Reinventing Policy Iteration under Time Inconsistency

Anonymous authors
Paper under double-blind review

Abstract

Policy iteration (PI) is a fundamental policy search algorithm in standard reinforcement learning (RL) setting, which can be shown to converge to an optimal policy by policy improvement theorems. However, under time-inconsistent (TIC) objectives, the use of standard PI has been marked with questions regarding the convergence of its policy improvement scheme and the optimality of its termination policy, often leading to its avoidance. In this paper, we consider infinite-horizon TIC RL setting and formally present a type of dynamic optimality: subgame perfect equilibrium that corresponds to the sophisticated behaviour of an economic agent in the face of TIC. We first analyze standard PI under this type of dynamic optimality, revealing its merits and insufficiencies. Drawing on these observations, we propose backward Q-learning (bwdQ), a new algorithm in the approximate PI family that targets SPE policy under general (non-exponential) discounting criteria. Finally, with two TIC gridworld environments, we demonstrate the implications of our theoretical findings on the behavior of the bwdQ and other approximate PI variants.

1 Introduction

Policy iteration (PI) has enjoyed a long history of success in standard reinforcement learning (RL), which can be attributed to standard PI that combines a dynamic programming (DP)-based policy evaluation and a greedy policy improvement; see Bellman (1957); Howard (1960). Standard PI has been the basis of many classical RL algorithms, such as the popular Q-learning (Watkins & Dayan (1992)), and it still inspires the design of modern RL algorithms. Despite its prominence in standard RL setting, standard PI has been deemed incompatible for time-inconsistent (TIC) objectives due to *non-monotonicity* and the implied *violation of Bellman's principle of optimality (BPO)*.

Time inconsistency (also abbreviated as TIC) can be defined loosely as a human's tendency to deviate from their current plan at a future time. In the context of RL, TIC often arises as an effort to more closely model human preferences and has been investigated through several major channels such as hyperbolic discounting and risk-sensitive RL. The idea of questioning the validity of standard PI under TIC was pioneered in risk-sensitive RL by Sobel (1982). In this seminal work, a counterexample to the monotonicity property (also referred to as consistent choice, temporal persistence, or stationarity across the literature) was posted and attention was raised in how this property is commonly exploited to prove the convergence of standard policy improvement scheme to an optimal policy. *Two puzzles are then left for answers*: (i) the optimality of a termination policy and (ii) the lack of guarantees on update monotonicity (a desirable algorithmic property that will lead to convergence).

In this paper, we focus on infinite-horizon TIC RL problems and formally present the subgame perfect equilibrium (SPE), which corresponds to how sophisticated agents behave in the face of TIC, as a dynamic optimum. We will then revisit the two questions above to highlight standard PI's merits and insufficiencies in achieving dynamic optimality.

The contribution of this paper can be summarized as follows:

- In terms of optimality, we establish that the termination policy of standard PI under TIC achieves SPE.

- We study the PIT failure issue and highlight some insufficiencies of standard PI update and the existing analysis tools, in the context of SPE policy search.
- TIC-adjusted DP formula is established to compute general-discounting TIC Q-function, addressing the insufficiency of standard DP formula.
- Based on the aforementioned analyses, we devise a new PI paradigm for generally (or non-exponentially) discounted rewards: backward Q-learning (bwdQ).
- We design toy Gridworld examples to demonstrate the implications of our findings on the behaviour of bwdQ and other approximate PI variants under TIC.
- The analyses relevant to the *backward conditioning* component in bwdQ is of independent interest: the characterization of its termination policy as SPE and its efficiency-related desirability as an SPE policy learner extend beyond general-discounting objectives.

Note that some lengthy proofs/justifications of our results are deferred to Appendix.

2 Related Works

Non-monotonicity in risk-sensitive RL and solutions. In risk-related context, several follow-up works since Sobel (1982) address the non-monotonicity issue following the line of reasoning that non-monotonic problems are computationally intractable such that new tractable solutions are required. For instance, Mannor & Tsitsiklis (2011) formally compares between several policy classes to reduce the search problem for globally-optimal policy (to a specific policy class) and proposes several tractable approximation algorithms. One important finding in their work is that randomization can improve control performance which inspires Di Castro et al. (2012); Tamar & Mannor (2013); Prashanth & Ghavamzadeh (2013) to propose gradient-based algorithms accustomed to mean-variance criteria, quoting parameterized stochastic policy as a manner to deal with non-monotonicity. The latter works are relevant to our case as they also use TIC adjustment terms to obtain temporal difference (TD)-based policy evaluation (PE) that resembles the one used in extended DP theory Björk et al. (2014). To distinguish our approach, we note our focus on using SPE policy itself to deal with non-monotonicity (by *modifying our optimality type*) as opposed to randomization or parameterization.

Non-monotonicity in hyperbolic-discounting RL and solutions. In hyperbolic-discounting context, non-monotonicity have also appeared, independent of Sobel (1982)'s work; see Kurth-Nelson & Redish (2010) for instance. In this work, several proposals towards computationally tractable models are reviewed, with varying action selection strategies drawn largely from behavioral or neuroscience point of view. A recent follow-up work by Fedus et al. (2019) extends their distributed micro-agents model (i.e. $\mu Agents$) to handle larger scale problems, utilizing deep neural network to model the different Q-values from a shared representation. Though such modifications in action selection may have implicitly addressed the non-monotonicity underlying PIT failure, to the best of our knowledge, an explicit connection between the two (as in Sobel (1982)) has never been made.

Time-consistent Planning and Control. The idea of locally optimal, time-consistent planning under TIC was pioneered by Strotz (1955); Pollak (1968). This type of planning corresponds to a sophisticated, rational agent's behavior who, when faced with TIC, compromises with their future selves by taking future disobedience as a constraint in their decision-making. The solution concept is developed as a game-theoretic framework that builds on backward inductive SPE search in games, thus coining the term SPE plan or policy. This then leads to an *intra-personal equilibria* formalism by Björk & Murgoci (2014) which unifies several task-specific TIC sources through extended DP theory and has attracted a wide array of literature in TIC stochastic control. The rise of SPE policy as a major contending solution to the globally optimal (precommitment) policy can then be attributed to two reasons: (i) as a controller, precommitment policy may lead to some undesirable outcomes since it may lose its optimality as time evolves (for instance, due to an unpredictable change in environment dynamics), (ii) computationally, there is lack of a pivotal tool to identify a globally-optimal policy that generalizes naturally to different TIC tasks (for instance, due to its disconnection to standard DP that requires BPO).

SPE Policy in TIC-RL. Some works in the general-discounting space have investigated TIC-RL from a purely behavioral lens, focusing particularly on the property of target policy rather than a computational aspect. For instance, Lattimore & Hutter (2014) proposes rational agents that act according to history-dependent SPE policies. In this work, the authors cover some theoretical aspects of policies such as characterization of different policy types, existence results connecting discounting and policy types, and comparative study in some example scenarios. In another work, Evans et al. (2016) proposes sophisticated agents that act according to Markovian SPE policies and are modelled with delay-augmented Q-learning algorithms. Though relevant, these algorithms are proposed in the context of generative models that aids human-like preference inferences; thus, algorithmic properties are not covered. A recent work by Lesmana & Pun (2021) considers the search of Markovian SPE policy under finite-horizon task-invariant TIC objectives. Drawing inspiration from the extended DP theory, the authors propose Backward Policy Iteration (BPI), which has lex-monotonicity guarantee in place of PIT. This work is the closest to ours where our *backward conditioning* can be viewed as an infinite-horizon extension to BPI. We distinguish our contribution by noting our main focus on *analyzing standard PI*, that motivates our *infinite-horizon, Markovian* SPE policy formalism and the corresponding drop of time-dependency, *shifting definition of players from times to states*. Relative to finite-horizon case, such formalism introduces technical challenges in both aspects of policy evaluation and improvement, which we will remark on the respective sections of this paper.

3 Problem Formulation and the SPE Concept

In this section, we introduce the class of TIC RL problems of our interest and formally present the solution concept of SPE policy. We then cast the general-discounting objective as a TIC RL problem and construct a few examples in this context that we will quote frequently throughout the paper.

3.1 TIC RL Problem Formulation

We consider the policy search in an infinite-horizon TIC-MDP, which consists of the standard MDP tuple $(\mathcal{S}, \mathcal{A}, P, \mathcal{R})$ and a specific TIC source. The state space $\mathcal{S}$ and action spaces $\mathcal{A}_s \subseteq \mathcal{A}, \forall s \in \mathcal{S}$, are assumed to be discrete with stationary probabilities $p_s^a(\cdot) \doteq P(R_{t+1} = \cdot, S_{t+1} = \cdot \,|\, S_t = s, A_t = a)$ governing the transitions from a current state $S_t = s$ to the next state S_{t+1} and reward R_{t+1} for $s \in \mathcal{S}$, given a particular action $A_t = a$. To define a stopping criterion, it is convenient to augment a so-called *absorbing* state, denoted by $\bar{s}_{\text{void}}$, which incurs no reward. Then, we define a *stopping* action $\bar{a}$ as an action that drives a transition to $\bar{s}_{\text{void}}$ from any states $s \in \mathcal{S}$ and *boundary* states $\bar{s} \in \bar{\mathcal{S}}$ as rewarding states with specific action space $\mathcal{A}_{\bar{s}}$ or $\bar{\mathcal{A}} \coloneqq \{\bar{a}\}$, i.e. once the boundary state is reached, we conclude with reward as there is only action $\bar{a}$ that will transit us from $\bar{s}$ to $\bar{s}_{\text{void}}$ and then make us stay at $\bar{s}_{\text{void}}$ forever. This setup is to complete the mathematical framework of the environment for analyses, where the problem of interest has certain stopping criteria, e.g., after receiving a target reward.

Let us denote by Π^{MD} the set of all Markovian, deterministic policies $\boldsymbol{\pi} \coloneqq \{\pi(s) : s \in \mathcal{S}\}$ with $\pi : \mathcal{S} \to \mathcal{A}_s$. To aid presentation in subsequent sections, we define $a \cdot \boldsymbol{\pi}$ as a policy that prescribes the use of action $a \in \mathcal{A}$ for a current decision and policy $\boldsymbol{\pi}$ for all remaining decisions. Similarly, we denote by $\boldsymbol{\delta}^\tau \cdot \boldsymbol{\pi}$ a policy that fixes the first $\tau > 0$ decisions to $\boldsymbol{\delta}^\tau \coloneqq \{\delta(S_w) : t \leq w < t + \tau\}$, with a map $\delta : \mathcal{S} \to \mathcal{A}_s$ and a current time t, and follows $\boldsymbol{\pi}$ afterwards.

TIC reward structures and criterion We first note that by our assumption on stationary transitions, we are limiting our TIC scope to those that arise from reward structures and criterion, described as follows. Let us consider a general criterion $V^{\boldsymbol{\pi}}(s)$ (with form not restricted at this point) for any $\boldsymbol{\pi} \in \Pi^{MD}$ and $s \in \mathcal{S}$. Given an initial state $s_0 \in \mathcal{S}$, a standard notion of optimality aims to solve the *global* problem $\mathcal{P}_{0,s_0} \doteq \max_{\boldsymbol{\pi}} V^{\boldsymbol{\pi}}(s_0)$ and obtain the corresponding globally-optimal (precommitment) policy denoted by $\boldsymbol{\pi}^{*0}$. Next, let us define for each delay $\tau > 0$, the *local* problem $\mathcal{P}_{\tau,s_\tau} \doteq \max_{\boldsymbol{\pi}} V^{\boldsymbol{\pi}}(s_\tau)$, where s_τ represents any realization of S_τ following the sequence of policies $\left\{\pi^{*0}(S_t) : 0 \leq t < \tau \,|\, S_0 = s_0\right\}$, and denote by $\pi^{*\tau}(s_\tau)$ its solution. **BPO** then states that

$$\forall \tau, s_\tau,\ \pi^{*\tau}(s_\tau) = \pi^{*0}(s_\tau) \tag{1}$$

By the definition above, the standard RL criteria belong to the time-consistent (TC) class that does not violate (1). In this paper, we consider criteria $V^{\boldsymbol{\pi}}(s)$ that violate (1); these include general-discounting, risk-related, and more; see Björk & Murgoci (2014). One can normally verify if $V^{\boldsymbol{\pi}}(s)$ belongs to TIC class through a counterexample, which will be illustrated with Example 3.6 below.

It is noteworthy that our way in defining TIC criterion is unlike most MDPs that specify the criterion $V^{\boldsymbol{\pi}}(s)$ up to the expectation of cumulative rewards. We aim to maintain generalities up to the formalism of SPE optimality (in Section 3.2) that is valid for different forms of reward structures and criterion, categorized with the TIC phenomena. For instance, while general-discounting objectives still admit an expectation form, risk-sensitive objectives involve non-linearity in expectations. That said, to fully define a TIC-MDP, we need exact the specifications of reward structures $\mathcal{R}$ and thus the corresponding TIC sources. We will further discuss on this topic in Section 3.3 with general-discounting specifications.

3.2 SPE Notion of Optimality

For any TIC criterion, we can define the corresponding action-value or Q-function $Q^{\boldsymbol{\pi}}(s,a) \coloneqq V^{a\cdot\boldsymbol{\pi}}(s)$. Under the *infinite-horizon* SPE notion of optimality, our aim is to find an SPE policy $\hat{\boldsymbol{\pi}}$ defined as follows.

Definition 3.1 (SPE Policy)**.** A policy $\hat{\boldsymbol{\pi}} \in \Pi^{MD}$ is an SPE policy if it satisfies

$$Q^{\hat{\boldsymbol{\pi}}}(s,\hat{\pi}(s)) \geq Q^{\hat{\boldsymbol{\pi}}}(s,a), \forall a \in \mathcal{A}_s, \forall s \in \mathcal{S} \tag{2}$$

From a game-theoretic perspective, Definition 3.1 means that any state s does not have incentive to deviate from its strategy $\hat{\pi}(s)$ at the current stage when in the continuation play, other states $s' \in \mathcal{S} \setminus \{s\}$ and s itself play $\hat{\boldsymbol{\pi}}$. In other words, the game consists of Players s, indexed by the *states* $s \in \mathcal{S}$, and we look for an SPE, where Player s takes into account the strategies of other Players $s' \in \mathcal{S} \setminus \{s\}$ in its decision making as s will be transited to an s'. Throughout this paper, we will adopt several technical assumptions[1] to address the technical challenges of such *infinite-horizon* SPE, particularly those that arise from the derivations of *TIC-adjusted DP* in Section 4.2 and *backward conditioning* update in Section 5.1.

Assumption 3.2. $\forall s \in \mathcal{S}, \forall \epsilon > 0, \exists \bar{T} < \infty$ s.t. $\forall \boldsymbol{\pi} \in \Pi^{MD}, \ddot{\boldsymbol{\pi}} : \mathcal{S} \to \bar{\mathcal{A}}$,

$$\left|\left| V^{\boldsymbol{\pi}}(s) - V^{\boldsymbol{\pi}^{\bar{T}}\cdot\ddot{\pi}}(s) \right|\right| \leq \epsilon. \tag{3}$$

Denote by $\bar{T}_{s,\epsilon}$ the *smallest* such $\bar{T}$.

Assumption 3.3. $\exists s_0 \in \mathcal{S}$ s.t. $\exists \hat{T}_{s_0} < \infty$

$$\forall s \in \mathcal{S} \setminus \{s_0\}, \exists \boldsymbol{\pi} \in \Pi^{MD}, \sum_{t=0}^{\hat{T}_{s_0}} \mathbb{P}[S_t^{\boldsymbol{\pi}} = s | s_0] > 0.$$

Intuitively, Assumption 3.2 asserts that starting from any s and following any policy $\boldsymbol{\pi}$, any rewards generated after $\bar{T}$ steps are negligible as the policy $\ddot{\boldsymbol{\pi}}$ incurs 0 rewards. Then, Assumption 3.3 ensures the existence of at least one state s_0 from which all other states $s \in \mathcal{S} \setminus \{s_0\}$ can be reached in finite time, with positive probability. Combining these two, we fix s_0 and set $\bar{T}_\epsilon \coloneqq \max\{\hat{T}_{s_0} + 1, \bar{T}_{s_0,\epsilon}\}$ to obtain our last assumption, which quantifies the negligibility of rewards when initiated at any intermediate states s_t against s_0.

