controls.

Remark 4.6. *Let $f(t,i,u)$ take the form (2.2) with $\delta(t) = e^{-\beta t}$ for some $\beta > 0$. As there is no time inconsistency under exponential discounting, the optimal value $\widetilde{J}^*(i) := \sup_{\pi\in\Pi} \widetilde{J}^\pi(i)$ fulfills the Bellman equation*

$$-\beta \widetilde{J}^*(i) + \sup_{\mu\in\mathcal{P}(U)} \int_U \left(g(i,u) + q_i^u \cdot \widetilde{J}^* \right) \mu(du) = 0.$$

Suppose that an optimal relaxed control $\pi^* \in \Pi$ exists, i.e., $\widetilde{J}^{\pi^*} = \widetilde{J}^*$. Then, arguments in Remark 4.4 can be adapted to the present continuous-time setting and show that $\widetilde{J}^\alpha = \widetilde{J}^{\pi^*}$ for any standard control α with $\alpha(i) \in \arg\max_{u\in U}\{g(i,u) + q_i^u \cdot \widetilde{J}^*\}$ for all $i \in \mathcal{S}$.

5 Convergence from Discrete Time to Continuous Time

In this section, we take up the same continuous-time setup as in Section 3. For a step size $h > 0$ small enough, we construct a discrete-time approximation as follows. Let $\{X_k^h\}_{k\in\mathbb{N}_0}$ be a discrete-time controlled Markov process as in Section 2, with its transition matrix given by

$$(p_h)_i^u := hq_i^u + e_i \quad \forall u \in U,\ i \in \mathcal{S}, \tag{5.1}$$

where $\{e_i\}_{i=1}^d$ is the standard basis of $\mathbb{R}^d$. Consider the reward function

$$f_h(k,i,u) := f(kh,i,u)h \quad \forall k \in \mathbb{N}_0, \tag{5.2}$$

where $f(t,i,u)$ is the reward function from Section 3. For any $\pi \in \Pi$, J^π in (2.1) now takes the form

$$J^{h,\pi}(i) := \mathbb{E}_i\left[\sum_{k=0}^\infty f_h^{\pi(X_k^h)}\left(k, X_k^h\right) \right] = h\mathbb{E}_i\left[\sum_{k=0}^\infty \int_U f(kh, X_k^h, u)(\pi(X_k^h))(du) \right], \quad \forall i \in \mathcal{S}. \tag{5.3}$$

Similarly, the auxiliary value function $V^{h,\pi}(i)$ is defined as in (2.19) with f and X^π therein replaced by f_h and X^h, respectively. In addition, for any $\lambda > 0$ and $\pi \in \Pi_r$, we recall J^π_λ in (2.4) and define

$$\begin{aligned} J^{h,\pi}_\lambda(i) := J^\pi_{h\lambda}(i) &= \mathbb{E}_i\bigg[\sum_{k=0}^\infty \Big(f_h^{\pi(X^h_k)}(k, X^h_k) + h\lambda\delta(k)\mathcal{H}\big(\pi(X^h_k)\big)\Big)\bigg] \\ &= h\mathbb{E}_i\bigg[\sum_{k=0}^\infty \Big(\int_U f(kh, X^h_k, u)(\pi(X^h_k))(du) + \lambda\delta(k)\mathcal{H}\big(\pi(X^h_k)\big)\Big)\bigg], \quad \forall i \in \mathcal{S}. \end{aligned} \tag{5.4}$$

Similarly, the auxiliary value function $V^{h,\pi}_\lambda(i)$ is defined as in (2.9) with f, X, and λ replaced by f_h, X^h, and $h\lambda$, respectively.

Let us now we study the convergence of value functions from discrete time to continuous time.

Lemma 5.1. *Let Assumptions 3.1, 2.2 (with p^u_i therein replaced by q^u_i), and 2.3 hold. Suppose that $f(\cdot, i, u)$ and $\delta(\cdot)$ are continuous on $[0,\infty)$. Take any sequence $\{h_n\}_{n\in\mathbb{N}}$ in $(0,1]$ with $h_n \downarrow 0$.*

(a) For any $\{\pi^n\}_{n\in\mathbb{N}}$ and π^∞ in Π such that $\pi^n(i) \to \pi^\infty(i)$ weakly for all $i \in \mathcal{S}$,

$$V^{h_n,\pi^n}(i) \to \widetilde{V}^{\pi^\infty}(0,i) = \widetilde{J}^{\pi^\infty}(i), \quad i \in \mathcal{S}.$$

(b) For any $\lambda > 0$ and $\{y^n\}_{n\in\mathbb{N}}$ with $y^n \to y^\infty$ for some $y^\infty \in \mathbb{R}^d$, we have $\Gamma_{h_n\lambda}(y^n, i) \to \Gamma_\lambda(y^\infty, i)$ for all $i \in \mathcal{S}$. Therefore,

$$V^{h_n,\Gamma_{h_n\lambda}(y^n,\cdot)}_\lambda(i) \to \widetilde{V}^{\widetilde{\Gamma}_\lambda(y^\infty,\cdot)}(0,i) = \widetilde{J}^{\widetilde{\Gamma}_\lambda(y^\infty,\cdot)}(i), \quad \forall i \in \mathcal{S}.$$

The proof of Lemma 5.1 is relegated to Appendix A.6.

We are ready to present our main convergence result from discrete time to continuous time: as the time step $h_n > 0$ tends to zero, a regular relaxed equilibrium (resp. a relaxed equilibrium) π^n in discrete time (with time step h_n) ultimately converges to a regular relaxed equilibrium (resp. a relaxed equilibrium) in continuous time.

Theorem 5.1. *Let Assumptions 3.1, 2.2 (with p^u_i therein replaced by q^u_i), and 2.3 hold. Suppose that $f(\cdot, i, u)$ and $\delta(\cdot)$ are continuous on $[0,\infty)$. Take any $\{h_n\}_{n\in\mathbb{N}}$ in $(0,1]$ with $h_n \downarrow 0$.*

(a) Given $\lambda > 0$, let $\pi^n \in \Pi_r$ be a regular relaxed equilibrium for $J^{h_n,\pi}_\lambda$ in (5.4) for all $n \in \mathbb{N}$. Then, for each $i \in \mathcal{S}$, $\pi^n(i) \in \mathcal{D}(U)$ converges pointwise to some $\pi^\infty(i) \in \mathcal{D}(U)$, up to a subsequence. Moreover, $\pi^\infty \in \Pi_r$ is a regular relaxed equilibrium for $\widetilde{J}^\pi_\lambda$ in (3.2).

(b) Let $\pi^n \in \Pi$ be a relaxed equilibrium for $J^{h_n,\pi}$ in (5.3) for all $n \in \mathbb{N}$. Then, for each $i \in \mathcal{S}$, $\pi^n(i) \in \mathcal{P}(U)$ converges weakly to some $\pi^\infty(i) \in \mathcal{P}(U)$, up to a subsequence. Moreover, $\pi^\infty \in \Pi$ is a relaxed equilibrium for $\widetilde{J}^\pi$ in (3.1).

The proof of Theorem 5.1 is relegated to Appendix A.7.

A Proofs

A.1 Proof of Lemma 2.1

Fix $i \in \mathcal{S}$ and $y \in \mathbb{R}^d$. For an arbitrary $\bar{u} \in U$, thanks to Assumption 2.2,

$$\big|(f(0,i,u) + p^u_i \cdot y) - (f(0,i,\bar{u}) + p^{\bar{u}}_i \cdot y)\big| \le \Theta(1+|y|)|u-\bar{u}|, \quad \forall u \in U. \tag{A.1}$$

This, along with (2.12), implies

$$\Gamma_\lambda(y,i)(\bar u)=\frac{\exp(\frac{1}{\lambda}[f(0,i,\bar u)+p_i^{\bar u}\cdot y])}{\int_U\exp(\frac{1}{\lambda}[f(0,i,u)+p_i^u\cdot y])du}\le\frac{1}{\int_U\exp\big(-\frac{\Theta}{\lambda}[(1+|y|)|u-\bar u|]\big)du},\quad\forall u\in U,\ \text{(A.2)}$$

where the inequality follows from dividing by $\exp(\frac{1}{\lambda}[f(0,i,\bar u)+p_i^{\bar u}\cdot y])$ the numerator and the denominator and then using the estimate (A.1).

