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

[*]Alphabetical order.

Abstract

We study the policy testing problem in discounted Markov decision processes (MDPs) in the fixed-confidence setting under a generative model with static sampling. The goal is to decide whether the value of a given policy exceeds a specified threshold while minimizing the number of samples. We first derive an instance-dependent lower bound that any reasonable algorithm must satisfy, characterized as the solution to an optimization problem with non-convex constraints. Guided by this formulation, we propose a new algorithm. While this design paradigm is common in pure exploration problems such as best-arm identification, the non-convex constraints that arise in MDPs introduce substantial difficulties. To address them, we reformulate the lower-bound problem by swapping the roles of the objective and the constraints, yielding an alternative problem with a non-convex objective but convex constraints. This reformulation admits an interpretation as a policy optimization task in a newly constructed *reversed MDP*. We further show that the global KL constraint can be decomposed exactly into a family of product-box subproblems, which are solved by projected policy gradient and combined through an outer budget search. Beyond policy testing, our reformulation and reversed MDP view suggest extensions to other pure exploration tasks in MDPs, including policy evaluation and best policy identification.

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

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 (MABs), 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 (§4). 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 computationally intractable in general. Common workarounds rely on convex relaxations (Al Marjani and Proutiere, 2021), which compromise statistical optimality.

In this paper, we address this challenge for the *policy testing* problem under discounted, tabular MDPs with access to a generative model and a static sampling allocation: given a confidence level δ, the agent must decide whether the value of a given policy exceeds a specified threshold with probability at least $1-\delta$. Our main contribution is a reformulation that turns the generally non-convex lower bound problem for policy testing into a tractable form without sacrificing statisti-

cal optimality. Specifically, we show that by swapping the roles of the objective and the constraints, the problem can be recast as policy optimization in a newly constructed *reversed MDP* (§5), yielding convex constraints and a non-convex objective. We then derive an exact decomposition of the global KL-feasible set into a family of product-box subproblems, each of which is solved by projected policy gradient, while an outer search optimizes the local KL-budget allocation (Theorem 2). Combining these pieces, we propose the PTST algorithm (§6) and prove that it attains asymptotically instance-optimal sample complexity (Theorem 3). To the best of our knowledge, this is the first computationally tractable algorithm to achieve instance optimality for pure exploration in MDPs. Beyond policy testing, our reformulation and the reversed MDP perspective suggest extensions to other pure exploration problems in MDPs, including policy evaluation and best policy identification (see Appendix B for a more detailed discussion).

Beyond these theoretical contributions, the problem is practically motivated. Policy testing asks whether a proposed policy meets a target (or improves on a deployed baseline) with high confidence while using as few samples as possible. For example, consider a recommendation system that leverages user history (states), where the algorithm driving recommendations is a policy. When a deployed policy has been used for a long period, its value—i.e., the utility achieved by the system—is typically well understood. Suppose we now design a new policy. Determining whether the value of this new policy exceeds that of the previously deployed one is crucial for maximizing overall benefit. A closely related scenario arises in healthcare, where a policy corresponds to a treatment strategy. More generally, policy testing can be viewed as a stateful analogue of thresholding bandits (Locatelli et al., 2016).

Contributions. Our contributions are as follows: (i) We derive an instance-specific lower bound on sample complexity (Theorem 1), revealing non-convex constraints. (ii) We reformulate the lower-bound optimization as an equivalent reversed MDP, derive an exact decomposition into product-box-constrained subproblems, and solve them by combining projected policy gradient with an outer budget search (Theorem 2). (iii) Building on this, we develop PTST and prove that it is asymptotically instance-optimal in sample complexity (Theorem 3). (iv) Empirically, across the evaluated instances, PTST requires fewer samples at a fixed confidence level than the adapted best policy identification baseline.

2 RELATED WORK

Pure exploration in MABs has 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, instance-specific lower bounds on sample complexity have been derived. These bounds, 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 line of analysis is feasible thanks to the relative simplicity of the optimization problems that underlie these lower bounds.

Similar challenges have been studied in the context of MDPs. Several works focus on characterizing the minimax sample complexity for best policy identification, typically under the generative model assumption (Gheshlaghi Azar et al., 2013; Agarwal et al., 2020; Li et al., 2024). Other efforts aim 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

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 $\pi(\cdot\,|s(t))$, col-

lects 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.[1] 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 $V_p^\pi(\boldsymbol{\rho})<0$.