Assumption 3.4. Let us define for all $t \in [0,\infty)$,

$$\mathcal{S}_t^{s_0,\Pi^{MD}} \coloneqq \{s \in \mathcal{S} : \exists \tau \in [0,t), s_\tau \in \mathcal{S}_\tau^{s_0,\Pi^{MD}} \text{ s.t. } \exists \boldsymbol{\pi} \in \Pi^{MD}, \mathbb{P}[S_{t-\tau}^{\boldsymbol{\pi}} = s | S_0 = s_\tau] > 0\} \tag{4}$$

with $\mathcal{S}_0^{s_0,\Pi^{MD}} \coloneqq \{s_0\}$. Then, $\forall \epsilon > 0, \forall t \in [0, \bar{T}_\epsilon], \forall s_t \in \mathcal{S}_t^{s_0,\Pi^{MD}}$, and $\forall \boldsymbol{\pi} \in \Pi^{MD}$ with $\mathbb{P}[S_t^{\boldsymbol{\pi}} = s_t | s_0] > 0$, $\exists \kappa = \kappa(\epsilon, t, s_t, \boldsymbol{\pi}; s_0) > 0$ s.t. $\left|\left| V^{\boldsymbol{\pi}}(s_t) - V^{\boldsymbol{\pi}^{\bar{T}_\epsilon - t}\cdot\ddot{\boldsymbol{\pi}}}(s_t) \right|\right| \leq \epsilon/\kappa$ and $\lim_{\epsilon\to 0} \epsilon/\kappa = 0$.

[1] Readers may refer to Appendix A for some MDP examples, in which Assumption 3.2-3.4 hold.

3.3 General-discounting Criterion

As a major concern of this paper, we consider the following infinite-horizon criterion

$$V^{\boldsymbol{\pi}}(s_\tau) \doteq \mathbb{E}_{s_\tau}\left[\sum_{t=\tau}^{\infty} \varphi(t-\tau) R(S_t^{\boldsymbol{\pi}}, \pi(S_t^{\boldsymbol{\pi}}))\right] \tag{5}$$

with a general discounting function $\varphi : \mathbb{N} \to [0, 1)$, defined for any $\tau \geq 0$. The intermediate (possibly random) reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$ follows the standard MDP formulation, with emphasis on its boundedness and non-negativity. We further note our use of notations $\mathbb{E}_{s_\tau}[\,\cdot\,]$ for $\mathbb{E}[\,\cdot\,|S_\tau = s_\tau]$ and $S_t^{\boldsymbol{\pi}}$ for the (random) state visited at time t on a trajectory generated by following policy $\boldsymbol{\pi}$ and initialized at $S_\tau = s_\tau$.

Next, we define action-value or Q-function that relates to the value function in (5). To emphasize on the stationarity of our problem, we avoid any explicit appearance of τ[2] and perform a change of parameter $m = t - \tau$.

Definition 3.5 (Q-function)**.** For each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ and a fixed policy $\boldsymbol{\pi} \in \Pi^{MD}$, we define *Q-function* as

$$Q^{\boldsymbol{\pi}}(s, a) \doteq \mathbb{E}_s\left[\sum_{m=0}^{\infty} \varphi(m) R(S_m^{a\cdot\boldsymbol{\pi}}, \boldsymbol{\pi}(S_m^{a\cdot\boldsymbol{\pi}}))\right] \tag{6}$$

We may now revisit the TIC concept by witnessing how criterion (5) violates BPO through a Gridworld counterexample.

Example 3.6 (BPO Violation). Consider a Gridworld environment as described in Figure 1(a) and a hyperbolic-discounting criterion by setting $\varphi(t-\tau) = 1/(1 + k(t-\tau))$ in (5). Given $s_0 = 21$, we can compute (by trajectory enumeration) the globally-optimal, precommitment policy $\boldsymbol{\pi}^{*0}$ as in Figure 1(b). After applying delay $\tau = 3$ and following the delaying policy $\boldsymbol{\delta}^\tau = \boldsymbol{\pi}^{*0}$, we reach $s_\tau = 9$, at which the locally-optimal policy suggests $\boldsymbol{\pi}^{*\tau} \doteq \{\pi^{*\tau}(9)\} = \{\leftarrow\}$ and accrues rewards $V^{\boldsymbol{\pi}^{*\tau}}(s_\tau) = 10/(1+1) > 19/(1+3) = V^{\boldsymbol{\pi}^{*0}}(s_\tau)$; see Figure 1(d). This violates BPO at $\tau = 3$ and $s_\tau = 9$.

Hereafter, we will use the general-discounting TIC value function and Q-function as defined in (5)-(6) for any appearance of $V^{\boldsymbol{\pi}}(s)$ and $Q^{\boldsymbol{\pi}}(s, a)$, unless specified otherwise. Correspondingly, we define exponential-discounting time-consistent (TC) value function and Q-function $V_{TC}^{\boldsymbol{\pi}}(s)$ and $Q_{TC}^{\boldsymbol{\pi}}(s, a)$ by letting $\varphi(m) = \gamma^m$ to exemplify any standard RL formulations in the subsequent sections.

4 An Analysis of Standard PI

In this section, we analyze standard PI under the SPE optimality type, revealing its merits and insufficiencies.

4.1 SPE Optimality of Termination Policy

We first present the standard PI update,

$$\pi'(s) \leftarrow \arg\max_{a \in \mathcal{A}_s} Q^{\boldsymbol{\pi}}(s, a), \forall s \in \mathcal{S} \tag{7}$$

where $\boldsymbol{\pi}', \boldsymbol{\pi}$ represent *new* and *old* policies in any two consecutive iterations. Next, we will show the merit of standard PI in Proposition 4.1: its termination policy achieves SPE optimality.

Proposition 4.1. *If $\boldsymbol{\pi}' = \boldsymbol{\pi}$ and update follows the rule in (7), then $\boldsymbol{\pi}, \boldsymbol{\pi}'$ are SPE policy.*

Proof. By (7) and $\boldsymbol{\pi}' = \boldsymbol{\pi}$, we obtain that $\forall a \in \mathcal{A}_s, \forall s \in \mathcal{S}$,

$$Q^{\boldsymbol{\pi}}(s, \pi'(s)) \geq Q^{\boldsymbol{\pi}}(s, a) \quad \Rightarrow \; Q^{\boldsymbol{\pi}}(s, \pi(s)) \geq Q^{\boldsymbol{\pi}}(s, a).$$

Thus, by Definition 3.1, $\boldsymbol{\pi}$, $\boldsymbol{\pi}'$ are SPE policy. □

[2]Later in Section 4.2, τ will be re-introduced as an in-training parameter of our agent that keeps track of nonstationarity changes.

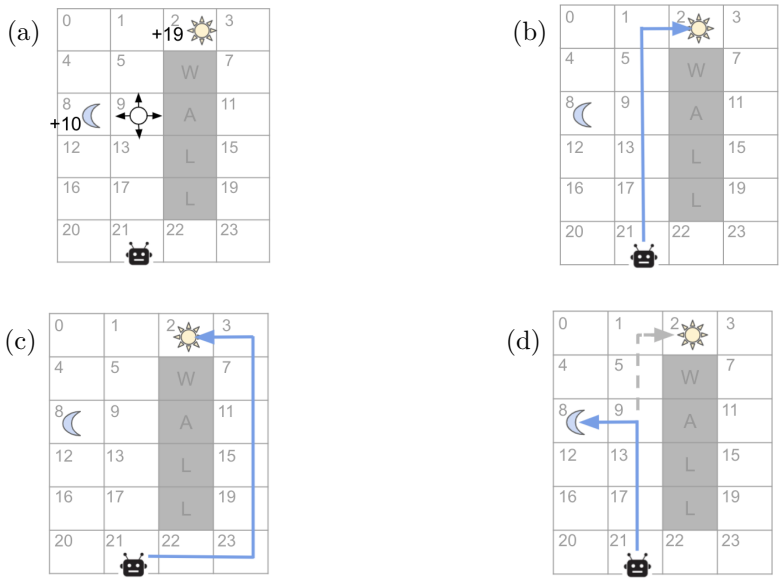

Figure 1: **(a) Deterministic, Hyperbolic ($k = 1$) Gridworld.** $\mathcal{S}$ comprises 2 absorbing states $\bar{S} = \{\bar{2}, \bar{8}\}$ emitting rewards $R(\bar{2}) = 19, R(\bar{8}) = 10$. Each action in $\mathcal{A} = \{\uparrow, \rightarrow, \downarrow, \leftarrow\}$ drives transition through deterministic P; transitions to WALL or outside the grid will spawn the agent back to its original location. **(b) Precommitment policy $\boldsymbol{\pi}^{*0}$** and its corresponding path, accruing accumulated rewards $V^{\boldsymbol{\pi}^{*0}}(s_0) = \frac{19}{1+6}$. This path exhibits TIC at $\tau = 3$ and $s_\tau = 9$ as shown in Example 3.6. **(c) SPE policy $\hat{\boldsymbol{\pi}}$** and its corresponding path, accruing rewards $V^{\hat{\boldsymbol{\pi}}}(s_0) = \frac{19}{1+8} < V^{\boldsymbol{\pi}^{*0}}(s_0)$. One could refer back to Definition 3.1 and verify that no states s_τ on this path have the incentive to deviate from its current policy $\hat{\pi}(s_\tau)$. **(d) Delusional policy $\boldsymbol{\delta}^\tau \cdot \boldsymbol{\pi}^{*\tau}$** and its corresponding path, with $\tau, \boldsymbol{\delta}^\tau$, and $\boldsymbol{\pi}^{*\tau}$ specified in Example 3.6, accruing rewards of $V^{\boldsymbol{\delta}^\tau \cdot \boldsymbol{\pi}^{*\tau}}(s_0) = \frac{10}{1+4} < V^{\hat{\boldsymbol{\pi}}}(s_0)$. The term 'delusional' is used to reflect how state 21 presumes 9 will go up, unaware of the TIC issue.

4.2 Policy Evaluation

The update rule (7) requires the computation of the true TIC Q-function $Q^{\boldsymbol{\pi}}(s, a)$, which is not straightforward. In standard RL setting, there is a DP formula to efficiently compute TC Q-function,

$$Q^{\boldsymbol{\pi}}_{TC}(s_t, a_t) = \mathbb{E}_{R,S' \sim p^{a_t}_{s_t}}[R(s_t, a_t) + \gamma V^{\boldsymbol{\pi}}_{TC}(S')], \tag{8}$$

where $V^{\boldsymbol{\pi}}_{TC}(s)$ (iteratively) solves (8) after substituting $\pi(s_t)$ into a_t. Under general discounting as in (5), (8) no longer holds. In this subsection, we present a recursive formula satisfied by our TIC Q-function (see (14) below) by leveraging the extended DP theory (Björk et al. (2014)).

TIC-adjusted DP Noting that in Section 3.2 we have assumed access to a fixed $\bar{T}_\epsilon < \infty$, we introduce a reward adjustment (or r-function) that our agent will use it to track the nonstationary changes (due to TIC) in Q-function.

Definition 4.2 (r-function)**.** For each $\tau \in \{0, \dots, \bar{T}_\epsilon\}$, $m \in \{\tau, \dots, \bar{T}_\epsilon\}$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, and a fixed policy $\boldsymbol{\pi} \in \Pi^{MD}$, we define r-function as

$$r^{\boldsymbol{\pi},\tau,m}(s, a) \doteq \mathbb{E}_s\left[\varphi(m - \tau) R\left(S^{a \cdot \boldsymbol{\pi}}_m, \pi(S^{a \cdot \boldsymbol{\pi}}_m)\right)\right] \tag{9}$$

where τ and m are fixed parameters.

Next, we will use the adjustment function above to obtain a formula that recursively computes our Q-function.

Theorem 4.3. *For any fixed $\boldsymbol{\pi} \in \Pi^{MD}, \tau \in \{0, \dots, \bar{T}_\epsilon\}$, $m \in \{\tau, \dots, \bar{T}_\epsilon\}$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, r-function satisfies that for $m = \tau$,*

$$r^{\boldsymbol{\pi},\tau,\tau}(s,a) = \mathbb{E}_{R\sim p_s^a}\left[\varphi(0)R(s,a)\right] \tag{10}$$

and for $m \geq \tau + 1$,

$$r^{\boldsymbol{\pi},\tau,m}(s,a) = \mathbb{E}_{S'\sim p_s^a}\left[\frac{\varphi(m-\tau)}{\varphi(m-(\tau+1))} r^{\boldsymbol{\pi},\tau+1,m}(S',\pi(S'))\right]. \tag{11}$$

$$\tag{12}$$

Then, by fixing the parameters τ, m on r-function accordingly, Q-function satisfies that $\forall s_t \in \bar{\mathcal{S}}, a_t \in \bar{\mathcal{A}}, \boldsymbol{\pi} \in \Pi^{MD}$,

$$Q^{\boldsymbol{\pi}}(s_t,a_t) = \mathbb{E}_{R\sim p_{s_t}^{a_t}}[\varphi(0)R(s_t,a_t)] \tag{13}$$

Moreover, under some technical conditions (see Assumption B.1), we have $\forall s_t \in \mathcal{S}\setminus\bar{\mathcal{S}}$ and $a_t \in \mathcal{A}_{s_t}$,

$$Q^{\boldsymbol{\pi}}(s_t,a_t) \overset{2\epsilon/\kappa}{\approx} \mathbb{E}_{R\sim p_{s_t}^{a_t}}[\varphi(0)R(s_t,a_t)] + \mathbb{E}_{S'\sim p_{s_t}^{a_t}}\left[Q^{\boldsymbol{\pi}}(S',\pi(S'))\right] - \Delta r_t^{\boldsymbol{\pi}}, \tag{14}$$

where $\Delta r_t^{\boldsymbol{\pi}} \doteq \sum_{m=t+1}^{\bar{T}_\epsilon}\left(\mathbb{E}_{S'\sim p_{s_t}^{a_t}}\left[r^{\boldsymbol{\pi},t+1,m}(S',\pi(S'))\right] - r^{\boldsymbol{\pi},t,m}(s_t,a_t)\right)$.

Remark 4.4. Both the proof of Theorem 4.3 and the technical assumptions for it are provided in Appendix B. Theorem 4.3 is an analog to Proposition 11 in Lesmana & Pun (2021), while the main technical difference is in the approximation ('$\approx$') step for deriving (14). Here, we have used Assumption 3.4 and some "relevance" conditions, whose details are deferred to Appendix B.1, to ensure our approximation error can be made arbitrarily small by choosing sufficiently large $\bar{T}_\epsilon$.

Remark 4.5. Theorem 4.3 has used the specific properties of general-discounting TIC source. For other types of TIC sources, recursive formulas need to be re-derived. In risk-sensitive case, for instance, readers may refer to Tamar & Mannor (2013); Sobel (1982).

Standard TD-based approximation algorithms such as Q-learning Watkins & Dayan (1992) are drawn from the standard formula (8) and thus, are insufficient for general-discounting factor. Theorem 4.3 provides a formula that addresses insufficiency of standard formula (8), which we will use to reinvent a new approximate PI algorithm for general-discounting objectives.