Now let us prove (2.17) for $\ell>1$. By Assumption 2.3,

$$\int_U e^{-\frac{\Theta}{\lambda}(1+|y|)|u-\bar u|}du\ge\int_{\mathrm{cone}(\bar u,\iota)\cap B_\vartheta(\bar u)}e^{-\frac{\Theta}{\lambda}(1+|y|)|u-\bar u|}du=\int_{\Delta_\iota\cap B_\vartheta(0)}e^{-\frac{\Theta}{\lambda}(1+|y|)|u|}du,\qquad\text{(A.3)}$$

where the equality follows from a translation from $\bar u$ to $0\in\mathbb{R}^\ell$ and an appropriate rotation about $0\in\mathbb{R}^\ell$. Let us estimate the right-hand side of (A.3) in the next two cases. If $\frac{\Theta}{\lambda}(1+|y|)\vartheta\le1$,

$$\int_{\Delta_\iota\cap B_\vartheta(0)}e^{-\frac{\Theta}{\lambda}(1+|y|)|u|}du\ge e^{-1}\,\mathtt{Leb}(\Delta_\iota\cap B_\vartheta(0))=:K_0.$$

If $\frac{\Theta}{\lambda}(1+|y|)\vartheta>1$, consider the two positive constants

$$K_1:=\int_0^\pi\sin^{\ell-2}(\varphi_1)d\varphi_1\cdots\int_0^\pi\sin(\varphi_{\ell-2})d\varphi_{\ell-2}\int_{-\iota}^{\iota}d\varphi_{\ell-1}\quad\text{and}\quad K_2:=\int_0^1 z^{\ell-1}e^{-z}dz.$$

By using the ℓ-dimensional spherical coordinates, we have

$$\begin{aligned}\int_{\Delta_\iota\cap B_\vartheta(0)}e^{-\frac{\Theta}{\lambda}(1+|y|)|u|}du&=K_1\int_0^\vartheta r^{\ell-1}e^{-\frac{\Theta}{\lambda}(1+|y|)r}dr\\&=K_1\bigg(\frac{\lambda}{\Theta(1+|y|)}\bigg)^\ell\int_0^{\frac{\Theta}{\lambda}(1+|y|)\vartheta}z^{\ell-1}e^{-z}dz\ge K_1K_2\bigg(\frac{\lambda}{\Theta(1+|y|)}\bigg)^\ell,\end{aligned}$$

where the second line follows from the change of variable $z=\frac{\Theta}{\lambda}(1+|y|)r$. Combining the above two cases, we conclude from (A.2) and (A.3) that

$$\Gamma_\lambda(y,i)(u)\le\max\bigg\{\frac{1}{K_0},\frac{1}{K_1K_2}\bigg(\frac{\Theta}{\lambda}(1+|y|)\bigg)^\ell\bigg\}\le C(1+|y|)^\ell,\qquad\text{(A.4)}$$

with $C:=\max\{\frac{1}{K_0},\frac{1}{K_1K_2}\big(\frac{\Theta}{\lambda}\big)^\ell\}>0$, which depends on ι, ϑ, Θ, λ and ℓ (Recall that K_0, K_1, and K_2 depend on ι, ϑ, and ℓ). It follows that $\ln(\Gamma_\lambda(y,i)(u))\le\ln C+\ell\ln(1+|y|)$ for all $(y,i,u)\in\mathbb{R}^d\times\mathcal{S}\times U$. As $\Gamma_\lambda(y,i)\in\mathcal{D}(U)$ by definition, this implies

$$\sup_{i\in\mathcal{S}}\int_U\ln(\Gamma_\lambda(y,i)(u))\Gamma_\lambda(y,i)(u)du\le\ln C+\ell\ln(1+|y|),\quad\forall y\in\mathbb{R}^d.$$

On the other hand, (2.5) readily shows that $\sup_{i\in\mathcal{S}}\int_U\ln(\Gamma_\lambda(y,i)(u))\Gamma_\lambda(y,i)(u)du\ge-\ln(\mathtt{Leb}(U))$. Hence, in view of (2.3), we conclude that

$$\begin{aligned}\sup_{i\in\mathcal{S}}|\mathcal{H}(\Gamma_\lambda(y,i))|&=\sup_{i\in\mathcal{S}}\bigg|\int_U\ln(\Gamma_\lambda(y,i)(u))\Gamma_\lambda(y,i)(u)du\bigg|\\&\le|\ln(\mathtt{Leb}(U))|+|\ln C|+\ell\ln(1+|y|),\quad\forall y\in\mathbb{R}^d,\end{aligned}\qquad\text{(A.5)}$$

which shows (2.17) holds. For $\ell=1$, by following arguments simiar to the above, with $\Delta_\iota\cap B_\vartheta(0)$, K_0,K_1, and K_2 replaced by $[0,\vartheta]$, $e^{-1}\vartheta,1$, and $\int_0^1e^{-z}dz=1-e^{-1}$, respectively, we can obtain the desired estimate in (2.17) for $\ell=1$.

Now, for $\lambda>0$ small enough, the constant $C>0$ in (A.4) simply becomes $\frac{1}{K_1K_2}\big(\frac{\Theta}{\lambda}\big)^\ell$ (for both the cases $\ell>1$ and $\ell=1$). The estimate (A.5) then directly implies (2.18).

A.2 Proof of Theorem 2.1

First, let us establish the continuity of $\Psi_\lambda : \mathbb{R}^d \to \mathbb{R}^d$. Take an arbitrary sequence $\{y^n\}_{n\in\mathbb{N}}$ in $\mathbb{R}^d$ such that $y^n \to y^\infty$ for some $y^\infty \in \mathbb{R}^d$. As $f(0,i,u)$ and p_i^u are continuous in u (Assumption 2.2), U is compact, and $\mathcal{S}$ is a finite set, we conclude from the dominated convergence theorem that

$$\sup_{i\in\mathcal{S}} \left| \int_U e^{\frac{1}{\lambda}[f(0,i,u)+p_i^u\cdot y^n]} du - \int_U e^{\frac{1}{\lambda}[f(0,i,u)+p_i^u\cdot y^\infty]} du \right| \to 0, \quad \text{as } n \to \infty. \tag{A.6}$$

In view of (2.12), this particularly implies that for any $i \in \mathcal{S}$, $\Gamma_\lambda(y^n,i)(u) \to \Gamma_\lambda(y^\infty,i)(u)$ for all $u \in U$. Moreover, for any $i \in \mathcal{S}$, observe that

$$\begin{aligned}
&\left|\mathcal{H}\big(\Gamma_\lambda(y^n,i)\big) - \mathcal{H}\big(\Gamma_\lambda(y^\infty,i)\big)\right| \\
\le& \int_U \left|\ln\big(\Gamma_\lambda(y^n,i)(u)\big) - \ln\big(\Gamma_\lambda(y^\infty,i)(u)\big)\right| \Gamma_\lambda(y^n,i)(u) du \\
&\quad + \int_U \left|\ln\big(\Gamma_\lambda(y^\infty,i)(u)\big)\right| \left|\Gamma_\lambda(y^n,i)(u) - \Gamma_\lambda(y^\infty,i)(u)\right| du \\
\le& \frac{1}{\lambda}\left(\sup_{u\in U}|p_i^u|\right)|y^n - y^\infty| + \left|\ln\left(\int_U e^{\frac{1}{\lambda}[f(0,i,v)+p_i^v\cdot y^n]} dv\right) - \ln\left(\int_U e^{\frac{1}{\lambda}[f(0,i,v)+p_i^v\cdot y^\infty]} dv\right)\right| \\
&\quad + \sup_{u\in U}\left|\ln\big(\Gamma_\lambda(y^\infty,i)(u)\big)\right| \int_U \left|\Gamma_\lambda(y^n,i)(u) - \Gamma_\lambda(y^\infty,i)(u)\right| du,
\end{aligned} \tag{A.7}$$

where the first inequality follows from (2.3) and the second inequality is due to (2.12). As $u \mapsto f(0,i,u)$ and $u \mapsto p_i^u$ are continuous on the compact set U, we get the boundedness of $u \mapsto |p_i^u|$ and $u \mapsto \Gamma_\lambda(y^\infty,i)(u)$. Particularly, in view of (2.12), $u \mapsto \Gamma_\lambda(y^\infty,i)(u)$ is bounded away from zero. This in turn implies that $u \mapsto \ln\big(\Gamma_\lambda(y^\infty,i)(u)\big)$ is bounded. Hence, as $n \to \infty$, we can now conclude from $y^n \to y^\infty$ and (A.6) that the right-hand side of (A.7) vanishes. That is, $\mathcal{H}\big(\Gamma_\lambda(y^n,i)\big) \to \mathcal{H}\big(\Gamma_\lambda(y^\infty,i)\big)$.