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). 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 $V_p^\pi(\boldsymbol{\rho})<0$).

Remark 1 (Generative Model)**.** *Much of the sample-complexity analysis for MDPs has been conducted under the assumption of access to a generative model (Ghesh-laghi Azar et al., 2013; Zanette et al., 2019; Al Marjani and Proutiere, 2021). While one can extend the analysis to the forward (online, single-trajectory) interaction model, such extensions typically require additional assumptions. See Appendix B for details.*

In addition to a sampling rule, 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 to design a δ-PC algorithm with minimal sample complexity.

3.3 Assumptions

To simplify the notation, we define $r_{\max} := \max_{s,a}|r(s,a)|$, $r^\pi(s) := \sum_{a\in\mathcal{A}}\pi(a|s)r(s,a)$ 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 the value maps all state-action pairs to the least rewarding state, $\arg\min_s r^\pi(s)$. That is,

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

$$\text{and } \min_p V_p^\pi(\boldsymbol{\rho}) = r^\pi(\boldsymbol{\rho}) + \frac{\gamma}{1-\gamma}\min_s r^\pi(s). \quad (2)$$

Throughout the paper we impose the following standing assumption, which ensures that both decision regions $\{q\in\mathcal{P}_{\text{Test}} : \text{Ans}(q)=-\}$ and $\{q\in\mathcal{P}_{\text{Test}} : \text{Ans}(q)=+\}$ are nonempty and simplifies the exposition.

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 I.1.

Lemma 1. *Under Assumption 1,*

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

[1] 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 $V_p^\pi(\boldsymbol{\rho}) > R$ is equivalent to testing $\tilde{V}_p^\pi(\boldsymbol{\rho})>0$.

Throughout the paper we also adopt the following assumption. We take the target policy π to have full support and require our static sampling rule to explore every state–action pair in the support of π. This is without loss of generality: if $\pi(a \mid s) = 0$ for some (s, a), then the transition law $p(\cdot \mid s, a)$ does not affect V_p^π and such pairs can be ignored.

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

4 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 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), $\mathrm{KL}_{sa}(p, q)$ 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 $\mathrm{Alt}(p) := \{q \in \mathcal{P}_{\mathrm{Test}} : \mathrm{Ans}(p) \neq \mathrm{Ans}(q)\}$. This set collects all the kernels for which the answer to the test differs from p. $\mathrm{Alt}(p)$ can also be written as follows: $\mathrm{Alt}(p) = \{q \in \mathcal{P}_{\mathrm{Test}} : V_q^\pi(\boldsymbol{\rho}) V_p^\pi(\boldsymbol{\rho}) < 0\}$.

Theorem 1. *Under Assumption 1, let $p \in \mathcal{P}_{\mathrm{Test}}$, $\boldsymbol{\omega} \in \Sigma$, and a δ-PC algorithm with sampling rule satisfying that for any $\varepsilon > 0$, there exists $c_\varepsilon > 0$ such that $\mathbb{E}[N_{sa}(t)] \leq t(\omega_{sa} + \varepsilon) + c_\varepsilon$, $\forall s \in \mathcal{S}, a \in \mathcal{A}$. Then,*

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_p[\tau]}{\log(1/\delta)} \geq T^\star_{\boldsymbol{\omega}}(p), \tag{3}$$

where

$$T^\star_{\boldsymbol{\omega}}(p)^{-1} := \inf_{q \in \mathrm{Alt}(p)} \sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p, q). \tag{4}$$

Theorem 1 is proved in Appendix C. The next result, proved in Appendix I.2, states that under Assumption 2, the characteristic time $T^\star_{\boldsymbol{\omega}}(p)$ 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 (4) (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. However, for MDPs, the optimization problem leading to the sample complexity lower bound is non-convex as shown below.

An example where $\mathrm{Alt}(p)$ is non-convex. 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

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

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

5 REVERSED MDP

In this section, we present novel ideas for constructing an optimal stopping rule for policy testing. We show that the problem can be viewed as policy optimization in a reversed MDP, where the roles of the transition kernel and the policy are interchanged.