4.3 Policy Improvement (Update Monotonicity)

In this subsection, we will highlight some insufficiencies of the standard PI's update and analysis tools in the fact of TIC by revisiting the unprovable Policy Improvement Theorem (PIT). To start off, we present the following proof of PIT in Sutton & Barto (2018): $\forall s \in \mathcal{S}$,

$$\begin{aligned} V_{TC}^{\boldsymbol{\pi}}(s) \leq Q_{TC}^{\boldsymbol{\pi}}(s,\pi'(s)) &= \mathbb{E}\left[R_{t+1}^{\pi'} + \gamma V_{TC}^{\boldsymbol{\pi}}(S_{t+1}^{\pi'})|S_t = s\right] \\ &\leq \mathbb{E}\left[R_{t+1}^{\pi'} + \gamma Q_{TC}^{\boldsymbol{\pi}}(S_{t+1}^{\pi'},\pi'(S_{t+1}^{\pi'}))|S_t = s\right] = \cdots \leq \cdots \\ &\leq V_{TC}^{\boldsymbol{\pi}'}(s). \end{aligned} \tag{15}$$

Note that in each alternating step of '$=$' and '$\leq$', two operations are performed: (i) a recursive expansion of TC Q-function, and (ii) substituting the **monotonicity** relation: $\forall s \in \mathcal{S}$,

$$V_{TC}^{\pi'\cdot\boldsymbol{\pi}}(s) \geq V_{TC}^{\pi\cdot\boldsymbol{\pi}}(s) \;\Rightarrow\; V_{TC}^{\boldsymbol{\delta}^\tau\cdot\pi'\cdot\boldsymbol{\pi}}(s) \geq V_{TC}^{\boldsymbol{\delta}^\tau\cdot\pi\cdot\boldsymbol{\pi}}(s) \tag{16}$$

for all delays $\tau \geq 1$ and $\boldsymbol{\delta}^\tau = \{\pi', \pi', \dots\}$. Let us pay attention to the monotonicity relation, particularly about how (16) fails under a TIC criterion. To this end, we recall Example 3.6 and focus on the states along the precommitment path in Figure 1(b). We can then counter (16) as follows:

Set $\boldsymbol{\pi} = \boldsymbol{\pi}^{*0}, \tau = 3, \boldsymbol{\delta}^{\tau} = \boldsymbol{\pi}^{*0}, \pi'(9) = \leftarrow$; then, $\frac{19}{1+3} = V^{\pi \cdot \boldsymbol{\pi}}(9) \leq V^{\pi' \cdot \boldsymbol{\pi}}(9) = \frac{10}{1+1}$ holds. However,

$$\frac{19}{1+6} = V^{\boldsymbol{\delta}^3 \cdot \pi \cdot \boldsymbol{\pi}}(21) = \mathbb{E}_{\boldsymbol{\delta}^3}[V^{\pi \cdot \boldsymbol{\pi}}(9)] > \mathbb{E}_{\boldsymbol{\delta}^3}[V^{\pi' \cdot \boldsymbol{\pi}}(9)] = V^{\boldsymbol{\delta}^3 \cdot \pi' \cdot \boldsymbol{\pi}}(21) = \frac{10}{1+4} \tag{17}$$

showing that at $s = 21$, the monotonicity relation (16) does not hold.

We make two observations here: (i) a PIT-like improvement (i.e. $\forall s \in \mathcal{S}, V^{\boldsymbol{\pi}'}(s) \geq V^{\boldsymbol{\pi}}(s)$) might not suffice as it targets optimal policies *not* SPE policies, (ii) the counterexample (17) suggests the existence of *priority ordering*[3] over $\mathcal{S}$ (i.e. 9 holds *priority* over 21) such that unordered (i.e. $\forall s \in \mathcal{S}$) update as in (7) might not suffice. To further probe on these issues, we will consider the following example.

Example 4.6 (Insufficiencies in Standard PI). Let us refer back our Hyperbolic Gridworld in Figure 1(a). We will keep our deterministic transition and reward functions, but restrict our state-space to:

$$\tilde{\mathcal{S}} = \{1, 2, 3, 5, 7, 8, 9, 11, 13, 15, 17, 19, 21, 22, 23\}$$

and action-spaces to:

$$\begin{aligned}
&\tilde{\mathcal{A}}_{21} = \{\uparrow, \rightarrow\}, \quad \tilde{\mathcal{A}}_9 = \{\uparrow, \leftarrow\},\\
&\mathcal{A}_3 = \{\leftarrow\},\\
&\mathcal{A}_s = \{\rightarrow\}, \quad \forall s \in \{1, 22\},\\
&\mathcal{A}_s = \{\uparrow\}, \quad \forall s \in \{17, 13, 5, 7, 11, 15, 19, 23\},\\
&\mathcal{A}_s = \{\bar{a}\}, \quad \forall s \in \{2, 8\}.
\end{aligned}$$

Letting $s_0 = 21, \epsilon = 0, \tilde{\Pi} = \{\boldsymbol{\pi} \in \Pi^{MD} | \boldsymbol{\pi} : \tilde{\mathcal{S}} \rightarrow \tilde{\mathcal{A}}_s\}$, we have a *priority-ordering* on $\tilde{\mathcal{S}}$, $\tilde{\mathcal{S}}_{0:\bar{T}} = \tilde{\mathcal{S}}^{s_0, \tilde{\Pi}}_{0:\bar{T}_0}$:

$$\tilde{\mathcal{S}}_0 = \{21\},\ \tilde{\mathcal{S}}_1 = \{17, 22\},\ \tilde{\mathcal{S}}_2 = \{13, 23\},\ \tilde{\mathcal{S}}_3 = \{9, 19\},\ \tilde{\mathcal{S}}_4 = \{5, 8, 15\},$$

$$\tilde{\mathcal{S}}_5 = \{1, 11\},\ \tilde{\mathcal{S}}_6 = \{2, 7\},\ \tilde{\mathcal{S}}_7 = \{3\},\ \tilde{\mathcal{S}}_8 = \{2\}. \tag{18}$$

To apply standard PI for SPE policy search from $s_0 = 21$, we can choose an initial policy $\boldsymbol{\pi}^{(0)}$ as illustrated in Figure 2(a). Following the standard PI's rule (7), policies at all states are updated conditional to the policy at previous iteration. Let us now focus on the two important states $9, 21$, in which decisions need to be made (i.e. $|\tilde{\mathcal{A}}_s| > 1$). First, note that after updated conditional to the old policy $\pi^{(0)}(9) = \uparrow$, $\pi^{(1)}(21) = \uparrow$ that incurs a higher reward; see Figure 2(b). Then, note that $\pi^{(1)}(9) = \leftarrow$. When combined, the current iteration ends up in $\boldsymbol{\pi}^{(1)}$ that corresponds to a *delusional* path; see Figure 2(c). We contend that such iterative update is insufficient as we would have found the desired SPE path if state 21 has knowledge of $\pi^{(1)}(9) = \leftarrow$. We will see how we can achieve this in the next section, by leveraging a known *priority-ordering* as in (18).

5 Backward Q-learning Algorithm

Drawing upon the analyses and observations in Section 4, we propose a new algorithm in the approximate PI family that targets SPE policy under a general-discounting criterion.

5.1 Backward Conditioning

To mitigate the insufficiencies surrounding update monotonicity, we build on a recent result in Lesmana & Pun (2021) and propose *backward conditioning*: to perform update *backward* from $\mathcal{S}_{\bar{T}}$ to $\mathcal{S}_0$ and *conditioning* the update of states with *lower* priority (happens *earlier*) on the *new* policy $\boldsymbol{\pi}'$ of states with *higher* priority (happens *later*). We formalize the above in the following update rule.

[3]By either the sophisticated agent's strategy Strotz (1955) or its SPE formalism Björk et al. (2014), *higher priority* here corresponds to a *later order of visitation* in a trajectory.

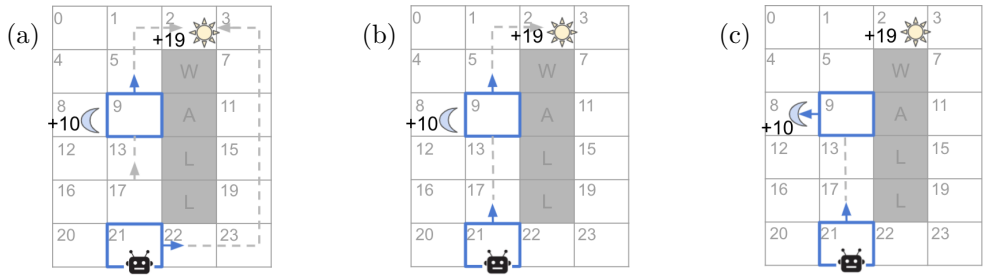

Figure 2: **2-Layered Correction with Standard PI.** (a) Initialization. (b) State 21 updates first. (c) State 9 updates last, after $21, 17, 13$ make their updates.

Definition 5.1 (*Backward Conditioning* Rule)**.** For any $\epsilon > 0$, set $\bar{T} \coloneqq \bar{T}_\epsilon$ and let $\mathcal{S}_{0:\bar{T}} \coloneqq \mathcal{S}_{0:\bar{T}}^{s_0,\Pi^{MD}}$ be a *priority-ordering* on $\mathcal{S}$. Then, for $t = \bar{T} - 1 : 0$:

$$\forall s \in \mathcal{S}_t, \pi'(s) \leftarrow \arg\max_{a \in \mathcal{A}_s} Q^{(\boldsymbol{\pi}')^{\bar{T}-1-t} \cdot \boldsymbol{\pi}}(s, a) \tag{19}$$

Remark 5.2. Note that in the Definition 5.1, we have assumed the existence of a *priority-ordering*: whatever actions the states in $\mathcal{S}_{0:t-1}$ are taking are assumed to have no effect on the choice of states in $\mathcal{S}_t$. This justifies (19): its *backward* order and conditioning the update of any $s \in \mathcal{S}_{\bar{T}-1}$ (with *highest* priority) on the *old* policy $\boldsymbol{\pi}$. We note however that even without such *priority-ordering*, the worst that can happen is $\mathcal{S}_t^{s_0,\Pi^{MD}} = \mathcal{S}, \forall t \in [0, \bar{T}]$, which is equivalent to performing standard PI in (7) $\bar{T}$ times.

Next, we will show that the backward conditioning rule preserves the SPE optimality of termination policy.

Proposition 5.3. *If $\boldsymbol{\pi}' = \boldsymbol{\pi}$ and update follows the rule in equation 19, then $\boldsymbol{\pi}, \boldsymbol{\pi}'$ are SPE policy.*

Proof. First, we will show that

$$\forall s \in \mathcal{S}, s \in \mathcal{S}_{0:\bar{T}-1}^{s_0,\Pi^{MD}} \tag{20}$$

Since $\bar{T}_\epsilon - 1 \geq \hat{T}_{s_0}$, by definition of $\hat{T}_{s_0}$, $\forall s \in \mathcal{S}$,

$$\begin{aligned}\exists \boldsymbol{\pi} \in \Pi^{MD}, \sum_{t=0}^{\bar{T}_\epsilon - 1} \mathbb{P}[S_t^{\boldsymbol{\pi}} = s | s_0] > 0 &\Rightarrow \exists t \in [0, \bar{T}_\epsilon - 1] \text{ s.t. } \exists \boldsymbol{\pi} \in \Pi^{MD}, \mathbb{P}[S_t^{\boldsymbol{\pi}} = s | s_0] > 0 \\ &\Rightarrow \exists t \in [0, \bar{T}_\epsilon - 1] \text{ s.t. } s \in \mathcal{S}_t \\ &\Rightarrow s \in \mathcal{S}_{0:\bar{T}_\epsilon - 1}.\end{aligned}$$

Now, let us consider arbitrary $s \in \mathcal{S}$. By (20), $\exists t \in [0, \bar{T} - 1]$ s.t. $s \in \mathcal{S}_t$. Using such t, we have $\forall a \in \mathcal{A}_s$,

$$Q^{(\boldsymbol{\pi}')^{\bar{T}-t-1} \cdot \boldsymbol{\pi}}(s, \pi'(s)) \geq Q^{(\boldsymbol{\pi}')^{\bar{T}-t-1} \cdot \boldsymbol{\pi}}(s, a) \Rightarrow Q^{\boldsymbol{\pi}'}(s, \pi'(s)) \geq Q^{\boldsymbol{\pi}'}(s, a)$$

by (19) and $\boldsymbol{\pi} = \boldsymbol{\pi}'$. □

Example 5.4 (Desirability of Backward Conditioning). To illustrate the difference between (19) and (7), we reconsider the setup in Example 4.6. Given the same initialization (see Figure 3(a)), *backward conditioning* will update state $9 \in \tilde{\mathcal{S}}_3$ earlier than state $21 \in \tilde{\mathcal{S}}_0$. Thus, by the time 21 is updated, we will have accounted for state 9's *new* policy $\pi^{(1)}(9) = \leftarrow$ (see Figure 3(b)) and get the desired $\pi^{(1)}(21) = \rightarrow$ (see Figure 3(c)). This iteration ends up in $\boldsymbol{\pi}^{(1)}$ that corresponds to the target SPE path $\hat{\boldsymbol{\pi}}$ by comparing the path extending from $s0 = 21$ in Figure 3(c) with Figure 1(c). In contrast to (7), such iterative update is desirable as it imposes that the choice of *later states* are directly propagated to *earlier states* in each policy iteration. This prevents an inefficient movement away from an SPE policy, as depicted by 21 in Figures 2(a)-(c).

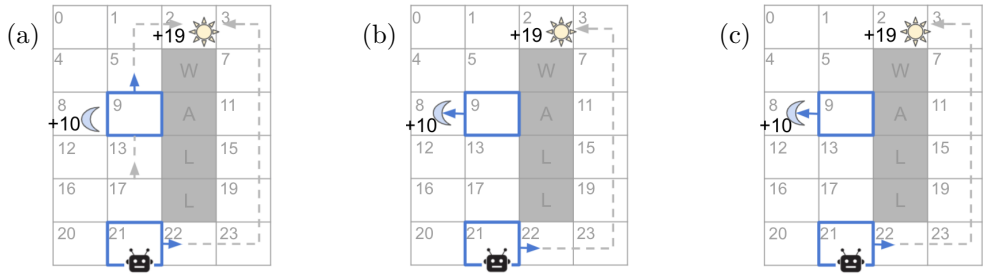

Figure 3: **2-Layered Correction with Backward Conditioning.** (a) Initialization. (b) State 9 updates first. (c) State 21 updates last, after $9, 13, 17$ make their updates.

Remark 5.5. In Theorem 25 of Lesmana & Pun (2021), a finite-horizon analog to the update rule (19) has been shown to exhibit lex-monotonicity (i.e. a weaker update monotonicity than PIT that reflects *closer to SPE*), by leveraging a policy-independent ordering on time-extended state-space due to $T < \infty$ (i.e. players are times). This prevents the cycling of policies, implying convergence. In $T = \infty$, we lose this order (i.e. *players are states*) and resort to use *visitation order* on a trajectory. This results in a lex-mono analog: $\forall t$, if $\boldsymbol{\pi}'_{S_{>t}}$ is *closer to SPE* $\hat{\boldsymbol{\pi}}_{S_{>t}}$ than $\boldsymbol{\pi}_{S_{>t}}$; so is π'_{S_t}. It was discussed through Example 5.4 when $\mathcal{S}_t = \{21\}$ and $\mathcal{S}_{>t} = \{17, 13, 9\}$. Convergence thus remains open as complications arise when $\mathcal{S}_t \cap \mathcal{S}_{t'} \neq \emptyset$ for some $t \neq t'$.

5.2 Approximate Backward Conditioning

In the previous subsection, we have presented a procedure to realize (19). Here, we are interested in replacing the the exact computation of $Q^{\boldsymbol{\pi}'}(s, a)$ with *prediction*. To this end, we will use our results in Theorem 4.3 and derive TIC-adjusted TD targets for predicting $r^{\boldsymbol{\pi}'}(s_t, a_t)$ from (10)-(11) and $Q^{\boldsymbol{\pi}'}(s_t, a_t)$ from (13)-(14),

$$\xi^r_t(m) = \begin{cases} \varphi(0) R(s_t, a_t), & m = t \\ \frac{\gamma(m-t)}{\gamma(m-(t+1))} r^{t+1,m}(S_{t+1}, \pi'(S_{t+1})), & m > t \end{cases} \tag{21}$$

$$\xi^Q_t = \begin{cases} \varphi(0) R(s_t, a_t) + Q(S_{t+1}, \pi'(S_{t+1})) \\ - \max(0, \Delta r_t), \quad t \leq T^* - 1 \\ \varphi(0) R(s_t, a_t), \quad t = T^* \text{ and } s_t \in \bar{\mathcal{S}} \end{cases} \tag{22}$$

where $\Delta r_t = \sum_{m=t+1}^{T^*} r^{t+1,m}(s_{t+1}, \pi'(s_{t+1})) - r^{t,m}(s_t, a_t)$.