For any $n \in \mathbb{N}$, consider the Markov chain $\bar{X}^n$ with transition matrix p^n given by $p_{ij}^n := \int_U p_{ij}^u \Gamma_\lambda(y^n,i)(u) du$ for all $i,j \in \mathcal{S}$, as well as the Markov chain $\bar{X}^\infty$ with transition matrix p^∞ given by $p_{ij}^\infty := \int_U p_{ij}^u \Gamma_\lambda(y^\infty,i)(u) du$ for all $i,j \in \mathcal{S}$. For each $i \in \mathcal{S}$, note that the pointwise convergence $\Gamma_\lambda(y^n,i) \to \Gamma_\lambda(y^\infty,i)$ in $\mathcal{D}(U)$ already implies the weak convergence of the corresponding probability measures. This, along with $u \mapsto p_i^u$ being continuous for all $i \in \mathcal{S}$, yields the convergence of the transition matrices, i.e., $p^n \to p^\infty$ component by component, which in turn implies that the law of $\bar{X}^n$ converges weakly to that of $\bar{X}^\infty$. In view of Remark 2.1, the law of $\bar{X}^n$ (resp. $\bar{X}^\infty$) coincides with that of $X^{\Gamma_\lambda(y^n,\cdot)}$ (resp. $X^{\Gamma_\lambda(y^\infty,\cdot)}$). That is, we actually have the law of $X^{\Gamma_\lambda(y^n,\cdot)}$ converging weakly to that of $X^{\Gamma_\lambda(y^\infty,\cdot)}$. Now, we can adapt an argument in the proof of Huang and Zhou [21, Theorem 3] to our present setting. Specifically, by Skorokhod's representation theorem, there exist $\mathcal{S}$-valued processes Y_n and Y, defined on some probability space $(\Omega,\mathcal{F},P)$, such that the law of Y_n coincides with that of $X^{\Gamma_\lambda(y^n,\cdot)}$, the law of Y coincides with that of $X^{\Gamma_\lambda(y^\infty,\cdot)}$, and $Y_k^n \to Y_k$ for all $k \in \mathbb{N}_0$ P-a.s. As $\mathcal{S}$ is a finite set, for each $k \in \mathbb{N}_0$, we in fact have $Y_k^n = Y_k$ for $n \in \mathbb{N}$ large enough. It follows that for any $i \in \mathcal{S}$,

$$\begin{aligned}
V_\lambda^{\Gamma_\lambda(y^n,\cdot)}(i) &= \mathbb{E}_i^P\left[\sum_{k=0}^\infty \left(\int_U f(1+k,Y_k^n,u)\Gamma_\lambda(y^n,Y_k^n)(u)du + \lambda\delta(1+k)\mathcal{H}\big(\Gamma_\lambda(y^n,Y_k^n)\big)\right)\right] \\
&\to \mathbb{E}_i^P\left[\sum_{k=0}^\infty \left(\int_U f(1+k,Y_k^\infty,u)\Gamma_\lambda(y^\infty,Y_k^\infty)(u)du + \lambda\delta(1+k)\mathcal{H}\big(\Gamma_\lambda(y^\infty,Y_k^\infty)\big)\right)\right] = V_\lambda^{\Gamma_\lambda(y^\infty,\cdot)}(i),
\end{aligned}$$

where the convergence follows from the dominated convergence theorem, $Y_k^n = Y_k^\infty$ for $n \in \mathbb{N}$ large enough, and $\Gamma_\lambda(y^n, i) \to \Gamma_\lambda(y^\infty, i)$ and $\mathcal{H}\big(\Gamma_\lambda(y^n, i)\big) \to \mathcal{H}\big(\Gamma_\lambda(y^\infty, i)\big)$ for all $i \in \mathcal{S}$. Let us stress that the dominated convergence theorem is applicable here thanks to Assumptions 2.1 and Lemma 2.1. We therefore conclude that $\Psi_\lambda : \mathbb{R}^d \to \mathbb{R}^d$ is continuous.

Now, we are ready to show that a fixed point of Ψ_λ exists. For any $y \in \mathbb{R}^d$,

$$\begin{aligned} |\Psi_\lambda(y)| = \Big|V_\lambda^{\Gamma_\lambda(y,\cdot)}(i)\Big| &\le \mathbb{E}_i\bigg[\sum_{t=0}^{\infty}\Big(\Big|f^{\Gamma_\lambda(y,X_t))}(t,i)\Big| + \lambda\delta(t)|\mathcal{H}(\Gamma_\lambda(y,X_t))|\Big)\bigg] \\ &\le M + \lambda\sum_{t=0}^{\infty}\delta(t)\phi(|y|) \le \big(1+\lambda\phi(|y|)\big)M, \end{aligned} \tag{A.8}$$

where the first inequality stems from (2.9), $M > 0$ is the constant in Assumption 2.1, and the second inequality follows from (2.17) in Lemma 2.1. Note from (2.17) that $\phi : \mathbb{R}_+ \to \mathbb{R}+$ grows sublinearly (i.e., $\phi(\alpha)/\alpha \to 0$ as $\alpha \to \infty$). Hence, $\alpha \mapsto (1+\lambda\phi(\alpha))M$ also grows sublinearly, such that

$$\alpha^* := \sup\{\alpha \ge 0 : \alpha \le (1+\lambda\phi(\alpha))M\} < \infty.$$

If $|y| > \alpha^*$, by (A.8) and the definition of α^*, $|\Psi_\lambda(y)| \le (1+\lambda\phi(|y|))M < |y|$. If $|y| \le \alpha^*$, by (A.8), ϕ being increasing, and the definition of α^*, we obtain $|\Psi_\lambda(y)| \le M(1+\lambda\phi(|y|) \le M(1+\lambda\phi(\alpha^*)) \le \alpha^*$. We then conclude $|\Psi_\lambda(y)| \le \max\{|y|, \alpha^*\}$ for all $y \in \mathbb{R}^d$. Hence, for any $r \ge \alpha^*$, $\Psi_\lambda(\overline{B_r(0)}) \subseteq \overline{B_r(0)}$. As $\Psi_\lambda : \mathbb{R}^d \to \mathbb{R}^d$ is continuous, this implies that Ψ_λ has a fixed point $y \in \overline{B_r(0)}$, thanks to Brouwer's fixed-point theorem. By Corollary 2.1, $\Gamma_\lambda(y,\cdot) \in \Pi_r$ is a regular relaxed equilibrium for (2.4).