5.1 Non-Convex Constraint in Stopping Rule

We begin with a standard fixed-confidence stopping rule and show that it leads to an optimization problem with non-convex constraints. Let $N_{sa}(t)$ denote the number of times the state–action pair (s, a) has been sampled up to round t. Let the threshold $\beta(t, \delta)$ be defined by:

$$\beta(t, \delta) := \log\left(\frac{1}{\delta}\right) + (|\mathcal{S}| - 1) \sum_{s,a} \log\left(e\left[1 + \frac{N_{sa}(t)}{|\mathcal{S}| - 1}\right]\right). \tag{5}$$

With the convention that $N_{sa}(t)\mathrm{KL}_{sa}(\hat{p}_t, p) = 0$ whenever $N_{sa}(t) = 0$, we claim that the stopping rule:

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

yields a δ-PC algorithm.

First, according to Proposition 1 in Jonsson et al. (2020) and Lemma 15 in Al Marjani and Proutiere (2021), we have: for each $p \in \mathcal{P}_{\mathrm{Test}}$,

$$\mathbb{P}_p\left[\exists t \geq 1, \sum_{s,a} N_{sa}(t) \mathrm{KL}_{sa}(\hat{p}_t, p) \geq \beta(t, \delta)\right] \leq \delta, \tag{7}$$

If the answer induced by the current empirical model $\hat{p}_t$ is incorrect, i.e., $\mathrm{Ans}(\hat{p}_t) \neq \mathrm{Ans}(p)$ (equivalently, $p \in \mathrm{Alt}(\hat{p}_t)$), the algorithm should continue sampling.

Moreover, if $p \in \mathrm{Alt}(\hat{p}_t)$ and (6) holds, then p is feasible for the infimum in (6), and therefore

$$\sum_{s,a} N_{sa}(t)\,\mathrm{KL}_{sa}(\hat{p}_t, p) \;\geq\; \beta(t,\delta).$$

By (7), this event occurs with probability at most δ. Hence, stopping at the first time t that satisfies (6) yields a δ-PC algorithm. Unfortunately, evaluating (6) is computationally difficult because $\mathrm{Alt}(\hat{p}_t)$ is generally non-convex.

5.2 From Non-Convex Constraint to Non-Convex Objective

In what follows, we transform the optimization problem into an equivalent one with a non-convex objective and convex constraints.

We first introduce a parameterized 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 denote its value by $\sigma_{\mathrm{NC}}(u,\boldsymbol{\omega},p)$. With this notation, the stopping rule (6) is

$$\sigma_{\mathrm{NC}}(0,\hat{\boldsymbol{\omega}}(t),\hat{p}_t) \geq \frac{\beta(t,\delta)}{t}.$$

Next, we define $u_{\mathrm{NO}}(\sigma,\boldsymbol{\omega},p)$ as the value of the following problem:

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

In (NO-$\sigma,\boldsymbol{\omega},p$), the objective is non-convex, while the constraint set is convex.

The next proposition, proved in Appendix D, formalizes the bijective 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. *Suppose that Assumption 1 holds and that $p \in \mathcal{P}_{\mathrm{Test}}$. Then: for all $\sigma \geq 0$ such that $u_{\mathrm{NO}}(\sigma,\boldsymbol{\omega},p) > \min_{q\in\mathcal{P}} V_p^\pi(\boldsymbol{\rho})V_q^\pi(\boldsymbol{\rho})$*

$$\sigma_{\mathrm{NC}}(u_{\mathrm{NO}}(\sigma,\boldsymbol{\omega},p),\boldsymbol{\omega},p) = \sigma,$$

for all $u \in (\min_{q\in\mathcal{P}} V_p^\pi(\boldsymbol{\rho})V_q^\pi(\boldsymbol{\rho}), u_{\mathrm{NO}}(0,\boldsymbol{\omega},p)]$,

$$u_{\mathrm{NO}}(\sigma_{\mathrm{NC}}(u,\boldsymbol{\omega},p),\boldsymbol{\omega},p) = u.$$