The full algorithm that implements (19) with approximation is described in Algorithm 1: lines 11, 18-20 capture *backward conditioning* improvement, while lines 12-17 capture *TIC-adjusted TD* evaluation[4].

Remark 5.6. While Algorithm 1 can be considered as a Q-learning's variant, standard convergence analysis such as in Bertsekas & Tsitsiklis (1996) does not apply to our case and is subject to future study.

6 Learning Performance: An Illustration

In this section, we illustrate the behaviour of bwdQ in two TIC Gridworld environments: (i) **Deterministic (D)**, by reusing our motivating example in Figure 1, which has been shown to exhibit perference reversals, and (ii) **Stochastic (S)**, by injecting some random noise into state 9's transition in (D). For the benchmarks, we consider two approximate PI variants that also target SPE policy under general-discounting objectives, namely *standard PI with Monte Carlo (MC)* and *sophisticated EU (sophEU)* from Evans et al. (2016), for a comparative study. Pseudocodes and training specifications are provided in Appendices D.2-D.3.

[4]For detailed derivations of Algorithm 1, readers can refer to Appendix C.

Algorithm 1 Backward Q-learning (bwdQ)

Parameters: exploration rate ϵ, episode length $\bar{T}$, learning rates α_Q, α_r
Init:
$Q(s,a) = 0, \forall s \in \mathcal{S} \setminus \bar{\mathcal{S}}, a \in \mathcal{A}$;
$Q(s,a) = \varphi(0)R(s,a), \forall s \in \bar{\mathcal{S}}, a \in \bar{\mathcal{A}}$;
$r^{\tau,m}(s,a) = 0, \forall \tau, m, s \in \mathcal{S}, a \in \mathcal{A}$;
$\pi'(s) \leftarrow \arg\max_a Q(s,a), \forall s \in \mathcal{S}, \boldsymbol{\pi} \leftarrow \emptyset$
repeat
  $\boldsymbol{\pi} \leftarrow \boldsymbol{\pi}'$;
  Choose S_0 randomly;
  Sample $S_0, A_0, \ldots, S_{T^*-1}, A_{T^*-1}, S_{T^*}, A_{T^*} = \bar{a} \sim \boldsymbol{\pi}^\epsilon$;
  for $t \leftarrow T^*$ **to** 0 **do**
    for $m \leftarrow t$ **to** T^* **do**
      Compute $\xi_t^r(m)$ according to (21);
      $r^{t,m}(S_t, A_t) \leftarrow r^{t,m}(S_t, A_t) + \alpha_r(\xi_t^r(m) - r^{t,m}(S_t, A_t))$;
    end for
    Compute ξ_t^Q according to (22);
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha_Q(\xi_t^Q - Q(S_t, A_t))$;
    if $Q(S_t, \pi(S_t)) < \max_a Q(S_t, a)$ **then**
      $\pi'(S_t) \leftarrow \arg\max_a Q(S_t, a)$
    else
      $\pi'(S_t) \leftarrow \pi(S_t)$
    end if
  end for
until stable ($\boldsymbol{\pi}' \neq \boldsymbol{\pi}$)

6.1 Results

Our results and evaluation can be segregated into three components: efficiency, value prediction, and termination policy, all of which are summarized into Table 1 and Figure 4.

Efficiency In Section 5.1, we have provided an intuition on the desirability of *backward conditioning*. From Table 1, we can see its implication to actual learning instances with approximation. In particular, we can observe that bwdQ demonstrates higher learning efficiency in both (D) and (S): it has significantly shorter Δi^* in average (mean) with lower standard deviation compared to the others.

Table 1: **Delusional period $\Delta i^* \doteq |i_{21}^* - i_9^*|$ statistics, presented as mean$_{(\textbf{stdev})}$ (in thousands).** This metric links with the 2-layered correction illustrated in Figures 2 and 3: Δi^* quantifies how many iterations 21 needs to reflect 9's move to SPE. Episode indexes i_9^* and i_{21}^* represent the first overtaking episodes of mean SPE Q-value at states 9 and 21, respectively; see Appendix D.4.1 for illustrative Q-value curves. For each algorithm and environment, 10 experiments are conducted and each consists of 50 random seeds.

	MC	SophEU	BwdQ
(D)	$15.39_{(3.69)}$	$69.97_{(1.81)}$	$2.37_{(0.73)}$
(S)	$14.55_{(4.83)}$	$97.56_{(2.17)}$	$3.68_{(0.51)}$

Value prediction From Figure 4(a), we can observe that in (D), the mean value of bwdQ matches closely the groundtruth (manually computed) upon convergence. On the contrary, sophEU and MC both converge at a value strictly smaller than the groundtruth. Similar conclusion can be drawn in (S), despite bwdQ produces higher variance than the rest; see Appendices D.4.2-D.4.3 for more results and discussions on value biases.

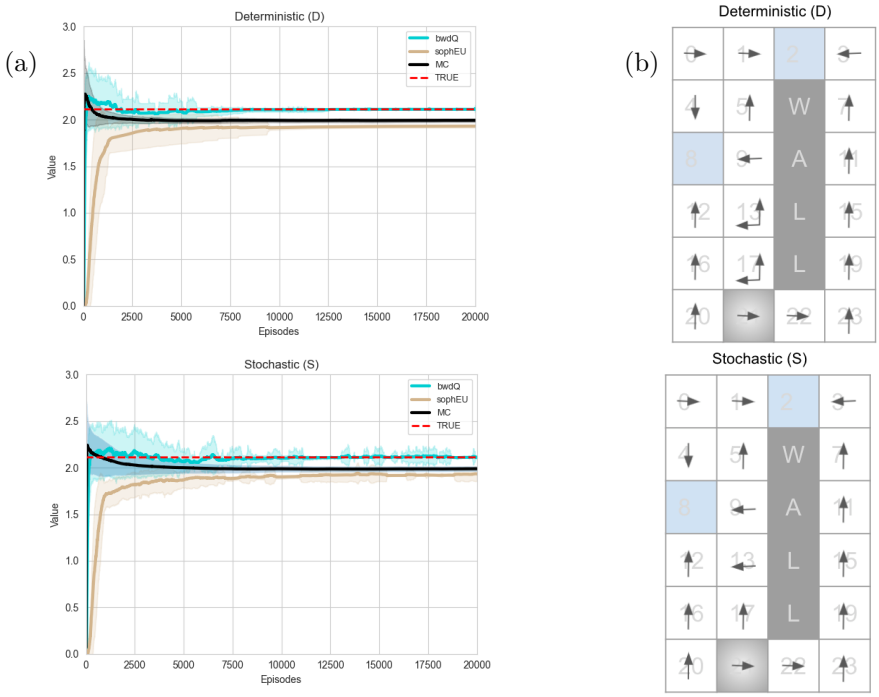

Figure 4: **(a) Value learning curves of** $s_0 = 21$.Groundtruth 'TRUE' is computed as the value of analytical SPE policy in (D) and the trained terminal policy in (S). **(b) Termination policies**.

Termination policy In both (D) and (S), *all algorithms* (i.e. MC, sophEU, and bwdQ) converge and the termination policies are plotted in Figure 4(b). While all algorithms converge to the same policies in (S), this is not true in (D): at $s = 13, 17$, MC and sophEU converge to $\{\uparrow, \uparrow\}$ when bwdQ converges to $\{\uparrow, \leftarrow\}$ or $\{\leftarrow, \leftarrow\}$. Thus, in Figure 4(b), we present together these three different termination policies. For the termination policies in (D), we can verify that they correspond to the groundtruth SPE policies (by Definition 3.1 and $Q^{\hat{\boldsymbol{\pi}}}$ computed manually from the reward specifications in Figure 1). This is consistent with our results in Propositions 4.1 (MC) and 5.3 (bwdQ) that guarantee SPE optimality if converged. For the termination policies in (S), we can see how the noise injected to 9 affects the SPE policy: $\hat{\pi}(13)$ shifts from $\{\leftarrow, \uparrow\}$ in (D) to $\{\leftarrow\}$ in (S) as $Q^{\hat{\boldsymbol{\pi}}}(13, \uparrow)$ in (S) is pulled down by random transitions of $9 \to 5$ and $9 \to 13$.

7 Conclusion and Future Works

Before this paper, it was unclear how PI will perform and whether it is sufficient in TIC settings, where BPO or DP becomes infeasible. This paper on TIC RL is of theoretical nature, while we managed to use a toy Gridworld example to demonstrate our claims. Specifically, we demonstrated how introducing SPE optimality can shed lights on the two fundamental questions surrounding the use of PI in TIC RL setting. In particular, we obtain positive results on PI (both standard PI and backward conditioning)'s capability to characterize SPE policies. Though we could not close the convergence of either standard PI or backward conditioning, we made progress towards it: on the importance of ordered policy iteration and improvement criteria.

From the perspective of policy evaluation, SPE optimality recovers the use of DP-like formulas resulting in familiar forms of algorithms, which is also important towards closing the analysis of SPE policy search. Formal convergence analyses are thus important future research directions. Another interesting future research is to extend the results of this paper to other TIC sources in different environments. For instance, whether the demonstrated behaviour of PI under TIC is generalizable and what SPE policy entails in these other settings. We should also anticipate more experiments on TIC RL are conducted in the future after our first attempt.

References

Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition, 1957.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-dynamic programming.*, volume 3 of *Optimization and neural computation series*. Athena Scientific, 1996. ISBN 1886529108.

Tomas Björk and Agatha Murgoci. A theory of Markovian time-inconsistent stochastic control in discrete time. *Finance and Stochastics*, 18(3):545–592, June 2014. doi: 10.1007/s00780-014-0234-y.

Tomas Björk, Agatha Murgoci, and Xun Yu Zhou. Mean-variance portfolio optimization with state-dependent risk aversion. *Mathematical Finance*, 24(1):1–24, January 2014. doi: 10.1111/j.1467-9965.2011.00515.x.

Dotan Di Castro, Aviv Tamar, and Shie Mannor. Policy gradients with variance related risk criteria. *arXiv preprint arXiv:1206.6404*, 2012.

Owain Evans, Andreas Stuhlmüller, and Noah D. Goodman. Learning the preferences of ignorant, inconsistent agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30 of *AAAI'16*, pp. 323–329, Phoenix, Arizona, February 2016. URL `https://dl.acm.org/doi/10.5555/3015812.3015860`.

William Fedus, Carles Gelada, Yoshua Bengio, Marc G. Bellemare, and Hugo Larochelle. Hyperbolic discounting and learning over multiple horizons. *arXiv: 1902.06865*, February 2019. URL `https://arxiv.org/abs/1902.06865`.

Ronald A Howard. Dynamic programming and markov processes. 1960.

Zeb Kurth-Nelson and A. David Redish. A reinforcement learning model of precommitment in decision making. *Frontiers in Behavioral Neuroscience*, 4:184, 2010. ISSN 1662-5153. doi: 10.3389/fnbeh.2010.00184. URL `https://www.frontiersin.org/article/10.3389/fnbeh.2010.00184`.

Tor Lattimore and Marcus Hutter. General time consistent discounting. *Theoretical Computer Science*, 519: 140–154, 2014.

Nixie S Lesmana and Chi Seng Pun. A subgame perfect equilibrium reinforcement learning approach to time-inconsistent problems. *Available at SSRN 3951936*, 2021.

Shie Mannor and John N. Tsitsiklis. Mean-variance optimization in markov decision processes. *CoRR*, abs/1104.5601, 2011. URL `http://arxiv.org/abs/1104.5601`.

Robert A. Pollak. Consistent planning. *The Review of Economic Studies*, 35(2):201, April 1968. doi: 10.2307/2296548.

LA Prashanth and Mohammad Ghavamzadeh. Actor-critic algorithms for risk-sensitive mdps. In *Advances in neural information processing systems*, pp. 252–260, 2013.

Matthew J Sobel. The variance of discounted Markov decision processes. *Journal of Applied Probability*, pp. 794–802, 1982.

Robert H. Strotz. Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies*, 23(3):165–180, December 1955. doi: 10.2307/2295722.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Aviv Tamar and Shie Mannor. Variance adjusted actor critic algorithms. *ArXiv*, abs/1310.3697, 2013.

Christopher J.C.H. Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, May 1992.

A Additional Details on Assumption 3.2-3.4

A.1 MDP Examples under General-Discounting Criterion

In this section, we derive several sufficient conditions for our assumptions in Section 3.2, in the context of general-discounting criterion.

Definition A.1 (*Boundary-only* Rewards)**.** The reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$ is *boundary-only* if it is non-zero only at *boundary* states, i.e. $R(s, a) > 0$ only if $s \in \bar{\mathcal{S}}$.

Lemma A.2. *Any MDP that has boundary-only rewards satisfies Assumption 3.2.*

Proof. Let $T_{\bar{S}}^{\boldsymbol{\pi}}|s$ defines the *minimum* hitting time of any *boundary* states $\bar{s} \in \bar{\mathcal{S}}$ when initiated at s and following $\boldsymbol{\pi}$. Thus, $\forall s \in \mathcal{S}, \forall \boldsymbol{\pi} \in \Pi^{MD}, \forall \bar{T} < \infty$,

$$V^{\boldsymbol{\pi}}(s) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \varphi(t) R_t^{\boldsymbol{\pi}} \mid s\right] = \sum_{\tau=0}^{\bar{T}} \mathbb{P}[T_{\bar{S}}^{\boldsymbol{\pi}} = \tau | s] \mathbb{E}\left[\sum_{t=0}^{\tau} \varphi(t) R_t^{\boldsymbol{\pi}} | s\right] + \sum_{\tau=\bar{T}+1}^{\infty} \mathbb{P}[T_{\bar{S}}^{\boldsymbol{\pi}} = \tau | s] \mathbb{E}\left[\sum_{t=0}^{\tau} \varphi(t) R_t^{\boldsymbol{\pi}} | s\right],$$

$$\begin{aligned}
V^{\boldsymbol{\pi}^{\bar{T}} \cdot \ddot{\boldsymbol{\pi}}}(s) &\coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \varphi(t) R_t^{\boldsymbol{\pi}^{\bar{T}} \cdot \ddot{\boldsymbol{\pi}}} \mid s\right] \\
&= \sum_{\tau=0}^{\bar{T}} \mathbb{P}[T_{\bar{S}}^{\boldsymbol{\pi}^{\bar{T}} \cdot \ddot{\boldsymbol{\pi}}} = \tau | s] \mathbb{E}\left[\sum_{t=0}^{\tau} \varphi(t) R_t^{\boldsymbol{\pi}^{\bar{T}} \cdot \ddot{\boldsymbol{\pi}}} | s\right] + \sum_{\tau=\bar{T}+1}^{\infty} \mathbb{P}[T_{\bar{S}}^{\boldsymbol{\pi}^{\bar{T}} \cdot \ddot{\boldsymbol{\pi}}} = \tau | s] \mathbb{E}\left[\sum_{t=0}^{\tau} \varphi(t) R_t^{\boldsymbol{\pi}^{\bar{T}} \cdot \ddot{\boldsymbol{\pi}}} | s\right] \\
&= \sum_{\tau=0}^{\bar{T}} \mathbb{P}[T_{\bar{S}}^{\boldsymbol{\pi}} = \tau | s] \mathbb{E}\left[\sum_{t=0}^{\tau} \varphi(t) R_t^{\boldsymbol{\pi}} | s\right]. \qquad (\text{since } \forall \tau \geq \bar{T}+1, S_\tau = \bar{s}_{\text{void}} \Rightarrow \mathbb{P}[T_{\bar{S}}^{\boldsymbol{\pi}^{\bar{T}} \cdot \ddot{\boldsymbol{\pi}}} = \tau | s] = 0.)
\end{aligned}$$

