Policy Testing in Markov Decision Processes

Kaito Ariu[1*] **Po-An Wang**[2*] **Alexandre Proutiere**[3] **Kenshi Abe**[1]
[1]CyberAgent [2]National Tsing Hua University [3]KTH, Digital Futures

Abstract

We study the policy testing problem in discounted Markov decision processes (MDPs) under the fixed-confidence setting. The goal is to determine whether the value of a given policy exceeds a specified threshold while minimizing the number of observations. We begin by deriving an instance-specific lower bound that any algorithm must satisfy. This lower bound is characterized as the solution to an optimization problem with non-convex constraints. We propose a policy testing algorithm inspired by this optimization problem—a common approach in pure exploration problems such as best-arm identification, where asymptotically optimal algorithms often stem from such optimization-based characterizations. As for other pure exploration tasks in MDPs, however, the non-convex constraints in the lower-bound problem present significant challenges, raising doubts about whether statistically optimal and computationally tractable algorithms can be designed. To address this, we reformulate the lower-bound problem by interchanging the roles of the objective and the constraints, yielding an alternative problem with a non-convex objective but convex constraints. Strikingly, this reformulated problem admits an interpretation as a policy optimization task in a newly constructed *reversed MDP*. Leveraging recent advances in policy gradient methods, we efficiently solve this problem and use it to design a policy testing algorithm that is statistically optimal—matching the instance-specific lower bound on sample complexity—while remaining computationally tractable. We validate our approach with numerical experiments.

1 Introduction

Reinforcement learning (RL) commonly models the interaction between a learning agent and its environment as a Markov Decision Process (MDP) (Puterman, 1994), due to its flexibility and wide applicability. Fundamental problems in RL, such as policy evaluation and best policy identification, have received significant attention, and the performance of learning algorithms on these pure exploration tasks is typically measured by their sample complexity. Ideally, we aim to design algorithms with instance-specific optimal sample complexity. This ensures that the algorithm adapts to the specific problem instance at hand, rather than to a worst-case scenario, and accurately reflects its true difficulty. In the context of multi-armed bandits (MAB), which can be interpreted as stateless RL, the design of algorithms for the best arm identification task is relatively well understood, and several instance-optimal algorithms exist (Garivier and Kaufmann, 2016; Degenne et al., 2019; Wang et al., 2021).

However, extending such guarantees to RL settings governed by MDPs is highly non-trivial. The primary challenge is that, unlike in bandits, the set of parameterizations that make two MDP instances hard to distinguish—the so-called *confusing parameters*—forms a non-convex set (Al Marjani and Proutiere, 2021). As a result, the optimization problems that characterize instance-specific complexity in RL, also referred to as the lower bound problems, are inherently non-convex and

[*]Alphabetical order. Emails: kaito_ariu@cyberagent.co.jp, po-an@stat.nthu.edu.tw.

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

computationally intractable in general. Common workarounds rely on convex relaxations, which compromise statistical optimality.

In this paper, we address this challenge for the *policy testing problem*: given a confidence level δ, the agent must determine whether the value of a given policy exceeds a specified threshold with confidence at least $1 - \delta$. Our main contribution is a new approach that transforms the generally non-convex lower bound problem for policy testing into a tractable form without sacrificing statistical optimality. Specifically, we prove that we can reformulate this problem as a policy optimization task in a newly constructed *reversed MDP*, where the usual roles of the agent's policy and the environment's transition dynamics are interchanged. By means of this reformulation, we obtain convex constraints and a non-convex objective function, which allows us to efficiently explore global solutions by leveraging and extending projected policy gradient methods. We carefully combine these results to propose the first algorithm for policy testing in MDPs that achieves instance-optimal sample complexity. As far as we are aware, this is the first computationally tractable algorithm to achieve instance-specific optimality for pure exploration in MDPs. Our framework can also be extended to other pure exploration problems in MDPs, such as policy evaluation and best policy identification, thus opening up new possibilities for developing instance-optimal and efficient reinforcement learning algorithms.

Contributions. Our contributions can be summarized as follows: (i) We derive an instance-specific lower bound on the sample complexity (Theorem 1), revealing that the problem involves non-convex constraints. (ii) To address this non-convexity, we reformulate the optimization problem as an equivalent reversed MDP and show that it can be solved using a projected policy gradient method (Theorem 3). (iii) Building on these results, we prove that our proposed method, PTST, is asymptotically instance-optimal in terms of sample complexity (Theorem 2). (iv) In experiments, we show that PTST achieves better sample complexity compared to existing methods.

2 Related work

Pure exploration and offline learning problems in MABs have been studied extensively. In particular, significant attention has been devoted to best-arm identification in both the fixed-confidence and fixed-budget settings (see e.g., (Audibert and Bubeck, 2010; Gabillon et al., 2012; Soare et al., 2014)). In the fixed-confidence setting, researchers have derived instance-specific lower bounds on sample complexity, which in turn have enabled the design of asymptotically optimal algorithms (Garivier and Kaufmann, 2016; Degenne et al., 2019; Jedra and Proutiere, 2020; Wang et al., 2021). This level of analysis is feasible due to the relative simplicity of the optimization problems underlying these lower bounds.

Similar challenges have been explored in the context of MDPs. Several works have focused on characterizing the minimax sample complexity for best-policy identification in offline RL, typically under the generative model assumption (Gheshlaghi Azar et al., 2013; Agarwal et al., 2020; Li et al., 2024). Other efforts have aimed at instance-optimality either in offline settings (Khamaru et al., 2021; Wang et al., 2024) or under adaptive sampling regimes (Zanette et al., 2019; Al Marjani and Proutiere, 2021; Al Marjani et al., 2021; Tirinzoni et al., 2022; Kitamura et al., 2023; Taupin et al., 2023; Russo and Vannella, 2024). However, none of these approaches achieves true instance-specific optimality. Even in the relatively simple case of tabular episodic MDPs, current results attain only near-optimal sample complexity (Tirinzoni et al., 2022; Al-Marjani et al., 2023; Narang et al., 2024). The core barrier to achieving instance-specific optimality in MDPs lies in the inherent complexity of the optimization problem that defines the sample complexity lower bound. In this paper, we provide the first approach to fully deal with this complexity in the context of the policy testing task. A more detailed discussion of related work can be found in Appendix A.

3 Preliminaries, objectives, and assumptions

3.1 Markov decision processes

We consider a discounted Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \boldsymbol{\rho}, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the finite state and action spaces, respectively. The (unknown) transition kernel is given by $p \in \mathcal{P} := \Delta(\mathcal{S})^{\mathcal{S}\times\mathcal{A}}$, where $\Delta(\mathcal{X})$ denotes the simplex over $\mathcal{X}$. $p(s'|s,a)$ denotes the probability to

move to state s' given the current state s and the selected action a. The reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is deterministic, $\boldsymbol{\rho}$ represents the known initial state distribution, and $\gamma \in (0, 1)$ is the discount factor. We denote the state-action pair at time t by $(s(t), a(t))$. At time t, the agent selects action $a(t)$ according to the distribution $\pi(\cdot \mid s(t))$, collects reward $r(s(t), a(t))$, and moves to the next state $s(t+1)$ according to the distribution $p(\cdot \mid s(t),\ a(t))$. The value function of a given policy $\pi \in \Pi := \Delta(\mathcal{A})^{\mathcal{S}}$ is defined by its average long-term discounted reward given any possible starting state s: $V_p^\pi(s) := \mathbb{E}_p^\pi\Big[\sum_{t\geq 0} \gamma^t r(s(t), a(t)) \mid s(0) = s\Big]$, where $\mathbb{E}_p^\pi$ represents the expectation taken with respect to randomness induced by π and p. Similarly, for each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, Q-function is defined as: $Q_p^\pi(s, a) := \mathbb{E}_p^\pi\Big[\sum_{t\geq 0} \gamma^t r(s(t), a(t)) \mid s(0) = s, a(0) = a\Big]$. The value of π is then defined as: $V_p^\pi(\boldsymbol{\rho}) := \sum_{s\in\mathcal{S}} \rho_s V_p^\pi(s)$. We also define the discounted state-visitation distribution:

$$d_{p,s,a}^\pi(s', a') = (1-\gamma)\mathbb{E}_p^\pi\Big[\sum_{t\geq 0} \gamma^t \mathbb{1}\{(s(t), a(t)) = (s', a') \mid (s(0), a(0)) = (s, a)\}\Big]$$

and $d_{p,s}^\pi(s') = \sum_{a\in\mathcal{A}} \pi(a|s) d_{p,s,a}^\pi(s', a')$. The state-visitation distribution initialized by $\boldsymbol{\rho}$, $d_{p,\boldsymbol{\rho}}^\pi \in \Delta(\mathcal{S})$, is defined with components $d_{p,\boldsymbol{\rho},s}^\pi = \sum_{s'\in\mathcal{S}} \rho_{s'} d_{p,s'}^\pi(s)$ for each $s \in \mathcal{S}$. For any $\boldsymbol{\rho}, \boldsymbol{\mu} \in \Delta(\mathcal{S})$, we define $\|\boldsymbol{\rho}/\boldsymbol{\mu}\|_\infty := \max_{s\in\mathcal{S}} \rho_s/\mu_s$, with $0/0 = 1$ by convention. We have $d_{p,\boldsymbol{\rho},s}^\pi \geq (1-\gamma)\rho_s$ and $\|d_{p,\boldsymbol{\rho}}^\pi / d_{p,\boldsymbol{\rho}}^{\pi'}\|_\infty \leq \|d_{p,\boldsymbol{\rho}}^\pi/\boldsymbol{\rho}\|_\infty/(1-\gamma)$ for all $\boldsymbol{\rho} \in \Delta(\mathcal{S})$, $\pi, \pi' \in \Pi$.

3.2 Policy testing

We aim to devise an algorithm that determines whether the value $V_p^\pi(\boldsymbol{\rho})$ of a given policy π exceeds some given threshold with a minimal number of samples. Without loss of generality, we can assume that this threshold is 0.[2] We assume that the kernel p should satisfy $V_p^\pi(\boldsymbol{\rho}) \neq 0$, i.e., the value is strictly positive or negative. Therefore, we write the set of problem instances: $\mathcal{P}_{\text{Test}} := \{q \in \mathcal{P} : V_q^\pi(\boldsymbol{\rho}) \neq 0\}$. For each $p \in \mathcal{P}_{\text{Test}}$, the answer $\text{Ans}(p)$ is $+$ if $V_p^\pi(\boldsymbol{\rho}) > 0$ and $-$ if $V_p^\pi(\boldsymbol{\rho}) < 0$. One can thus divide $\mathcal{P}_{\text{Test}}$ into two disjoint sets, $\mathcal{P}_{\text{Test}} = \mathcal{P}_{\text{Test}}^+ \cup \mathcal{P}_{\text{Test}}^-$, where $\mathcal{P}_{\text{Test}}^+ := \{p \in \mathcal{P}_{\text{Test}} : \text{Ans}(p) = +\}$, $\mathcal{P}_{\text{Test}}^- := \{p \in \mathcal{P}_{\text{Test}} : \text{Ans}(p) = -\}$.

We assume that the agent has access to a generative model. In each step, the agent selects a state-action pair, from which the transition to the next state is observed. We consider the case where the agent uses a static sampling rule, targeting fixed proportions of state-action draws $\boldsymbol{\omega} \in \Sigma := \{\boldsymbol{\omega}' \in [0, 1]^{|\mathcal{S}|\times|\mathcal{A}|} : \sum_{s,a} \omega'_{sa} = 1\}$ (ω_{sa} denotes the proportion of time state-action pair is sampled). We leave the case of adaptive sampling rules for future work. Our goal is to design an algorithm that, with a fixed confidence level of $1-\delta$ (where $\delta \in (0, 1)$ is a predefined parameter), determines as quickly as possible whether $V_p^\pi(\boldsymbol{\rho})$ exceeds the given threshold (i.e., whether $V_p^\pi(\boldsymbol{\rho}) > 0$ or $V_p^\pi(\boldsymbol{\rho}) < 0$).

The algorithm consists of a stopping rule and a decision rule. The stopping rule is defined through a stopping time τ w.r.t. the natural filtration $\mathcal{F} = (\mathcal{F}_t)_{t\geq 1}$, where $\mathcal{F}_t$ denotes the σ-field generated by all the observations collected up to and including round t. In round τ, after stopping, the algorithm returns a $\mathcal{F}_\tau$-measurable decision $\hat{\imath} \in \{+, -\}$, corresponding to the answer which is believed to be correct. The sample complexity of an algorithm is defined as $\mathbb{E}_p[\tau]$ where the expectation is with respect to the sampling process, the observations, and the stopping rule.

Definition 1. *An algorithm is δ-Probably Correct (δ-PC) if for all $p \in \mathcal{P}_{\text{Test}}$, (i) it stops almost surely, $\mathbb{P}_p[\tau < \infty] = 1$ and (ii) $\mathbb{P}_p[\hat{\imath} \neq \text{Ans}(p)] \leq \delta$.*

We aim at devising a δ-PC algorithm with minimal sample complexity.

3.3 Assumptions

To simplify the notation, we define $r^\pi(s) := \sum_{a\in\mathcal{A}} \pi(a|s) r(s, a)$ and $r^\pi(\boldsymbol{\rho}) := \sum_{s\in\mathcal{S}} \rho_s r^\pi(s)$.

As $V_p^\pi(\boldsymbol{\rho}) = r^\pi(\boldsymbol{\rho}) + \sum_{t=1}^\infty \mathbb{E}_p^\pi[\gamma^t r(s(t), a(t))]$, the transition kernel p maximizing the value maps all state-action pairs to the most rewarding state, $\arg\max_s r^\pi(s)$. In contrast, the kernel minimizing

[2] If the threshold is R, we can instead use the shifted reward function $\tilde{r} = r - (1-\gamma)R$, and the new value function is $\tilde{V}_p^\pi(s) = V_p^\pi(s) - R$. Therefore, testing $V_p^\pi(s) > R$ is equivalent to testing $\tilde{V}_p^\pi(s) > 0$.

the value maps all state-action pairs to the least rewarding state, $\arg\min_s r^\pi(s)$. That is,

$$\max_p V_p^\pi(\boldsymbol{\rho}) = r^\pi(\boldsymbol{\rho}) + \frac{\gamma}{1-\gamma}\max_s r^\pi(s) \text{ and } \min_p V_p^\pi(\boldsymbol{\rho}) = r^\pi(\boldsymbol{\rho}) + \frac{\gamma}{1-\gamma}\min_s r^\pi(s). \quad (1)$$

Throughout the paper, we consider the following Assumption 1 holds, which ensures $\{q \in \mathcal{P}_{\text{Test}} : \text{Ans}(q) = -\}$ and $\{q \in \mathcal{P}_{\text{Test}} : \text{Ans}(q) = +\}$ are nonempty sets and simplifies the presentation.

Assumption 1. *$\rho_s > 0$ for all $s \in \mathcal{S}$. r and $\boldsymbol{\rho}$ satisfy: $\frac{-\gamma}{1-\gamma}\min_s r^\pi(s) > r^\pi(\boldsymbol{\rho}) > \frac{-\gamma}{1-\gamma}\max_s r^\pi(s)$.*

This assumption also implies that for any transition kernel, the state value function is not constant (it varies across states). This is formalized in the following lemma, proved in Appendix H.1.

Lemma 1. *Under Assumption 1, $\min_{q\in\mathcal{P}}\max_{s,s'\in\mathcal{S}} V_q^\pi(s) - V_q^\pi(s') > 0$.*

Throughout the paper, we also make the following assumption, stating that all the studied policy, π, is full-supported and all the actions played under π must be explored under our static sampling rule $\boldsymbol{\omega}$. Notice that if there exists (s,a) such that $\pi(a \mid s) = 0$, the dynamic on this pair, $p(\cdot \mid s,a)$ does not affect the value of π.

Assumption 2. *$\pi(a \mid s) > 0$ and $\omega_{sa} > 0$ for all $s, a \in \mathcal{S} \times \mathcal{A}$.*

4 Sample complexity lower bound

We derive sample complexity lower bounds satisfied by any δ-PC algorithm. To this aim, we leverage the classical change-of-measure arguments in multi-armed bandit (MAB) (Lai and Robbins, 1985; Garivier and Kaufmann, 2016). To state our lower bound, we need the following notation. For any state-action pair (s,a), $\text{KL}_{sa}(p,q)$ denotes the KL divergence between the distributions $p(\cdot \mid s,a)$ and $q(\cdot \mid s,a)$. Finally, for $t \in \mathbb{N}$, $(s,a) \in \mathcal{S}\times\mathcal{A}$, $N_{sa}(t)$ denotes the number of times (s,a) is sampled up to t.

We introduce the set of *alternative or confusing* kernels as $\text{Alt}(p) := \{q \in \mathcal{P}_{\text{Test}} : \text{Ans}(p) \neq \text{Ans}(q)\}$. This set collects all the kernels for which the answer to the test differs from p. $\text{Alt}(p)$ can also be written as follows: $\text{Alt}(p) = \{q \in \mathcal{P}_{\text{Test}} : V_q^\pi(\boldsymbol{\rho})V_p^\pi(\boldsymbol{\rho}) < 0\}$.

Theorem 1. *Under Assumption 1, 2, let $p \in \mathcal{P}_{\text{Test}}$, and a δ-PC algorithm with sampling rule satisfying that $N_{sa}(t)/t = t\omega_{sa} + c$ for some constant $c > 0$. Then,*

$$\liminf_{\delta\to 0} \frac{\mathbb{E}_p[\tau]}{\log(1/\delta)} \geq T^\star_{\boldsymbol{\omega}}(p), \quad (2)$$

where

$$T^\star_{\boldsymbol{\omega}}(p)^{-1} := \inf_{q\in\text{Alt}(p)} \sum_{s,a} \omega_{sa}\text{KL}_{sa}(p,q). \quad (3)$$

Theorem 1 is proved in Appendix B. The next result, proved in Appendix H.2, states that under Assumption 2, the lower bound is finite.

Proposition 1. *If Assumption 2 holds, $T^\star_{\boldsymbol{\omega}}(p)$ is finite.*

Most existing asymptotic optimal algorithms for pure exploration in MAB involve solving the optimization problem (3) (Garivier and Kaufmann, 2016; Degenne et al., 2019; Wang et al., 2021). Using a certain threshold parameter $\beta(t,\delta)$, determining whether this optimization problem exceeds $\beta(t,\delta)/t$ becomes the key to deriving the optimal stopping rule. This is the case for the celebrated Track-and-Stop algorithm (Garivier and Kaufmann, 2016) in unstructured MAB. Here, we deal with MDPs, and unfortunately, the optimization problem leading to the sample complexity lower bound is nonconvex. More precisely, the constraint set in (3) is non-convex as shown in the example below.