Proof Sketch of Proposition 2. We can replace the infimum and strict inequality in (NC-$u,\boldsymbol{\omega},p$) with a minimum and a non-strict inequality, respectively. From this, the result follows: (i) If $\sigma_{\mathrm{NC}}(u,\boldsymbol{\omega},p) \leq \sigma$, then there exists q such that $V_p^\pi(\boldsymbol{\rho})V_q^\pi(\boldsymbol{\rho}) \leq u$ 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 is equivalent to stating that if $\sigma_{\mathrm{NC}}(u,\boldsymbol{\omega},p) > \sigma$, then $u_{\mathrm{NO}}(\sigma,\boldsymbol{\omega},p) > u$. (ii) We further show that if $\sigma_{\mathrm{NC}}(u,\boldsymbol{\omega},p) \geq \sigma$, then $u_{\mathrm{NO}}(\sigma,\boldsymbol{\omega},p) \geq u$. Combining (i) and (ii) directly implies that if $\sigma_{\mathrm{NC}}(u,\boldsymbol{\omega},p) = \sigma$, then $u_{\mathrm{NO}}(\sigma,\boldsymbol{\omega},p) = u$. □

This proposition shows 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 may appear intuitive, it does not generally hold in non-convex settings. We state general conditions in Assumption 4 (in Appendix D) on the objective and the constraint set that ensure this inverse relationship, and we verify that these conditions are met in our setting.

Using Proposition 2, the stopping rule (6)—namely, $\sigma_{\mathrm{NC}}(0,\hat{\boldsymbol{\omega}}(t),\hat{p}_t) \geq \beta(t,\delta)/t$—is equivalent to

$$u_{\mathrm{NO}}(\beta(t,\delta)/t,\hat{\boldsymbol{\omega}}(t),\hat{p}_t) \;\geq\; 0,$$

since $u_{\mathrm{NO}}(\cdot,\boldsymbol{\omega},p)$ is nonincreasing in its first argument for any fixed $\boldsymbol{\omega}$ and p.

Remark 2. *Proposition 2 is a central component of our approach to solving the optimization problem using the reversed MDP formulation, and thereby to developing an instance-optimal algorithm. While we establish this result for policy testing, we also show that it holds for policy evaluation. Extending this result to other pure exploration tasks, such as best policy identification, remains an interesting direction for future work. See Appendix B for details.*

5.3 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.

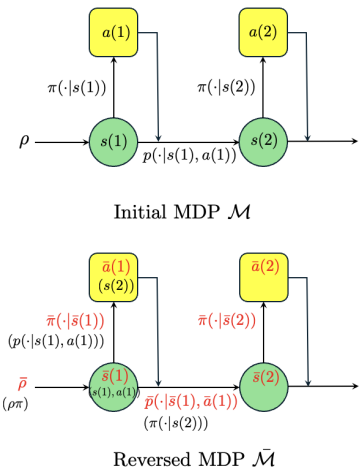

Figure 1: From the original MDP (top) to the reversed MDP (bottom). In the reversed MDP, reversed variables are shown in red and the original counterparts in black.

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

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

$$\bar{Q}^{\bar{\pi}}_{\bar{p}}(\bar{s},\bar{a}) := \mathbb{E}^{\pi}_{p}[\sum_{t=0}^{\infty}\gamma^t \bar{r}(\bar{s}(t),\bar{a}(t))|(\bar{s}(0),\bar{a}(0))=(\bar{s},\bar{a})]. \tag{9}$$

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:

$$\min_{\bar{\pi}\in\mathcal{P}} \bar{V}^{\bar{\pi}}_{\bar{p}}(\bar{\boldsymbol{\rho}}) V^{\pi}_{p}(\boldsymbol{\rho}) \quad \text{s.t.} \quad \sum_{s,a}\omega_{sa}\mathrm{KL}_{sa}(p,\bar{\pi}) \le \sigma. \tag{10}$$

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 present several results for the reversed MDP that support the analysis of constrained policy gradient methods.

5.4 Preliminary Results for Reversed MDP

We provide preliminary results that aid our analysis and may be of independent interest. These results are the counterparts, for the 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 standard MDPs.

We begin with the performance difference lemma, which for the reversed MDP coincides with the celebrated simulation lemma (Kearns and Singh, 2002) (see also Lemma A.1 in Vemula et al. (2023)).

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

$$Q^{\pi}_{p}(s,a) - Q^{\pi}_{\tilde{p}}(s,a) = \frac{\gamma}{1-\gamma}\sum_{s'\in\mathcal{S},a'\in\mathcal{A}} d^{\pi}_{p,s,a}(s',a')D,$$

$$\text{where } D = \sum_{s''\in\mathcal{S}} V^{\pi}_{\tilde{p}}(s'')\left(p(s''\mid s',a') - \tilde{p}(s''\mid s',a')\right).$$

Lemma 2 directly implies that $Q^{\pi}_{p}(s,a)$ is continuous in p, a property that is used in the proof of Proposition 2.