By bounded reward function,

$$R_{max} \coloneqq \max\{||R(s, a)|| : s \in \mathcal{S}, a \in \mathcal{A}\} \tag{23}$$

exists. Then, $\forall \epsilon > 0, \forall s \in \mathcal{S}$, we can set $\bar{T} < \infty$ s.t.

$$|R_{\max}| \varphi(\bar{T} + 1) \leq \epsilon \tag{24}$$

and the following holds,

$$\begin{aligned}
\sup_{\boldsymbol{\pi} \in \Pi^{MD}} ||V^{\boldsymbol{\pi}^{\bar{T}} \cdot \boldsymbol{\pi}}(s) - V^{\boldsymbol{\pi}^{\bar{T}} \cdot \ddot{\boldsymbol{\pi}}}(s)|| &= \sup_{\boldsymbol{\pi} \in \Pi^{MD}} \left|\left| \sum_{\tau=\bar{T}+1}^{\infty} \mathbb{P}[T_{\bar{S}}^{\boldsymbol{\pi}} = \tau | s] \mathbb{E}\left[\varphi(\tau) R_\tau^{\boldsymbol{\pi}} | s\right] \right|\right| && \text{(by } \textit{boundary-only} \text{ rewards)} \\
&\leq \sup_{\boldsymbol{\pi} \in \Pi^{MD}} |R_{\max}| \sum_{\tau=\bar{T}+1}^{\infty} \mathbb{P}[T_{\bar{S}}^{\boldsymbol{\pi}} = \tau | s] \varphi(\tau) && \text{(by (23))} \\
&\leq \sup_{\boldsymbol{\pi} \in \Pi^{MD}} |R_{\max}| \varphi(\bar{T}+1) \mathbb{P}[T_{\bar{S}}^{\boldsymbol{\pi}} > \bar{T} | s] && \text{(by } \varphi(\cdot) \text{ decreasing)} \\
&\leq |R_{\max}| \varphi(\bar{T}+1) \leq \epsilon && \text{(by (24))}
\end{aligned}$$

□

Lemma A.3. *Suppose an MDP has boundary-only rewards, s_0 that satisfies Assumption 3.3 such that $\bar{T}_{s_0,0} < \infty$, and a discounting factor $\varphi(\cdot)$ that satisfies*

$$\forall t \geq 0, \frac{\varphi(\tau + t)}{\varphi(\tau)} \text{ is increasing in } \tau, \tau \geq 1 \tag{25}$$

with $\varphi(1) = 1$. Then, Assumption 3.4 holds.

Proof. Suppose otherwise, $\exists \epsilon^* > 0, t^* \in [0, \bar{T}_{\epsilon^*}], s^* \in \mathcal{S}_{t^*}^{s_0,\Pi^{MD}}, \boldsymbol{\pi}_* \in \Pi^{MD}$ s.t.

$$\mathbb{P}[S_{t^*}^{\boldsymbol{\pi}_*} = s^* | s_0] > 0 \tag{26}$$

and $\forall \kappa > 0$,

$$\frac{\epsilon^*}{\kappa} < \|V^{\boldsymbol{\pi}_*}(s^*) - V^{\boldsymbol{\pi}_*^{\bar{T}_{\epsilon^*}-t^*} \cdot \ddot{\boldsymbol{\pi}}}(s_{t^*})\| = \sum_{\tau=\bar{T}_{\epsilon^*}-t^*+1}^{\infty} \mathbb{P}[T_{\bar{\mathcal{S}}}^{\boldsymbol{\pi}_*} = \tau | s^*] \mathbb{E}[\varphi(\tau) R_\tau^{\boldsymbol{\pi}_*} | s^*] \tag{27}$$

Let us fix $\boldsymbol{\pi} \coloneqq \boldsymbol{\pi}_*$ and set

$$\kappa^* \coloneqq \mathbb{P}[S_{t^*}^{\boldsymbol{\pi}_*} = s^* | s_0] \varphi(\bar{T}_{\epsilon^*} + 1) \tag{28}$$

Note that $\kappa^* > 0$ by (26) and $\bar{T}_{\epsilon^*} < \infty$ (by Assumption 3.2, $\bar{T}_{s_0,\epsilon^*} < \infty$, and by Assumption 3.3, $\hat{T}_{s_0} < \infty$). Then,

$$\begin{aligned}
\left\| V^{\boldsymbol{\pi}_*}(s_0) - V^{\boldsymbol{\pi}_*^{\bar{T}_{s_0,\epsilon^*}} \cdot \ddot{\boldsymbol{\pi}}}(s_0) \right\| &= \sum_{\tau=\bar{T}_{s_0,\epsilon^*}+1} \mathbb{P}[T_{\bar{\mathcal{S}}}^{\boldsymbol{\pi}_*} = \tau | s_0] \mathbb{E}[\varphi(\tau) R_\tau^{\boldsymbol{\pi}_*} | s_0] \\
&\qquad \text{(by boundary-only rewards; see Lemma A.2's proof)} \\
&\geq \sum_{\tau=\bar{T}_{\epsilon^*}+1} \mathbb{P}[T_{\bar{\mathcal{S}}}^{\boldsymbol{\pi}_*} = \tau | s_0] \mathbb{E}[\varphi(\tau) R_\tau^{\boldsymbol{\pi}_*} | s_0] \qquad (\text{by } \bar{T}_{s_0,\epsilon^*} \leq \bar{T}_{\epsilon^*}) \\
&= \sum_{\tau=\bar{T}_{\epsilon^*}+1} \sum_{s \in \mathcal{S}} \mathbb{P}[S_{t^*}^{\boldsymbol{\pi}_*} = s | s_0] \mathbb{P}[T_{\bar{\mathcal{S}}}^{\boldsymbol{\pi}_*} = \tau - t^* | s] \mathbb{E}[\varphi(\tau) R_{\tau-t^*}^{\boldsymbol{\pi}_*} | s] \\
&\geq \mathbb{P}[S_{t^*}^{\boldsymbol{\pi}_*} = s^* | s_0] \sum_{\tau=\bar{T}_{\epsilon^*}-t^*+1} \mathbb{P}[T_{\bar{\mathcal{S}}}^{\boldsymbol{\pi}_*} = \tau | s^*] \mathbb{E}[\varphi(\tau + t^*) R_\tau^{\boldsymbol{\pi}_*} | s^*] \\
&\qquad \text{(by non-negative rewards and probabilities)} \\
&\geq \mathbb{P}[S_{t^*}^{\boldsymbol{\pi}_*} = s^* | s_0] \varphi(\bar{T}_{\epsilon^*} + 1) \sum_{\tau=\bar{T}_{\epsilon^*}-t^*+1} \mathbb{P}[T_{\bar{\mathcal{S}}}^{\boldsymbol{\pi}_*} = \tau | s^*] \mathbb{E}[\varphi(\tau) R_\tau^{\boldsymbol{\pi}_*} | s^*] \\
&\qquad \text{(by (25) and } t^* \in [0, \bar{T}_{\epsilon^*}]) \\
&> \mathbb{P}[S_{t^*}^{\boldsymbol{\pi}_*} = s^* | s_0] \varphi(\bar{T}_{\epsilon^*} + 1) \frac{\epsilon^*}{\kappa^*} = \epsilon^* \qquad (\text{by (27) and (28)})
\end{aligned}$$

This contradicts definition of $\bar{T}_{s_0,\epsilon^*}$ (see Assumption 3.2), implying that our supposition is false.

With $\kappa(\epsilon, t, s_t, \boldsymbol{\pi}; s_0) \coloneqq \mathbb{P}[S_t^{\boldsymbol{\pi}} = s_t | s_0] \varphi(\bar{T}_\epsilon + 1)$, we will now show that

$$\lim_{\epsilon \to 0} \frac{\epsilon}{\mathbb{P}[S_t^{\boldsymbol{\pi}} = s_t | s_0] \varphi(\bar{T}_\epsilon + 1)} = 0 \tag{29}$$

For any fixed $\epsilon > 0$, let us define

$$G(\bar{T}_\epsilon; s_0) \coloneqq \min\{\mathbb{P}[S_t^{\boldsymbol{\pi}} = s_t | s_0] > 0 : t \in [0, \bar{T}_\epsilon], s_t \in \mathcal{S}, \boldsymbol{\pi} \in \Pi^{MD}\}. \tag{30}$$

Then, $\forall \boldsymbol{\pi} \in \Pi^{MD}, \forall t \in [0, \bar{T}_\epsilon], \forall s_t \in \mathcal{S}_t^{s_0,\Pi^{MD}}$,

$$0 \leq \frac{\epsilon}{\mathbb{P}[S_t^{\boldsymbol{\pi}} = s_t | s_0] \varphi(\bar{T}_\epsilon + 1)} \leq \frac{\epsilon}{G(\bar{T}_\epsilon; s_0) \varphi(\bar{T}_\epsilon + 1)} \tag{31}$$

Since $\bar{T}_{s_0,0} < \infty$, we have

$$\bar{T}_0 \doteq \max\{\hat{T}_{s_0}, \bar{T}_{s_0,0}\} < \infty. \tag{32}$$

Let us fix arbitrarily $\boldsymbol{\pi} \in \Pi^{MD}, t \in [0, \bar{T}_0], s_t \in \mathcal{S}_t^{s_0,\Pi^{MD}}$. By (32), $\lim_{\epsilon\to 0} G(\bar{T}_\epsilon; s_0) = G(\bar{T}_0; s_0) > 0$ and $\lim_{\epsilon\to 0} \varphi(\bar{T}_\epsilon + 1) = \varphi(\bar{T}_0 + 1) > 0$. Thus, we can take $\lim_{\epsilon\to 0}$ on the upper and lower bound in (31) and have shown

$$\lim_{\epsilon\to 0} \frac{\epsilon}{\mathbb{P}[S_t^{\boldsymbol{\pi}} = s_t|s_0]\varphi(\bar{T}_\epsilon + 1)} = 0$$

□

Finally, it's straightforward to verify that our hyperbolic Gridworld in Figure 1(a) has *boundary-only* rewards and $s_0 = 21$ that satisfies Assumption 3.3. Moreover, due to the existence of $\tau^* \coloneqq \max\{T_{\bar{\mathcal{S}}}^{\boldsymbol{\pi}} < \infty : \boldsymbol{\pi} \in \Pi^{MD}\} < \infty$ by its deterministic transition and $|\Pi^{MD}| < \infty$, we have $\forall \bar{T} \geq \tau^*$,

$$\begin{aligned}\sup_{\boldsymbol{\pi}\in\Pi^{MD}} ||V^{\boldsymbol{\pi}}(s_0) - V^{\boldsymbol{\pi}^{\bar{T}}\cdot\bar{\boldsymbol{\pi}}}(s_0)|| &= \sup_{\boldsymbol{\pi}\in\Pi^{MD}: T_{\bar{\mathcal{S}}}^{\boldsymbol{\pi}}<\infty} ||V^{\boldsymbol{\pi}}(s_0) - V^{\boldsymbol{\pi}^{\bar{T}}\cdot\bar{\boldsymbol{\pi}}}(s_0)|| \qquad \text{(by boundary-only rewards)} \\ &= 0 \qquad (\text{ by } \forall \tau > \bar{T} \geq \tau^*, \forall \boldsymbol{\pi} \in \Pi^{MD} \text{ with } \bar{T}_{\bar{\mathcal{S}}}^{\boldsymbol{\pi}} < \infty, \mathbb{P}[T_{\bar{\mathcal{S}}}^{\boldsymbol{\pi}} = \tau|s_0] = 0)\end{aligned}$$

and thus, $\bar{T}_{s_0,0} \leq \tau^* < \infty$. For a more concrete example, we can refer to the *restricted* Hyperbolic Gridworld in Example 4.6, where we can compute manually $\hat{T}_{s_0} = 7$ and $\bar{T}_{s_0,0} = 8$.

A.2 Implied Bounded Value Functions

Through the following lemma, we can link Assumption 3.2 to the standard well-posedness condition of bounded value functions that ensures the existence of optimal policy.

Lemma A.4. *If Assumption 3.2 holds, then* $\forall s \in \mathcal{S}, \forall \boldsymbol{\pi} \in \Pi^{MD}, V^{\boldsymbol{\pi}}(s) < \infty$.

Proof. Suppose $\exists \boldsymbol{\pi}_*, s_*$ s.t. $V^{\boldsymbol{\pi}_*}(s_*) = \infty$. Then, we can set $s, \boldsymbol{\pi} \leftarrow s_*, \boldsymbol{\pi}_*$ and arbitrary $\epsilon^* > 0$ s.t. $\forall \bar{T} < \infty$, $||V^{\boldsymbol{\pi}_*^{\bar{T}}\cdot\tilde{\boldsymbol{\pi}}}(s_*) - V^{\boldsymbol{\pi}_*}(s_*)|| > \epsilon^*$ since $V^{\boldsymbol{\pi}_*^{\bar{T}}\cdot\tilde{\boldsymbol{\pi}}}(s_*) < \infty$. □

B Theorem 4.3

B.1 Technical Assumptions

Assumption B.1 ("Relevant at t under $\boldsymbol{\pi}$")**.** If $t = 0$,

$$\mathbb{P}[S_t^{\boldsymbol{\pi}} = s_0|s_0] = 0, \forall t \geq 1 \wedge \mathbb{P}[S_1 = s_0|S_0 = s_0, A_0 = a_0] = 0 \tag{33}$$

If $t > 0$, $\exists (s_{t-1}, a_{t-1})$ "relevant at $t-1$ under $\boldsymbol{\pi}$" s.t.

$$\mathbb{P}[S_t = s_t|S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}] > 0 \wedge a_t = \pi(s_t). \tag{34}$$

Intuitively, for $t > 0$, (34) exhausts the use instances of (s_t, a_t) in PE updates $Q^{\boldsymbol{\pi}}(s_{t-1}, a_{t-1}) \leftarrow \mathbb{E}[Q^{\boldsymbol{\pi}}(s_t, \pi(s_t))] + \dots$ and thus, it must hold. Whereas for $t = 0$, some restrictions on the MDP (e.g., $\forall t' \neq t, \mathcal{S}_t^{s_0,\Pi^{MD}} \cap \mathcal{S}_{t'}^{s_0,\Pi^{MD}} = \emptyset$ as in Example 4.6) can be imposed to ensure that (33) holds $\forall \boldsymbol{\pi} \in \Pi^{MD}, s_0 \in \mathcal{S} \setminus \{\bar{\mathcal{S}}\}, a_0 \in \mathcal{A}_{s_0}$. Note however that in actual use instances, (33) only need to hold for the $\boldsymbol{\pi}$ encountered in the PI updates (instead of $\forall \boldsymbol{\pi} \in \Pi^{MD}$). This may relax the need for such MDP restrictions: as we can observe from our experiments (see Section 6), our algorithm still performs plausibly well even when $\forall t' \neq t, \mathcal{S}_t^{s_0,\Pi^{MD}} \cap \mathcal{S}_{t'}^{s_0,\Pi^{MD}} = \emptyset$ does not hold. In what follows, we present several intermediate results that link the conditions in Assumption B.1 to the "approximation" (14) in Theorem 4.3.

Lemma B.2. *At any $t \geq 0$, if (s_t, a_t) is "relevant at t under $\boldsymbol{\pi}$", then $\exists s_0, a_0$ "relevant at 0 under $\boldsymbol{\pi}$" s.t.*

$$\mathbb{P}[S_t^{\boldsymbol{\pi}_{s_0,a_0}} = s_t|s_0] > 0 \wedge a_t = \pi_{s_0,a_0}(s_t)$$

with $\boldsymbol{\pi}_{s_0,a_0}$ *defined as follows*

$$\pi_{s_0,a_0}(s) = \begin{cases} a_0, & \text{if } s = s_0 \\ \pi(s), & \text{otherwise} \end{cases} \tag{35}$$

Proof. (Base case: $t = 0$.) Note that for any (s_0, a_0) that is "relevant at 0 under $\boldsymbol{\pi}$", we have $\mathbb{P}[S_0^{\boldsymbol{\pi}_{s_0,a_0}}|s_0] = 1 > 0$. Moreover, $a_0 = \pi_{s_0,a_0}(s_0)$ holds by definition in (35).