An example where the confusing set $\text{Alt}(p)$ is nonconvex. Let $\mathcal{M}$ be a MDP which consists of three states s_1, s_2, s_3, and π be a deterministic policy such that $\pi(a|s_i) = 1$ for all $i = 1,2,3$. The initial distribution, discount factor, and reward function are set as: $\boldsymbol{\rho} = (1/3, 1/3, 1/3)$, $\gamma = 0.9$, $r(a|s_1) = -0.88$, $r(a|s_2) = r(a|s_3) = 0.12$. We define the transition kernels $p, q^{(1)}, q^{(2)}$ as

$$[p_{ij}] = \begin{pmatrix} 0 & 0.75 & 0.25 \\ 0.5 & 0.5 & 0 \\ 0 & 0 & 1 \end{pmatrix}, [q^{(1)}_{ij}] = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, [q^{(2)}_{ij}] = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where $p_{i,j}$ is the abbreviation for $(p(s_j|s_i,a))$, and likewise for $q_{ij}^{(1)}, q_{ij}^{(2)}$. One can see that $p = (q^{(1)} + q^{(2)})/2$ and $V_p^\pi(\boldsymbol{\rho}) \approx -0.15 < 0, V_{q^{(1)}}^\pi(\boldsymbol{\rho}) \approx 0.87 > 0, V_{q^{(2)}}^\pi(\boldsymbol{\rho}) \approx 0.13 > 0$. Hence $q^{(1)}, q^{(2)} \in \text{Alt}(p)$ but $(q^{(1)} + q^{(2)})/2 = p \notin \text{Alt}(p)$.

5 Testing algorithm and its sample complexity

The pseudo-code of Policy Testing with Static Sampling (PTST) is presented in Algorithm 1. It has two main components. (i) The first makes sure that the algorithm samples the state-action pairs with a predefined allocation $\boldsymbol{\omega} \in \Sigma$. In the t-th round, we track $\boldsymbol{\omega}$ by sampling the pair minimizing the $\hat{\omega}_{sa}(t)/\omega_{sa}$, where $\hat{\omega}_{sa}(t) := N_{sa}(t)/t$ is the fraction of time (s,a) has been sampled so far, and $N_{sa}(t)$ denotes the number of times the state-action pair (s,a) has been sampled up to the t-th round.

(ii) The second component is the stopping rule. It is inspired by the following result. Introduce the threshold $\beta(t,\delta)$:

$$\beta(t,\delta) := \log(1/\delta) + (|\mathcal{S}|-1)\sum_{s,a} \log\left(e[1 + N_{sa}(t)/(|\mathcal{S}|-1)]\right). \tag{4}$$

Then according to Proposition 1 in Jonsson et al. (2020) and Lemma 15 in Al Marjani and Proutiere (2021), we have:

$$\mathbb{P}_p\left[\exists t \geq 1, \sum_{s,a} N_{sa}(t)\text{KL}_{sa}(\hat{p}_t, p) \geq \beta(t,\delta)\right] \leq \delta, \tag{5}$$

with the convention that $N_{sa}(t)\text{KL}_{sa}(\hat{p}_t(\cdot \mid s,a), p(\cdot \mid s,a)) = 0$ whenever $N_{sa}(t) = 0$. To define the stopping rule, we observe that when $\text{Ans}(\hat{p}_t) \neq \text{Ans}(p)$ or equivalently when $p \in \text{Alt}(\hat{p}_t)$, then the algorithm should not stop sampling. But if $p \in \text{Alt}(\hat{p}_t)$ and

$$\inf_{q \in \text{Alt}(\hat{p}_t)} \sum_{s,a} N_{sa}(t)\text{KL}_{sa}(\hat{p}_t, q) \geq \beta(t,\delta), \tag{6}$$

then one has $\sum_{s,a} N_{sa}(t)\text{KL}_{sa}(\hat{p}_t, p) \geq \beta(t,\delta)$, which occurs with probability less than δ in view of (5). Thus, stopping as soon as (6) holds yields a δ-PC algorithm. Unfortunately, solving the optimization problem involved in evaluating (6) is difficult due to the non-convexity of $\text{Alt}(\hat{p}_t)$. To circumvent this difficulty, we will show in the next section that (6) is equivalent to $u_{\text{NO}}(\beta(t,\delta)/t, \hat{\boldsymbol{\omega}}(t), \hat{p}_t) \geq 0$, where $u_{\text{NO}}(\sigma, \boldsymbol{\omega}, p)$ is the value of the optimization problem (NO-$\sigma, \boldsymbol{\omega}, p$):

$$\min_{q \in \mathcal{P}} V_q^\pi(\boldsymbol{\rho}) V_p^\pi(\boldsymbol{\rho}) \quad \text{s.t.} \sum_{s,a} \omega_{sa}\text{KL}_{sa}(p,q) \leq \sigma. \tag{NO-$\sigma, \boldsymbol{\omega}, p$}$$

In (NO-$\sigma, \boldsymbol{\omega}, p$), the objective function is non-convex, but the constraint defines a convex set. We will leverage this observation to solve it. To this aim, we will treat the variable q, the kernel defining the dynamics, as a *policy* in a new MDP referred to as the *reversed MDP*. This will allows us to use policy gradient algorithms to approximately solve (NO-$\sigma, \boldsymbol{\omega}, p$). Specifically, Algorithm 2 presented in the next section, will output u_{ζ_t} such that $u_{\zeta_t} \geq u_{\text{NO}}(\beta(t,\delta)/t, \hat{\boldsymbol{\omega}}(t), \hat{p}_t) \geq u_{\zeta_t} - \zeta_t$. In view of the above analysis, u_{ζ_t} can be used as our stopping rule.

Algorithm 1 Policy Testing with Static Sampling (PTST)

Input: $\pi \in \Pi, \delta \in (0,1), \boldsymbol{\omega} \in \Sigma, \{\zeta_t\}_{t \geq 1}$
Initialization Sample $(s,a) \in \mathcal{S} \times \mathcal{A}$ once if $\omega_{sa} > 0$. $t \leftarrow \sum_{s,a} \mathbb{1}\{\omega_{sa} > 0\}$.
while $u_{\zeta_t} - \zeta_t < 0$ **do**
    Sample $(s_t, a_t) \leftarrow \arg\min_{(s,a):\omega_{sa}>0} N_{sa}(t-1)/\omega_{sa}$ (tie-broken arbitrarily)
    $t \leftarrow t+1$
    Update $\hat{p}_t$, $N_{sa}(t)$, and $\hat{\boldsymbol{\omega}}(t)$
    Run Algorithm 2) with inputs $(\hat{p}_t, \zeta_t, \beta(t,\delta)/t, \boldsymbol{\omega}(t))$, and let u_{ζ_t} be its output.
end while
$\tau \leftarrow t$
Output: $\hat{\imath} \leftarrow \text{Ans}(\hat{p}_\tau)$

The following theorem, proved in Appendix E, establishes the asymptotic optimality of the PSTS algorithm.

Theorem 2. *Suppose Assumption 1, 2 hold. For any positive sequence* $\{\zeta_t\}_{t=1}^{\infty}$ *with* $\lim_{t\to\infty}\zeta_t = 0$*, Algorithm 1 satisfies that* $\mathbb{P}_p[\hat{\imath} \neq \mathrm{Ans}(p)] \leq \delta$*, and*

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_p[\tau]}{\log(1/\delta)} \leq \left(\inf_{q\in \mathrm{Alt}(p)} \sum_{s,a} \omega_{sa}\mathrm{KL}_{sa}(p,q)\right)^{-1}.$$

The proof of the theorem relies on combining concentration results with a sensitivity analysis of u_{NO}. We outline the main ideas of the proof below.

1. First, leveraging the concentration inequalities and the fact that PTST tracks a fixed allocation $\boldsymbol{\omega} \in \Sigma$, we can define, for a round T, a "good" event $\mathcal{C}_T(\xi)$ under which empirical estimates $\{\hat{p}_t\}_{t\geq\sqrt{T}}^{T}$ (resp. empirical allocation $\{\hat{\boldsymbol{\omega}}(t)\}_{t\geq\sqrt{T}}^{T}$) are very close to p (resp. $\boldsymbol{\omega}$ resp.) and such that $\mathbb{P}_p[\mathcal{C}_T(\xi)^c] < \infty$ for large enough T.

2. Next, we can show that under event $\mathcal{C}_T(\xi)$, the "ideal" stopping rule (6) will be activated when $\sigma_{\mathrm{NC}}(0,\boldsymbol{\omega},p) \approx \beta(T,\delta)/T$. Our approximate stopping rule is more conservative. To measure its conservativeness, we conduct a sensitivity analysis of u_{NO}: we prove that $c(\sigma_2-\sigma_1) \leq u_{\mathrm{NO}}(\sigma_1,\boldsymbol{\omega},p) - u_{\mathrm{NO}}(\sigma_2,\boldsymbol{\omega},p)$ for some $c>0$ (Theorem 5 in Appendix F). This result is a consequence of a series of theorems in parameterized optimization and real analysis.

3. We finally establish that if $\sigma_2 - \sigma_1 \geq \zeta_T/c$ with $\sigma_1 = \beta(T,\delta)/T$ and $\sigma_2 = \sigma_{\mathrm{NC}}(0,\boldsymbol{\omega},p)$, then Proposition 2 (see the next section) implies that $\zeta_T \leq u_{\mathrm{NO}}(\beta(T,\delta)/T) \leq u_{\zeta_T}$. Thus, PTST stops when $\sigma_{\mathrm{NC}}(0,\boldsymbol{\omega},p) \approx \beta(T,\delta)/T + \zeta_T/c \approx \log(1/\delta)T$. Or equivalently $T \approx \sigma_{\mathrm{NC}}(0,\boldsymbol{\omega},p)^{-1}\log(1/\delta) \approx T^{\star}_{\boldsymbol{\omega}}(p)\log(1/\delta)$, which completes the proof.

6 Reversed MDP and projected policy gradient

In this section, we formalize the equivalence between the optimization problem in (6) whose value is used in our stopping rule and the optimization problem (NO-$\sigma,\boldsymbol{\omega},p$). To solve the latter, we introduce the reversed MDP and analyze the convergence of a projected policy gradient method applied to this new MDP.

6.1 From non-convex constraint to non-convex objective

We first introduce an extension of the optimization problem defining our sample complexity lower bound (Theorem 1):

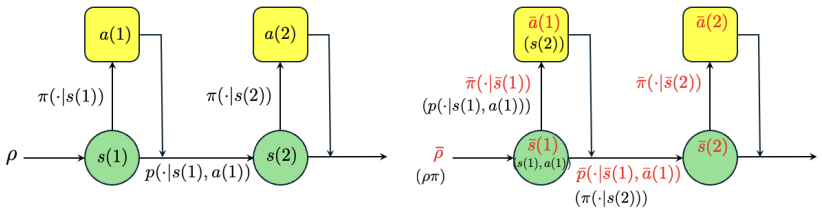
$$\inf_{q} \sum_{s,a} \omega_{sa}\mathrm{KL}_{sa}(p,q) \quad \text{s.t. } V_q^{\pi}(\boldsymbol{\rho})V_p^{\pi}(\boldsymbol{\rho}) < u, \tag{NC-$u,\boldsymbol{\omega},p$}$$

where $u \in \mathbb{R}$. The problem (NC-$u,\boldsymbol{\omega},p$) has a non-convex constraint, and we define $\sigma_{\mathrm{NC}}(u,\boldsymbol{\omega},p)$ as its value. From these definitions, the value of the optimization problem in (6) underlies the stopping rule corresponds to $\sigma_{\mathrm{NC}}(0,\hat{\boldsymbol{\omega}}(t),\hat{p}_t)$. The following proposition, proved in Appendix C, formalizes the relationship between the values $u_{\mathrm{NO}}(\sigma,\boldsymbol{\omega},p)$ and $\sigma_{\mathrm{NC}}(u,\boldsymbol{\omega},p)$ associated with the problems (NO-$\sigma,\boldsymbol{\omega},p$) and (NC-$u,\boldsymbol{\omega},p$), respectively.

Proposition 2. *Assume that Assumption 1 holds and that* $p \in \mathcal{P}_{\mathrm{Test}}$*. We have:*

$$\textit{for all } \sigma \geq 0 \ \textit{ such that } \ u_{\mathrm{NO}}(\sigma,\boldsymbol{\omega},p) > \min_{q\in\mathcal{P}} V_p^{\pi}(\boldsymbol{\rho})V_q^{\pi}(\boldsymbol{\rho}), \quad \sigma_{\mathrm{NC}}(u_{\mathrm{NO}}(\sigma,\boldsymbol{\omega},p),\boldsymbol{\omega},p) = \sigma,$$
$$\textit{for all } u \in (\min_{q\in\mathcal{P}} V_p^{\pi}(\boldsymbol{\rho})V_q^{\pi}(\boldsymbol{\rho}), u_{\mathrm{NO}}(0,\boldsymbol{\omega},p)], \quad u_{\mathrm{NO}}(\sigma_{\mathrm{NC}}(u,\boldsymbol{\omega},p),\boldsymbol{\omega},p) = u.$$

This proposition establishes that the mappings $u_{\mathrm{NO}}(\cdot,\boldsymbol{\omega},p)$ and $\sigma_{\mathrm{NC}}(\cdot,\boldsymbol{\omega},p)$ are inverses of each other. While this result may seem intuitive, it does not generally hold in non-convex settings. We identify general conditions, outlined in Assumption 3 in Appendix C, on the objective function and constraint set that ensure the validity of this inverse relationship. We verify that these conditions are satisfied in our setting. Under these conditions, we can replace the infimum and strict inequality in (NC-$u,\boldsymbol{\omega},p$) with a minimum and a weak inequality, respectively. From this, the result follows: (i) If $\sigma_{\mathrm{NC}}(u,\boldsymbol{\omega},p) \leq \sigma$, then there exists q such that $V_p^{\pi}(\boldsymbol{\rho})V_q^{\pi} \leq u$ and $\sum_{sa}\omega_{sa}\mathrm{KL}_{sa}(p,q) \leq \sigma$, implying $u_{\mathrm{NO}}(\sigma,\boldsymbol{\omega},p) \leq u$. Conversely, if $u_{\mathrm{NO}}(\sigma,\boldsymbol{\omega},p) \leq u$, then $\sigma_{\mathrm{NC}}(u,\boldsymbol{\omega},p) \leq \sigma$, which



Initial MDP $\mathcal{M}$ Reversed MDP $\bar{\mathcal{M}}$

Figure 1: From the initial MDP (left) to the reversed MDP (right): In the reversed MDP, variables are shown in red; their initial MDP counterparts are shown in black.

is equivalent to stating that if $\sigma_{\text{NC}}(u, \boldsymbol{\omega}, p) > \sigma$, then $u_{\text{NO}}(\sigma, \boldsymbol{\omega}, p) > u$. (ii) We further show that if $\sigma_{\text{NC}}(u, \boldsymbol{\omega}, p) \geq \sigma$, then $u_{\text{NO}}(\sigma, \boldsymbol{\omega}, p) \geq u$. Combining (i) and (ii) directly implies that if $\sigma_{\text{NC}}(u, \boldsymbol{\omega}, p) = \sigma$, then $u_{\text{NO}}(\sigma, \boldsymbol{\omega}, p) = u$.

Proposition 2 is a central component of our approach to solving the lower-bound optimization problem and thus to developing an instance-optimal algorithm. While we establish that the required conditions hold for the policy testing task, we can also prove their validity for the policy evaluation task (see Appendix C.2). Extending this result to other pure exploration tasks, such as the best policy identification, remains an interesting direction for future work.

6.2 The reversed MDP

We can interpret the dual optimization problem (NO-$\sigma, \boldsymbol{\omega}, p$) as a policy optimization problem in a new MDP. This MDP $\bar{\mathcal{M}} = (\bar{\mathcal{S}}, \bar{\mathcal{A}}, \bar{p}, \bar{r}, \bar{\boldsymbol{\rho}}, \gamma)$ is referred to as reversed MDP, since the roles of policy π and transition kernel p are swapped. $\bar{\mathcal{M}}$ is constructed as follows. The state and action spaces are $\bar{\mathcal{S}} := \mathcal{S} \times \mathcal{A}$ and $\bar{\mathcal{A}} = \mathcal{S}$. The initial state distribution $\bar{\boldsymbol{\rho}}$ is such that for all (s, a), $\bar{\boldsymbol{\rho}}(s, a) = \rho_s \pi(a \mid s)$. In state $\bar{s} = (s, a) \in \bar{\mathcal{S}}$, a policy $\bar{\pi}$ takes an action $\bar{a} = s' \in \bar{\mathcal{A}}$ with probability $p(s'|s, a)$. Given an action $\bar{a} = s'$ selected in $\bar{s}$, the system moves to state $\bar{s}' = (s', a')$ with probability $\pi(a'|s')$ (all other transitions occur with probability 0), so that $\bar{p}(\bar{s}' = (s'', a')|\bar{s}, \bar{a} = s') = \pi(a'|s')1_{\{s''=s'\}}$. The reward function, $\bar{r} : \bar{\mathcal{S}} \times \bar{\mathcal{A}} \to \mathbb{R}$ is defined as $\bar{r}(\bar{s}, \bar{a}) = r(s, a)$ if $\bar{s} = (s, a)$. The reversed MDP $\bar{\mathcal{M}}$ is illustrated in Figure 1.

For the reversed MDP, the discounted state-visitation distribution starting at $\bar{s} \in \bar{\mathcal{S}}$ is defined as:

$$\forall \bar{s}' \in \bar{\mathcal{S}},\ d^{\bar{\pi}}_{\bar{p}, \bar{s}}(\bar{s}') := (1-\gamma)\mathbb{E}^{\pi}_{p}[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}\{\bar{s}(t) = \bar{s}'\} \mid \bar{s}(0) = \bar{s}],$$

which is equal to $d^{\pi}_{p,s,a}(s', a')$ when $\bar{s} = (s, a)$ and $\bar{s}' = (s', a')$. For any $\boldsymbol{\mu} \in \Delta(\bar{\mathcal{S}})$, we define $\bar{d}^{\pi}_{p,\boldsymbol{\mu}}(\bar{s}') := \sum_{\bar{s} \in \bar{\mathcal{S}}} \mu_{\bar{s}} \bar{d}^{\pi}_{p,\bar{s}}(\bar{s}')$. The state and state-action value functions of $\bar{\mathcal{M}}$ are defined as: for all $(\bar{s}, \bar{a})$:

$$\bar{V}^{\bar{\pi}}_{\bar{p}}(\bar{s}) := \mathbb{E}^{\pi}_{p}[\sum_{t=0}^{\infty} \gamma^t \bar{r}(\bar{s}(t), \bar{a}(t))|\bar{s}(0) = \bar{s}], \tag{7}$$

$$\bar{Q}^{\bar{\pi}}_{\bar{p}}(\bar{s}, \bar{a}) := \mathbb{E}^{\pi}_{p}[\sum_{t=0}^{\infty} \gamma^t \bar{r}(\bar{s}(t), \bar{a}(t))|(\bar{s}(0), \bar{a}(0)) = (\bar{s}, \bar{a})]. \tag{8}$$

For any $\boldsymbol{\mu} \in \Delta(\bar{\mathcal{S}})$, we define $\bar{V}^{\pi}_{p}(\boldsymbol{\mu}) := \sum_{\bar{s} \in \bar{\mathcal{S}}} \mu_{\bar{s}} \bar{V}^{\pi}_{p}(\bar{s})$. Observe that $\bar{V}^{\bar{\pi}}_{\bar{p}}(\bar{s}) = Q^{\pi}_{p}(s, a)$ if $\bar{s} = (s, a)$, and $\bar{Q}^{\bar{\pi}}_{\bar{p}}(\bar{s}, \bar{a}) = r(s, a) + \gamma V^{\pi}_{p}(s')$ if $(\bar{s}, \bar{a}) = (s, a, s')$. We simply deduce that for each $s \in \mathcal{S}$, $V^{\pi}_{p}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \bar{V}^{\bar{\pi}}_{\bar{p}}(s, a)$, and $V^{\pi}_{p}(\boldsymbol{\rho}) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \rho_s \pi(a|s) \bar{V}^{\bar{\pi}}_{\bar{p}}(s, a) = \bar{V}^{\bar{\pi}}_{\bar{p}}(\bar{\rho})$. Thus, optimizing transition kernel q in (NO-$\sigma, \boldsymbol{\omega}, p$) is equivalent to optimizing the policy against in the reversed MDP. More precisely, (NO-$\sigma, \boldsymbol{\omega}, p$) is equivalent to:

$$\min_{\bar{\pi} \in \mathcal{P}} \bar{V}^{\bar{\pi}}_{\bar{p}}(\bar{\boldsymbol{\rho}}) V^{\pi}_{p}(\boldsymbol{\rho}) \quad \text{s.t.} \sum_{s,a} \omega_{sa} \text{KL}_{sa}(p, \bar{\pi}) \leq \sigma. \tag{9}$$

Reformulating (NO-$\sigma, \boldsymbol{\omega}, p$) as a policy optimization problem in the reversed MDP offers a key advantage: it allows us to leverage recent advances in the convergence analysis of policy gradient methods. In the next subsection, we build on this reformulation to analyze the resulting optimization algorithm.