A.3 Proof of Lemma 2.2

Note that the existence of the sequence $\{y^n\}_{n\in\mathbb{N}}$ is guaranteed by Theorem 2.1. By using (2.18) in Lemma 2.1 in the calculation (A.8), we get $|\Psi_\lambda(y)| \le \big(1+\lambda\varphi(\lambda,y)\big)M$ for all $y \in \mathbb{R}^d$, where φ is specified in (2.18) and $M > 0$ is the constant in Assumption 2.1. In view of (2.18),

$$\lambda\varphi(\lambda,y) = \kappa_1\lambda + \kappa_2|\lambda\ln\lambda| + \ell\lambda\ln(1+|y|),$$

where $\kappa_1, \kappa_2 > 0$ are constants independent of λ. Since $|\lambda\ln\lambda| \to 0$ as $\lambda \downarrow 0$, the above equation implies that for all $\lambda \in (0,1]$, $\lambda\varphi(\lambda,y) \le \eta(|y|)$, with $\eta(z) := K(1+\ln(1+z))$ for some $K > 0$ independent of $\lambda \in (0,1]$. We therefore obtain

$$|\Psi_\lambda(y)| \le \big(1+\eta(|y|)\big)M, \quad \forall y \in \mathbb{R}^d \text{ and } \lambda \in (0,1]. \tag{A.9}$$

As $\eta : \mathbb{R}_+ \to \mathbb{R}+$ grows sublinearly (i.e., $\eta(\alpha)/\alpha \to 0$ as $\alpha \to \infty$), $\alpha \mapsto (1+\eta(\alpha))M$ also grows sublinearly, such that $\alpha^* := \sup\{\alpha \ge 0 : \alpha \le (1+\eta(\alpha))M\} < \infty$. By using (A.9) and the definition of α^*, we may repeat the argument in the last paragraph of the proof of Theorem 2.1 and obtain

$$|\Psi_\lambda(y)| \begin{cases} < |y|, & \text{if } |y| > \alpha^*, \\ \le \alpha^*, & \text{if } |y| \le \alpha^*, \end{cases} \quad \forall \lambda \in (0,1].$$

Now, for each $n \in \mathbb{N}$, as $y^n = \Psi_{\lambda_n}(y^n)$, the above inequality entails $|y^n| \le \alpha^*$. This readily implies $\sup_{n\in\mathbb{N}}|y^n| \le \alpha^* < \infty$.

A.4 Proof of Theorem 2.2

Take any sequence $\{\lambda_n\}_{n\in\mathbb{N}}$ in $(0,1]$ with $\lambda_n \downarrow 0$. By Theorem 2.1, for each $n \in \mathbb{N}$, there exists $y^n \in \mathbb{R}^d$ such that $\Psi_{\lambda_n}(y^n) = y^n$ and $\pi^n := \Gamma_{\lambda_n}(y^n, \cdot) \in \Pi_r$ is a regular relaxed equilibrium for (2.4) (with $\lambda = \lambda_n$ therein). As the sequence $\{y^n\}_{n\in\mathbb{N}}$ is bounded in $\mathbb{R}^d$ (Lemma 2.2), it has a subsequence (without relabeling) that converges to some $y^\infty \in \mathbb{R}^d$. On the other hand, as U is compact, $\mathcal{P}(U)$ is compact under the topology of weak convergence of probability measures. Hence, for each $i \in \mathcal{S}$, $\{\pi^n(i)\}_{n\in\mathbb{N}}$ in $\mathcal{P}(U)$ has a subsequence (without relabeling) that converges weakly to some $\pi^*(i) \in \mathcal{P}(U)$. This gives rise to $\pi^* \in \Pi$ (which is not necessarily regular).

Now, we claim that $\pi^* \in \Gamma(y^\infty)$. In view of (2.21), we need to show that $\pi^*(i) \in \mathcal{P}(U)$ is supported by the closed set $E(y^\infty, i) \subseteq U$ in (2.20) for all $i \in \mathcal{S}$. As this holds trivially when $E(y^\infty, i) = U$, we assume $E(y^\infty, i) \subsetneq U$ in the following. Our goal is to prove that for any $i \in \mathcal{S}$, $(\pi^*(i))\big(\overline{B_r(u_0)}\big) = 0$ for all $u_0 \in U$ and $r > 0$ such that

$$\overline{B_r(u_0)} \cap E(y^\infty, i) = \emptyset \quad \text{and} \quad B_r(u_0) \subseteq U.$$

Note that this readily implies $\mathrm{supp}(\pi^*(i)) \subseteq E(y^\infty, i)$ for all $i \in \mathcal{S}$, as desired. To this end, take a continuous and bounded function $h : U \to [0,1]$ such that

$$h(u) \equiv 1 \quad \text{for } u \in B_r(u_0) \quad \text{and} \quad h(u) \equiv 0 \quad \text{for } u \notin \overline{B_{r+d/2}(u_0)}, \tag{A.10}$$

where $d := \mathrm{dist}\big(\overline{B_r(u_0)}, E(v^\infty, i)\big)$ is strictly positive, as $\overline{B_r(u_0)}$ and $E(v^\infty, i)$ are disjoint closed sets. Consider $A := \max_{u\in U}\{f(0,i,u) + p_i^u \cdot y^\infty\} < \infty$. Note that $d > 0$ implies

$$\varepsilon := A - \sup\Big\{f(0,i,u) + p_i^u \cdot y^\infty : u \in \overline{B_r(u_0)}\Big\} > 0.$$

Also, by the continuity of $u \mapsto f(0,i,u) + p_i^u \cdot y^\infty$,

$$L := \mathtt{Leb}(\{u \in U : f(0,i,u) + p_i^u \cdot y^\infty > A - \varepsilon/2\}) > 0.$$

As $y^n \to y^\infty$, for all $n \in \mathbb{N}$ large enough, we have

$$A - \sup\Big\{f(0,i,u) + p_i^u \cdot y^n : u \in \overline{B_r(u_0)}\Big\} > \frac{\varepsilon}{2}, \quad \mathtt{Leb}\left(\left\{u \in U : f(0,i,u) + p_i^u \cdot y^n > A - \frac{\varepsilon}{2}\right\}\right) \geq \frac{L}{2}.$$

This, along with the definition of $\Gamma_{\lambda_n}(y^n, i)$ in (2.12), yields

$$\begin{aligned}
\limsup_{n\to\infty} \int_U h(u)\Gamma_{\lambda_n}(y^n, i)(u)du &\leq \limsup_{n\to\infty} \int_{\overline{B_{r+d/2}(u_0)}} \frac{e^{\frac{1}{\lambda_n}(f(0,i,u)+p_i^u\cdot y^n)}}{\int_U e^{\frac{1}{\lambda_n}(f(0,i,u)+p_i^u\cdot y^n)}du}du \\
&= \limsup_{n\to\infty} \int_{\overline{B_{r+d/2}(u_0)}} \frac{e^{\frac{1}{\lambda_n}(f(0,i,u)+p_i^u\cdot y^n-(A-\varepsilon/2))}}{\int_U e^{\frac{1}{\lambda_n}(f(0,i,u)+p_i^u\cdot y^n-(A-\varepsilon/2))}du}du \\
&\leq \limsup_{n\to\infty} \int_{\overline{B_{r+d/2}(u_0)}} \frac{e^{-\frac{\varepsilon/2}{\lambda_n}}}{\int_{\{u\in U: f(0,i,u)+p_i^u\cdot y^n \geq A-\frac{1}{2}\varepsilon\}} 1du}du \\
&\leq \limsup_{n\to\infty} \frac{2}{L} e^{-\frac{\varepsilon/2}{\lambda_n}}\, \mathtt{Leb}\left(\overline{B_{r+d/2}(u_0)}\right) = 0.
\end{aligned} \tag{A.11}$$

Now, as h is continuous, bounded, and satisfies (A.10),

$$\begin{aligned}
(\pi^*(i))\left(\overline{B_r(u_0)}\right) \leq \int_U h(u)(\pi^*(i))(du) &= \lim_{n\to\infty} \int_U h(u)(\pi^n(i))(du) \\
&= \lim_{n\to\infty} \int_U h(u)\Gamma_{\lambda_n}(y^n, i)(u)du = 0,
\end{aligned}$$

where the first equality follows from $\pi^n(i) \to \pi^*(i)$ weakly in $\mathcal{P}(U)$ and the last equality is due to (A.11). The claim "$\pi^* \in \Gamma(y^\infty)$" is then established.