($t > 0$.) Proof by induction. Suppose that the relation holds at $t = t' - 1$, we will show that it also holds at $t = t'$. By $(s_{t'}, a_{t'})$'s "relevance at t' under $\boldsymbol{\pi}$", $\exists(s_{t'-1}, a_{t'-1})$ "relevant at $t' - 1$ under $\boldsymbol{\pi}$" s.t.

$$\mathbb{P}[S_{t'} = s_{t'}|S_{t'-1} = s_{t'-1}, A_{t'-1} = a_{t'-1}] > 0 \wedge a_{t'} = \pi(s_{t'}) \tag{36}$$

Moreover, by assumption (that at $t = t' - 1$ the relation holds), the above $(s_{t'-1}, a_{t'-1})$ satisfies: $\exists s_0, a_0$ "relevant at 0 under $\boldsymbol{\pi}$ s.t.

$$\mathbb{P}[S_{t'-1}^{\boldsymbol{\pi}_{s_0,a_0}} = s_{t'-1}|s_0] > 0 \wedge a_{t'-1} = \pi_{s_0,a_0}(s_{t'-1}). \tag{37}$$

Therefore,

$$\begin{aligned} \mathbb{P}[S_{t'}^{\boldsymbol{\pi}_{s_0,a0}} = s_{t'}|s_0] &\geq \mathbb{P}[S_{t'}^{\boldsymbol{\pi}_{s_0,a0}} = s_{t'}|S_{t'-1} = s_{t'-1}]\mathbb{P}[S_{t'-1}^{\boldsymbol{\pi}_{s_0,a0}} = s_{t'-1}|s_0] \\ &= \mathbb{P}[S_{t'} = s_{t'}|S_{t'-1} = s_{t'-1}, A_{t'-1} = a_{t'-1}]\mathbb{P}[S_{t'-1}^{\boldsymbol{\pi}_{s_0,a0}} = s_{t'-1}|s_0] > 0 \qquad \text{(by (37))} \end{aligned}$$

Moreover, by (s_0, a_0)'s "relevance at 0 under $\boldsymbol{\pi}$" and $t' > 0$, we must have $s_{t'} \neq s_0$ which then implies

$$a_{t'} = \pi_{s_0,a_0}(s_{t'}) \qquad \text{(by (36))}$$

□

Lemma B.3. *For any* $\boldsymbol{\pi} \in \Pi^{MD}$, $t \geq 0$, *and* (s_t, a_t) *"relevant at* t *under* $\boldsymbol{\pi}$*", $\exists \kappa > 0$ s.t.*

$$\forall s_{t+1} \sim p_{s_t}^{a_t}, \left|\left|V^{\boldsymbol{\pi}}(s_{t+1}) - V^{\boldsymbol{\pi}^{\bar{T}_\epsilon-(t+1)}\cdot\ddot{\boldsymbol{\pi}}}(s_{t+1})\right|\right| \leq \frac{\epsilon}{\kappa} \tag{38}$$

Proof. Let us first fix arbitrarily $(s_t, a_t, \boldsymbol{\pi})$. By Lemma B.2, $\exists s_0, a_0$ and $\tilde{\boldsymbol{\pi}} \coloneqq \boldsymbol{\pi}_{s_0,a_0}$ s.t.

$$\mathbb{P}[S_t^{\tilde{\boldsymbol{\pi}}} = s_t|s_0] > 0 \wedge a_t = \tilde{\pi}(s_t) \tag{39}$$

Next, for any arbitrary choice of $s_{t+1} \sim p_{s_t}^{a_t}$, we have

$$\mathbb{P}[S_{t+1} = s_{t+1}|S_t = s_t, A_t = a_t] > 0 \tag{40}$$

Therefore,

$$\begin{aligned} \mathbb{P}[S_{t+1}^{\tilde{\boldsymbol{\pi}}} = s_{t+1}|s_0] &\geq \mathbb{P}[S_{t+1}^{\tilde{\boldsymbol{\pi}}} = s_{t+1}|S_t = s_t]\mathbb{P}[S_t^{\tilde{\boldsymbol{\pi}}} = s_t|s_0] \\ &= \mathbb{P}[S_{t+1} = s_{t+1}|S_t = s_t, A_t = a_t]\mathbb{P}[S_t^{\tilde{\boldsymbol{\pi}}} = s_t|s_0] \qquad \text{(by (39))} \\ &> 0 \qquad \text{(by (39) and (40))} \end{aligned}$$

which by Assumption 3.4, implies

$$\begin{aligned} \frac{\epsilon}{\kappa(\epsilon, t+1, s_{t+1}, \tilde{\boldsymbol{\pi}}; s_0)} &\geq \left|\left|V^{\tilde{\boldsymbol{\pi}}}(s_{t+1}) - V^{\tilde{\boldsymbol{\pi}}^{\bar{T}_\epsilon-(t+1)}\cdot\ddot{\boldsymbol{\pi}}}(s_{t+1})\right|\right| \\ &= \left|\left|V^{\boldsymbol{\pi}}(s_{t+1}) - V^{\boldsymbol{\pi}^{\bar{T}-(t+1)}\cdot\ddot{\boldsymbol{\pi}}}(s_{t+1})\right|\right| \\ &\qquad \text{(by } (s_0, a_0)\text{'s "relevance at 0 under } \boldsymbol{\pi}\text{" and } t+1 > 0,\ s_{t+1} \neq s_0.) \end{aligned}$$

Finally, we can set $\kappa \coloneqq \min\{\kappa(\epsilon, t+1, s_{t+1}, \tilde{\boldsymbol{\pi}}; s_0) : s_{t+1} \sim p_{s_t}^{a_t}\}$ and (38) directly holds. □

B.2 Proof of Theorem 4.3

For any $m \geq \tau + 1$, we can derive r-function recursion as follows

$$
\begin{aligned}
r^{\boldsymbol{\pi},\tau,m}(s,a) &\doteq \mathbb{E}_s\left[\varphi(m-\tau)R\left(S_m^{a\cdot\boldsymbol{\pi}}, \pi(S_m^{a\cdot\boldsymbol{\pi}})\right)\right] && (41)\\
&= \mathbb{E}_{S'\sim p_s^a}\left[\mathbb{E}_{S'}\left[\varphi(m-\tau)R\left(S_m^{\boldsymbol{\pi}}, \pi(S_m^{\boldsymbol{\pi}})\right)\right]\right] && (42)\\
&= \mathbb{E}_{S'\sim p_s^a}\left[\frac{\varphi(m-\tau)}{\varphi(m-(\tau+1))}\mathbb{E}_{S'}\left[\varphi(m-(\tau+1))R\left(S_m^{\boldsymbol{\pi}}, \pi(S_m^{\boldsymbol{\pi}})\right)\right]\right] && (43)\\
&= \mathbb{E}_{S'\sim p_s^a}\left[\frac{\varphi(m-\tau)}{\varphi(m-(\tau+1))}r^{\boldsymbol{\pi},\tau+1,m}(S', \pi(S')\right] && (44)
\end{aligned}
$$

For $m = \tau$, by Definition 4.2, we have

$$
\begin{aligned}
r^{\boldsymbol{\pi},\tau,\tau}(s,a) &\doteq \mathbb{E}_s\left[\varphi(m-\tau)R\left(S_m^{a\cdot\boldsymbol{\pi}}, \boldsymbol{\pi}(S_m^{a\cdot\boldsymbol{\pi}})\right)\right] && (45)\\
&= \mathbb{E}_{R\sim p_s^a}[\varphi(0)R(s,a)] && (46)
\end{aligned}
$$

Next, we derive Q-function recursion,

$$
\begin{aligned}
Q^{\boldsymbol{\pi}}(s_t,a_t) &\doteq \mathbb{E}_{s_t}\left[\varphi(0)R(s_t,a_t) + \varphi(1)R(S_{t+1}^{a_t\cdot\boldsymbol{\pi}}, \boldsymbol{\pi}(S_{t+1}^{a_t\cdot\boldsymbol{\pi}})) + \dots\right] && (47a)\\
&= \mathbb{E}_{R\sim p_{s_t}^{a_t}}[\varphi(0)R(s_t,a_t)] + \mathbb{E}_{s_t}\left[\varphi(0)R(S_{t+1}^{a_t\cdot\boldsymbol{\pi}}, \boldsymbol{\pi}(S_{t+1}^{a_t\cdot\boldsymbol{\pi}})) + \varphi(1)R(S_{t+2}^{a_t\cdot\boldsymbol{\pi}}, \boldsymbol{\pi}(S_{t+2}^{a_t\cdot\boldsymbol{\pi}})) + \dots\right]\\
&\quad - \left\{\mathbb{E}_{s_t}\left[\varphi(0)R(S_{t+1}^{a_t\cdot\boldsymbol{\pi}}, \boldsymbol{\pi}(S_{t+1}^{a_t\cdot\boldsymbol{\pi}})) + \varphi(1)R(S_{t+2}^{a_t\cdot\boldsymbol{\pi}}, \boldsymbol{\pi}(S_{t+2}^{a_t\cdot\boldsymbol{\pi}})) + \dots\right]\right.\\
&\quad \left. - \mathbb{E}_{s_t}\left[\varphi(1)R(S_{t+1}^{a_t\cdot\boldsymbol{\pi}}, \boldsymbol{\pi}(S_{t+1}^{a_t\cdot\boldsymbol{\pi}})) + \varphi(2)R(S_{t+2}^{a_t\cdot\boldsymbol{\pi}}, \boldsymbol{\pi}(S_{t+2}^{a_t\cdot\boldsymbol{\pi}})) + \dots\right]\right\} && (47b)\\
&= \mathbb{E}_{R\sim p_{s_t}^{a_t}}[\varphi(0)R(s_t,a_t)] + \mathbb{E}_{S_{t+1}\sim p_{s_t}^{a_t}}\left[Q^{\boldsymbol{\pi}}(S_{t+1}, \boldsymbol{\pi}(S_{t+1}))\right]\\
&\quad - \left\{\mathbb{E}_{S_{t+1}\sim p_{s_t}^{a_t}}\left[\mathbb{E}_{S_{t+1}}\left[\varphi(0)R(S_{t+1}^{\boldsymbol{\pi}}, \boldsymbol{\pi}(S_{t+1}^{\boldsymbol{\pi}})) + \varphi(1)R(S_{t+2}^{\boldsymbol{\pi}}, \boldsymbol{\pi}(S_{t+2}^{\boldsymbol{\pi}})) + \dots\right]\right]\right.\\
&\quad \left. - \mathbb{E}_{S_{t+1}\sim p_{s_t}^{a_t}}\left[\mathbb{E}_{S_{t+1}}\left[\varphi(1)R(S_{t+1}^{\boldsymbol{\pi}}, \boldsymbol{\pi}(S_{t+1}^{\boldsymbol{\pi}})) + \varphi(2)R(S_{t+2}^{\boldsymbol{\pi}}, \boldsymbol{\pi}(S_{t+2}^{\boldsymbol{\pi}})) + \dots\right]\right]\right\} && (47c)\\
&= \mathbb{E}_{R\sim p_{s_t}^{a_t}}[\varphi(0)R(s_t,a_t)] + \mathbb{E}_{S_{t+1}\sim p_{s_t}^{a_t}}\left[Q^{\boldsymbol{\pi}}(S_{t+1}, \boldsymbol{\pi}(S_{t+1}))\right]\\
&\quad - \left\{\mathbb{E}_{S_{t+1}\sim p_{s_t}^{a_t}}\left[\mathbb{E}_{S_{t+1}}\left[\sum_{m=t+1}^{\infty}\varphi(m-(t+1))R(S_m^{\boldsymbol{\pi}}, \boldsymbol{\pi}(S_m^{\boldsymbol{\pi}}))\right]\right]\right.\\
&\quad \left. - \mathbb{E}_{S_{t+1}\sim p_{s_t}^{a_t}}\left[\mathbb{E}_{S_{t+1}}\left[\sum_{m=t+1}^{\infty}\varphi(m-t)R(S_m^{\boldsymbol{\pi}}, \boldsymbol{\pi}(S_m^{\boldsymbol{\pi}}))\right]\right]\right\} && (47d)
\end{aligned}
$$

On the 2nd line, we apply

$$
\begin{aligned}
&\left|\left|\mathbb{E}\left[\sum_{m=t+1}^{\infty}\varphi(m-(t+1))R_m^{\boldsymbol{\pi}}||S_{t+1}=s\right] - \mathbb{E}\left[\sum_{m=t+1}^{\bar{T}_\epsilon}\varphi(m-(t+1))R_m^{\boldsymbol{\pi}}|S_{t+1}=s\right]\right|\right|\\
&= \left|\left|\mathbb{E}\left[\sum_{m=0}^{\infty}\varphi(m)R_m^{\boldsymbol{\pi}}||S_0=s\right] - \mathbb{E}\left[\sum_{m=0}^{\bar{T}_\epsilon-(t+1)}\varphi(m)R_m^{\boldsymbol{\pi}}|S_0=s\right]\right|\right|\\
&= ||V^{\boldsymbol{\pi}}(s) - V^{\boldsymbol{\pi}^{\bar{T}_\epsilon-(t+1)}\cdot\ddot{\boldsymbol{\pi}}}(s)||\\
&\leq \frac{\epsilon}{\kappa} \qquad \text{(by Lemma B.3)}
\end{aligned}
$$

On the 3rd line, we apply

$$\begin{aligned}
&\left\|\mathbb{E}\left[\sum_{m=t+1}^{\infty}\varphi(m-t)R_m^{\boldsymbol{\pi}}||S_{t+1}=s\right]-\mathbb{E}\left[\sum_{m=t+1}^{\bar{T}_\epsilon}\varphi(m-t)R_m^{\boldsymbol{\pi}}|S_{t+1}=s\right]\right\| \\
&=\left\|\mathbb{E}\left[\sum_{m=1}^{\infty}\varphi(m)R_{m-1}^{\boldsymbol{\pi}}|S_0=s\right]-\mathbb{E}\left[\sum_{m=1}^{\bar{T}_\epsilon-t}\varphi(m)R_{m-1}^{\boldsymbol{\pi}}|S_0=s\right]\right\| \\
&\leq\left\|\mathbb{E}\left[\sum_{m=1}^{\infty}\varphi(m-1)R_{m-1}^{\boldsymbol{\pi}}|S_0=s\right]-\mathbb{E}\left[\sum_{m=1}^{\bar{T}_\epsilon-t}\varphi(m-1)R_{m-1}^{\boldsymbol{\pi}}|S_0=s\right]\right\| && \text{(by } \varphi(.) \text{ decreasing)} \\
&=\left\|V^{\boldsymbol{\pi}}(s)-V^{\boldsymbol{\pi}^{\bar{T}_\epsilon-(t+1)}\cdot\ddot{\boldsymbol{\pi}}}(s)\right\| && \text{(by (5))} \\
&\leq\frac{\epsilon}{\kappa} && \text{(by Lemma B.3)}
\end{aligned}$$

Therefore, continuing from (47d), we can perform approximation with $\bar{T}_\epsilon<\infty$ as follows,

$$\begin{aligned}
&\overset{2\epsilon/\kappa}{\approx}\mathbb{E}_{R\sim p_{s_t}^{a_t}}[\varphi(0)R(s_t,a_t)]+\mathbb{E}_{S_{t+1}\sim p_{s_t}^{a_t}}\left[Q^{\boldsymbol{\pi}}(S_{t+1},\boldsymbol{\pi}(S_{t+1}))\right] \\
&\quad-\left\{\mathbb{E}_{S_{t+1}\sim p_{s_t}^{a_t}}\left[\sum_{m=t+1}^{\bar{T}_\epsilon}\mathbb{E}_{S_{t+1}}\left[\varphi(m-(t+1))R(S_m^{\boldsymbol{\pi}},\boldsymbol{\pi}(S_m^{\boldsymbol{\pi}}))\right]\right]\right. \\
&\quad\left.-\mathbb{E}_{S_{t+1}\sim p_{s_t}^{a_t}}\left[\sum_{m=t+1}^{\bar{T}_\epsilon}\frac{\varphi(m-t)}{\varphi(m-(t+1))}\mathbb{E}_{S_{t+1}}\left[\varphi(m-(t+1))R(S_m^{\boldsymbol{\pi}},\boldsymbol{\pi}(S_m^{\boldsymbol{\pi}}))\right]\right]\right\} \qquad (47e)
\end{aligned}$$

By applying (46), Definition 4.2, and (44) on the 1st, 2nd, and 3rd line, respectively, we can then obtain

$$\begin{aligned}
&=\mathbb{E}_{R\sim p_{s_t}^{a_t}}[\varphi(0)R(s_t,a_t)]+\mathbb{E}_{S_{t+1}\sim p_{s_t}^{a_t}}\left[Q^{\boldsymbol{\pi}}(S_{t+1},\boldsymbol{\pi}(S_{t+1}))\right] \\
&\quad-\left\{\sum_{m=t+1}^{\bar{T}_\epsilon}\left(\mathbb{E}_{S_{t+1}\sim p_{s_t}^{a_t}}\left[r^{\boldsymbol{\pi},t+1,m}(S_{t+1},\boldsymbol{\pi}(S_{t+1}))\right]-r^{\boldsymbol{\pi},t,m}(s_t,a_t)\right)\right\} \qquad (47f)
\end{aligned}$$

Finally, based on the Definition 3.5, we will set our boundary conditions (when we are at some *boundary* states),

$$Q^{\boldsymbol{\pi}}(s_t,a_t)=\mathbb{E}_{R\sim p_{s_t}^{a_t}}[\varphi(0)R(s_t,a_t)],\forall s_t\in\bar{\mathcal{S}},a_t\in\bar{\mathcal{A}} \qquad (48)$$

C Backward Q-learning Algorithm

In this section, we detail the derivations of our Backward Q-learning in Section 5.2 from Theorem 4.3.