6.3 Projected policy gradient

We apply policy gradient methods to solve (9). This problem is equivalent to (NO-$\sigma, \boldsymbol{\omega}, p$) and consists in optimizing the policy of the reversed MDP. This policy corresponds to the transition kernel in the initial MDP. Before we present our algorithm, we provide preliminary results that will help its analysis. These results are the equivalent for our reversed MDP of the performance difference lemma (Kakade and Langford, 2002), the policy gradient theorem (Sutton et al., 1999), and the smoothness lemma (Agarwal et al., 2021) for regular MDPs. Since as shown above for the reversed MDP, $\bar{V}_{\bar{p}}^{\bar{\pi}}(\bar{s}) = Q_p^{\pi}(s,a)$ and the objective function $\bar{V}_{\bar{p}}^{\bar{\pi}}(\bar{\rho}) = V_p^{\pi}(\boldsymbol{\rho})$ can be expressed using $Q_p^{\pi}(s,a)$, these results will essentially describe how $Q_p^{\pi}(s,a)$ evolves w.r.t. p. We start with the performance difference lemma, which interestingly, for the reversed MDP, corresponds to the celebrated simulation lemma (Kearns and Singh, 2002) (also see Lemma A.1 in Vemula et al. (2023)).

Lemma 2 (Simulation / performance difference lemma)**.** *For any $p, \tilde{p} \in \mathcal{P}$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$Q_p^{\pi}(s,a) - Q_{\tilde{p}}^{\pi}(s,a) = \frac{\gamma}{1-\gamma} \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} d_{p,s,a}^{\pi}(s',a') \sum_{s'' \in \mathcal{S}} V_{\tilde{p}}^{\pi}(s'') \left(p(s'' \mid s',a') - \tilde{p}(s'' \mid s',a')\right).$$

Lemma 2 is proved in Appendix D.4 for completeness. It directly implies that $Q_p^{\pi}(s,a)$ is continuous in p, a property that is used in the proof of Proposition 2.

Lemma 3 (Policy gradient)**.** *For each $s, s', s'' \in \mathcal{S}$, $a, a' \in \mathcal{A}$, we have*

$$\frac{\partial Q_p^{\pi}(s,a)}{\partial p(s'' \mid s',a')} = \frac{1}{(1-\gamma)} d_{p,s,a}^{\pi}(s',a') \left(r(s',a') + \gamma V_p^{\pi}(s'')\right), \tag{10}$$

$$\frac{\partial V_p^{\pi}(\boldsymbol{\rho})}{\partial p(s' \mid s,a)} = \frac{1}{(1-\gamma)} d_{p,\boldsymbol{\rho}}^{\pi}(s,a) \left(r(s,a) + \gamma V_p^{\pi}(s')\right). \tag{11}$$

The proof of this lemma is presented in Appendix D.5. It provides an explicit expression of the gradient $\nabla_{\bar{\pi}} \bar{V}_{\bar{p}}^{\bar{\pi}}(\bar{\boldsymbol{\rho}})$ used in our policy gradient algorithm. The final result concerns the smoothness of the gradient, and it can be established using tools from the theory of the policy gradient (Agarwal et al., 2021). It will be useful when assessing the convergence rate of our algorithm. Refer to Appendix D.6 for a proof.

Lemma 4 (Smoothness)**.** *For any $p, \tilde{p} \in \mathcal{P}$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\left\| \nabla_p Q_p^{\pi}(s,a) - \nabla_p Q_{\tilde{p}}^{\pi}(s,a) \right\|_2 \leq \frac{2\gamma |\mathcal{S}|}{(1-\gamma)^3} \|p - \tilde{p}\|_2 .$$

Algorithm 2 Projected Policy Gradient

Input: $(p, \zeta, \sigma, \boldsymbol{\omega})$
Initialization:
Define $\bar{\mathcal{S}}, \bar{\mathcal{A}}, \bar{s}, \bar{a}, \bar{r}, \bar{p}$, and $\bar{V}_{\bar{p}}^{\bar{\pi}}(\bar{\boldsymbol{\rho}})$ as in Section 6.2.
$M \leftarrow \lceil \|1/\bar{\boldsymbol{\rho}}\|_{\infty}^2 (128\gamma |\bar{\mathcal{S}}||\bar{\mathcal{A}}|)/((1-\gamma)^5 \zeta) \rceil$, $L \leftarrow 2\gamma \left|\bar{\mathcal{A}}\right| / (1-\gamma)^3 |V_p^{\pi}(\boldsymbol{\rho})|$
$\bar{\pi}^{(0)} \leftarrow \text{Unif}(\bar{\mathcal{A}})$
for $k = 0, 1, \ldots, M-1$ **do**
    $\bar{\pi}^{(k+1)} \leftarrow \mathbf{proj}_{\bar{\Pi}_{\sigma}^{p}} \left(\bar{\pi}^{(k)} - \frac{V_{\tilde{p}}^{\pi}(\boldsymbol{\rho})}{L} \nabla_{\bar{\pi}} \bar{V}_{\bar{p}}^{\bar{\pi}^{(k)}}(\bar{\boldsymbol{\rho}}) \right)$
end for
Output: $V_p^{\pi}(\boldsymbol{\rho}) \bar{V}_{\bar{p}}^{\bar{\pi}^{(M)}}(\bar{\boldsymbol{\rho}})$

We are ready to present the policy gradient algorithm used to solve (9). In this problem, the constraint set $\Pi_{\sigma}^{p} := \{\bar{\pi} \in \mathcal{P} : \sum_{s,a} \omega_{sa} \text{KL}_{sa}(p, \bar{\pi}) \leq \sigma\}$ is closed and convex. The algorithm, whose pseudo-code is presented in Algorithm 2, is hence a projected policy gradient algorithm, where in each iteration we make sure that the constraint is satisfied by projecting on Π_{σ}^{p}. Specifically, the policy updates of the algorithm are:

$$\bar{\pi}^{(k+1)} = \mathbf{proj}_{\Pi_{\sigma}^{p}} \left(\bar{\pi}^{(k)} - \frac{V_p^{\pi}(\boldsymbol{\rho})}{L} \nabla_{\bar{\pi}} \bar{V}_{\bar{p}}^{\bar{\pi}^{(k)}}(\bar{\boldsymbol{\rho}}) \right), \tag{12}$$

where $\mathbf{proj}_{\Pi_\sigma^p}(x)$ denotes the projection of x onto Π_σ^p in the Euclidean norm and $L = \frac{2\gamma|\bar{\mathcal{A}}|}{(1-\gamma)^3|V_p^\pi(\boldsymbol{\rho})|}$. Observe that the value of (9) is $u_{\mathrm{NO}}(\sigma, \boldsymbol{\omega}, p)$. The following theorem provides a finite time analysis of the projected gradient algorithm.

Theorem 3. *Under Assumption 1, 2, the projected policy gradient method* (12) *satisfies, for all* $k \geq 1$, $\bar{V}_{\bar{p}}^{\bar{\pi}^{(k)}}(\bar{\boldsymbol{\rho}})V_p^\pi(\boldsymbol{\rho}) - u_{\mathrm{NO}}(\sigma, \boldsymbol{\omega}, p)$ *is upper bounded by*

$$\max\left\{\frac{128\gamma|\bar{\mathcal{S}}||\bar{\mathcal{A}}|}{(1-\gamma)^5 k}\left\|\frac{1}{\bar{\boldsymbol{\rho}}}\right\|_\infty^2, \left(\frac{1}{\sqrt{2}}\right)^k\left(\bar{V}_{\bar{p}}^{\bar{\pi}^{(0)}}(\bar{\boldsymbol{\rho}})V_p^\pi(\boldsymbol{\rho}) - u_{\mathrm{NO}}(\sigma, \boldsymbol{\omega}, p)\right)\right\}.$$

The proof of Theorem 3 is provided in Appendix D. This theorem motivates the design of Algorithm 2, which aims to approximate $u_{\mathrm{NO}}(\sigma, \boldsymbol{\omega}, p)$ to within a specified accuracy level ζ.

Finally, we note that when used in our policy testing algorithm (Algorithm 1), the unknown transition kernel p in the input of Algorithm 2 is replaced by its empirical estimator $\hat{p}_t$, ζ is replaced by ζ_t, $\boldsymbol{\omega}$ by $\hat{\boldsymbol{\omega}}(t)$, and σ by the threshold $\beta(t, \delta)/t$.

7 Experiments

In this section, we test the proposed method in several scenarios. As a comparative method, we consider the KLB-TS algorithm proposed by Al Marjani and Proutiere (2021). Unlike the setting of this paper, their approach aims to identify the best policy from among multiple candidate policies. To adapt their method to our setting, we prepared two policies: one identical to π and another policy π' such that $V_p^{\pi'}(\boldsymbol{\rho}) = 0$. We enforce uniform sampling for the sampling rule. The stopping rule of KLB-TS is based on an upper bound derived from a convexification of the original minimax optimization problem. Notably, this stopping rule does not explicitly utilize the fact that $V_p^{\pi'}(\boldsymbol{\rho}) = 0$.

We conduct experiments using three MDP settings: $|\mathcal{S}| = |\mathcal{A}| = 2$, $|\mathcal{S}| = |\mathcal{A}| = 3$, and $|\mathcal{S}| = |\mathcal{A}| = 5$. In all cases, the discount factor is set to $\gamma = 0.9$ and the initial state distribution is uniform over all states. The reward function $r(s, a)$, the transition kernel $p(\cdot \mid s, a)$, and the policy π are specified in Table 1, Table 2, and Table 3 in Appendix I for the respective settings. For each setting, we vary δ from 10^{-15} to 10^{-2}. The results for $|\mathcal{S}| = |\mathcal{A}| = 2$ are shown in the left panel of Figure 2, those for $|\mathcal{S}| = |\mathcal{A}| = 3$ in the middle panel, and those for $|\mathcal{S}| = |\mathcal{A}| = 5$ in the right panel. In all three cases, PTST outperforms KLB-TS for all values of δ. For further details, please refer to Appendix I.

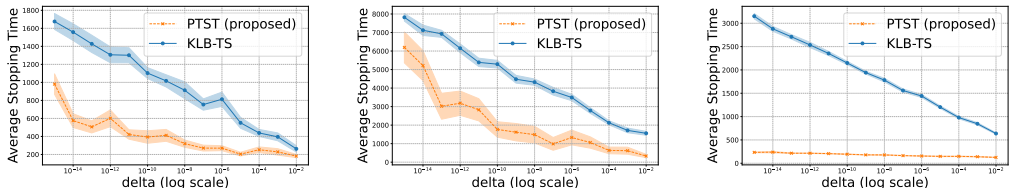


Figure 2: Comparison of average stopping times and delta for the proposed algorithm and KLB-TS. The left, center, and right panels correspond to $|\mathcal{S}| = |\mathcal{A}| = 2, 3, 5$, respectively. Results are averaged over 30 instances. Error bars indicate the standard error of the mean.

8 Conclusion

In this paper, we formulated the policy testing problem in MDPs and characterized its instance-specific complexity. To the best of our knowledge, this is the first instance-specific optimal and computationally tractable algorithm for pure exploration in MDPs. The key components of our approach include a novel transformation of a non-convex problem and the use of policy gradients in the reversed MDP. Although our work specifically focuses on policy testing, we anticipate that our method can be extended to other pure exploration tasks, such as policy evaluation. This potential is discussed in Appendix C.2, where we outline the corresponding assumptions and minimization problems. Currently, our results are limited to the static sampling setting; extending them to adaptive sampling is an interesting direction for future work. Our findings could serve as a foundation for developing more efficient algorithms for pure exploration in MDPs.

Acknowledgements

Kaito Ariu is supported by JSPS KAKENHI Grant No. 25K21291.

References

Acemoglu, D. (2008). *Introduction to modern economic growth*. Princeton university press.

Agarwal, A., Kakade, S., and Yang, L. F. (2020). Model-based reinforcement learning with a generative model is minimax optimal. In Abernethy, J. and Agarwal, S., editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 67–83. PMLR.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76.

Al Marjani, A., Garivier, A., and Proutiere, A. (2021). Navigating to the best policy in markov decision processes. *Advances in Neural Information Processing Systems*, 34:25852–25864.

Al Marjani, A. and Proutiere, A. (2021). Adaptive sampling for best policy identification in markov decision processes. In *International Conference on Machine Learning*, pages 7459–7468. PMLR.

Al-Marjani, A., Tirinzoni, A., and Kaufmann, E. (2023). Towards instance-optimality in online pac reinforcement learning. *arXiv preprint arXiv:2311.05638*.

Audibert, J.-Y. and Bubeck, S. (2010). Best arm identification in multi-armed bandits. In *COLT-23th Conference on learning theory-2010*, pages 13–p.

Berge, C. (1877). *Topological spaces: Including a treatment of multi-valued functions, vector spaces and convexity*. Oliver & Boyd.

Bhandari, J. and Russo, D. (2024). Global optimality guarantees for policy gradient methods. *Operations Research*, 72(5):1906–1927.

Chen, L. and Li, J. (2015). On the optimal sample complexity for best arm identification. *arXiv preprint arXiv:1511.03774*.

Degenne, R. and Koolen, W. M. (2019). Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems*, 32.

Degenne, R., Koolen, W. M., and Ménard, P. (2019). Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems*, 32.

Ding, D., Wei, C.-Y., Zhang, K., and Ribeiro, A. (2023). Last-iterate convergent policy gradient primal-dual methods for constrained mdps. *Advances in Neural Information Processing Systems*, 36:66138–66200.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. (2012). Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in neural information processing systems*, 25.

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR.

Gheshlaghi Azar, M., Munos, R., and Kappen, H. J. (2013). Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91:325–349.

Jedra, Y. and Proutiere, A. (2020). Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017.

Jonsson, A., Kaufmann, E., Ménard, P., Darwiche Domingues, O., Leurent, E., and Valko, M. (2020). Planning in markov decision processes with gap-dependent sample complexity. *Advances in Neural Information Processing Systems*, 33:1253–1263.

Kakade, S. and Langford, J. (2002). Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 267–274.

Kano, H., Honda, J., Sakamaki, K., Matsuura, K., Nakamura, A., and Sugiyama, M. (2019). Good arm identification via bandit feedback. *Machine Learning*, 108:721–745.

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42.

Kearns, M. and Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49:209–232.

Khamaru, K., Xia, E., Wainwright, M. J., and Jordan, M. I. (2021). Instance-optimality in optimal value estimation: Adaptivity via variance-reduced q-learning. *arXiv preprint arXiv:2106.14352*.

Kitamura, T., Kozuno, T., Tang, Y., Vieillard, N., Valko, M., Yang, W., Mei, J., Ménard, P., Azar, M. G., Munos, R., et al. (2023). Regularization and variance-weighted regression achieves minimax optimality in linear mdps: Theory and practice. In *International Conference on Machine Learning*, pages 17135–17175. PMLR.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. (2024). Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics*, 52(1):233 – 260.

Locatelli, A., Gutzeit, M., and Carpentier, A. (2016). An optimal algorithm for the thresholding bandit problem. In *International Conference on Machine Learning*, pages 1690–1698. PMLR.

Montenegro, A., Mussi, M., Papini, M., and Metelli, A. M. (2024). Last-iterate global convergence of policy gradients for constrained reinforcement learning. *Advances in Neural Information Processing Systems*, 37:126363–126416.

Narang, A., Wagenmaker, A., Ratliff, L., and Jamieson, K. G. (2024). Sample complexity reduction via policy difference estimation in tabular reinforcement learning. *Advances in Neural Information Processing Systems*, 37:22772–22826.

Nesterov, Y. (2013). Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition.

Reverdy, P., Srivastava, V., and Leonard, N. E. (2016). Satisficing in multi-armed bandit problems. *IEEE Transactions on Automatic Control*, 62(8):3788–3803.

Russo, A. and Pacchiano, A. (2025). Adaptive exploration for multi-reward multi-policy evaluation. *arXiv preprint arXiv:2502.02516*.

Russo, A. and Vannella, F. (2024). Multi-reward best policy identification. *Advances in Neural Information Processing Systems*, 37:105583–105662.

Russo, D. and Van Roy, B. (2022). Satisficing in time-sensitive bandit learning. *Mathematics of Operations Research*, 47(4):2815–2839.

Soare, M., Lazaric, A., and Munos, R. (2014). Best-arm identification in linear bandits. *Advances in neural information processing systems*, 27.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.

Tabata, K., Nakamura, A., Honda, J., and Komatsuzaki, T. (2020). A bad arm existence checking problem: How to utilize asymmetric problem structure? *Machine learning*, 109(2):327–372.

Tao, T. (2011). *An introduction to measure theory*, volume 126. American Mathematical Soc.

Taupin, J., Jedra, Y., and Proutiere, A. (2023). Best policy identification in linear mdps. In *2023 59th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1–8. IEEE.

Tirinzoni, A., Al Marjani, A., and Kaufmann, E. (2022). Near instance-optimal pac reinforcement learning for deterministic mdps. *Advances in neural information processing systems*, 35:8785–8798.

Uehara, M., Shi, C., and Kallus, N. (2022). A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*.

Vemula, A., Song, Y., Singh, A., Bagnell, D., and Choudhury, S. (2023). The virtues of laziness in model-based rl: A unified objective and algorithms. In *International Conference on Machine Learning*, pages 34978–35005. PMLR.

Wang, P.-A., Tzeng, R.-C., and Proutiere, A. (2021). Fast pure exploration via frank-wolfe. *Advances in Neural Information Processing Systems*, 34:5810–5821.

Wang, S., Blanchet, J., and Glynn, P. (2024). Optimal sample complexity for average reward markov decision processes. In *The Twelfth International Conference on Learning Representations*.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. (2003). Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, page 125.

Xiao, L. (2022). On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36.

Zalinescu, C. (2002). *Convex analysis in general vector spaces*. World scientific.

Zanette, A., Kochenderfer, M. J., and Brunskill, E. (2019). Almost horizon-free structure-aware best policy identification with a generative model. *Advances in Neural Information Processing Systems*, 32.