Now, we set out to prove $V_{\lambda_n}^{\Gamma_{\lambda_n}(y^n,\cdot)} \to V^{\pi^*}$. Consider the Markov chain $\bar{X}^n$ with transition matrix p^n given by $p^n_{ij} := \int_U p^u_{ij} \Gamma_{\lambda_n}(y^n, i)(u)du$ for all $i, j \in \mathcal{S}$, as well as the Markov chain $\bar{X}^*$ with transition matrix p^* given by $p^*_{ij} := \int_U p^u_{ij}(\pi^*(i))(du)$ for all $i, j \in \mathcal{S}$. Similarly to the discussion in the last paragraph of the proof of Theorem 2.1, $p^n \to p^*$ component by component (thanks to $\Gamma_{\lambda_n}(y^n, i) \to \pi^*(i)$ weakly for all $i \in \mathcal{S}$), whence the law of $\bar{X}^n$ converges weakly to that of $\bar{X}^*$, which in turn implies that the law of $X^{\Gamma_{\lambda_n}(y^n,\cdot)}$ converges weakly to that of X^{π^*} (by Remark 2.1). By Skorokhod's representation theorem, there exist $\mathcal{S}$-valued processes Y_n and Y, defined on some probability space $(\Omega, \mathcal{F}, P)$, such that the law of Y_n coincides with that of $X^{\Gamma_{\lambda_n}(y^n,\cdot)}$, the law of Y coincides with that of X^{π^*}, and $Y^n_k \to Y_k$ for all $k \in \mathbb{N}_0$ P-a.s. As $\mathcal{S}$ is a finite set, for each $k \in \mathbb{N}_0$, we in fact have $Y^n_k = Y_k$ for $n \in \mathbb{N}$ large enough. It follows that for any $i \in \mathcal{S}$,

$$\begin{aligned} &V_{\lambda_n}^{\Gamma_{\lambda_n}(y^n,\cdot)}(i) - \mathbb{E}^P_i\left[\sum_{k=0}^\infty \delta(1+k)\lambda_n \mathcal{H}\left(\Gamma_\lambda(y^n, Y^n_k)\right)\right] \\ &\quad = \mathbb{E}^P_i\left[\sum_{k=0}^\infty \left(\int_U f(1+k, Y^n_k, u)\Gamma_\lambda(y^n, Y^n_k)(u)du\right)\right] \\ &\quad \to \mathbb{E}^P_i\left[\sum_{k=0}^\infty \left(\int_U f(1+k, Y_k, u)\pi^*(Y_k)(u)du\right)\right] = V^{\pi^*}(i), \end{aligned} \tag{A.12}$$

where the convergence follows from $Y^n_k = Y_k$ for $n \in \mathbb{N}$ large enough, $\Gamma_\lambda(y^n, i) \to \pi^*(i)$ weakly for all $i \in \mathcal{S}$, and $u \mapsto f(t, i, u)$ being continuous on the compact set U. Moreover, by (2.18) in Lemma 2.1, the boundedness of $\{y^n\}_{n\in\mathbb{N}}$, and $\sum_{k=0}^\infty \delta(1+k) < \infty$ (Assumption 2.1), there exist constants $C_1, C_2 > 0$ independent of $\{\lambda_n\}_{n\in\mathbb{N}}$ such that

$$\sum_{k=0}^\infty \delta(1+k)\lambda_n \sup_{i\in\mathcal{S}} |\mathcal{H}(\Gamma_{\lambda_n}(y^n, i))| = \left(C_1\lambda_n + C_2\lambda_n|\ln\lambda_n|\right)\sum_{k=0}^\infty \delta(1+k) \to 0 \quad \text{as } n \to \infty,$$

which implies that $\mathbb{E}^P_i\left[\sum_{k=0}^\infty \delta(1+k)\lambda_n \mathcal{H}\left(\Gamma_\lambda(y^n, Y^n_k)\right)\right] \to 0$. We then conclude from (A.12) that $V_{\lambda_n}^{\Gamma_{\lambda_n}(y^n,\cdot)}(i) \to V^{\pi^*}(i)$ for all $i \in \mathcal{S}$.

Finally, recall that $\Psi_{\lambda_n}(y^n) = y^n$ means $y^n = V_{\lambda_n}^{\Gamma_{\lambda_n}(y^n,\cdot)}$. As a result,

$$y^\infty = \lim_{n\to\infty} y^n = \lim_{n\to\infty} V_{\lambda_n}^{\Gamma_{\lambda_n}(y^n,\cdot)} = V^{\pi^*} \in \Psi(y^\infty), \tag{A.13}$$

where the inclusion follows from $\pi^* \in \Gamma(y^\infty)$. In view of (2.22) and (2.23), the above relation implies $\pi^* \in \Phi(\pi^*)$. Hence, by Proposition 2.2, π^* is a relaxed equilibrium for (2.1).

A.5 Proof of Proposition 3.1

Fix $\pi \in \Pi_r$. By taking $\pi' = \pi$ in (3.6) and noting $\widetilde{J}^{\pi\otimes_\varepsilon\pi}_\lambda(i) = \widetilde{J}^\pi_\lambda(i) = \widetilde{V}^\pi_\lambda(0, i)$, we get

$$\widetilde{V}^\pi_\lambda(\varepsilon, i) - \widetilde{V}^\pi_\lambda(0, i) = -\left(f^{\pi(i)}(0, i) + \lambda\mathcal{H}(\pi(i)) + Q^{\pi(i)}_i \cdot \widetilde{V}^\pi_\lambda(\varepsilon)\right)\varepsilon, \quad \forall i \in \mathcal{S}. \tag{A.14}$$

This implies that $t \mapsto \widetilde{V}^\pi_\lambda(t, i)$ is continuous. Moreover, when we divide both sides by $\varepsilon > 0$ and take $\varepsilon \downarrow 0$, since $t \mapsto \widetilde{V}^\pi_\lambda(t, i)$ is continuous for all $i \in \mathcal{S}$, we get

$$\partial_t \widetilde{V}^\pi_\lambda(0, i) + f^{\pi(i)}(0, i) + \lambda\mathcal{H}(\pi(i)) + Q^{\pi(i)}_i \cdot \widetilde{V}^\pi_\lambda(0) = 0, \quad \forall i \in \mathcal{S}. \tag{A.15}$$

Now, for any $\pi' \in \Pi_r$, thanks to (3.6) and (A.14),

$$\widetilde{V}_\lambda^{\pi' \otimes_\varepsilon \pi}(0,i) - \widetilde{V}_\lambda^{\pi}(0,i) = \Big([f^{\pi'(i)}(0,i) + \lambda \mathcal{H}(\pi'(i)) + Q_i^{\pi'(i)} \cdot \widetilde{V}_\lambda^{\pi}(\varepsilon)] \\ - [f^{\pi(i)}(0,i) + \lambda \mathcal{H}(\pi(i)) + Q_i^{\pi(i)} \cdot \widetilde{V}_\lambda^{\pi}(\varepsilon)] \Big) \varepsilon + o(\varepsilon).$$

It follows that

$$\begin{aligned} &\lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left(\widetilde{V}_\lambda^{\pi' \otimes_\varepsilon \pi}(0,i) - \widetilde{V}_\lambda^{\pi}(0,i) \right) \\ &= \left(f^{\pi'(i)}(0,i) + \lambda \mathcal{H}(\pi'(i)) + Q_i^{\pi'(i)} \cdot \widetilde{V}_\lambda^{\pi}(\varepsilon) \right) - \left(f^{\pi(i)}(0,i) + \lambda \mathcal{H}(\pi(i)) + Q_i^{\pi(i)} \cdot \widetilde{V}_\lambda^{\pi}(\varepsilon) \right) \\ &= f^{\pi'(i)}(0,i) + \lambda \mathcal{H}(\pi'(i)) + Q_i^{\pi'(i)} \cdot \widetilde{V}_\lambda^{\pi}(0) + \partial_t \widetilde{V}_\lambda^{\pi}(0,i), \quad \forall i \in \mathcal{S}, \end{aligned} \tag{A.16}$$

where the last line follows from the continuity of $t \mapsto \widetilde{V}_\lambda^{\pi}(t,i)$ for all $i \in \mathcal{S}$ and (A.15). Hence, π is a regular relaxed equilibrium for (3.2) if and only if

$$\partial_t \widetilde{V}_\lambda^{\pi}(0,i) + \sup_{\rho \in \mathcal{D}(U)} \left(f^{\rho}(0,i) + \lambda \mathcal{H}(\rho) + Q_i^{\rho} \cdot \widetilde{V}_\lambda^{\pi}(0) \right) \leq 0, \quad \forall i \in \mathcal{S}.$$