C.1 r-table Update

Based on the r-function recursion, i.e. (44) and (46), we obtain bootstrap targets that corresponds to (21) in the main paper,

$$\xi_t^r(m)\leftarrow\varphi(0)R(s_t,a_t),\quad\text{for } m=t \qquad (49)$$

$$\xi_t^r(m)\leftarrow\frac{\varphi(m-t)}{\varphi(m-(t+1))}r^{t+1,m}(S_{t+1},\pi'(S_{t+1})),\quad\text{for } m=t+1:T^* \qquad (50)$$

Then, updates to r-table are made as follows,

$$r^{t,m}(s_t,a_t)\leftarrow(1-\alpha_r)r^{t,m}(s_t,a_t)+\alpha_r\xi_t^r(m),\quad\text{for } m=t:T^* \qquad (51)$$

given learning rate $\alpha_r>0$.

C.2 Q-table Update

Based on the Q-function recursion, i.e. (47f) and (48), we obtain bootstrap targets that corresponds to (22) in the main paper,

$$\xi_t^Q \leftarrow \gamma(0)R(s_t, a_t), \quad \text{for } t = T^* \text{ and } s_t \in \bar{\mathcal{S}} \tag{52}$$

$$\xi_t^Q \leftarrow \gamma(0)R(s_t, a_t) + Q(S_{t+1}, \pi'(S_{t+1})) - \max(0, \Delta r_t), \quad \text{for } t \leq T^* - 1 \tag{53}$$

where $\Delta r_t = \sum_{m=t+1}^{T^*} r^{t+1,m}(s_{t+1}, \pi'(s_{t+1})) - r^{t,m}(s_t, a_t)$. Then, updates to Q-table can be done as follows,

$$Q(s_t, a_t) \leftarrow (1 - \alpha_Q)Q(s_t, a_t) + \alpha_Q \xi_t^Q, \quad \text{for } t \leq T^* \tag{54}$$

given learning rate $\alpha_Q > 0$.

Truncation from $\bar{T}$ to T^*. For our implementation, instead of keeping track of all values up to $\bar{T}$, we use the variable length T^* of each trajectory sampled following a current policy $\boldsymbol{\pi}$. However, we will still set a sufficiently large $\bar{T}$ as a proxy for $\bar{T}_\epsilon$ to ensure that all trajectory terminates.

Clipping of adjustment terms. Let us denote by $\Delta r_t^{\boldsymbol{\pi}}$ the adjustment terms in the 2nd row of (47f) as in the main paper. Referring to (53), we note that the clipped function $\max(0, \Delta r_t)$ has been used in place of Δr_t. This is done to slow down the accumulation of error relevant to $\Delta r_t \approx \Delta r_t^{\boldsymbol{\pi}}$. In particular, we note that $\Delta r_t^{\boldsymbol{\pi}} \geq 0$:

$$\begin{aligned}
\Delta r_t^{\boldsymbol{\pi}} &\doteq \sum_{m=t+1}^{\bar{T}-1} \left(\mathbb{E}_{S' \sim p_{s_t}^{a_t}} \left[r^{\boldsymbol{\pi},t+1,m}(S', \pi(S')) \right] - r^{\boldsymbol{\pi},t,m}(s_t, a_t) \right) \\
&= \sum_{m=t+1}^{\bar{T}-1} \left(\mathbb{E}_{S' \sim p_{s_t}^{a_t}} \left[\left(1 - \frac{\varphi(m-t)}{\varphi(m-(t+1))} \right) r^{\boldsymbol{\pi},t+1,m}(S', \pi(S') \right] \right) && \text{(by (44))} \\
&\geq 0 && \text{(by } \varphi(\cdot) \text{ discount factor and } R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+ \text{ s.t. (9) is non-negative)}
\end{aligned}$$

But without clipping, $\Delta r_t < 0$ may happen in subsequent iterations, inflating Q-values at some states past a certain threshold such that their neighboring states prefer transition to these inflated states than moving towards goal states, when the latter clearly results in a fewer steps. This then creates a looping behaviour which eventually lead to divergence.

C.3 Policy-table Update

We note our use of policy-table separate from the arg max of a Q-table to represent *greedy* policy. This is due to the possibilities of *non-unique* actions realizing $\arg\max_a Q(s, a)$ for some $s \in \mathcal{S}$ which may cause non-unique r-function related values, i.e. the components in $\sum_{m=t+1}^{T^*} \Delta r_t^m$, after substituting *different* global optima actions. Specifically, we follow the *consistent tie-break* rule proposed in Section 3.3 of Lesmana & Pun (2021); see line 18-22 in Algorithm 1.

D NUMERICAL EXAMPLES

This section provides some missing details on Section 6.

D.1 Environment Setup

We review 3 important considerations in our Gridworld designs: (i) existence of actual *preference reversal* (i.e. if states like $(s_0, s_\tau) = (21, 9)$ exist, where the optimality of 21's action is constrained by 9's action such that we have *priority ordering* on $\mathcal{S}$), (ii) $\pi^{*0}(s_0) \neq \hat{\pi}(s_0)$ where the value of following SPE path $\hat{\boldsymbol{\pi}}$ is strictly less than following precommitment path $\boldsymbol{\pi}^{*0}$, and (iii) initialization to TIC, precommitment policy (that is

Algorithm 2 On-policy Monte Carlo Control (MC)

Input: Hyperbolic ($k = 1$) Gridworld, Hyperparameters $(\epsilon, \bar{T})$
Output: Approximate SPE Q-function $Q^{\hat{\boldsymbol{\pi}}}(s, a) \forall s \in \mathcal{S}, a \in \mathcal{A}$
Initialize: $Q(s, a) \leftarrow 0, \forall s \in \mathcal{S} \setminus \bar{\mathcal{S}}, a \in \mathcal{A}$; $Q(s, a) \leftarrow \varphi(0)R(s, a), \forall s \in \bar{\mathcal{S}}, a \in \bar{\mathcal{A}}$; $Returns(s, a) \leftarrow \emptyset, \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$; $\pi'(s) \leftarrow \arg\max_a Q(s, a), \forall s \in \mathcal{S}$; $\boldsymbol{\pi} \leftarrow \emptyset$;
repeat
  Update $\boldsymbol{\pi} \leftarrow \boldsymbol{\pi}'$
  Choose S_0 randomly
  Generate *trajectory* $\omega_{0:T^*} \doteq S_0, A_0, \ldots, S_{T^*}, A_{T^*}$ following $\boldsymbol{\pi}^\epsilon$
  Set $G \leftarrow 0$
  for $t \leftarrow 0$ **to** T^* **do**
    if the pair (S_t, A_t) does not appear in $\omega_{0:t-1}$ **then**
      Compute $G \leftarrow \varphi(T^* - t)R(S_{T^*}, A_{T^*})$
      Append G to $Returns(S_t, A_t)$
      Update $Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$
      Update $\pi'(S_t) \leftarrow \arg\max_a Q(S_t, a)$
    end if
  end for
until stable $(\boldsymbol{\pi}' \neq \boldsymbol{\pi})$

necessary to invoke the insufficiency of standard PI as illustrated in Example 4.6). For our stochastic (S) example, we inject noise to the deterministic transitions $p_9^a(\cdot)$ of state 9 in (D) such that

$$\forall a \in \{\leftarrow, \uparrow, \rightarrow, \downarrow\},\ P(s'|s = 9, a) = \begin{cases} .9, & \text{if } p_9^a(s') = 1 \text{ in (D)} \\ \frac{1-.9}{3}, & \text{else} \end{cases}$$

D.2 Benchmark Algorithms

Following, we describe the two benchmark algorithms that we use in our experiments: `MC` and `sophEU`. For our `MC` implementation, we use the fist-visit variant on-policy MC control[5] as described in Algorithm 2. For the `sophEU`, we adapt the `sophEU` algorithm proposed in Evans et al. (2016) by modifying the exploration technique to ϵ-greedy; see Algorithm 3. This is done for fair comparison with the other two methods, i.e. `MC` and `bwdQ`.

D.3 Training Setup

For each pair of algorithm and environment, hyperparameters are informally selected from the sets $\alpha, \alpha_Q \in \{.2, .3, .4, .5\}, \alpha_r \in \{.7, .8, .9, 1.0\}, \epsilon \in \{.01, .03, .05, .07, .1\}$ with the following criteria in mind: (i) small overtaking-mean i_{21}^*, (ii) small stdev-shade on the Q-value learning curves at $s = 9, 21$, and (iii) identifiable i_9^*, i_{21}^* (i.e. reducing the overlapping frequencies between two contending actions' mean Q-value learning curves); see Figure 8(a)-(b) for relatively bad instances. For all environments and algorithms, we set $\bar{T} = 100$; larger episode truncation does not affect much our experiment results. We summarize our final choice of hyperparameters in Table 2.

Table 2: Hyperparameters

$(\epsilon, \bar{T}, \alpha_Q/\alpha, \alpha_r)$	`MC`	`sophEU`	`bwdQ`
(D)	(.07, 100, -, -)	(.07, 100, .4, -)	(.07, 100, .4, 1.0)
(S)	(.07, 100, -, -)	(.07, 100, .4, -)	(.07, 100, .4, .9)

[5]We refer to the sourcecode in https://github.com/dennybritz/reinforcement-learning prior to our hyperbolic-discounting modification.

Algorithm 3 Sophisticated Expected-Utility Agent (sophEU)

1: **Input:** Hyperbolic ($k = 1$) Gridworld, Hyperparameters$(\epsilon, \bar{T}, \alpha)$
2: **Output:** Approximate SPE Q-function $Q^{\hat{\boldsymbol{\pi}}}(s, a) = Q(s, a, 0), \forall s \in \mathcal{S}, a \in \mathcal{A}$
3: **Initialize:** $Q(s, a, d) \leftarrow 0, \forall d, s \in \mathcal{S} \setminus \bar{\mathcal{S}}, a \in \mathcal{A}$; $Q(s, a, d) \leftarrow \varphi(0)R(s, a), \forall d, s \in \bar{\mathcal{S}}, a \in \mathcal{A}$; $\pi'_d(s) \leftarrow \arg\max_a Q(s, a, d), \forall d, s \in \mathcal{S}, \boldsymbol{\pi} \leftarrow \emptyset$
4: **repeat**
5: Update $\boldsymbol{\pi} \leftarrow \boldsymbol{\pi}'$
6: Choose S_0 randomly
7: **for** $t \leftarrow 0$ **to** $\bar{T} - 1$ **do**
8: Sample action $A_t \sim \pi_0^{\epsilon}(.|S_t)$
9: Observe reward $R_{t+1} \doteq R(S_t, A_t)$ and next state S_{t+1}
10: Set $d \leftarrow t$
11: Compute utility $U \leftarrow \varphi(d) \cdot R(S_t, A_t)$
12: Compute expectation $E \leftarrow \sum_{a' \sim \mathcal{A}} \pi_0^{\epsilon}(a'|S_{t+1}) Q(S_{t+1}, a', d + 1)$
13: Update $Q(S_t, A_t, d) \leftarrow Q(S_t, A_t, d) + \alpha(U + E - Q(S_t, A_t, d))$
14: Update $\pi'_d(S_t) \leftarrow \arg\max_a Q(S_t, a, d)$
15: **end for**
16: **until** stable $(\boldsymbol{\pi}'_0 \neq \boldsymbol{\pi}_0)$

D.4 Additional Results and Evaluation

This subsection expands the results and evaluation subsection in the main paper.

D.4.1 Q-value Learning Curves

To illustrate how we record the overtaking indexes i^*_9, i^*_{21} used to compute Δi^* in Table 1, we plot in Figure 5-8 the Q-value learning curves that correspond to Figure 4.

D.4.2 Terminal Policies vs Groundtruth Value Comparisons

In Figure 4(b) of the main paper, we have shown that all algorithms will eventually terminate at SPE policy $\hat{\boldsymbol{\pi}}(s_0)$ for $s_0 = 21$. However, Figure 4(a) shows that both `MC` and `sophEU` do not flatten to the groundtruth SPE value function $V^{\hat{\boldsymbol{\pi}}}(s_0) = Q^{\hat{\boldsymbol{\pi}}}(s_0, \hat{\pi}(s_0))$. Now that we have Q-value learning curves in Figure 5-8, it becomes clearer that the source of this discrepancy lies on the mis-evaluated Q-values; see $Q(21, \rightarrow)$ in Figure 5(a) for instance. This is explainable for a few reasons. Firstly, in the case of `MC`, the magnitude of exploratory rate ϵ causes Q-values to evaluate the exploratory policy $\boldsymbol{\pi}^{\epsilon}$ consisting paths of extended lengths, which correspondingly lead to an underestimated cumulative discounted reward. In the case of `sophEU`, similar undervaluation of $\boldsymbol{\pi}$ happens due to the action-taking probabilities being included in the Q-table updates; see line 12-13 in Algorithm 3. While making ϵ smaller intuitively fixes this issue, learning performance deteriorates (i.e. highly variable across seeds) once we decrease ϵ up to certain threshold; our final choice of $\epsilon = .07$ has taken this into consideration. Secondly, `MC` observes some kind of smoothening effect across updates, which if combined with the delayed reflecting of information (i.e. prolonged Δi^*) exacerbate the early flattening of policy values. Such smoothening concurrently explains how `MC` appears to have lesser variance as compared to `bwdQ` or `sophEU` at later iterations; see Figure 4(a)-Stochastic (S) in the main paper.

D.4.3 Ablation Study: Reversed Backward Q-learning

Since both benchmark algorithms suffer from similar undervaluation of policy issue, we construct an additional benchmark: Reversed Backward Q-learning (`bwdQ-rev`), that is based on our own algorithm `bwdQ`. Here, we only retain the extended DP-based policy evaluation component of `bwdQ` (that resembles TD-based methods in standard RL literature) and apply standard conditioning by reversing the backward order of policy update in line 11, Algorithm 1. This benchmarking can also be seen as an ablation study to see how backward conditioning alone can reduce delusionality and improve learning performance. Figure 9 displays the value and Q-value learning curves of `bwdQ-rev` against `bwdQ` in both (D) and (S), under the same learning rates.