A Additional related work

Comparison with Thresholding Bandits. When compared with the bandit literature, the policy testing problem can be regarded as a generalization of the thresholding bandit problem (Chen and Li, 2015; Locatelli et al., 2016; Degenne and Koolen, 2019; Wang et al., 2021), where the goal is to adaptively sample each arm and identify those whose mean reward exceeds a given threshold; our work generalizes this concept to the value function setting in MDPs. Similar settings using known thresholds in MABs have also led to other important variants, such as the good arm identification (Kano et al., 2019) and bad arm existence checking problems (Tabata et al., 2020). Such a setting has applications in scenarios with practical constraints on the horizon, such as in timely recommendations (Kano et al., 2019). It can also be viewed as an application of the concept of a satisficing objective (Russo and Van Roy, 2022; Reverdy et al., 2016).

Analysis of Policy Gradient Methods. Our analysis of the projected policy gradient method (Section 6.3) is partly based on recent advances in the analysis of convergence rates for policy gradient methods (Xiao, 2022; Agarwal et al., 2021). While studies on the convergence rate of policy gradients with constraints exist (Ding et al., 2023; Montenegro et al., 2024), most focus on linear constraints on the state visitation distribution. In contrast, our results derive convergence rates under (convex but) nonlinear constraints, and provide rigorous sensitivity analysis of solutions with respect to these constraints (see Sections 5 and F), which is of independent technical interest.

Policy Evaluation. Furthermore, our technique suggests that the instance-specific optimality is likely to be achievable for policy evaluation problems, including off-policy evaluation (Uehara et al., 2022) and more generally (Russo and Pacchiano, 2025). We discuss the possibility of extending our results to policy evaluation in Section C.2.

B Instance-specific sample complexity lower bound–Proof of Theorem 1

Proof of Theorem 1. Consider two cases, (i) $\inf_{q\in \mathrm{Alt}(p)} \sum_{s,a} \mathrm{KL}_{sa}(p,q) = \infty$; (ii) $\inf_{q\in \mathrm{Alt}(p)} \sum_{s,a} \mathrm{KL}_{sa}(p,q) < \infty$. Observe that our theorem holds directly in case (i), one only needs to focus on the case (ii). Together with Assumption 2, case (ii) implies that there exists $\tilde{q} \in \mathrm{Alt}(p)$ such that $\forall s, a,\ \mathrm{KL}_{sa}(p,\tilde{q}) < \infty$.

Let $q \in \mathrm{Alt}(p)$, which is a nonempty set thanks to Assumption 1. Let $\mathbb{P}_p$ and $\mathbb{P}_q$ denote the probability measure generated by p and q respectively. According to property (i) of the δ-PC algorithm definition, the stopping time τ is almost surely finite. Using Lemma 1 in Al Marjani and Proutiere (2021) and the classical data processing inequality (see e.g. Lemma 1 in Kaufmann et al. (2016)), we derive that for any $\mathcal{F}_\tau$-measurable event E,

$$\sum_{s,a} \mathbb{E}_p[N_{sa}(\tau)]\mathrm{KL}_{sa}(p,q) \geq \mathrm{kl}(\mathbb{P}_p[E], \mathbb{P}_q[E]), \tag{13}$$

where $\mathrm{kl}(a,b)$ denotes the Kullback-Leibler (KL) divergence between two Bernoulli distributions with means a and b. With choice $E = \{\hat{\imath} = \mathrm{Ans}(p)\}$, the definition of δ -PC algorithm (Definition 1) and the assumption that $q \in \mathrm{Alt}(p)$ yield that $\mathbb{P}_p[E] \geq 1-\delta$ and $\mathbb{P}_q[E] \leq \delta$. After applying the monotonicity of KL divergence, we obtain $\mathrm{kl}(\mathbb{P}_p[E], \mathbb{P}_q[E]) \geq \mathrm{kl}(\delta, 1-\delta)$. As (13) holds for any $q \in \mathrm{Alt}(p)$,

$$\begin{aligned}
\mathrm{kl}(\delta,1-\delta) &\leq \inf_{q\in\mathrm{Alt}(p)} \mathbb{E}_p[\tau] \sum_{s,a} \frac{\mathbb{E}_p[N_{sa}(\tau)]}{\mathbb{E}_p[\tau]} \mathrm{KL}_{sa}(p,q) \\
&\leq \mathbb{E}_p[\tau] \left(\inf_{q\in\mathrm{Alt}(p)} \sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p,q) + \frac{c}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]} \right) \qquad (14) \\
&\leq \mathbb{E}_p[\tau] \left(\sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p,\tilde{q}) + c \right), \qquad (15)
\end{aligned}$$

where the last inequality follows as $\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq 1$ and $\tilde{q} \in \text{Alt}(p)$. Since $\left(\sum_{s,a} \omega_{sa} \text{KL}_{sa}(p, \tilde{q}) + c\right)$ is finite, we conclude $\mathbb{E}_{\boldsymbol{\mu}}[\tau] \to \infty$ as $\delta \to 0$ from (15). Rearranging (14) yields that

$$\left(\inf_{q \in \text{Alt}(p)} \sum_{s,a} \omega_{sa} \text{KL}_{sa}(p, q) + \frac{c}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}\right)^{-1} \leq \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}{\text{kl}(\delta, 1-\delta)} \tag{16}$$

Using the fact that $\text{kl}(\delta, 1-\delta) \sim \log(1/\delta)$ and $\mathbb{E}_{\boldsymbol{\mu}}[\tau] \to \infty$ when δ goes to zero, one can conclude the theorem by taking the limit inferior on both sides of (16). □

C Dual interpretation of the non-convex problems–Proof of Proposition 2

Here, the optimization problems (NC-$u, \boldsymbol{\omega}, p$) and (NO-$\sigma, \boldsymbol{\omega}, p$) are abstracted as the following two optimization problems, respectively.

$$\inf_{x \in \mathcal{X}} h(x) \quad \text{s.t. } g(x) < u, \tag{NC-u}$$

and

$$\min_{x \in \mathcal{X}} g(x) \quad \text{s.t. } h(x) \leq \sigma, \tag{NO-σ}$$

where $h, g : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$, $\mathcal{X} \subset \mathbb{R}^d$ is a nonempty set, $d \in \mathbb{N}$, and $u, \sigma \in \mathbb{R}$. Let $\sigma_{\text{NC}}(u)$ and $u_{\text{NO}}(\sigma)$ denote the value of (NC-u) and that of (NO-σ), respectively. As one can see, in Section 6.1, we make the substitutions $\mathcal{X} = \mathcal{P}$, $x = q$, $h(q) = \sum_{sa} \omega_{sa} \text{KL}_{sa}(p, q)$, and $g(q) = V_p^{\pi}(\boldsymbol{\rho}) V_q^{\pi}(\boldsymbol{\rho})$. As stated in Proposition 3, the desired bijection holds under the mild assumptions in Assumption 3. In Appendix C.1, we verify that these assumptions hold for the aforementioned substitution.

Assumption 3. *The following conditions hold.*

(a) $\exists \underline{x} \in \mathcal{X}$ *such that* $h(\underline{x}) \leq 0$

(b) $\min_{x \in \mathcal{X}} g(x) < 0$.

(c) All local minimums of g *(*h *resp.) in* $\mathcal{X}$ *are global minimums of* g *(*h *resp.).*

(d) h, g *are continuous mappings.*

The following proposition shows that, under Assumption 3, there exists a bijection between problems NC-u and NO-σ.

Proposition 3. *Under Assumption 3, we have*

$$\textit{for all } \sigma \geq 0 \textit{ such that } u_{\text{NO}}(\sigma) > \min_{x \in \mathcal{X}} g(x), \quad \sigma_{\text{NC}}(u_{\text{NO}}(\sigma)) = \sigma, \tag{17}$$

$$\textit{for all } u \in (\min_{x \in \mathcal{X}} g(x), u_{\text{NO}}(0)] \quad u_{\text{NO}}(\sigma_{\text{NC}}(u)) = u. \tag{18}$$

In words, $\sigma_{\text{NC}}(u)$ and $u_{\text{NO}}(\sigma)$ are one-to-one mappings as decreasing functions, and knowing the value of $u_{\text{NO}}(\sigma)$ would be equivalent to knowing the value of $\sigma_{\text{NC}}(u)$.

Proof of Proposition 3. We first show (17). Let $\sigma \geq 0$ and $u = u_{\text{NO}}(\sigma) > \min_{x \in \mathcal{X}} g(x)$. By Lemma 5 and Lemma 6, we get $\sigma_{\text{NC}}(u) \leq \sigma$ and $\sigma_{\text{NC}}(u) \geq \sigma$ respectively, which directly yield (17). For proving (18), we consider $u \in (\min_{x \in \mathcal{X}} g(x), u_{\text{NO}}(0)]$. Using intermediate value theorem and the continuity of $u_{\text{NO}}(\cdot)$ proved in Lemma 20 given in Appendix G.3, we deduce that there is $\sigma \in [0, \infty)$ such that $u_{\text{NO}}(\sigma) = u$. As a consequence,

$$u_{\text{NO}}(\sigma_{\text{NC}}(u)) = u_{\text{NO}}(\sigma_{\text{NC}}(u_{\text{NO}}(\sigma))) = u_{\text{NO}}(\sigma) = u,$$

where the second equation is the application of (17).

□

Lemma 5. *Under Assumption 3, whenever $\sigma \geq 0, u \in (\min_{x\in\mathcal{X}} g(x), \infty)$, and $u_{\text{NO}}(\sigma) \leq u$, then $\sigma_{\text{NC}}(u) \leq \sigma$ holds.*

Proof of Lemma 5. As $u_{\text{NO}}(\sigma) \leq u$, $\exists x' \in \mathcal{X}$, such that $h(x') \leq \sigma$, and $g(x') \leq u$. If $g(x') < u$ (case i), then x' is a feasible point for (NC-u), and therefore $\sigma_{\text{NC}}(u) \leq h(x') \leq \sigma$. We next consider the case where $g(x') = u$ (case ii).
Since $g(x') = u > \min_{x\in\mathcal{X}} g(x)$, x' is not a global minimum of g in $\mathcal{X}$. Hence, by Assumption 3-(c), we deduce that x' is not a local minimum either. Thus, there is a sequence of $\{x_n\}_{n=1}^{\infty}$ such that $x_n \xrightarrow{n\to\infty} x'$ and $g(x_n) < u$, $\forall n$. As a consequence of Assumption 3-(d), $\lim_{n\to\infty} h(x_n) = h(x') \leq \sigma$, which yields $\sigma_{\text{NC}}(u) \leq \sigma$. □

Lemma 6. *Whenever $\sigma \geq 0, u \in \mathbb{R}$ and $u_{\text{NO}}(\sigma) \geq u$, $\sigma_{\text{NC}}(u) \geq \sigma$ holds.*

Proof of Lemma 6. Suppose in contrast, $\sigma_{\text{NC}}(u) < \sigma$. There exist $x \in \mathcal{X}$ such that $g(x) < u$, $h(x) < \sigma$. Hence $u_{\text{NO}}(\sigma) < u$, which contradicts that $u_{\text{NO}}(x) \geq u$. □

C.1 Proof of Proposition 2

Proof of Proposition 2. Thanks to Proposition 3, the proof is completed by verifying that the following substitutions satisfy Assumption 3 when $p \in \mathcal{P}_{\text{Test}}$.

$$\mathcal{X} = \mathcal{P}, \qquad x = q,$$
$$h(q) = \sum_{s,a} \omega_{sa} \text{KL}_{sa}(p, q), \qquad g(q) = V_q^{\pi}(\boldsymbol{\rho}) V_p^{\pi}(\boldsymbol{\rho}).$$

(a) Let $\underline{x} = p$, one has that $h(\underline{x}) = \sum_{sa} w_{sa} \text{KL}_{sa}(p, p) = 0$.
(b) Due to Assumption 1 and inequalities (1), there exists $q \in \mathcal{P}$ such that $\text{sign}(V_q^{\pi}(\boldsymbol{\rho})) \neq \text{sign}(V_p^{\pi}(\boldsymbol{\rho}))$, we have $\min_{x\in\mathcal{X}} g(x) \leq V_q^{\pi}(\boldsymbol{\rho}) V_p^{\pi}(\boldsymbol{\rho}) < 0$.
(c) The local minimum of h is the global minimum since h is a convex function. As for g, we reverse the policy and transition kernel as in Section 6.2. Since the reversed MDP is still a tabular MDP, Theorem 1 in Bhandari and Russo (2024) has verified that all the stationary points for the value function on the policy space are global optima. As a consequence, the local minimum of g is the global minimum.
(d) It is clear that h is a continuous function. The continuity of g directly follows from the simulation lemma (Lemma 2).

□

C.2 On the policy evaluation

Policy evaluation is the task where we aim to approximate the value of a given policy up to a predetermined constant ε with a certain confidence. Specifically, if $\hat{v}$ denotes the approximation of $V_p^{\pi}(\boldsymbol{\rho})$, the goal is to minimize the number of samples $\mathbb{E}_p[\tau]$ while satisfying $\mathbb{P}_p[\left|\hat{v} - V_p^{\pi}(\boldsymbol{\rho})\right| > \varepsilon] < \delta$. Similar to Assumption 1 considered for the policy testing, we present Assumption 4 for policy evaluation. Notice that if Assumption 4 does not hold, returning $\hat{v}$ as an arbitrary value between $r^{\pi}(\boldsymbol{\rho}) + \min_s \frac{\gamma}{1-\gamma} r^{\pi}(s)$ and $r^{\pi}(\boldsymbol{\rho}) + \max_s \frac{\gamma}{1-\gamma} r^{\pi}(s)$ satisfies that $\left|\hat{v} - V_p^{\pi}(\boldsymbol{\rho})\right| \leq \varepsilon$.

Assumption 4. *$\rho_s > 0$ for all $s \in \mathcal{S}$. r and $\boldsymbol{\rho}$ satisfy:*

$$\max_s \frac{\gamma}{1-\gamma} r^{\pi}(s) - \min_s \frac{\gamma}{1-\gamma} r^{\pi}(s) > \varepsilon.$$

As discussed in Section 5 and 6.1, solving the stopping condition boils down to identify whether the minimal value of the following two optimization problems is larger than $\beta(t, \delta)/t$.

$$\inf_q \sum_{s,a} \omega_{sa} \text{KL}_{sa}(p, q) \quad \text{s.t. } V_q^{\pi}(\boldsymbol{\rho}) - V_p^{\pi}(\boldsymbol{\rho}) + \varepsilon < 0, \tag{19}$$

and

$$\inf_q \sum_{s,a} \omega_{sa} \text{KL}_{sa}(p, q) \quad \text{s.t. } V_p^{\pi}(\boldsymbol{\rho}) - V_q^{\pi}(\boldsymbol{\rho}) + \varepsilon < 0. \tag{20}$$

Under Assumption 4, either $\{q \in \mathcal{P} : V_q^\pi(\boldsymbol{\rho}) - V_p^\pi(\boldsymbol{\rho}) + \varepsilon < 0\}$ or $\{q \in \mathcal{P} : V_p^\pi(\boldsymbol{\rho}) - V_q^\pi(\boldsymbol{\rho}) + \varepsilon < 0\}$ will be nonempty. For simplicity, we restrict our attention to solving (20) and assume $\{q \in \mathcal{P} : V_p^\pi(\boldsymbol{\rho}) - V_q^\pi(\boldsymbol{\rho}) + \varepsilon < 0\} \neq \emptyset$. In the following, we prove Assumption 3 holds with the corresponding substitution in Lemma 7, then one can implement a projected policy gradient method to approximate the value of its dual problem, as described in Section 6.3.

$$\min_q V_p^\pi(\boldsymbol{\rho}) - V_q^\pi(\boldsymbol{\rho}) + \varepsilon \quad \text{s.t.} \sum_{sa} \omega_{sa} \mathrm{KL}_{sa}(p, q) \leq \frac{\beta(t, \delta)}{t}. \tag{21}$$

Lemma 7. *When $\{q \in \mathcal{P} : V_p^\pi(\boldsymbol{\rho}) - V_q^\pi(\boldsymbol{\rho}) + \varepsilon < 0\} \neq \emptyset$, Assumption 3 holds with the following substitution.*

$$\mathcal{X} = \mathcal{P}, \qquad x = q,$$
$$h(q) = \sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p, q), \qquad g(q) = V_p^\pi(\boldsymbol{\rho}) - V_q^\pi(\boldsymbol{\rho}) + \varepsilon.$$

Proof. (a) Let $\underline{x} = p$, one has that $h(\underline{x}) = \sum_{sa} w_{sa} \mathrm{KL}_{sa}(p, p) = 0$.
(b) is a direct consequence of the assumption that $\{q \in \mathcal{P} : V_p^\pi(\boldsymbol{\rho}) - V_q^\pi(\boldsymbol{\rho}) + \varepsilon < 0\} \neq \emptyset$.
(c) and (d) hold as the proof in Proposition 2. □

D Convergence analysis of constrained policy gradient–Proof of Theorem 3

For notational simplicity and clarity in the following description, we assume $V_p^\pi(\boldsymbol{\rho}) = 1$ and omit the $\bar{\square}$ notation. Moreover, we write Π_σ instead of Π_σ^p for brevity.

Proof of Theorem 3. Consider the minimization problem of a smooth function as follows.

$$\min_{x \in Q} f(x),$$

where Q is a closed, convex subset of $\mathbb{R}^n$ and $f(x)$ is L-smooth, i.e., there exists a constant $L > 0$ such that

$$\|\nabla_x f(x) - \nabla_x f(y)\|_2 \leq L \|x - y\|_2, \qquad \forall x, y \in Q.$$

We consider the *projected gradient descent* algorithm for this optimization problem:

$$x^{(k+1)} = \mathbf{proj}_Q \left(x^{(k)} - \frac{1}{L} \nabla_x f(x^{(k)}) \right). \tag{22}$$

We first employ the following (weak) gradient-mapping condition introduced in Xiao (2022).

Definition 2 (gradient-mapping domination Xiao (2022))**.** *Consider the L-smooth objective function $f(x)$ and let Q be a compact convex set. $f(x)$ satisfies the gradient-mapping dominance condition if for some constant $\omega > 0$:*

$$\|G_L(x)\|_2 \geq \sqrt{2\omega} \left(f(T_L(x) - f^\star) \right), \qquad \forall x \in Q,$$

where $f^\star = \min_{x \in Q} f(x)$, and $T_L(x), G_L(x)$ are defined as

$$T_L(x) := \mathbf{proj}_Q \left(x - \frac{1}{L} \nabla_x f(x) \right), \qquad G_L(x) := L(x - T_L(x)).$$

Under the assumption that gradient-mapping domination condition holds, the following (global) convergence rate is obtained.

Theorem 4 (Xiao (2022))**.** *Consider the minimization problem of an L-smooth function $f(x)$ over a convex compact set Q. Denote $f^\star = \min_{x \in Q} f(x)$. Suppose that $f(x)$ satisfies the gradient-mapping domination condition (with constant $\omega > 0$ as defined above). Then, the sequence $(x^{(k)})_{k \geq 0}$ generated by the projected gradient descent algorithm* (22) *satisfies, for all $k \geq 1$,*

$$f(x^{(k)}) - f^\star \leq \max \left\{ \frac{4L}{\omega k}, \left(\frac{\sqrt{2}}{2} \right)^k (f(x^{(0)}) - f^\star) \right\}.$$

We present the proof of Theorem 4 in Appendix D.1 for completeness.