In view of (A.15), this holds if and only if

$$\begin{aligned} \pi(i) &\in \arg\max_{\rho \in \mathcal{D}(U)} \left(f^{\rho}(0,i) + \lambda \mathcal{H}(\rho) + Q_i^{\rho} \cdot \widetilde{V}_\lambda^{\pi}(0) \right) \\ &= \arg\max_{\rho \in \mathcal{D}(U)} \int_U \left(f(0,i,u) - \lambda \ln \rho(u) + q_i^u \cdot \widetilde{V}_\lambda^{\pi}(0) \right) \rho(u) du, \quad \forall i \in \mathcal{S}. \end{aligned} \tag{A.17}$$

As the set on the right-hand side above is a singleton that contains the density

$$\rho^*(u) = \frac{e^{\frac{1}{\lambda}\left(f(0,i,u) + q_i^u \cdot \widetilde{V}_\lambda^{\pi}(0)\right)}}{\int_U e^{\frac{1}{\lambda}\left(f(0,i,v) + q_i^v \cdot \widetilde{V}_\lambda^{\pi}(0)\right)} dv} = \widetilde{\Gamma}_\lambda(\widetilde{V}_\lambda^{\pi}(0), i)(u) = \widetilde{\Gamma}_\lambda(\widetilde{J}_\lambda^{\pi}, i)(u), \quad u \in U,$$

the relation (A.17) amounts to $\pi(i) = \widetilde{\Gamma}_\lambda(\widetilde{J}_\lambda^{\pi}, i)$ for all $i \in \mathcal{S}$, which is equivalent to $\pi = \widetilde{\Phi}_\lambda(\pi)$.

A.6 Proof of Lemma 5.1

(a) For notational convenience, we will write X^{h_n} for the discrete-time Markov chain whose transition matrix is $P^n = \{P^n(i)\}_{i \in \mathcal{S}}$ with its i^{th}-row given by

$$P^n(i) := \int_U (p_{h_n})_i^u (\pi^n(i))(du) = \int_U \left(h_n q_i^u + e_i \right) (\pi^n(i))(du) = Q_i^{\pi^n(i)} h_n + e_i,$$

where the second equality follows from (5.1). We will also write X^n and X^∞ for the continuous-time Markov chains with generators $\{Q_i^{\pi^n(i)}\}_{i \in \mathcal{S}}$ and $\{Q_i^{\pi^\infty(i)}\}_{i \in \mathcal{S}}$, respectively. For each $n \in \mathbb{N}$, note that the transition matrix of the discrete-time Markov chain $\left\{X^n_{kh_n}\right\}_{k \in \mathbb{N}}$ is $\widetilde{P}^n = \{\widetilde{P}^n(i)\}_{i \in \mathcal{S}}$ with its i^{th}-row given by

$$\widetilde{P^n}(i) := e_i + Q_i^{\pi^n(i)} h_n + o(h_n) = P^n(i) + o(h_n). \tag{A.18}$$

In the following, we will adapt the arguments in the proof of Huang and Zhou [21, Lemma 3] to the present setting. For any $i \in \mathcal{S}$, observe that

$$\begin{aligned}
&\left|V^{h_n,\pi^n}(i) - \widetilde{V}^{\pi^\infty}(0,i)\right| \\
&\leq h_n \Bigg| \mathbb{E}_i \left[\sum_{k=0}^{\infty} \int_U f\left((k+1)h_n, X_k^{h_n}, u\right) \left(\pi^n(X_k^{h_n})\right)(du) \right] \\
&\qquad - \mathbb{E}_i \left[\sum_{k=0}^{\infty} \int_U f\left((k+1)h_n, X_{kh_n}^n, u\right) (\pi^n(X_{kh_n}^n))(du) \right] \Bigg| \\
&\quad + \left| h_n \mathbb{E}_i \left[\sum_{k=0}^{\infty} \int_U f\left((k+1)h_n, X_{kh_n}^n, u\right) (\pi^n(X_{kh_n}^n))(du) \right] \right. \\
&\qquad \left. - \mathbb{E}_i \left[\int_0^{\infty} \int_U f(t, X_t^n, u)(\pi^n(X_t^n))(du)dt \right] \right| \\
&\quad + \left| \mathbb{E}_i \left[\int_0^{\infty} \int_U f(t, X_t^n, u)(\pi^n(X_t^n))(du)dt \right] - \mathbb{E}_i \left[\int_0^{\infty} \int_U f(t, X_t^{\infty}, u)(\pi^{\infty}(X_t^{\infty}))(du)dt \right] \right|.
\end{aligned} \tag{A.19}$$

Let I_1^n, I_2^n, and I_3^n denote the second, third, and fourth lines, respectively, in the above inequality.

Let us first deal with I_1^n. By Assumption 3.1, for any $\varepsilon > 0$, we can take $T > 0$ such that

$$\int_T^{\infty} \sup_{i,u} |f(t,i,u)| dt < \varepsilon. \tag{A.20}$$

It follows that

$$\begin{aligned}
I_1^n \leq h_n \Bigg| \sum_{k=0}^{T/h_n} \mathbb{E}_i \Bigg[&\int_U f\left((k+1)h_n, X_k^{h_n}, u\right) \left(\pi^n(X_k^{h_n})\right)(du) \\
&- \int_U f\left((k+1)h_n, X_{kh_n}^n, u\right) \left(\pi^n(X_{kh_n}^n)\right)(du) \Bigg] \Bigg| + 2\varepsilon
\end{aligned} \tag{A.21}$$

In addition, (A.18) implies $(\widetilde{P^n})^k(i) = (P^n)^k(i) + ko(h_n)(1+o(h_n))^k$, where $(\widetilde{P^n})^k(i)$ (resp. $(P^n)^k(i)$) denotes the i^{th}-column of the matrix $(\widetilde{P^n})^k$ (resp. $(P^n)^k$). By writing $\int_U f(kh_n, \cdot, u)(\pi^n(\cdot))(du)$ for the vector $\left(\int_U f(kh_n, 1, u)(\pi^n(1))(du), ..., \int_U f(kh_n, d, u)(\pi^n(d))(du)\right) \in \mathbb{R}^d$, we observe that

$$\begin{aligned}
\mathbb{E}_i \left[\int_U f\left((k+1)h_n, X_k^{h_n}, u\right) \pi^n\left(X_k^{h_n}\right)(du) \right] &= (P^n)^k(i) \cdot \left(\int_U f\left((k+1)h_n, \cdot, u\right) (\pi^n(\cdot))(du) \right) \\
\mathbb{E}_i \left[\int_U f\left((k+1)h_n, X_{kh_n}^n, u\right) \pi^n\left(X_{kh_n}^n\right)(du) \right] &= (\widetilde{P^n})^k(i) \cdot \left(\int_U f((k+1)h_n, \cdot, u)(\pi^n(\cdot))(du) \right).
\end{aligned} \tag{A.22}$$

It then follows from that

$$\begin{aligned}
I_1^n \leq h_n \widetilde{M} \sum_{k=0}^{T/h_n} ko(h_n)(1+o(h_n))^k + 2\varepsilon &\leq h_n \widetilde{M} \sum_{k=0}^{T/h_n} \frac{T}{h_n} o(h_n)(1+h_n)^{T/h_n} + 2\varepsilon \\
&= \sum_{k=0}^{T/h_n} o(h_n) = \left(\frac{T}{h_n} + 1\right) o(h_n) + 2\varepsilon = o(1) + 2\varepsilon.
\end{aligned}$$

We then obtain $\lim_{n\to\infty} I_1^n \le 2\varepsilon$. As $\varepsilon > 0$ is arbitrary, we conclude $\lim_{n\to\infty} I_1^n = 0$.