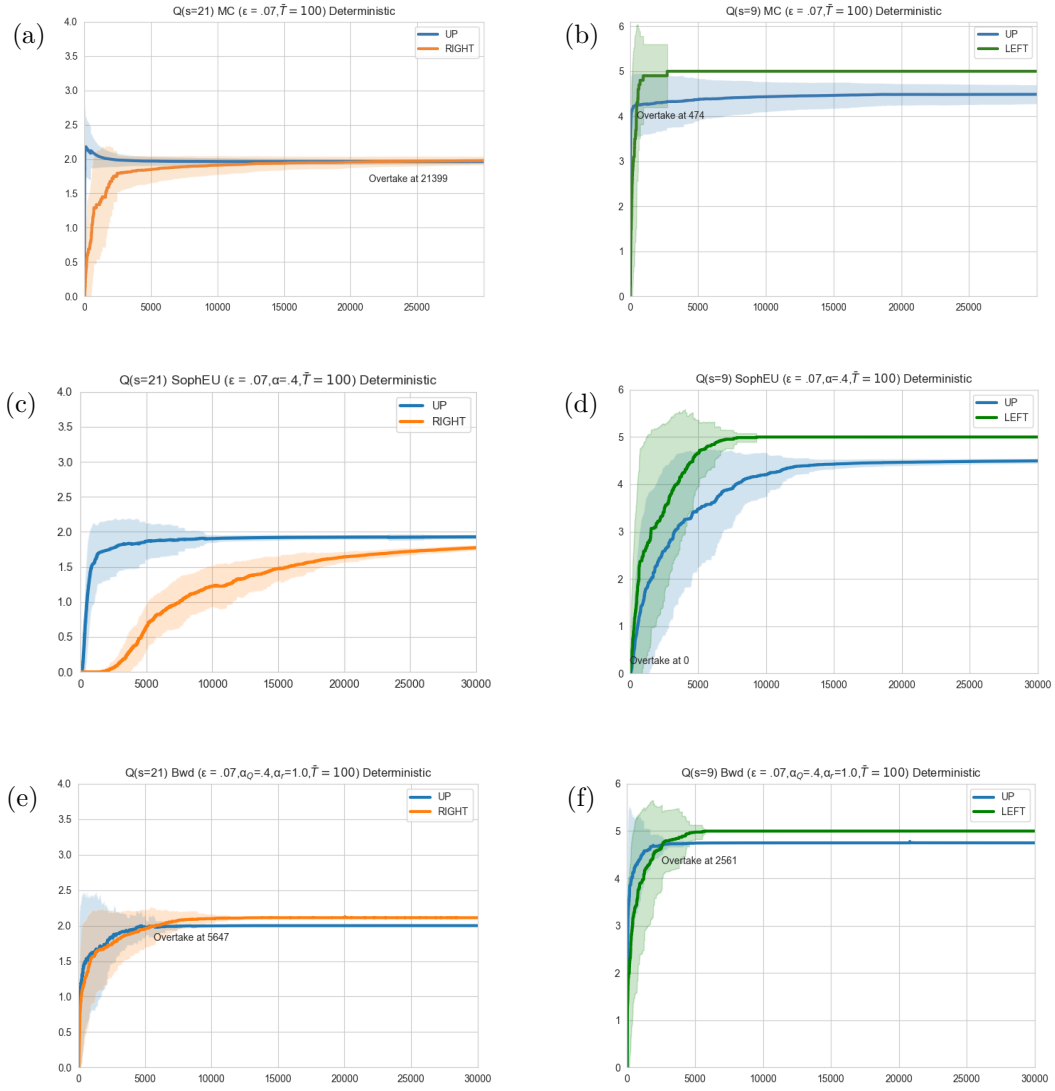

Figure 5: Q-value Learning Curves for `MC`, `sophEU`, and `bwdQS` at $s = 21, 9$ in (D).

Value learning curves. From Figure 9(a), it can be seen that `bwdQ-rev` also manages to match the groundtruth values at about the same speed of `bwdQ`. However, we can observe a large swing at earlier iterations which indicate `bwdQ-rev`'s degree of delusionality. In particular, such a swing is caused by 21's late update about 9's strategy of going '←', resulting in delusional prediction targets $Q(21, \uparrow; \pi(9) = \uparrow)$ and an inflation of Q-values. `BwdQ-rev`'s speed of correction towards the groundtruth here is then made possible by its large learning rates[6], which we will show to have some disadvantages next.

[6]Some comparisons can be made with `MC`'s degree of delusionality in Figure 4(a) of the main paper, that is milder for its smaller (smoothened) learning rate. It is then natural to ask how `sophEU` does not seem to exhibit such (Q-)value inflation. This can be explained by `sophEU`'s *delay-augmentation*, in which the rate of value propagation from delays $d > 0$ to $d = 0$ may match the speed of delusionality correction. To illustrate, we can observe how in Figure 5(c), `sophEU`'s $Q(21, \uparrow)$ climbs up slowly from 0 instead of jumping to near 2.0 like most others.

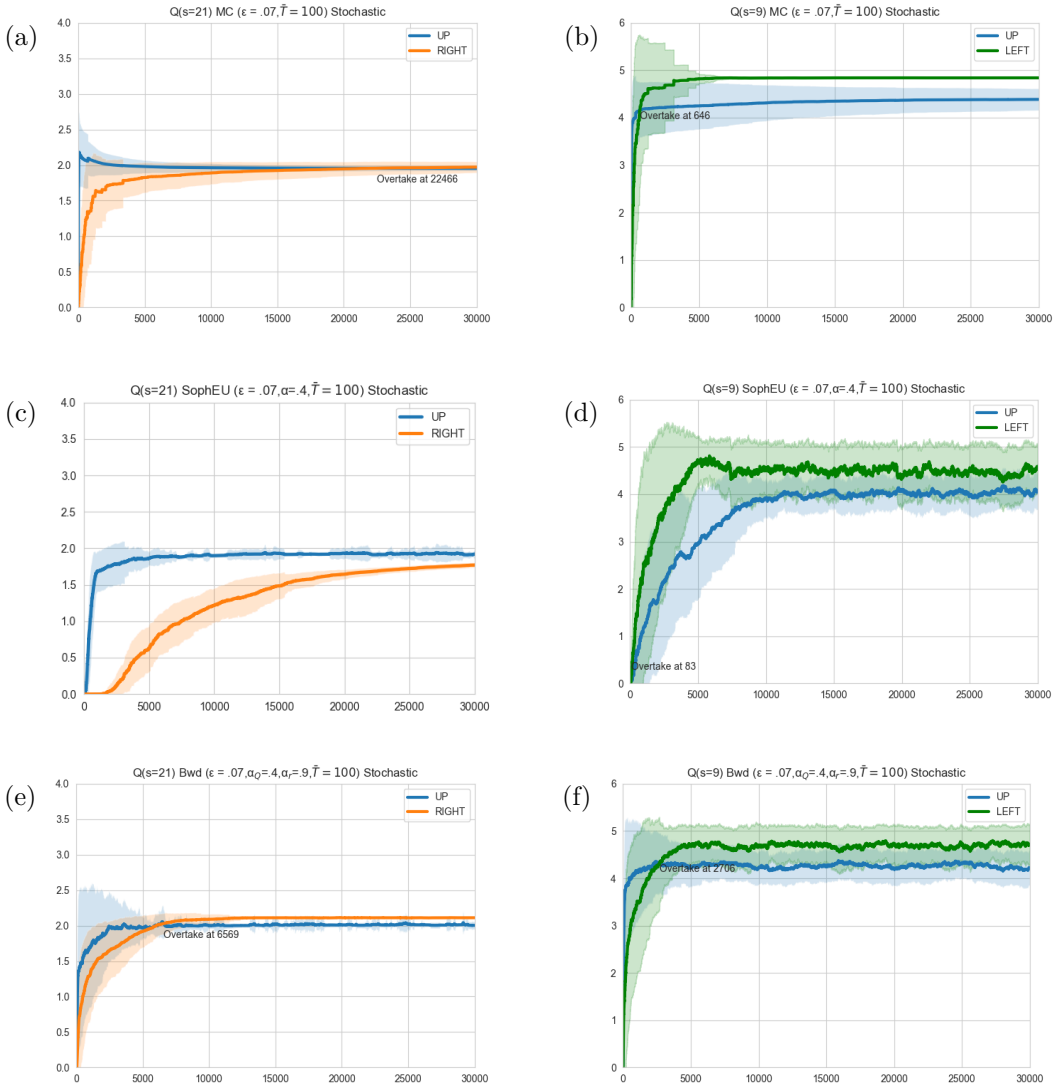

Figure 6: Q-value Learning Curves for `MC`, `sophEU`, and `bwdQ` at $s = 21, 9$ in (S).

Q-value learning curves. From Figure 9(e), while i^*_{21} of `bwdQ-rev` seems to match `bwdQ`, we observe wide stdev-shades for two contending actions '$\rightarrow$' and '$\uparrow$' that overlap throughout training episodes, indicating indecisive behaviour i.e. high variability of trained policy at convergence across random seeds. We note that this phenomenon happens in 5 out of 10 `bwdQ-rev` experiments we conducted under (S) setup, while never happening in `bwdQ`. Moreover, in Figure 9(f), the mean Q-curves of the two contending actions '$\leftarrow$' and '$\uparrow$' are relatively unstable and overlap frequently; see how `bwdQ` behaves in Figure 6(f) for comparison. These evidences suggest that reversing backward conditioning to standard impedes learning, particularly impairing `bwdQ`'s ability to handle larger learning rates. The results for both algorithms under (D) setup are largely similar, except for `bwdQ-rev`'s inflated Q-values at earlier iterations that has been covered previously.

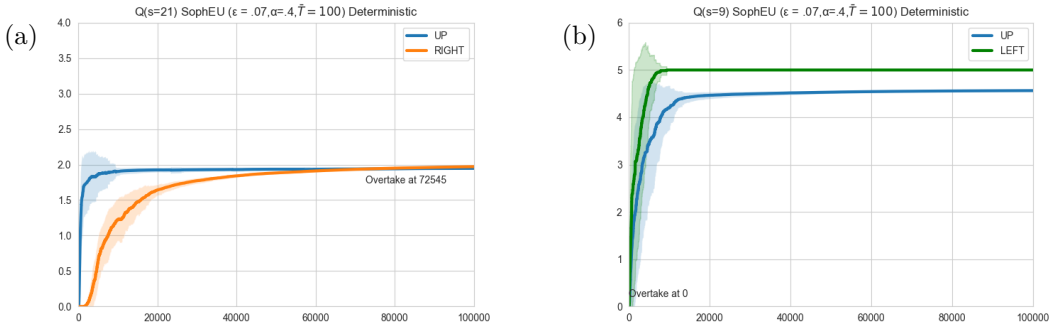

Figure 7: Q-value Learning Curves for `sophEU` (Ext.) at $s = 21, 9$ in (D).

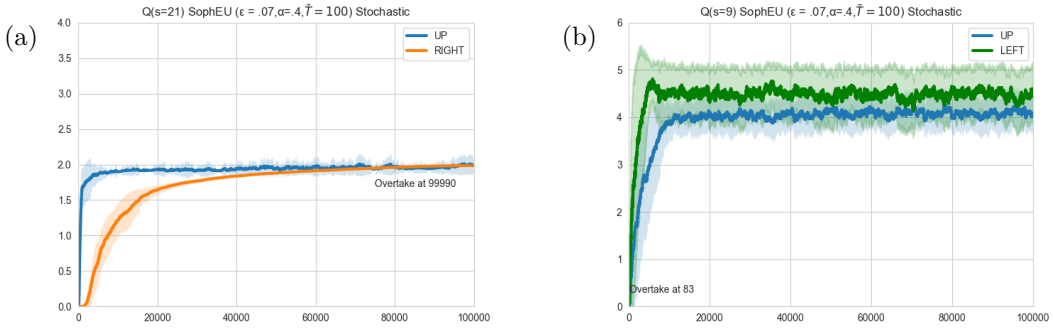

Figure 8: Q-value Learning Curves for `sophEU` (Ext.) at $s = 21, 9$ in (S).

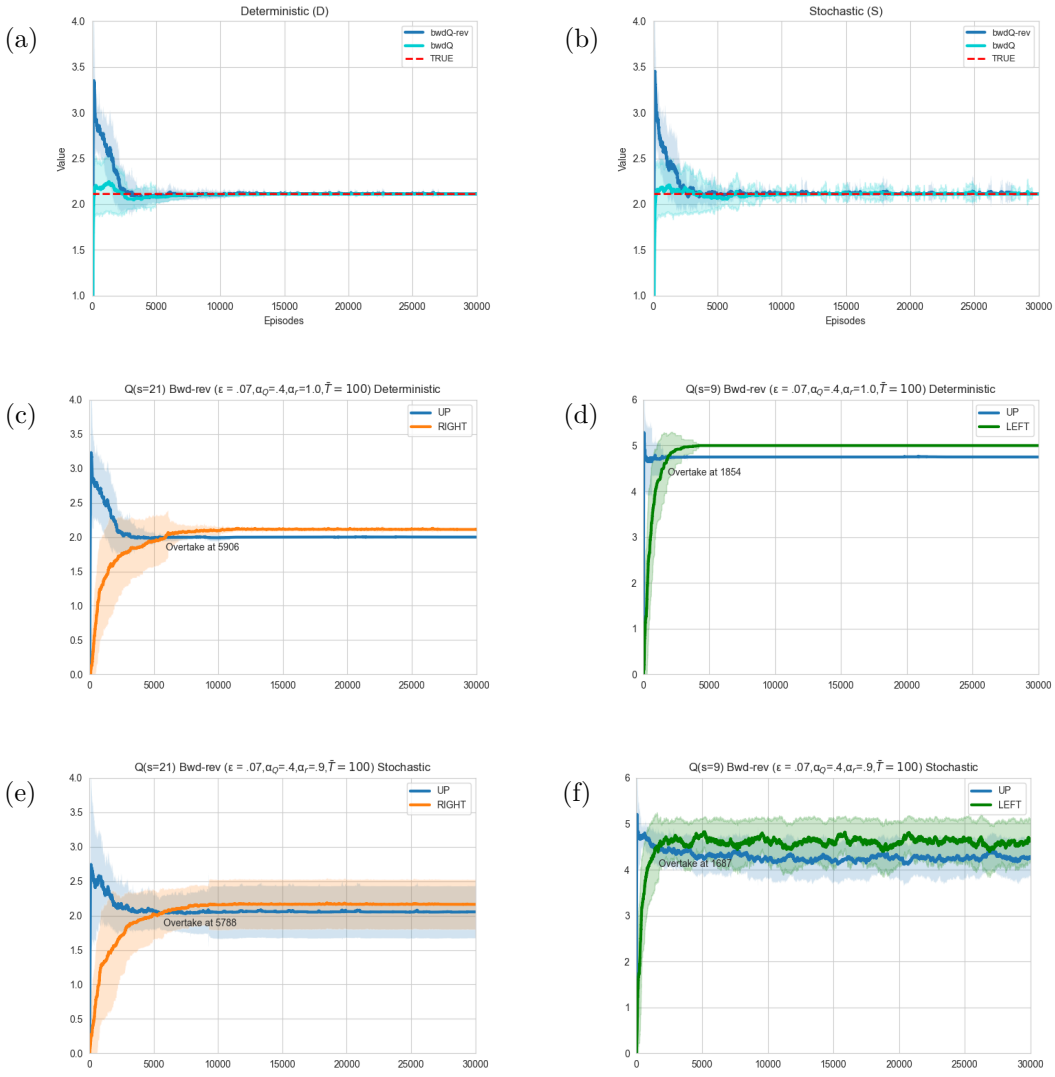

Figure 9: **Value and Q-value Learning Curve at $s = 21, 9$ of `bwdQ-rev` in (D) and (S).** Experiment ID for `bwdQ-rev` is similarly set to the one exhibiting slowest termination at 21 as indicated by the largest i^*_{21}.