We now present the convergence rate of the projected policy gradient method over the convex compact subset Π_c:

$$\pi^{(k+1)} = \mathbf{proj}_{\Pi_\sigma}\left(\pi^{(k)} - \frac{1}{L}\nabla_\pi V_p^{\pi^{(k)}}(\boldsymbol{\rho})\right). \tag{23}$$

We now consider the convergence rate under the L-smoothness assumption of the value function:

Lemma 8. *Suppose that $V_p^\pi(\boldsymbol{\rho})$ satisfies L-smoothness with respect to π. Then, the constrained projected policy gradient method* (23) *satisfies, for all $k \geq 1$,*

$$V_p^{\pi^{(k)}}(\boldsymbol{\rho}) - V_p^{\sigma\star}(\boldsymbol{\rho}) \leq \max\left\{\frac{64L|\mathcal{S}|}{(1-\gamma)^2 k}\left\|\frac{1}{\boldsymbol{\rho}}\right\|_\infty^2, \left(\frac{\sqrt{2}}{2}\right)^k (V_p^{\pi^{(0)}}(\boldsymbol{\rho}) - V_p^{\sigma\star}(\boldsymbol{\rho}))\right\}.$$

The proof of Lemma 8 is presented in Appendix D.2.

In Lemma 54 in Agarwal et al. (2021), the authors show the smoothness of the value function, that is,

$$\left\|\nabla_\pi V_p^\pi(\boldsymbol{\rho}) - \nabla_\pi V_p^{\tilde{\pi}}(\boldsymbol{\rho})\right\|_2 \leq \frac{2\gamma\,|\mathcal{A}|}{(1-\gamma)^3}\left\|\pi - \tilde{\pi}\right\|_2, \qquad \forall \pi, \tilde{\pi} \in \Pi.$$

Therefore, $V_p^\pi(\boldsymbol{\rho})$ satisfies L-smoothness by taking

$$L = \frac{2\gamma\,|\mathcal{A}|}{(1-\gamma)^3}. \tag{24}$$

Theorem 3 is an immediate consequence from Lemma 8 and (24).

□

D.1 Proof of Theorem 4

Proof of Theorem 4. We obtain, for any $x \in Q$,

$$\begin{aligned} f(x) - f(T_L(x)) &\geq \frac{1}{2L}\|G_L(x)\|_2^2 \\ &\geq \frac{\omega}{L}(f(T_L(x)) - f^\star)^2. \end{aligned}$$

where for the first inequality, we used Theorem 1 of Nesterov (2013), and for the second inequality, the gradient-mapping domination condition is used. We obtain, for each $s \geq 0$,

$$f(x^{(s)}) - f(x^{(s+1)}) \geq \frac{\omega}{L}(f(x^{(s+1)}) - f^\star)^2. \tag{25}$$

Denote $\delta_s = f(x^{(s)}) - f^\star$; note that $\delta_s \geq 0$. We obtain:

$$f(x^{(s)}) - f(x^{(s+1)}) = \delta_s - \delta_{s+1} \geq \frac{\omega}{L}\delta_{s+1}^2$$

$$\text{which is equivalent to } \frac{1}{\delta_{s+1}} - \frac{1}{\delta_s} \geq \frac{\omega}{L}\frac{\delta_{s+1}}{\delta_s}.$$

Summing up the inequality from $s = 0$ to $s = k-1$, we obtain:

$$\frac{1}{\delta_k} - \frac{1}{\delta_0} \geq \frac{\omega}{L}\sum_{s=0}^{k-1}\frac{\delta_{s+1}}{\delta_s}.$$

Using a constant $r \in (0,1)$, we define $n(k,r)$ as the number of times the ratio δ_{s+1}/δ_s is at least r among the first k iterations. Let $c \in (0,1)$ be a constant. Suppose $n(k,r) \geq ck$; then $\delta_{s+1}/\delta_s \geq r$ for at least $\lceil ck \rceil$ values of s in $\{0,\ldots,k-1\}$. In this case,

$$\frac{\omega}{L}rck \leq \frac{1}{\delta_k} - \frac{1}{\delta_0} \leq \frac{1}{\delta_k}.$$

Then, we derive that

$$\delta_k \leq \frac{L}{\omega r c k}.$$

Otherwise, when $n(k,r) < ck$, it holds that $\delta_{s+1}/\delta_s < r$ at least $\lceil (1-c)k \rceil$ times. From the descent property (25), we further obtain $\delta_{s+1} \leq \delta_s \leq 1$ for each $s \in \{0, \ldots, k-1\}$. Therefore, we get

$$\delta_k = \frac{\delta_k}{\delta_{k-1}} \frac{\delta_{k-1}}{\delta_{k-2}} \cdots \frac{\delta_1}{\delta_0} \delta_0 < \delta_0 r^{(1-c)k}.$$

Therefore, by taking $c = r = 1/2$, we obtain,

$$\delta_k \leq \max\left\{ \frac{4L}{\omega k}, \left(\frac{1}{\sqrt{2}}\right)^k \delta_0 \right\}.$$

This concludes the proof. □

D.2 Proof of Lemma 8

Proof of Lemma 8. Define

$$T_L(\pi) := \mathbf{proj}_{\Pi_\sigma}\left(\pi - \frac{1}{L}\nabla_\pi V_p^\pi(\boldsymbol{\rho})\right), \qquad G_L(\pi) := L(\pi - T_L(\pi)).$$

The following gradient mapping domination condition for the projected policy gradient can be proved.

Lemma 9 (Weak gradient-mapping domination)**.** *Let Π_σ be a closed and convex subset of Π, $V_p^{\sigma\star}(\boldsymbol{\rho}) := \min_{\pi\in\Pi_\sigma} V_p^\pi(\boldsymbol{\rho})$, and $\pi^\star \in \arg\min_{\pi'\in\Pi} V_p^{\pi'}(\boldsymbol{\rho})$. Assume that $V_p^\pi(\boldsymbol{\rho})$ satisfies L-smoothness with respect to π, i.e., the following holds:*

$$\|\nabla_\pi V_p^\pi(\boldsymbol{\rho}) - \nabla_\pi V_p^{\pi'}(\boldsymbol{\rho})\|_2 \leq L\|\pi - \pi'\|_2, \qquad \forall \pi, \pi' \in \Pi_\sigma.$$

Then, we have, for all $\pi \in \Pi_\sigma$,

$$V_p^{T_L(\pi)}(\boldsymbol{\rho}) - V_p^{\sigma\star}(\boldsymbol{\rho}) \leq \frac{2\sqrt{2|\mathcal{S}|}}{1-\gamma} \left\| \frac{1}{\boldsymbol{\rho}} \right\|_\infty \|G_L(\pi)\|_2,$$

where

$$T_L(\pi) := \mathbf{proj}_{\Pi_\sigma}\left(\pi - \frac{1}{L}\nabla_\pi V_p^\pi(\boldsymbol{\rho})\right), \qquad G_L(\pi) := L(\pi - T_L(\pi)).$$

The proof of Lemma 9 is presented in Appendix D.3.

From Lemma 9, for all $\pi \in \Pi_\sigma$,

$$V_p^{T_L(\pi)}(\boldsymbol{\rho}) - V_p^{\sigma\star}(\boldsymbol{\rho}) \leq \frac{2\sqrt{2|\mathcal{S}|}}{1-\gamma} \left\| \frac{1}{\boldsymbol{\rho}} \right\|_\infty \|G_L(\pi)\|_2.$$

Therefore, the gradient-mapping dominantion condition holds with

$$\omega = \frac{(1-\gamma)^2}{16|\mathcal{S}| \left\| \frac{1}{\boldsymbol{\rho}} \right\|_\infty^2}.$$

From Theorem 4, we obtain

$$V_p^{\pi^{(k)}}(\boldsymbol{\rho}) - V_p^{\sigma\star}(\boldsymbol{\rho}) \leq \max\left\{ \frac{64L|\mathcal{S}|}{(1-\gamma)^2 k} \left\| \frac{1}{\boldsymbol{\rho}} \right\|_\infty^2, \left(\frac{\sqrt{2}}{2}\right)^k (V_p^{\pi^{(0)}}(\boldsymbol{\rho}) - V_p^{\sigma\star}(\boldsymbol{\rho})) \right\}$$

This concludes the proof. □

D.3 Proof of Lemma 9

Proof of Lemma 9. We apply Theorem 1 from Nesterov (2013). By setting $Q = \Pi$ and Ψ as the indicator function of Π_σ, we deduce that

$$\langle \nabla_\pi V_p^\pi(\boldsymbol{\rho}), T_L(\pi) - \pi' \rangle \leq 2\|G_L(\pi)\|_2 \|T_L(\pi) - \pi'\|_2, \qquad \forall \pi, \pi' \in \Pi.$$

From $T_L(\pi) \in \Pi$, we obtain $\|T_L(\pi) - \pi'\| \leq \sqrt{2|\mathcal{S}|}$. We further deduce that

$$\max_{\pi' \in \Pi} \langle \nabla_\pi V_p^\pi(\boldsymbol{\rho}), T_L(\pi) - \pi' \rangle \leq 2\sqrt{2|\mathcal{S}|} \|G_L(\pi)\|_2.$$

Let $V_p^\star(\boldsymbol{\rho}) = \min_{\pi' \in \Pi} V_p^{\pi'}(\boldsymbol{\rho})$ and $\pi^\star \in \arg\min_{\pi' \in \Pi} V_p^{\pi'}(\boldsymbol{\rho})$. As $\Pi_\sigma \subseteq \Pi$, we obtain

$$\begin{aligned} V_p^{T_L(\pi)}(\boldsymbol{\rho}) - V_p^{\sigma\star}(\boldsymbol{\rho}) \leq V_p^{T_L(\pi)}(\boldsymbol{\rho}) - V_p^\star(\boldsymbol{\rho}) &\leq \frac{1}{1-\gamma} \left\| \frac{1}{\boldsymbol{\rho}} \right\|_\infty \max_{\pi' \in \Pi} \langle \nabla_\pi V_p^\pi(\boldsymbol{\rho}), T_L(\pi) - \pi' \rangle \\ &\leq \frac{2\sqrt{2|\mathcal{S}|}}{1-\gamma} \left\| \frac{1}{\boldsymbol{\rho}} \right\|_\infty \|G_L(\pi)\|_2, \end{aligned}$$

where the second inequality follows from Lemma 10 below.

Lemma 10 (Variational gradient domination, Lemma 4 in Agarwal et al. (2021))**.** *Let $V_p^\star(\boldsymbol{\rho}) = \min_{\pi' \in \Pi} V_p^{\pi'}(\boldsymbol{\rho})$ and $\pi^\star \in \arg\min_{\pi' \in \Pi} V_p^{\pi'}(\boldsymbol{\rho})$. We have, for any $\pi \in \Pi$,*

$$\begin{aligned} V_p^\pi(\boldsymbol{\rho}) - V_p^\star(\boldsymbol{\rho}) &\leq \left\| \frac{d_{p,\boldsymbol{\rho}}^{\pi^\star}}{d_{p,\boldsymbol{\rho}}^\pi} \right\|_\infty \max_{\pi' \in \Pi} \langle \nabla_\pi V_p^\pi(\boldsymbol{\rho}), \pi - \pi' \rangle \\ &\leq \frac{1}{1-\gamma} \left\| \frac{d_{p,\boldsymbol{\rho}}^{\pi^\star}}{\boldsymbol{\rho}} \right\|_\infty \max_{\pi' \in \Pi} \langle \nabla_\pi V_p^\pi(\boldsymbol{\rho}), \pi - \pi' \rangle \\ &\leq \frac{1}{1-\gamma} \left\| \frac{1}{\boldsymbol{\rho}} \right\|_\infty \max_{\pi' \in \Pi} \langle \nabla_\pi V_p^\pi(\boldsymbol{\rho}), \pi - \pi' \rangle. \end{aligned}$$

This concludes the proof. □

D.4 Proof of Lemma 2

Proof of Lemma 2. Let $\bar{s} = (s, a)$, $\bar{p}$ is the transition kernel yielded by π, and $\bar{\pi}$ ($\tilde{\bar{\pi}}$ resp.) is the policy yielded by p ($\tilde{p}$ resp.). Invoke the performance difference lemma, we have

$$\bar{V}_{\bar{p}}^{\bar{\pi}}(\bar{s}) - \bar{V}_{\bar{p}}^{\tilde{\bar{\pi}}}(\bar{s}) = \frac{1}{1-\gamma} \sum_{\bar{s}' \in \bar{\mathcal{S}}} d_{\bar{p},\bar{s}}^{\bar{\pi}}(\bar{s}') \sum_{\bar{a}'' \in \bar{\mathcal{A}}} \bar{Q}_{\bar{p}}^{\tilde{\bar{\pi}}}(\bar{s}', \bar{a}'') \left(\bar{\pi}(\bar{a}'' \mid \bar{s}') - \tilde{\bar{\pi}}(\bar{a}'' \mid \bar{s}') \right). \tag{26}$$

Observe that the L.H.S. of (26) is exactly $Q_p^\pi(s,a) - Q_{\tilde{p}}^\pi(s,a)$. Next denote $\bar{s}' = (s', a')$ and $\bar{a}'' = s''$, then the proof is completed by substituting the R.H.S of (26) with $d_{\bar{p},\bar{s}}^{\bar{\pi}}(\bar{s}') = d_{p,s,a}^\pi(s', a')$,

$$\sum_{\bar{a}'' \in \bar{\mathcal{A}}} \bar{Q}_{\bar{p}}^{\tilde{\bar{\pi}}}(\bar{s}', \bar{a}'') \bar{\pi}(\bar{a}'' \mid \bar{s}') = r(s', a') + \gamma \sum_{s'' \in \mathcal{S}} p(s'' \mid s', a') V_{\tilde{p}}^{\tilde{\pi}}(s''),$$

and

$$\sum_{\bar{a}'' \in \bar{\mathcal{A}}} \bar{Q}_{\bar{p}}^{\tilde{\bar{\pi}}}(\bar{s}', \bar{a}'') \tilde{\bar{\pi}}(\bar{a}'' \mid \bar{s}') = r(s', a') + \gamma \sum_{s'' \in \mathcal{S}} \tilde{p}(s'' \mid s', a') V_{\tilde{p}}^{\tilde{\pi}}(s'').$$

This concludes the proof. □

D.5 Proof of Lemma 3

Proof of Lemma 3. Show (10). Let $\bar{s} = (s, a)$, $\bar{s}' = (s', a')$, $\bar{a}'' = s''$. The policy gradient on $\bar{V}_p^{\bar{\pi}}(\bar{s})$, which equals to $Q_p^\pi(s,a)$, with respect to $\bar{\pi}(\bar{a}'' \mid \bar{s}')$, is

$$\frac{\partial \bar{V}_p^{\bar{\pi}}(\bar{s})}{\partial \bar{\pi}(\bar{a}'' \mid \bar{s}')} = \frac{1}{1-\gamma} d_{\bar{p},\bar{s}}^{\bar{\pi}}(\bar{s}') \bar{Q}_{\bar{p}}^{\bar{\pi}}(\bar{s}', \bar{a}'').$$

The proof of (10) is completed as: $\bar{d}^{\bar{\pi}}_{\bar{p},\bar{s}}(\bar{s}') = d^{\pi}_{p,s,a}(s',a')$, and

$$\bar{Q}^{\bar{\pi}}_{\bar{p}}(\bar{s}',\bar{a}'') = r(s',a') + \gamma V^{\pi}_{p}(s'').$$

Show (11). As $V^{\pi}_{p}(\boldsymbol{\rho}) = \sum_{\tilde{s},\tilde{a}} \rho_{\tilde{s}}\pi(\tilde{a} \mid \tilde{s}) Q^{\pi}_{p}(\tilde{s},\tilde{a})$. Using (10), one obtains

$$\begin{aligned}\frac{\partial V^{\pi}_{p}(\boldsymbol{\rho})}{\partial p(s' \mid s,a)} &= \sum_{\tilde{s},\tilde{a}} \rho_{\tilde{s}}\pi(\tilde{a} \mid \tilde{s}) \frac{1}{1-\gamma} d^{\pi}_{p,\tilde{s},\tilde{a}}(s,a)\left(r(s,a) + \gamma V^{\pi}_{p}(s')\right) \\ &= \frac{1}{(1-\gamma)} d^{\pi}_{p,\boldsymbol{\rho}}(s,a)\left(r(s,a) + \gamma V^{\pi}_{p}(s')\right).\end{aligned}$$

□

D.6 Proof of Lemma 4

Proof of Lemma 4. Let $\bar{s} = (s,a)$, and $\bar{\pi}$ ($\tilde{\bar{\pi}}$ resp.) denotes the policy yielded by p ($\tilde{p}$ resp.). Applying Lemma 54 in Agarwal et al. (2021) implies that

$$\left\| \nabla_{\bar{\pi}} \bar{V}^{\bar{\pi}}_{\bar{p}}(\bar{s}) - \nabla_{\bar{\pi}} \bar{V}^{\tilde{\bar{\pi}}}_{\bar{p}}(\bar{s}) \right\|_2 \leq \frac{2\gamma \left|\bar{\mathcal{A}}\right|}{(1-\gamma)^3} \left\|\bar{\pi} - \tilde{\bar{\pi}}\right\|_2.$$

The lemma directly stems from the facts that

$$\nabla_{\bar{\pi}} \bar{V}^{\bar{\pi}}_{\bar{p}}(\bar{s}) = \nabla_{p} Q^{\pi}_{p}(s,a), \nabla_{\bar{\pi}} \bar{V}^{\tilde{\bar{\pi}}}_{\bar{p}}(\bar{s}) = \nabla_{p} Q^{\pi}_{\tilde{p}}(s,a), \text{ and } \bar{\mathcal{A}} = \mathcal{S}.$$

□

E Upper bound on the sample complexity–Proof of Theorem 2

Convention. Throughout this section, for simplicity of presentation, we assume $\text{Ans}(p) = +$.

Proof of Theorem 2. Show δ-PC. Recall that Algorithm 1 stops in round τ only if $u_{\zeta_\tau} - \zeta_\tau > 0$. Thanks to Theorem 3 and the number of iterations M chosen in Algorithm 2, we have

$$u_{\text{NO}}(\beta(\tau,\delta)/\tau, \hat{\boldsymbol{\omega}}(\tau), \hat{p}_\tau) \geq u_{\zeta_\tau} - \zeta_\tau > 0. \tag{27}$$

As Assumption 1 implies that $\min_q V^{\pi}_{q}(\boldsymbol{\rho}) V^{\pi}_{\hat{p}_\tau}(\boldsymbol{\rho}) < 0$, inequality (27) yields that

$$u_{\text{NO}}(\beta(\tau,\delta)/\tau, \hat{\boldsymbol{\omega}}(\tau), \hat{p}_\tau) > \min_{q} V^{\pi}_{q}(\boldsymbol{\rho}) V^{\pi}_{\hat{p}_\tau}(\boldsymbol{\rho}). \tag{28}$$

Observing that $\sigma_{\text{NC}}(\cdot, \hat{\boldsymbol{\omega}}(\tau), \hat{p}_\tau)$ is a decreasing function, (27) also yields that

$$\sigma_{\text{NC}}(0, \hat{\boldsymbol{\omega}}(\tau), \hat{p}_\tau) \geq \sigma_{\text{NC}}(u_{\text{NO}}(\beta(\tau,\delta)/\tau, \hat{\boldsymbol{\omega}}(\tau), \hat{p}_\tau), \hat{\boldsymbol{\omega}}(\tau), \hat{p}_\tau) = \beta(\tau,\delta)/\tau, \tag{29}$$

where the last equality follows from Proposition 2 with $\sigma = \beta(\tau,\delta)/\tau \geq 0$ and condition (28). One can notice that (29) is equivalent to (6). Hence, if $\text{Ans}(\hat{p}_\tau) \neq \text{Ans}(p)$ (in other words, if $p \in \text{Alt}(\hat{p}_\tau)$), then

$$\sum_{s,a} N_{sa}(\tau) \text{KL}_{sa}(\hat{p}_\tau, p) \geq \beta(\tau,\delta).$$

By Proposition 1 in Jonsson et al. (2020) (see (5)), we deduce that $\mathbb{P}_p[\text{Ans}(\hat{p}_\tau) \neq \text{Ans}(p)] \leq \delta$.