We now deal with I_2^n. For any $\varepsilon > 0$, consider $T > 0$ as in (A.20). Then, observe that $I_2^n \le \sum_{k=0}^{T/h_n} \mathbb{E}_i[\eta_k] + 2\varepsilon$, with

$$\eta_k := \left| h_n \int_U f\left((k+1)h_n, X^n_{kh_n}, u\right)(\pi^n(X^n_{kh_n}))(du) - \int_{kh_n}^{(k+1)h_n} \int_U f(t, X^n_t, u)(\pi^n(X^n_t))(du)dt \right|.$$

Set $A_k := \{\text{there is no jump for } X^n \text{ in the time interval } (kh_n, (k+1)h_n]\}$. As $f(\cdot, i, \cdot)$ is continuous, it is uniformly continuous on the compact set $[0,T] \times U$. Hence, there exists a modulus of continuity L, independent of i and u, such that $|f(t,i,u) - f(s,i,u)| \le L(|t-s|)$ for all $t, s \in [0,T]$. It follows that

$$\mathbb{E}_i[\eta_k] \le \mathbb{E}_i[\eta_k \mid A_k]\mathbb{P}(A_k) + \mathbb{E}_i[\eta_k \mid A_k^c]\mathbb{P}(A_k^c) \le L(h_n)h_n(1 - o(1)) + O(h_n)o(1) = o(h_n).$$

Hence, $I_2^n \le \sum_{k=0}^{T/h_n} o(h_n) + 2\varepsilon = o(1) + 2\varepsilon$, which implies $\lim_{n\to\infty} I_2^n \le 2\varepsilon$. As $\varepsilon > 0$ is arbitrary, we conclude $\lim_{n\to\infty} I_2^n = 0$.

Finally, we deal with I_3^n. For any $i \in \mathcal{S}$, as $u \mapsto q_u^i$ is continuous and U is compact, the fact "$\pi^n(i) \to \pi^\infty(i)$ weakly" readily implies $Q_i^{\pi_n(i)} \to Q_i^{\pi^\infty(i)}$. That is, the rate matrix of X^n converges to that of X^∞ (component by component). Then, we may follow the argument in the proof of Huang and Zhou [21, Theorem 3] (particularly, from the third last line of p. 448 to the fourth line on p. 449) to obtain $\lim_{n\to\infty} I_3^n = 0$. As I_1^n, I_2^n, and I_3^n all converge to zero, we conclude from (A.19) that $V^{h_n,\pi^n}(i) \to \widetilde{V}^{\pi^\infty}(0,i)$.

(b) For any $i \in \mathcal{S}$ and $u \in U$, thanks to (2.12), (5.2), and (5.1),

$$\begin{aligned}\Gamma_{h_n\lambda}(y^n, i)(u) &= \frac{\exp[\frac{1}{h_n\lambda}(f_{h_n}(0,i,u) + (p_h)_i^u \cdot y^n)]}{\int_U \exp[\frac{1}{h_n\lambda}(f_{h_n}(0,i,u) + (p_h)_i^u \cdot y^n)]du} = \frac{\exp[\frac{1}{\lambda}(f(0,i,u) + q_i^u \cdot y^n)]}{\int_U \exp[\frac{1}{\lambda}(f(0,i,u) + q_i^u \cdot y^n)]du} \\ &\to \frac{\exp\left[\frac{1}{\lambda}(f(0,i,u) + q_i^u \cdot y^\infty)\right]}{\int_U \exp[\frac{1}{\lambda}(f(0,i,u) + q_i^u \cdot y^\infty)]du} = \widetilde{\Gamma}_\lambda(y^\infty, i)(u), \quad \text{as } n \to \infty, \end{aligned} \tag{A.23}$$

where the convergence follows from $y^n \to y^\infty$. In view of (3.8), the above implies $\Gamma_{h_n\lambda}(y^n,i)(u) = \widetilde{\Gamma}_\lambda(y^n,i)(u) \to \widetilde{\Gamma}_\lambda(y^\infty,i)(u)$. Hence, $\mathcal{H}(\Gamma_{h_n\lambda}(y^n,i)) = \mathcal{H}(\widetilde{\Gamma}_\lambda(y^n,i)) \to \mathcal{H}(\widetilde{\Gamma}_\lambda(y^\infty,i))$, where the convergence follows from an argument similar to (A.7). By taking $\pi^n := \Gamma_{h_n\lambda}(y^n,\cdot)$, the desired result follows from the arguments in part (a) and $\mathcal{H}(\Gamma_{h_n\lambda}(y^n,i)) \to \mathcal{H}(\widetilde{\Gamma}_\lambda(y^\infty,i))$ for all $i \in \mathcal{S}$.

A.7 Proof of Theorem 5.1

(a) For any $n \in \mathbb{N}$, set $y^n := V_\lambda^{h_n,\pi^n} \in \mathbb{R}^d$. As $\pi^n \in \Pi_r$ is a regular relaxed equilibrium for (5.4), Proposition 2.1 implies $\pi^n = \Phi_{h_n\lambda}(\pi^n)$, i.e., $\pi^n(i) = \Gamma_{h_n\lambda}(y^n,i)$. In addition, Corollary 2.1 implies $\Psi_{h_n\lambda}(y^n) = y^n$. Hence, by Lemma 2.2, $\{y^n\}_{n\in\mathbb{N}}$ is bounded in $\mathbb{R}^d$. For any subsequence of $\{y^n\}_{n\in\mathbb{N}}$ (without relabeling) that converges to some $y^\infty \in \mathbb{R}^d$, Lemma 5.1 (b) asserts $\Gamma_{h_n\lambda}(y^n,i) \to \Gamma_\lambda(y^\infty,i)$. Thus, $\pi^\infty(i) := \lim_{n\to\infty} \pi^n(i) = \widetilde{\Gamma}_\lambda(y^\infty,i)$ is well-defined for all $i \in \mathcal{S}$. Now, note that

$$y^\infty = \lim_{n\to\infty} y^n = \lim_{n\to\infty} V_\lambda^{h_n,\pi^n} = \lim_{n\to\infty} V_\lambda^{h_n,\Gamma_{h_n\lambda}(y^n,\cdot)} = \widetilde{J}_\lambda^{\widetilde{\Gamma}_\lambda(y^\infty,\cdot)},$$

where the last equality follows from Lemma 5.1 (b). That is, we have $y^\infty = \widetilde{\Psi}_\lambda(y^\infty)$. By Corollary 3.1, this implies $\pi^\infty = \widetilde{\Gamma}_\lambda(y^\infty,\cdot)$ is a regular relaxed equilibrium for (3.2).

(b) As U is compact, $\mathcal{P}(U)$ is compact under the topology of weak convergence of probability measures. Hence, for each $i \in \mathcal{S}$, $\{\pi^n(i)\}_{n\in\mathbb{N}}$ in $\mathcal{P}(U)$ has a subsequence (without relabeling) that converges weakly to some $\pi^*(i) \in \mathcal{P}(U)$. By Proposition 2.2, for each $n \in \mathbb{N}$,

$$\begin{aligned}\operatorname{supp}(\pi^n(i)) &\subseteq \arg\max_{u\in U}\left\{f_h(0,i,u) + (p_h)_i^u \cdot V^{h_n,\pi^n}\right\}\\ &= \arg\max_{u\in U}\left\{f(0,i,u) + q_i^u \cdot V^{h_n,\pi^n}\right\} = \widetilde{E}\left(V^{h_n,\pi^n}, i\right), \quad \forall i \in \mathcal{S}, \end{aligned} \tag{A.24}$$

where the first equality follows from (5.2) and (5.1) and the last equality holds in view of (3.9). By Proposition 3.2, to show that π^∞ is a relaxed equilibrium, it suffices to prove $\pi^\infty \in \widetilde{\Phi}(\pi^\infty)$, i.e., $\operatorname{supp}(\pi^\infty(i)) \subseteq \widetilde{E}(\widetilde{J}^{\pi^\infty}, i)$ for all $i \in \mathcal{S}$. By contradiction, suppose that for some $i \in \mathcal{S}$, there exist $u_0 \in U$ and $r > 0$ such that

$$\operatorname{dist}\left(B_r(u_0), \widetilde{E}(\widetilde{J}^{\pi^\infty}, i)\right) > 0 \quad \text{and} \quad (\pi^\infty(i))(B_r(u_0)) > 0 \tag{A.25}$$