The next result provides an explicit expression of the gradient $\nabla_{\bar{\pi}}\bar{V}^{\bar{\pi}}_{\bar{p}}(\bar{\boldsymbol{\rho}})$ used in our policy gradient algorithm.

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

$$\frac{\partial Q^{\pi}_{p}(s,a)}{\partial p(s''\mid s',a')} = \frac{d^{\pi}_{p,s,a}(s',a')}{(1-\gamma)}\left(r(s',a') + \gamma V^{\pi}_{p}(s'')\right), \tag{11}$$

$$\frac{\partial V^{\pi}_{p}(\boldsymbol{\rho})}{\partial p(s'\mid s,a)} = \frac{d^{\pi}_{p,\boldsymbol{\rho}}(s,a)}{(1-\gamma)}\left(r(s,a) + \gamma V^{\pi}_{p}(s')\right). \tag{12}$$

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.

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

$$\left\|\nabla_p Q^{\pi}_{p}(s,a) - \nabla_p Q^{\pi}_{\tilde{p}}(s,a)\right\|_2 \le \frac{2\gamma\left|\mathcal{S}\right| r_{\max}}{(1-\gamma)^3}\left\|p-\tilde{p}\right\|_2.$$

6 OPTIMAL ALGORITHM

In this section, we present an asymptotically optimal algorithm, Policy Testing with Static Sampling (PTST).

Algorithm 1 provides the pseudocode for PTST. It has three main components.

(i) Static sampling rule (line 4). It ensures that the algorithm samples state–action pairs according to a predefined allocation vector $\boldsymbol{\omega} \in \Sigma$. At round t, the algorithm tracks $\boldsymbol{\omega}$ by sampling the pair that minimizes $\hat{\omega}_{sa}(t)/\omega_{sa}$, where $\hat{\omega}_{sa}(t) := N_{sa}(t)/t$ is the fraction of rounds in which (s,a) has been sampled up to time t.

(ii) Stopping rule (lines 3 and 7). This component is inspired by the reversed-MDP formulation (§5). We use the nested projected-gradient descent method of Algorithm 2 to approximately solve (NO-$\sigma, \boldsymbol{\omega}, p$) at time t, with $p = \hat{p}_t$, $\boldsymbol{\omega} = \hat{\boldsymbol{\omega}}(t)$, and $\sigma = \beta(t,\delta)/t$. Specifically, given a tolerance $\zeta_t > 0$, Algorithm 2 returns a value u_{ζ_t} such that

$$u_{\zeta_t} \;\geq\; u_{\mathrm{NO}}\Big(\tfrac{\beta(t,\delta)}{t},\, \hat{\boldsymbol{\omega}}(t),\, \hat{p}_t\Big) \;\geq\; u_{\zeta_t} - \zeta_t.$$

The solver searches over local KL-budget vectors and, for each fixed budget vector, runs exact projected gradient descent on the corresponding product feasible set in the reversed MDP. The algorithm stops when $u_{\zeta_t} \geq \zeta_t$.

(iii) Decision rule (line 10). Upon stopping, PTST outputs $\mathrm{Ans}(\hat{p}_\tau)$ as its decision.

Algorithm 1 Policy Testing with Static Sampling (PTST)

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

6.1 Projected Gradient Descent

In PTST, the value of (NO-$\sigma, \boldsymbol{\omega}, p$) is approximated by the nested projected-gradient descent method of Algorithm 2. For $\sigma > 0$, define the weighted budget simplex

$$\mathcal{B}_\sigma(\boldsymbol{\omega}) := \left\{ b \in \mathbb{R}_+^{|\mathcal{S}||\mathcal{A}|} : \sum_{s,a} \omega_{sa} b_{sa} \leq \sigma \right\},$$

and, for each $b \in \mathcal{B}_\sigma(\boldsymbol{\omega})$,

$$Q(b) := \{ q \in \mathcal{P} : \mathrm{KL}_{sa}(p,q) \leq b_{sa},\ \forall (s,a) \in \mathcal{S} \times \mathcal{A} \}.$$

The key observation is that the global KL ball is exactly the union of these product boxes.