Show the upper bound of sample complexity. For simplicity of presentation, we assume $\text{Ans}(p) = +$ in this proof; the case where $\text{Ans}(p) = -$ can be derived analogously. We first introduce the function

$$F(\boldsymbol{\omega}', p') := \inf_{q \in \text{Alt}(p)} \sum_{s,a} \omega'_{sa} \text{KL}_{sa}(p', q), \quad \forall \boldsymbol{\omega}' \in \Sigma,\ p' \in \mathcal{P}^{+}_{\text{Test}},$$

and $\varepsilon \in (0, F(\boldsymbol{\omega}, p)/2)$. As shown in Lemma 18 (Appendix G.1), F is a continuous function on $\Sigma \times \mathcal{P}^{+}_{\text{Test}}$; thus, there exists $\xi_1 \in (0,1)$ such that

$$|F(\boldsymbol{\omega}, p) - F(\boldsymbol{\omega}', p')| < \varepsilon \quad \text{if } \max\{\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_1, \|p' - p\|_1\} \leq \xi_1. \tag{30}$$

Moreover, an application of Theorem 5 in Appendix F with $u = 0$ and $p = p$ implies that there exist $\xi_2 \in (0,1)$ and $c > 0$ such that $u_{\text{NO}}(\cdot, \hat{\boldsymbol{\omega}}(t), \hat{p}_t)$ decays faster than a linear function $f(\sigma) = -c\sigma$ if $\|\hat{p}_t - p\|_1 < \xi_2$ and $\|\hat{\boldsymbol{\omega}}(t) - \boldsymbol{\omega}\|_1 < \xi_2$. We introduce $\xi = \min\{\xi_1, \xi_2\}$ and define the 'good event'

$$\mathcal{C}_T(\xi) = \bigcap_{\sqrt{T} \le t \le T} \left\{ \max\left\{ \|\hat{\boldsymbol{\omega}}(t) - \boldsymbol{\omega}\|_1, \|\hat{p}_t - p\|_1 \right\} \le \xi \right\}. \tag{31}$$

By Proposition 4 (proved later in this section), there exists $T_1(\xi)$ such that for $T \ge T_1(\xi)$, the event $\mathcal{C}_T(\xi)$ occurs with high probability. Moreover, as $\beta(T,\delta) + \zeta_T T/c = \log(1/\delta) + o(T)$ and $F(\boldsymbol{\omega}, p) - \varepsilon > 0$, one can find an integer $T_2(\xi) \in \mathbb{N}$ such that if $T \ge T_2(\xi)$,

$$\beta(T,\delta) + \frac{\zeta_T T}{c} \le \log(1/\delta) + (F(\boldsymbol{\omega}, p) - \varepsilon)\xi T. \tag{32}$$

Finally, we define

$$T_3(\xi, \varepsilon, \delta) = \frac{(F(\boldsymbol{\omega}, p) - \varepsilon)^{-1} \log(1/\delta)}{1 - \xi}. \tag{33}$$

With these definitions, if $T \ge \max\{T_1(\xi), T_2(\xi), T_3(\xi,\delta)\}$, then conditional on $\mathcal{C}_T(\xi)$, we have

$$\begin{aligned} \frac{\zeta_T T}{c} + \beta(T,\delta) &\le \log(1/\delta) + (F(\boldsymbol{\omega}, p) - \varepsilon)\xi T \\ &\le (F(\boldsymbol{\omega}, p) - \varepsilon)T \\ &\le F(\hat{\boldsymbol{\omega}}(T), \hat{p}_T)T, \end{aligned} \tag{34}$$

where the first inequality is (32); the second one follows from (33); the last one is a consequence of (31) and $T \ge T_1(\xi)$. Recall that $F(\hat{\boldsymbol{\omega}}(T), \hat{p}_T)$ is exactly $\sigma_{\text{NC}}(0, \hat{\boldsymbol{\omega}}(T), \hat{p}_T)T$. Applying Theorem 5 with $u = 0$, $\hat{\boldsymbol{\omega}} = \hat{\boldsymbol{\omega}}(T)$, $\hat{p} = \hat{p}_T$, $\sigma_1 = \beta(T,\delta)/T$, $\sigma_2 = \sigma_{\text{NC}}(0, \hat{\boldsymbol{\omega}}(T), \hat{p}_T)$, (34) implies that

$$\begin{aligned} -\zeta_T &\ge u_{\text{NO}}(\sigma_{\text{NC}}(0, \hat{\boldsymbol{\omega}}(T), \hat{p}_T), \hat{\boldsymbol{\omega}}(T), \hat{p}_T) - u_{\text{NO}}(\beta(T,\delta)/T, \hat{\boldsymbol{\omega}}(T), \hat{p}_T) \\ &= -u_{\text{NO}}(\beta(T,\delta)/T, \hat{\boldsymbol{\omega}}(T), \hat{p}_T) \ge u_{\zeta_T}, \end{aligned} \tag{35}$$

where last equation follow from Proposition 3. Notice that (35) is our approximate stopping rule used in Algorithm, hence $\tau \le T$. Namely, $\forall T \ge \max\{T_1(\xi), T_2(\xi), T_3(\xi,\varepsilon,\delta)\}$, $\mathcal{C}_T(\xi) \subseteq \{\tau \le T\}$. We can conclude that

$$\begin{aligned} \mathbb{E}_p[\tau] &\le \max\{T_1(\xi), T_2(\xi), T_3(\xi,\varepsilon,\delta)\} + \sum_{T=\max\{T_1(\xi),T_2(\xi),T_3(\xi,\varepsilon,\delta)\}+1}^{\infty} \mathbb{P}_p[\tau > T] \\ &\le \max\{T_1(\xi), T_2(\xi), T_3(\xi,\varepsilon,\delta)\} + \sum_{T=\lceil T_1(\xi)\rceil}^{\infty} \mathbb{P}_p[\mathcal{C}_T(\xi)^c]. \end{aligned} \tag{36}$$

From Proposition 4, $\sum_{T=\lceil T_1(\xi)\rceil}^{\infty} \mathbb{P}_p[\mathcal{C}_T(\xi)^c] \le 8\,|\mathcal{S}|^4\,|\mathcal{A}|^3 / \xi^2 \min_{s,a} w_{sa}$. As a consequence of (36),

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_p[\tau]}{\log(1/\delta)} \le \limsup_{\delta \to 0} \frac{T_3(\xi,\varepsilon,\delta)}{\log(1/\delta)} \le \frac{(F(\boldsymbol{\omega}, p) - \varepsilon)^{-1}}{1 - \xi}.$$

As ε, ξ can be taken arbitrarily small, the proof is completed.

□

Proposition 4. *Under Assumption 2, in Algorithm 1, for any $\xi \in (0,1), \boldsymbol{\omega} \in \Sigma$, there exists $T_1(\xi) > 0$ such that*

$$\sum_{T=\lceil T_1(\xi)\rceil}^{\infty} \mathbb{P}_p\left[\mathcal{C}_T(\xi)^c\right] < \frac{8\,|\mathcal{S}|^4\,|\mathcal{A}|^3}{\xi^2 \min_{s,a} w_{sa}},$$

where $\mathcal{C}_T(\xi)$ is introduced in (31).

Proof. Due Assumption 2, $\mathcal{W} := \{(s,a) : \omega_{sa} > 0\} = \mathcal{S} \times \mathcal{A}$. By Lemma 12, $\|\hat{\boldsymbol{\omega}}(t) - \boldsymbol{\omega}\|_1 \le \sum_{s,a} |\mathcal{S}|\,|\mathcal{A}|\,/t \le |\mathcal{S}|^2\,|\mathcal{A}|^2\,/t$. We then derive that when $T \ge |\mathcal{S}|^4\,|\mathcal{A}|^4\,/\xi^2$ and $t \ge \sqrt{T}$, one can

deduce that $\|\hat{\boldsymbol{\omega}}(t) - \boldsymbol{\omega}\|_1 \le \xi$.
As for the estimate on p, we apply Lemma 12 again to have that for each (s,a),

$$N_{sa}(t) \ge t \min_{s,a} \omega_{sa} - |\mathcal{S}|\,|\mathcal{A}| \ge t \min_{s,a} \omega_{sa}/2 \tag{37}$$

if $T \ge 4\,|\mathcal{S}|^2\,|\mathcal{A}|^2 / \min_{s,a} w_{sa}^2$ and $t \ge \sqrt{T}$. Using the union bound yields that

$$\begin{aligned}\mathbb{P}_p\left[\|\hat{p}_t - p\|_1 \ge \xi\right] &\le \sum_{s,a} \mathbb{P}_p\left[\|\hat{p}_t(\cdot \mid s,a) - p(\cdot \mid s,a)\|_1 \ge \frac{\xi}{|\mathcal{S}|\,|\mathcal{A}|}\right] \\ &\le 2\,|\mathcal{S}|\,|\mathcal{A}| \exp\left(-\frac{t\xi^2 \min_{s,a} w_{sa}}{2\,|\mathcal{S}|^3\,|\mathcal{A}|^2}\right),\end{aligned}$$

where the last inequality follows from Lemma 11 and (37). By introducing $T_1(\xi) = \max\{\frac{|\mathcal{S}|^4|\mathcal{A}|^4}{\xi^2}, \frac{4|\mathcal{S}|^2|\mathcal{A}|^2}{\min_{(s,a)\in\mathcal{W}} w_{sa}^2}\}$, union bound yields that

$$\begin{aligned}\sum_{T=\lceil T_1(\xi)\rceil}^{\infty} \mathbb{P}_p[\mathcal{C}_T(\xi)^c] &\le \sum_{T=\lceil T_1(\xi)\rceil}^{\infty} \sum_{\sqrt{T}\le t\le T} 2\,|\mathcal{S}|\,|\mathcal{A}| \exp\left(-\frac{t\xi^2 \min_{s,a} w_{sa}}{2\,|\mathcal{S}|^3\,|\mathcal{A}|^2}\right) \\ &\le \int_1^{\infty}\int_{\sqrt{T}}^{T} 2\,|\mathcal{S}|\,|\mathcal{A}| \exp\left(-\frac{t\xi^2 \min_{s,a} w_{sa}}{2\,|\mathcal{S}|^3\,|\mathcal{A}|^2}\right) dt dT.\end{aligned}$$

The proof is completed by applying Lemma 13 with $A = \frac{t\xi^2 \min_{s,a} w_{sa}}{2|\mathcal{S}|^3|\mathcal{A}|^2}$, $\alpha = 1/2$, $\beta = 1$. □

Lemma 11 (Proposition 1 in Weissman et al. (2003))**.** *Suppose one has samples the state-action pair (s,a) for $n \ge 1$ times, then the empirical estimate on $p(\cdot \mid s,a)$, $\hat{p}_n(\cdot \mid s,a)$ satisfies that*

$$\mathbb{P}\left[\|\hat{p}_n(\cdot \mid s,a) - p(\cdot \mid s,a)\|_1 \ge \varepsilon\right] \le 2e^{-\frac{n\varepsilon^2}{|\mathcal{S}|}}, \quad \forall \varepsilon \in (0,1).$$

Lemma 12. *Let $\boldsymbol{\omega} \in \Sigma$ and define $\mathcal{W} = \{(s,a) : \omega_{sa} > 0\}$. A sampling rule does*

$$\begin{aligned} &A_t \leftarrow (s,a), && \text{if } (s,a) \in \mathcal{W} \text{ and } N_{sa}(t) = 0, \\ &A_t \leftarrow \arg\min_{(s,a)} N_{sa}(t-1)/\omega_{sa} \text{ (tie-broken arbitrarily)}, && \text{otherwise.}\end{aligned}$$

Then for each $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $t \ge |\mathcal{W}|$, one has

$$t\omega_{sa} - |\mathcal{W}| \le N_{sa}(t) \le t\omega_{sa} + 1. \tag{38}$$

Proof. If $(s,a) \notin \mathcal{W}$, $N_{sa}(t) = 0$ for all $t \in \mathbb{N}$, (38) holds directly.
For a fixed $(s,a) \in \mathcal{W}$, we prove the upper bound in (38) by induction. When $t = |\mathcal{W}|$, $N_{sa}(t) = 1 \le t\omega_{sa} + 1$. Now suppose $N_{sa}(t-1) \le (t-1)\omega_{sa} + 1$, and consider two following cases, (i) $A_t \ne (s,a)$; (ii) $A_t = (s,a)$.
When (i) $A_t \ne (s,a)$, using the inductive hypothesis yields that

$$N_{sa}(t) = N_{sa}(t-1) \le (t-1)\omega_{sa} + 1 \le t\omega_{sa} + 1.$$

As for (ii) $A_t = (s,a)$, one can observe that

$$\min_{s',a'} \frac{N_{s'a'}(t-1)}{\omega_{s'a'}} \le \frac{\min_{s'a'} N_{s'a'}(t-1)}{\max_{s'a'} \omega_{s'a'}} \le \frac{\frac{t-1}{K}}{\frac{1}{K}} \le t - 1 \le t.$$

Since $(s'a')$ is the minimizer,

$$\frac{N_{sa}(t)}{\omega_{sa}} = \frac{N_{sa}(t-1)}{\omega_{sa}} + \frac{1}{\omega_{sa}} \le 1 + \frac{1}{\omega_{sa}}.$$

Thus the upper bound in (38) is obtained by multiplying $t\omega_{sa}$ on the the both sides of the above inequality.

We now prove the lower bound in (38). Notice that

$$N_{sa}(t) = t - \sum_{s' \neq s, a' \neq a} N_{s'a'} \geq t - \sum_{s' \neq s, a' \neq a} (t\omega_{sa'} + 1) \geq t\omega_{sa} + |\mathcal{W}|,$$

where the second inequality is due to the upper bound in (38).

□

Lemma 13 (Lemma 5 in Wang et al. (2021))**.** *Let $\alpha, \beta \in (0,1)$ and $A > 0$.*

$$\int_0^\infty \left(\int_{T^\alpha}^\infty \exp(-At^\beta) dt \right) dT = \frac{\Gamma\left(\frac{1}{\alpha\beta} + \frac{1}{\beta}\right)}{\beta A^{\frac{1}{\alpha\beta} + \frac{1}{\beta}}}.$$

F Sensitivity analysis on u_{NO}

Theorem 5. *Suppose Assumptions 1 and 2 hold. For any $p \in \mathcal{P}$, $u \in \mathbb{R}$, there exist constants $c > 0$, $\xi \in (0, \min_{sa} \omega_{sa}/2)$ such that if $\hat{p} \in \{q \in \mathcal{P} : \|p - q\|_1 \leq \xi\}$, $\hat{\boldsymbol{\omega}} \in \{\boldsymbol{\omega}' \in \Sigma : \|\boldsymbol{\omega}' - \boldsymbol{\omega}\|_1 \leq \xi\}$ and $0 < \sigma_1 < \sigma_2 \leq \bar{\sigma}$, where $\bar{\sigma} = \max_{\|\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}\|_1 \leq \xi} \max_{\|\hat{p} - p\|_1 \leq \xi} \{\sigma_{\text{NC}}(u, \hat{\boldsymbol{\omega}}, \hat{p})\}$, then*

$$u_{\text{NO}}(\sigma_2, \hat{\boldsymbol{\omega}}, \hat{p}) - u_{\text{NO}}(\sigma_1, \hat{\boldsymbol{\omega}}, \hat{p}) \leq -c(\sigma_2 - \sigma_1).$$

Proof. One can assume $\hat{p}$ is full-supported. Otherwise, due to the continuity of $u_{\text{NO}}(\sigma, \hat{\boldsymbol{\omega}}, \cdot)$ with respect to its third argument (the kernel), as shown in Lemma 19 (Appendix G.2), for an arbitrary $\varepsilon > 0$, one can always find a full-supported kernel $\tilde{p}$ sufficiently close to $\hat{p}$ such that

$$|u_{\text{NO}}(\sigma_2, \hat{\boldsymbol{\omega}}, \tilde{p}) - u_{\text{NO}}(\sigma_1, \hat{\boldsymbol{\omega}}, \tilde{p}) - u_{\text{NO}}(\sigma_2, \hat{\boldsymbol{\omega}}, \hat{p}) - u_{\text{NO}}(\sigma_1, \hat{\boldsymbol{\omega}}, \hat{p})| \leq \varepsilon.$$

Because $u_{\text{NO}}(\sigma, \hat{\boldsymbol{\omega}}, \hat{p})$ is a decreasing function of σ, an application of Monotone difference lemma (14) implies that $u_{\text{NO}}(\sigma, \hat{\boldsymbol{\omega}}, \hat{p})$ as a function of σ is differentiable almost everywhere. Let $\sigma \in [\sigma_1, \sigma_2]$ be a point at which $u_{\text{NO}}(\sigma, \hat{\boldsymbol{\omega}}, \hat{p})$ is differentiable and q_σ be the solution of (NO-$\sigma, \hat{\boldsymbol{\omega}}, \hat{p}$). Let $\eta_\sigma \in \mathbb{R}_+, \lambda_{s'sa} \in \mathbb{R}_+, \mu_{sa} \in \mathbb{R}$ be the Lagrange multipliers associated with the constraints $\sum_{s,a} \omega_{sa} \text{KL}_{sa}(\hat{p}, q) - \sigma \leq 0$, $-q(s' \mid s, a) \leq 0$, $\sum_{s' \in \mathcal{S}} q(s' \mid s, a) - 1 = 0$ respectively. An application of Envelope Theorem (Theorem 7) yields that

$$\begin{aligned}\frac{\partial}{\partial \sigma} u_{\text{NO}}(\sigma, \hat{\boldsymbol{\omega}}, \hat{p}) = \frac{\partial}{\partial \sigma} \left(V_{\hat{p}}^\pi(\boldsymbol{\rho}) V_{q_\sigma}^\pi(\boldsymbol{\rho}) \right) + \eta_\sigma \frac{\partial}{\partial \sigma} \left(\sum_{s,a} \hat{\omega}_{sa} \text{KL}_{sa}(\hat{p}, q_\sigma) - \sigma \right) \\ + \frac{\partial}{\partial \sigma} \sum_{s',s,a} \lambda_{s'sa}(-q_\sigma(s' \mid s, a)) + \frac{\partial}{\partial \sigma} \sum_{s,a} \mu_{sa} \left(\sum_{s' \in \mathcal{S}} q_\sigma(s' \mid s, a) - 1 \right) = -\eta_\sigma\end{aligned}$$