By the weak convergence of $\pi^n(i)$ to $\pi^\infty(i)$, $\liminf_{n\to\infty}(\pi^n(i))(B_r(u_0)) \geq (\pi^\infty(i))(B_r(u_0)) > 0$. This, along with (A.24), implies $B_r(u_0) \cap \widetilde{E}(V^{h_n,\pi^n}, i) \neq \emptyset$ for $n \in \mathbb{N}$ large enough. Take $u_n \in B_r(u_0) \cap \widetilde{E}(V^{h_n,\pi^n}, i) \in U$ for $n \in \mathbb{N}$ large enough. As $\{u_n\}_{n\in\mathbb{N}}$ is a bounded sequence, it converges up to a subsequence to some $u^* \in \overline{B_r(u_0)} \cap U$. Now, by Lemma 5.1 (a),

$$\begin{aligned} f(0,i,u^*) + q_i^{u^*} \cdot \widetilde{J}^{\pi^\infty} &= \lim_{n\to\infty}\left\{f(0,i,u_n) + q_i^{u_n} \cdot V^{h_n,\pi^n}\right\}\\ &= \lim_{n\to\infty}\max_{u\in U}\left\{f(0,i,u) + q_i^u \cdot V^{h_n,\pi^n}\right\} \geq \max_{u\in U}\left\{f(0,i,u) + q_i^u \cdot \widetilde{J}^{\pi^\infty}\right\}, \end{aligned}$$

where the second equality follows from $u_n \in \widetilde{E}(V^{h_n,\pi^n}, i)$ and the inequality holds by exchanging the limit and maximization. As $u^* \in U$, the above inequality is in fact an equality, which implies $u^* \in \widetilde{E}(\widetilde{J}^{\pi^\infty}, i)$. The fact $u^* \in \overline{B_r(u_0)} \cap \widetilde{E}(\widetilde{J}^{\pi^\infty}, i)$ readily contradicts the first condition in (A.25).

References

[1] Alexander, W. H. and Brown, J. W. [2010], 'Hyperbolically discounted temporal difference learning', *Neural Comput.* **22**(6), 1511–1527.

[2] Balbus, Ł., Jaśkiewicz, A. and Nowak, A. S. [2020], 'Markov perfect equilibria in a dynamic decision model with quasi-hyperbolic discounting', *Ann. Oper. Res.* **287**(2), 573–591.

[3] Balbus, Ł., Reffett, K. and Woźny, Ł. [2015], 'Time consistent markov policies in dynamic economies with quasi-hyperbolic consumers', *Internat. J. Game Theory* **44**(1), 83–112.

[4] Balbus, Ł., Reffett, K. and Woźny, Ł. [2018], 'On uniqueness of time-consistent markov policies for quasi-hyperbolic consumers under uncertainty', *J. Econom. Theory* **176**, 293–310.

[5] Balbus, Ł., Reffett, K. and Woźny, Ł. [2022], 'Time-consistent equilibria in dynamic models with recursive payoffs and behavioral discounting', *J. Econom. Theory* **204**, 105493.

[6] Bayraktar, E. and Han, B. [2023], 'Existence of markov equilibrium control in discrete time', *SIAM J. Financial Math.* **14**(4), SC60–SC71.

[7] Bayraktar, E., Wang, Z. and Zhou, Z. [2022], 'Short communication: stability of time-inconsistent stopping for one-dimensional diffusions', *SIAM J. Financial Math.* **13**(4), SC123–SC135.

[8] Bayraktar, E., Wang, Z. and Zhou, Z. [2023], 'Equilibria of time-inconsistent stopping for one-dimensional diffusion processes', *Math. Finance* **33**(3), 797–841.

[9] Björk, T., Khapko, M. and Murgoci, A. [2017], 'On time-inconsistent stochastic control in continuous time', *Finance Stoch.* **21**(2), 331–360.

[10] Chatterjee, S. and Eyigungor, B. [2016], 'Continuous markov equilibria with quasi-geometric discounting', *J. Econom. Theory* **163**, 467–494.

[11] Dai, M., Dong, Y. and Jia, Y. [2023], 'Learning equilibrium mean-variance strategy', *Math. Finance* **33**(4), 1166–1212.

[12] Ekeland, I. and Lazrak, A. [2006], Being serious about non-commitment: subgame perfect equilibrium in continuous time, Technical report, University of British Columbia. Available at http://arxiv.org/abs/math/0604264.

[13] Ekeland, I., Mbodji, O. and Pirvu, T. A. [2012], 'Time-consistent portfolio management', *SIAM J. Financial Math.* **3**(1), 1–32.

[14] Ekeland, I. and Pirvu, T. A. [2008], 'Investment and consumption without commitment', *Math. Financ. Econ.* **2**(1), 57–86.

[15] Fedus, W., Gelada, C., Bengio, Y., Bellemare, M. G. and Larochelle, H. [2020], Hyperbolic discounting and learning over multiple horizons, *in* 'International Conference on Machine Learning'.

[16] Fox, R., Pakman, A. and Tishby, N. [2016], Taming the noise in reinforcement learning via soft updates, *in* 'Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, UAI 2016, June 25-29, 2016, New York City, NY, USA'.

[17] Huang, Y.-J. and Nguyen-Huu, A. [2018], 'Time-consistent stopping under decreasing impatience', *Finance Stoch.* **22**(1), 69–95.

[18] Huang, Y.-J., Wang, Z. and Zhou, Z. [2022], 'Convergence of policy iteration for entropy-regularized stochastic control problems'. Preprint, available at https://arxiv.org/abs/2209.07059.

[19] Huang, Y.-J. and Zhou, Z. [2019], 'The optimal equilibrium for time-inconsistent stopping problems—the discrete-time case', *SIAM J. Control Optim.* **57**(1), 590–609.

[20] Huang, Y.-J. and Zhou, Z. [2020], 'Optimal equilibria for time-inconsistent stopping problems in continuous time', *Math. Finance* **30**(3), 1103–1134.

[21] Huang, Y.-J. and Zhou, Z. [2021], 'Strong and weak equilibria for time-inconsistent stochastic control in continuous time', *Math. Oper. Res.* **46**(2), 428–451.

[22] Jaśkiewicz, A. and Nowak, A. S. [2021], 'Markov decision processes with quasi-hyperbolic discounting', *Finance Stoch.* **25**(2), 189–229.

[23] Laibson, D. [1997], 'Golden eggs and hyperbolic discounting', *Q. J. Econ* **112**(2), 443–477.

[24] Loewenstein, G. and Thaler, R. [1989], 'Anomalies: Intertemporal choice', *J. Econ. Perspect.* **3**, 181–193.

[25] Schultheis, M., Rothkopf, C. A. and Koeppl, H. [2022], Reinforcement learning with non-exponential discounting, *in* 'Advances in Neural Information Processing Systems'.

[26] Strotz, R. H. [1955], 'Myopia and inconsistency in dynamic utility maximization', *Rev. Econ. Stud.* **23**(3), 165–180.

[27] Thaler, R. [1981], 'Some empirical evidence on dynamic inconsistency', *Econ. Lett.* **8**, 201–207.

[28] Wang, H., Zariphopoulou, T. and Zhou, X. [2020], 'Reinforcement learning in continuous time and space: A stochastic control approach', *J. Mach. Learn. Res.* **21**(198), 1–34.

[29] Yong, J. [2012], 'Time-inconsistent optimal control problems and the equilibrium HJB equation', *Math. Control Relat. Fields* **2**(3), 271–329.

[30] Ziebart, B. D., Maas, A. L., Bagnell, J. A. and Dey, A. K. [2008], Maximum entropy inverse reinforcement learning, *in* 'Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008', pp. 1433–1438.