By Lemma 15, we know

$$\eta_\sigma = \frac{\gamma V_{\hat{p}}^\pi(\boldsymbol{\rho}) d_{q_\sigma, \boldsymbol{\rho}}(s, a) \left(V_{q_\sigma}^\pi(s_\sigma^M) - V_{q_\sigma}^\pi(s_\sigma^m) \right)}{\hat{\omega}_{sa}(1-\gamma) \left(\frac{\hat{p}(s_\sigma^M \mid s,a)}{q_\sigma(s_\sigma^M \mid s,a)} - \frac{\hat{p}(s_\sigma^m \mid s,a)}{q_\sigma(s_\sigma^m \mid s,a)} \right)}, \quad \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{39}$$

where $s_\sigma^M \in \arg\max_s V_{\hat{p}}^\pi(\boldsymbol{\rho}) V_{q_\sigma}^\pi(s)$ and $s_\sigma^m \in \arg\min_s V_{\hat{p}}^\pi(\boldsymbol{\rho}) V_{q_\sigma}^\pi(s)$. By invoking the fundamental theorem of calculus, we have

$$u_{\text{NO}}(\sigma_2, \hat{\boldsymbol{\omega}}, \hat{p}) - u_{\text{NO}}(\sigma_1, \hat{\boldsymbol{\omega}}, \hat{p}) = \int_{\sigma_1}^{\sigma_2} -\eta_\sigma d\sigma.$$

It suffices to show $\eta_\sigma > c$ for some $c > 0$. Notice that if r and $\boldsymbol{\rho}$ satisfy Assumption 1, so do $rV_p^\pi(\boldsymbol{\rho})$ and $\boldsymbol{\rho}$. Lemma 1 then implies that $\min_{q \in \mathcal{P}} \max_{s,s'} V_p^\pi(\boldsymbol{\rho}) V_q^\pi(s) - V_p^\pi(\boldsymbol{\rho}) V_q^\pi(s') > 0$. As $V_\cdot^\pi(\boldsymbol{\rho})$ is a continuous function, there exists $c_1 > 0$, $\xi \in (0, \min_{sa} \omega_{sa}/2)$ such that

$$\min_{q \in \mathcal{P}} \max_{s,s'} V_{\hat{p}}^\pi(\boldsymbol{\rho}) V_q^\pi(s) - V_{\hat{p}}^\pi(\boldsymbol{\rho}) V_q^\pi(s') \geq c_1, \quad \forall \|\hat{p} - p\|_1 < \xi. \tag{40}$$

Further observe that

$$\frac{\hat{p}(s^M \mid s, a)}{q(s^m \mid s, a)} - \frac{\hat{p}(s^M \mid s, a)}{q(s^m \mid s, a)} \leq 2 \max_{s',s,a} \frac{\hat{p}(s' \mid s, a)}{q(s' \mid s, a)} \leq 2 \max_{q: \sum_{sa} w_{sa} \text{KL}_{sa}(\hat{p}, q) \leq \bar{\sigma}} \max_{s',s,a} \frac{\hat{p}(s' \mid s, a)}{q(s' \mid s, a)},$$

which is upper bounded by some $c_2 > 0$ for any $\|\hat{p} - p\|_1 \leq \xi$. Hence the proof is completed by setting $c = \frac{\gamma c_1}{(1-\gamma)c_2} \min_{sa} \rho(s)\pi(a \mid s)$, where $\min_{sa} \rho(s)\pi(a \mid s) > 0$ thanks to Assumptions 1 and 2. □

F.1 Technical lemmas

Lemma 14 (Monotone difference lemma, see e.g. Theorem 1.6.25 in Tao (2011))**.** *Any function $F : \mathbb{R} \mapsto \mathbb{R}$ which is monotone is differentiable almost everywhere.*

Theorem 6 (Kuhn-Tucker Theorem, Theorem A.30 in Acemoglu (2008))**.** *Consider the constrained minimization problem*

$$\inf_{x \in \mathbb{R}^K} f(x)$$
$$\text{s.t. } g(x) \leq 0 \quad \text{and} \quad h(x) = 0,$$

where $f : x \in X \to \mathbb{R}$, $g : x \in X \to \mathbb{R}^N$, $h : x \in X \to \mathbb{R}^M$ (for some $K, N, M \in \mathbb{N}$) and $X \subset \mathbb{R}^K$ is a vector space. Let $x^* \in X$ be a solution to this minimization problem, and suppose that $N_1 \leq N$ of the inequality constraints are active, in the sense that they hold as equality at x^*. Define $\tilde{h} : X \to \mathbb{R}^{M+N_1}$ to be the mapping of these N_1 active constraints stacked with $h(x)$ (so that $\tilde{h}(x^*) = 0$). Suppose that the following constraint qualification condition is satisfied: the Jacobian matrix $D_x(\tilde{h}(x^*))$ has rank $N_1 + M$. Then the following Kuhn-Tucker condition is satisfied: there exist Lagrange multipliers $\boldsymbol{\lambda}^* \in \mathbb{R}^{N_1}$ and $\boldsymbol{\mu}^* \in \mathbb{R}^M$ such that

$$D_x f(x^*) + \boldsymbol{\lambda}^* \cdot D_x g(x^*) + \mu^* \cdot D_x h(x^*) = 0,$$

and the complementary slackness condition

$$\boldsymbol{\lambda}^* \cdot g(x^*) = 0$$

holds

Theorem 7 (Envelope Theorem for constrained optimization problem, Theorem A.31 in Acemoglu (2008))**.** *Consider the constrained minimization problem*

$$v(p) = \min_{x \in X} f(x, p)$$
$$\text{s.t.} \quad g(x, p) \leq 0, \text{ and } h(x, p) = 0,$$

where $X \subset \mathbb{R}^K$ is a vector space, $p \in \mathbb{R}$; and $f : X \times \mathbb{R} \to \mathbb{R}, g : X \times \mathbb{R}^N \to \mathbb{R}$, and $h : X \times \mathbb{R}^M \to \mathbb{R}$ are differentiable ($K, N, M \in \mathbb{N}$). Let $x^*(p) \in \text{Int}(X)$ be a solution to the problem. Denote the Lagrangian multipliers associated with the inequality and equality by $\boldsymbol{\lambda}^* \in \mathbb{R}^N_+$ and $\boldsymbol{\mu}^* \in \mathbb{R}^M$. Suppose also $v(p)$ is differentiable at $\bar{p}$. Then we have

$$\frac{dv(\bar{p})}{dp} = \frac{\partial f(x^*(\bar{p}), \bar{p})}{\partial p} + \boldsymbol{\lambda}^* D_p g(x^*(\bar{p}), \bar{p}) + \boldsymbol{\mu}^* D_p h(x^*(\bar{p}), \bar{p}).$$

F.2 The value of the Lagrangian multiplier

Lemma 15. *Suppose Assumption 1 and 2 hold. Let $\sigma > 0$ and $p \in \mathcal{P}$ is full-supported. Denote q_σ as the solution to (NO-$\sigma, \boldsymbol{\omega}, p$). Then the Lagrange multiplier associated with the inequality $\sum_{s,a} \omega_{sa} \text{KL}_{sa}(p, q) \leq \sigma$ is*

$$\eta_\sigma = \frac{\gamma V_p^\pi(\boldsymbol{\rho}) d_{q_\sigma, \boldsymbol{\rho}}(s, a) \left(V_{q_\sigma}^\pi(s_\sigma^M) - V_{q_\sigma}^\pi(s_\sigma^m) \right)}{\omega_{sa}(1-\gamma) \left(\frac{p(s_\sigma^M|s,a)}{q_\sigma(s_\sigma^M|s,a)} - \frac{p(s_\sigma^m|s,a)}{q_\sigma(s_\sigma^m|s,a)} \right)} \quad \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{41}$$

where $s_\sigma^M \in \arg\max_s V_p^\pi(\boldsymbol{\rho}) V_{q_\sigma}^\pi(\boldsymbol{\rho})$ and $s_\sigma^m \in \arg\min_s V_p^\pi(\boldsymbol{\rho}) V_{q_\sigma}^\pi(\boldsymbol{\rho})$.

Proof. Let $\sigma > 0$. The Lagrangian function of the optimization problem (NO-$\sigma, \boldsymbol{\omega}, p$) is

$$L(q, \eta, \boldsymbol{\lambda}, \boldsymbol{\mu}) = V_p^\pi(\boldsymbol{\rho}) V_q^\pi(\boldsymbol{\rho}) + \eta \left(\sum_{s,a} \omega_{sa} \text{KL}_{sa}(p, q) - \sigma \right)$$
$$+ \sum_{s',s,a} \lambda_{s'sa} \left(-q(s'|s, a) \right) + \sum_{s,a} \mu_{sa} \left(\sum_{s' \in \mathcal{S}} q(s'|s, a) - 1 \right),$$

where $\eta \geq 0$, $\boldsymbol{\lambda} \in \mathbb{R}_+^{|\mathcal{S}|^2|\mathcal{A}|}$ and $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. As p is full-supported, q_σ is full-supported as well (otherwise, it violates the constraint that $\sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p, q_\sigma) \leq \sigma$). That is, $q_\sigma(s'|s,a) > 0$, $\forall s, s', a$. Further using Corollary 1 in Appendix F.3, we conclude that $\sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p, q_\sigma) = \sigma$. In other words, the upper bound of the weighted KL-divergence is the only active inequality. We now prove that $\{D_q \sum_{s' \in \mathcal{S}} q_\sigma(s'|s,a) - 1\}_{s \in \mathcal{S}, a \in \mathcal{A}} \cup \{D_q \left(\sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p, q_\sigma) - \sigma \right)\}$ is linear independent. Suppose on the contrary, $D_q \left(\sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p, q_\sigma) - \sigma \right)$ is spanned by $\{D_q \sum_{s' \in \mathcal{S}} q_\sigma(s'|s,a) - 1\}_{s \in \mathcal{S}, a \in \mathcal{A}}$. As

$$\frac{\partial}{\partial q(s'|s,a)} \sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p, q_\sigma) - \sigma = -\frac{\omega_{sa} p(s'|s,a)}{q_\sigma(s'|s,a)} \quad \forall s', s, a,$$

$$\text{and } \frac{\partial}{\partial q(s'|s,a)} \sum_{s' \in \mathcal{S}} q_\sigma(s'|s,a) - 1 = 1, \quad \forall s', s, a.$$

we deduce that $q_\sigma = p$ which contradicts that $\sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p, q_\sigma) = \sigma$. Hence, we can apply Kuhn-Tucker Theorem (Theorem 6) and obtain that there exists $\eta_\sigma \geq 0, \boldsymbol{\lambda} \in \mathbb{R}_+^{|\mathcal{S}|^2|\mathcal{A}|}$ and $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ such that

$$\frac{\partial}{\partial q(s'|s,a)} V_p^\pi(\boldsymbol{\rho}) V_{q_\sigma}^\pi(\boldsymbol{\rho}) - \frac{\eta \omega_{sa} p(s'|s,a)}{q_\sigma(s'|s,a)} + \mu_{sa} - \lambda_{s'sa} \quad = 0, \quad \forall s', s, a, \quad \text{(Stationarity)}$$

$$\text{and } \eta \left(\sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p, q_\sigma) - \sigma \right) = 0, \ \lambda_{s'sa}(-q_\sigma(s'|s,a)) \qquad = 0, \quad \forall s', s, a,$$

(Complementary slackness)

As q_σ is full-supported, we derive $\lambda_{s'sa} = 0, \forall s', s, a$, from Complementary slackness. From (11) in Lemma 3, Stationarity can be rewritten as:

$$\frac{V_p^\pi(\boldsymbol{\rho})}{1-\gamma} d_{q_\sigma, \boldsymbol{\rho}}(s,a) \left(r(s,a) + \gamma V_{q_\sigma}^\pi(s') \right) - \frac{\eta \omega_{sa} p(s' \mid s,a)}{q_\sigma(s' \mid s,a)} = -\mu_{sa}, \ \forall s', s, a. \tag{42}$$

By taking difference of the equations (42) with $s' = s_\sigma^M$ and $s' = s_\sigma^m$, we obtain

$$\frac{\gamma V_p^\pi(\boldsymbol{\rho}) d_{q_\sigma, \boldsymbol{\rho}}(s,a)}{1-\gamma} \left(V_{q_\sigma}^\pi(s_\sigma^M) - V_{q_\sigma}^\pi(s_\sigma^m) \right) - \eta \omega_{sa} \left(\frac{p(s_\sigma^M \mid s,a)}{q_\sigma(s_\sigma^M \mid s,a)} - \frac{p(s_\sigma^m \mid s,a)}{q_\sigma(s_\sigma^m \mid s,a)} \right) = 0.$$

(41) follows from a simple rearrangement on the above equation. □

F.3 Properties for the stationary points

For the clarity of presentation, here we fix some $p \in \mathcal{P}_{\text{Test}}$, $\boldsymbol{\omega} \in \Sigma$ and introduce the constrained set,

$$\mathcal{P}_\sigma := \left\{ q \in \mathcal{P} : \sum_{sa} \omega_{sa} \mathrm{KL}_{sa}(p, q) \leq \sigma \right\}.$$

The goal of this subsection is to prove Corollary 1, where we show the minimizer $q_\sigma \in \arg\min_{q \in \mathcal{P}_\sigma} V_p^\pi(\boldsymbol{\rho}) V_q^\pi(\boldsymbol{\rho})$ satisfies that $\sum_{sa} \omega_{sa} \mathrm{KL}_{sa}(p, q_\sigma) = \sigma$. For this purpose, we firstly consider the stationary points in Lemma 16.

Lemma 16. *Consider the optimization problem,* $\min_{q \in \mathcal{P}_\sigma} V_p^\pi(\boldsymbol{\rho}) V_q^\pi(\boldsymbol{\rho})$ *under Assumption 1, 2. All the stationary points will be on the boundary of* $\mathcal{P}_\sigma$[3].

Proof. Suppose on the contrary, there is a stationary point, say q_o, at the interior of $\mathcal{P}_\sigma$. As q_o is a stationary point, one has $\langle q - q_o, \nabla V_{q_o}^\pi(\boldsymbol{\rho}) V_p^\pi(\boldsymbol{\rho}) \rangle \geq 0$ for all $q \in \mathcal{P}_\sigma$. By invoking Lemma 17, we derive that $\forall (s', a') \in \mathcal{S} \times \mathcal{A}$, $\exists \alpha_{s'a'} \in \mathbb{R}$ such that for each $s'' \in \mathcal{S}$,

$$\alpha_{s'a'} = \frac{\partial V_{q_o}^\pi(\boldsymbol{\rho}) V_p^\pi(\boldsymbol{\rho})}{\partial q(s''|s',a')} = \frac{V_p^\pi(\boldsymbol{\rho})}{1-\gamma} \sum_{s,a} \rho(s) \pi(a|s) d_{q_o,s,a}^\pi(s',a') (r(s,a) + \gamma V_{q_o}^\pi(s'')), \tag{43}$$

[3]The interior (boundary resp.) is referred to relatively interior, i.e. the topological interior (boundary) relative to the affine hull of the simplex. Interested readers are referred to Zalinescu (2002).

where the last equation stems directly from Lemma 3. Let $\alpha := \sum_{s',a'} \alpha_{s'a'}/V_p^\pi(\boldsymbol{\rho})$ and sum (43) over all $s', a' \in \mathcal{S} \times \mathcal{A}$, one has

$$r^\pi(\boldsymbol{\rho}) + \gamma V_{q_o}^\pi(s'') = \alpha,\ \forall s'' \in \mathcal{S},$$

which yields that $\forall s \in \mathcal{S},\ V_{q_o}^\pi(s) = \alpha' := (\alpha - r_{\boldsymbol{\rho}}^\pi)/\gamma$. However, it contradicts Lemma 1, hence the stationary points are on the boundary of $\mathcal{P}_\sigma$. □

Corollary 1. *Consider the minimizer $q_\sigma \in \arg\min_{q \in \mathcal{P}_\sigma} V_p^\pi(\boldsymbol{\rho}) V_q^\pi(\boldsymbol{\rho})$, one has*

$$\sum_{sa} \omega_{sa} \mathrm{KL}_{sa}(p, q_\sigma) = \sigma \tag{44}$$

if p is full-supported and Assumption 1, 2 hold.

Proof. Observe that $p(s' \mid s, a) > 0$ for all s', s, a, hence $q_\sigma(s' \mid s, a) > 0$ for all s', s, a. Otherwise, it violates $\sum_{sa} \omega_{sa} \mathrm{KL}_{sa}(p, q_\sigma) \leq \sigma$. Moreover, as q_σ is the stationary point, together with the conclusion from Lemma 16, q_σ is on the boundary, we deduce (44). □

Lemma 17. *Consider $v \in \mathbb{R}^{|\mathcal{S}|^2|\mathcal{A}|}$ and an interior point of $\mathcal{P}_\sigma$, denoted by q_o. If $\langle q - q_o, v\rangle := \sum_{s,a} \sum_{s'} (q(s'|s,a) - q_o(s'|s,a)) v_{s',s,a} \geq 0$ for all $q \in \mathcal{P}_\sigma$, then for each $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\exists \alpha_{sa}$ such that $v_{s',s,a} = \alpha_{sa},\ \forall s', s \in \mathcal{S}, a \in \mathcal{A}$. In other words, for each $(s,a) \in \mathcal{S} \times \mathcal{A}$ $v_{\cdot,s,a}$ is parallel to a $|\mathcal{S}|$-dimensional vector whose components are all 1's.*

Proof. Let $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $u \in \mathbb{R}^{|\mathcal{S}|^2|\mathcal{A}|}$ such that $u_{s'',s',a'} = 0$ if $(s',a') \neq (s,a)$ for each $s'' \in \mathcal{S}$, and $\sum_{s' \in \mathcal{S}} u_{s',s,a} = 0$. As p_o is an interior point in $\mathcal{P}_\sigma$, one can find a small constant $c > 0$ such that $q^+(s'|s,a) := q_o(s'|s,a) + c u_{s',s,a}$ and $q^- := q_o(s'|s,a) - c u_{s',s,a},\ \forall s, s' \in \mathcal{S},\ a \in \mathcal{A}$ be two policies in $\mathcal{P}_\sigma$. From the assumption on v, we have $c\langle u, v\rangle = \langle q^+ - q_o, v\rangle \geq 0$ and $-c\langle u, v\rangle = \langle q^- - q_o, v\rangle \geq 0$, which implies that $\langle u, v\rangle = 0$. As the only constraint of $u_{\cdot,s,a}$ is $\sum_{s' \in \mathcal{S}} u_{s',s,a} = 0$, we deduce that $v_{\cdot,s,a}$ is parallel to a $|\mathcal{S}|$-dimensional vector whose components are all 1's □

G Maximal theorem and its applications

Suppose $\mathbb{X}$ and $\mathbb{Y}$ are Hausdorff topological spaces. Let $\psi : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ be a function and $\Phi : \mathbb{X} \rightrightarrows \mathbb{S}(\mathbb{Y})$ be a set-valued function, where $\mathbb{S}(\mathbb{Y})$ is the set of non-empty subsets of $\mathbb{Y}$. Furthermore, we introduce $\mathbb{K}(\mathbb{X}) = \{F \in \mathbb{S}(\mathbb{X}) : F \text{ is compact}\}$. We are interested in a minimization problem of the form:

$$\begin{aligned} v(x) &= \inf_{y \in \Phi(x)} \psi(x, y)\,, \\ \Phi^*(x) &= \{y \in \Phi(x)\ :\ \psi(x,y) = v(x)\}\,. \end{aligned}$$

For $U \subset \mathbb{X}$, let the graph of Φ restricted to U be $Gr_U(\Phi) = \{(x,y) \in U \times \mathbb{Y}\ :\ y \in \Phi(x)\}$.

Theorem 8 (Maximal theorem Berge (1877))**.** *Let $\mathbb{X}$ and $\mathbb{Y}$ be Hausdorff topological spaces. Assume that*

- *$\Phi : \mathbb{X} \rightrightarrows \mathbb{K}(\mathbb{X})$ is continuous (i.e. both lower and upper hemicontinous),*
- *$\psi : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ is continuous.*

Then the function $v : \mathbb{X} \to \mathbb{R}$ is continuous and the solution multifunction $\Phi^* : \mathbb{X} \to \mathbb{S}(\mathbb{Y})$ is upper hemicontinuous and compact valued.

G.1 Proof of Lemma 18

Lemma 18. *A function $F : \Sigma \times \mathcal{P}_{\text{Test}}^+ \to \mathbb{R}$ defined as $F(\boldsymbol{\omega}, q) := \inf_{q \in \text{Alt}(p)} \sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p, q)$ is a continuous function.*

Proof. The proof is established by invoking Theorem 8 with the following substitution:

$$\mathbb{X} = \Sigma \times \mathcal{P}^{+}_{\text{Test}}, \qquad \psi(\boldsymbol{\omega}, p, q) = \sum_{s,a} \omega_{sa} \text{KL}_{sa}(p, q),$$
$$\mathbb{Y} = \mathcal{P}, \qquad \Phi(\boldsymbol{\omega}, p) = \text{cl}(\text{Alt}(p)).$$

Observe that objective function, $\sum_{s,a} \omega_{sa} \text{KL}_{sa}(p, q)$, is a continuous function on $\Sigma \times \mathcal{P}^{+}_{\text{Test}} \times \mathcal{P}$, and the corresponding $\Phi(\boldsymbol{\omega}, p) = \text{cl}(\text{Alt}(p))$ is always a constant. □

G.2 Proof of Lemma 19

Lemma 19. *Let $\boldsymbol{\omega} \in \Sigma$ such that $\omega_{sa} > 0$, $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$. $u_{\text{NO}}(\cdot, \boldsymbol{\omega}, \cdot)$ is a continuous function on $\mathbb{R}_+ \times \mathcal{P}$.*

Proof. The proof is established by invoking Theorem 8 with the following substitution:

$$\mathbb{X} = \mathbb{R}_+ \times \mathcal{P}, \qquad \psi(\sigma, p, q) = V^{\pi}_{p}(\boldsymbol{\rho}) V^{\pi}_{q}(\boldsymbol{\rho}),$$
$$\mathbb{Y} = \mathcal{P}, \qquad \Phi(\sigma, p) = \{q \in \mathcal{P} : \sum_{sa} \omega_{sa} \text{KL}_{sa}(p, q) \leq \sigma\}.$$

As the objective function, ψ, is a continuous function on $\mathbb{R}_+ \times \mathcal{P}$, it suffices to show the corresponding $\Phi(\sigma, p) = \{q \in \mathcal{P} : \sum_{s,a} \omega_{sa} \text{KL}_{sa}(p, q) \leq \sigma\}$ is hemi-continuous.

Show the upper hemi continuous of Φ. Let $\{p_n\}_{n=1}^{\infty} \subset \mathcal{P}$ and $\{\sigma_n\}_{n=1}^{\infty} \subset \mathbb{R}_+$ such that $p_n \xrightarrow{n\to\infty} p$ and $\sigma_n \xrightarrow{n\to\infty} \sigma$. And consider a sequence $\{q_n\}_{n=1}^{\infty} \subset \mathcal{P}$ such that $q_n \in \Phi(\sigma_n, p_n)$, which is equivalent to $\sum_{s,a} \omega_{sa} \text{KL}_{sa}(p_n, q_n) - \sigma_n \leq 0$, and $q_n \xrightarrow{n\to\infty} q$. By continuity of KL-divergence, one has

$$\sum_{s,a} \omega_{sa} \text{KL}_{sa}(p, q) - \sigma = \lim_{n\to\infty} \sum_{s,a} \omega_{sa} \text{KL}_{sa}(p_n, q_n) - \sigma_n \leq 0,$$

which yields that $q \in \Phi(\sigma, p)$ and hence Φ is upper hemi-continuous.
Show the lower hemi continuous of Φ. Let $\{p_n\}_{n=1}^{\infty} \subset \mathcal{P}$ and $\{\sigma_n\}_{n=1}^{\infty} \subset \mathbb{R}_+$ be the sequences such that $p_n \xrightarrow{n\to\infty} p$ and $\sigma_n \xrightarrow{n\to\infty} \sigma$ for some $p \in \mathcal{P}$ and $\sigma \in \mathbb{R}$. Consider $q \in \Phi(\sigma, p)$, we now aim to show that $\exists \{q_n\}_{n=1}^{\infty}$ such that $q_n \in \Phi(\sigma_n, p_n)$ and $q_n \xrightarrow{n\to\infty} q$. The proof is separated into two cases (i) $\sum_{s,a} \omega_{sa} \text{KL}_{sa}(p, q) < \sigma$ and (ii) $\sum_{s,a} \omega_{sa} \text{KL}_{sa}(p, q) = \sigma$.
Case (i) As $(\sigma_n, p_n) \xrightarrow{n\to\infty} (\sigma, p)$ and the continuity of KL-divergence, $\exists N > 0$ such that $\sum_{s,a} \omega_{sa} \text{KL}_{sa}(p_n, q) - \sigma_n < 0$ for all $n \geq N$. Choosing $q_n = q$ yields the conclusion.
Case (ii) For each $n \in \mathbb{N}$, we define $\alpha_n = \max\{\alpha \in [0, 1] : (1 - \alpha)p_n + \alpha q \in \Phi(\sigma_n, p_n)\}$ and $q_n = (1 - \alpha_n)p_n + \alpha_n q$, which directly implies $q_n \in \Phi(\sigma_n, p_n)$. From the definition of α_n, when $\sum_{s,a} \omega_{sa} \text{KL}_{sa}(p_n, q) = 0$, $\alpha_n = 1$. When $\sum_{s,a} \omega_{sa} \text{KL}_{sa}(p_n, q) \neq 0$ Due to the joint convexity, one has

$$\sum_{s,a} \omega_{sa} \text{KL}_{sa}(p_n, (1 - \alpha)p_n + \alpha q) \leq \alpha \sum_{s,a} \omega_{sa} \text{KL}_{sa}(p_n, q).$$

Hence $\alpha \geq \frac{\sigma_n}{\sum_{s,a} \omega_{sa} \text{KL}_{sa}(p_n, q)}$. In summary,

$$\alpha_n \begin{cases} \geq \frac{\sigma_n}{\sum_{s,a} \omega_{sa} \text{KL}_{sa}(p_n, q)}, & \text{if } \sum_{s,a} \omega_{sa} \text{KL}_{sa}(p_n, q) \neq 0, \\ = 1, & \text{otherwise.} \end{cases}$$

Using $(\sigma_n, p_n) \xrightarrow{n\to\infty} (\sigma, p)$ and the continuity of KL-divergence again, we have $\alpha_n \to 1$ and $q_n \to q$ as $n \to \infty$.

□

G.3 Proof of Lemma 20

Lemma 20. *Under Assumption 3 and the notion in Appendix C, $u_{\text{NO}}(\cdot)$ is continuous in $[0, \infty)$.*

Proof. We prove it by applying Theorem 8 with the following substitution:

$$\mathbb{X} = [0, \infty), \qquad \psi(\sigma, x) = g(x),$$
$$\mathbb{Y} = \mathcal{X}, \qquad \Phi(\sigma) = \{x \in \mathcal{X} : h(x) \leq \sigma\}.$$

As $\mathcal{X}$ is compact and ψ is continuous according to Assumption 3-(c), $\Phi(\sigma)$ is always a compact set. Additionally, $h(\underline{x}) \leq 0$ from Assumption 3-(a), $\Phi(\sigma) \neq \emptyset$ for all $\sigma \geq 0$. It remains to show $\Phi(\cdot)$ is a continuous corresponding.

Upper hemicontinuity. Let $\{\sigma_n\}_{n=1}^{\infty} \subset \mathbb{X}$ and $\{x_n\}_{n=1}^{\infty} \subset \mathcal{X}$ be the sequences such that $x_n \in \Phi(\sigma_n)$, or equivalently $h(x_n) \leq \sigma_n$, $\forall n \in \mathbb{N}$, $\lim_{n\to\infty} \sigma_n = \sigma^\star$, and $\lim_{n\to\infty} x_n = x^\star$. Since the continuity of h is assumed in Assumption 3-(d), we derive

$$h(x^\star) = \limsup_{n\to\infty} h(x_n) \leq \limsup_{n\to\infty} \sigma_n = \sigma^\star,$$

i.e. $x^\star \in \Phi(\sigma^\star)$, and hence $\Phi(\cdot)$ is upper hemicontinuous.
Lower hemicontinuity. Let $\{\sigma_n\}_{n=1}^{\infty} \subset \mathbb{X}$ be a sequence converging to $\sigma^\star \geq 0$ as $n \to \infty$, and $x^\star \in \Phi(\sigma^\star)$, or equivalently $h(x^\star) \leq \sigma^\star$. We claim there exists $\{\sigma_{n_m}\}_{m=1}^{\infty} \subseteq \{\sigma_n\}_{n=1}^{\infty}$ and $\{x_m\}_{m=1}^{\infty}$ such that $x_m \in \Phi(\sigma_{n_m})$ and $x_m \xrightarrow{m\to\infty} x^\star$.

We first consider case $h(x^\star) \leq 0$. As $h(x^\star) \leq 0$, $x^\star \in \Phi(\sigma)$ for any $\sigma \geq 0$. We choose whatever subsequence $\{\sigma_{n_m}\}_{m=1}^{\infty} \subseteq \{\sigma_n\}_{n=1}^{\infty}$ and $x_m = x^\star \in \Phi(\sigma_{n_m})$, $\forall m \in \mathbb{N}$, the claim is satisfied. As for the case $h(x^\star) > 0$, Assumption 3-(b)(c) implies that $x^\star$ is not local minimum of h. Hence for any $m \in \mathbb{N}$, $\exists x_m \in \mathcal{X}$ such that $|x_m - x^\star| \leq 1/m$ and $h(x_m) < h(x^\star)$. As $\sigma_n \xrightarrow{n\to\infty} \sigma^\star$, there is a subsequence $\{\sigma_{n_m}\}_{m=1}^{\infty}$ such that $n_m < n_{m+1}$ and $h(x_m) \leq \sigma_m$, or equivalently $x_m \in \Phi(\sigma_{n_m})$, and hence $\Phi(\cdot)$ is lower hemicontinuous.

□

H Proof of the remaining lemma and proposition

H.1 Proof of Lemma 1

Proof. As $\mathcal{P}$ is a compact set and $V_\cdot^\pi(s)$ is a continuous function for each $s \in \mathcal{S}$, it suffices to show that for all $q \in \mathcal{P}$, $\max_{s,s' \in \mathcal{S}} V_q^\pi(s) - V_q^\pi(s') > 0$. Suppose on the contrary, there is $q \in \mathcal{P}$ such that $\max_{s,s' \in \mathcal{S}} V_q^\pi(s) - V_q^\pi(s') = 0$, then $\forall s \in \mathcal{S}$, $V_q^\pi(s) = \alpha$ for some constant $\alpha \in \mathbb{R}$. As

$$Q_q^\pi(s, a) = r(s, a) + \gamma \sum_{s'} q(s'|s, a) V_q^\pi(s') = r(s, a) + \gamma\alpha,$$

the definition of $r^\pi(s)$ yields that

$$\begin{aligned} r^\pi(s) &= \sum_{a \in \mathcal{A}} \pi(a|s) r(s, a) = \sum_{a \in \mathcal{A}} \pi(a|s)(Q_q^\pi(s, a) - \gamma\alpha) \\ &= V_q^\pi(s) - \gamma\alpha = (1 - \gamma)\alpha. \end{aligned}$$

However, this contradicts Assumption 1, where $\max_s r^\pi(s) > \min_s r^\pi(s)$. □

H.2 Proof of Proposition 1

Proof. Suppose not, there exists a fully supported $\boldsymbol{\omega} \in \Sigma$ such that $T_{\boldsymbol{\omega}}(p) = \infty$, or equivalently $T_{\boldsymbol{\omega}}^{-1}(p) = 0$, then one has $\sum_{sa} \mathrm{KL}_{sa}(p, q^\star) = 0$ for some $q^\star \in \mathrm{cl}(\mathrm{Alt}(p))$, where $\mathrm{cl}(S)$ denotes the closure of a set S. As for each s, a, $\omega_{sa} > 0$, and hence $\mathrm{KL}_{sa}(p, q^\star) = 0$. This yields that $p(\cdot \mid s, a) = q^\star(\cdot \mid s, a)$ for all s, a. Since the transition probability under π and p is the same as the one under π and $q^\star$, $V_p^\pi(\boldsymbol{\rho}) = V_{q^\star}^\pi(\boldsymbol{\rho})$, which however contradicts $q^\star \in \mathrm{cl}(\mathrm{Alt}(p))$ and the assumption $p \in \mathcal{P}_{\text{Test}}$. □

I Experimental details

The simulations presented in this paper were conducted using the following computational environment:

- Operating system: macOS Sonoma
- Programming language: Python
- Processor: Apple M1 Max
- Memory: 64 GB

Uniform sampling was used as the sampling rule. We define the sequence ζ_t as $\zeta_t = \frac{5}{t^{3/2}}$. We used the SLSQP optimization method to perform the projection. We fixed $L = 400.0$ and capped the maximum value of M at 20.

Table 1: Reward function $r(s, a)$, transition kernel $p(\cdot|s, a)$, and policies $\pi(a|s)$ and $\pi'(a|s)$ for all state-action pairs in the 2-state, 2-action case ($|\mathcal{S}| = 2$, $|\mathcal{A}| = 2$).

Reward function $r(s, a)$

$s \backslash a$	0	1
0	0.50	-0.175
1	-0.775	1.00

Transition kernel $p(\cdot|s, a)$

(s, a)	$p(0\|s, a)$	$p(1\|s, a)$
(0, 0)	0.700	0.300
(0, 1)	0.400	0.600
(1, 0)	0.800	0.200
(1, 1)	0.100	0.900

Policy $\pi(a|s)$

$s \backslash a$	0	1
0	0.150	0.850
1	0.507	0.493

Another policy $\pi'(a|s)$

$s \backslash a$	0	1
0	0.3848	0.6152
1	0.6152	0.3848

Table 2: Reward function $r(s, a)$, transition kernel $p(\cdot|s, a)$, and policies $\pi(a|s)$ and $\pi'(a|s)$ for all state-action pairs in the 3-state, 3-action case ($|\mathcal{S}| = 3, |\mathcal{A}| = 3$).

Reward function $r(s, a)$

$s\backslash a$	0	1	2
0	-0.20	0.02	-0.01
1	-0.50	-0.01	0.50
2	-0.01	-0.05	0.20

Transition kernel $p(\cdot|s, a)$

(s, a)	$p(0\|s, a)$	$p(1\|s, a)$	$p(2\|s, a)$
(0, 0)	0.3460	0.5027	0.1513
(0, 1)	0.2230	0.7014	0.0756
(0, 2)	0.4077	0.3005	0.2919
(1, 0)	0.2711	0.5011	0.2277
(1, 1)	0.1711	0.6011	0.2277
(1, 2)	0.1711	0.1011	0.7277
(2, 0)	0.2433	0.5999	0.1568
(2, 1)	0.1867	0.2998	0.5135
(2, 2)	0.4033	0.0993	0.4974

Policy $\pi(a|s)$

$s\backslash a$	0	1	2
0	0.6	0.3	0.1
1	0.333	0.333	0.333
2	0.1	0.2	0.7

Another policy $\pi'(a|s)$

$s\backslash a$	0	1	2
0	0.329963	0.335487	0.334550
1	0.329790	0.329798	0.340412
2	0.331231	0.330005	0.338764

Table 3: Reward function $r(s, a)$, transition kernel $p(\cdot|s, a)$, and two policies $\pi(a|s)$ and $\pi'(a|s)$ for all state-action pairs in the 5-state, 5-action case ($|\mathcal{S}| = 5, |\mathcal{A}| = 5$).

Reward function $r(s, a)$

$s\backslash a$	0	1	2	3	4
0	0.11596	-0.10323	0.07086	-0.14514	0.01885
1	-0.08898	0.18378	0.20909	0.18429	-0.00352
2	-0.11392	0.23644	-0.15099	-0.20320	-0.23474
3	0.10058	0.08980	0.00906	0.19939	0.02957
4	0.11086	0.02878	-0.12984	0.17238	0.03751

Transition kernel $p(s'|s, a)$

(s, a)	$p(0\|s, a)$	$p(1\|s, a)$	$p(2\|s, a)$	$p(3\|s, a)$	$p(4\|s, a)$
(0, 0)	0.0191	0.2797	0.3241	0.0813	0.2958
(0, 1)	0.2279	0.2631	0.0458	0.2566	0.2066
(0, 2)	0.1418	0.2505	0.2561	0.2799	0.0718
(0, 3)	0.3117	0.1916	0.0851	0.1691	0.2424
(0, 4)	0.1199	0.6589	0.2133	0.0040	0.0038
(1, 0)	0.1452	0.3076	0.0715	0.1816	0.2941
(1, 1)	0.4654	0.0252	0.2148	0.2654	0.0292
(1, 2)	0.2123	0.0780	0.2095	0.2257	0.2745
(1, 3)	0.2350	0.1905	0.1488	0.1254	0.3003
(1, 4)	0.0091	0.3348	0.0134	0.1328	0.5099
(2, 0)	0.2699	0.3663	0.2291	0.0208	0.1139
(2, 1)	0.2535	0.2019	0.1512	0.2041	0.1893
(2, 2)	0.3340	0.2574	0.1303	0.1418	0.1365
(2, 3)	0.1428	0.1237	0.1114	0.0747	0.5474
(2, 4)	0.1530	0.3078	0.1651	0.3379	0.0362
(3, 0)	0.0043	0.3403	0.1235	0.0826	0.4493
(3, 1)	0.0870	0.3120	0.0742	0.2682	0.2587
(3, 2)	0.1755	0.2717	0.1635	0.1257	0.2637
(3, 3)	0.2272	0.1819	0.2460	0.0933	0.2516
(3, 4)	0.2717	0.1775	0.0811	0.1830	0.2868
(4, 0)	0.2812	0.0261	0.0534	0.4150	0.2243
(4, 1)	0.2381	0.2541	0.1767	0.2693	0.0617
(4, 2)	0.4520	0.1074	0.0020	0.1489	0.2897
(4, 3)	0.3384	0.0184	0.1746	0.3144	0.1541
(4, 4)	0.0686	0.1741	0.2139	0.1872	0.3563

Policy $\pi(a|s)$

$s\backslash a$	0	1	2	3	4
0	0.1535	0.2298	0.0998	0.2521	0.2648
1	0.2159	0.2917	0.1054	0.0903	0.2967
2	0.0452	0.0699	0.1839	0.3681	0.3329
3	0.2078	0.3493	0.0826	0.2214	0.1389
4	0.2311	0.1292	0.2522	0.2173	0.1701

Another policy $\pi'(a|s)$

$s\backslash a$	0	1	2	3	4
0	0.1387	0.2651	0.1637	0.3034	0.1291
1	0.2705	0.1384	0.1378	0.1367	0.3167
2	0.1471	0.1155	0.1624	0.1891	0.3859
3	0.1489	0.1512	0.2145	0.1346	0.3508
4	0.1398	0.2038	0.3177	0.1403	0.1984