Proposition 3 (Exact reduction to product-box budgets)**.** *For every $\sigma > 0$, $\boldsymbol{\omega} \in \Sigma$, and $p \in \mathcal{P}$,*

$$u_{\mathrm{NO}}(\sigma, \boldsymbol{\omega}, p) = \min_{b \in \mathcal{B}_\sigma(\boldsymbol{\omega})} \; \min_{q \in Q(b)} V_p^\pi(\boldsymbol{\rho}) V_q^\pi(\boldsymbol{\rho}).$$

Indeed, if $q \in \Pi_\sigma^p$, then setting $b_{sa} := \mathrm{KL}_{sa}(p,q)$ gives $b \in \mathcal{B}_\sigma(\boldsymbol{\omega})$ and $q \in Q(b)$. Conversely, if $q \in Q(b)$ for some $b \in \mathcal{B}_\sigma(\boldsymbol{\omega})$, then $\sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p,q) \leq \sum_{s,a} \omega_{sa} b_{sa} \leq \sigma$, so $q \in \Pi_\sigma^p$. The proof of Proposition 3 is deferred to Appendix E.2.

Thus, solving (NO-$\sigma, \boldsymbol{\omega}, p$) can be decomposed into an outer search over $b \in \mathcal{B}_\sigma(\boldsymbol{\omega})$ and an inner exact projected-gradient method over the product set $Q(b)$. The inner projection factorizes across state–action pairs:

$$\mathbf{proj}_{Q(b)}(x) = \Big(\mathbf{proj}_{\{q \in \Delta(\mathcal{S}) : \frac{\mathrm{KL}_{sa}(p,q)}{b_{sa}} \leq 1\}}(x_{sa}) \Big)_{s \in \mathcal{S}, a \in \mathcal{A}}$$

so each block projection reduces to a one-dimensional dual root-finding problem on a KL-ball-constrained simplex.

Algorithm 2 Nested Projected Gradient Descent

Input: $(p, \zeta, \sigma, \boldsymbol{\omega})$.
if $V_p^\pi(\boldsymbol{\rho}) = 0$ **then**
  Output: 0
end if
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 5.3
$r_{\max} \leftarrow \max_{s,a} |r(s,a)|$
$L \leftarrow 2(\gamma|\bar{\mathcal{A}}|+1) r_{\max}/(1-\gamma)^3$
$h_\zeta \leftarrow \zeta^2(1-\gamma)^4 / \big(18|\mathcal{S}|^2|\mathcal{A}|^2 |V_p^\pi(\boldsymbol{\rho})|^2 r_{\max}^2\big)$
Define $\mathcal{G}_{h_\zeta} := \{ b \in \mathcal{B}_\sigma(\boldsymbol{\omega}) : b_{sa} \in h_\zeta \mathbb{Z}_+, \forall (s,a) \}$
$M \leftarrow \left\lceil \frac{384(\gamma|\bar{\mathcal{A}}|+1)|\bar{\mathcal{S}}| r_{\max} |V_p^\pi(\boldsymbol{\rho})| \|1/\bar{\boldsymbol{\rho}}\|_\infty^2}{(1-\gamma)^5 \zeta} \right\rceil$
for each $b \in \mathcal{G}_{h_\zeta}$ **do**
  $\bar{\pi}_b^{(0)} \leftarrow p$
  for $k = 0, 1, \ldots, M-1$ **do**
    $\bar{\pi}_b^{(k+1)} \leftarrow$ $\mathbf{proj}_{Q(b)} \left(\bar{\pi}_b^{(k)} - \frac{\mathrm{sign}(V_p^\pi(\boldsymbol{\rho}))}{L} \nabla_{\bar{\pi}} \bar{V}_{\bar{p}}^{\bar{\pi}_b^{(k)}}(\bar{\boldsymbol{\rho}}) \right)$
  end for
  $u_b \leftarrow V_p^\pi(\boldsymbol{\rho}) \bar{V}_{\bar{p}}^{\bar{\pi}_b^{(M)}}(\bar{\boldsymbol{\rho}})$
end for
Output: $u_\zeta \leftarrow \min_{b \in \mathcal{G}_{h_\zeta}} u_b$

Line 8 fixes the grid resolution explicitly. By Lemma 12, this choice guarantees that coordinatewise flooring of an optimal budget vector to the grid $\mathcal{G}_{h_\zeta}$ incurs an outer discretization error of at most $\zeta/3$.

The following theorem provides the fixed-budget inner convergence guarantee together with the global ap-

proximation guarantee of the nested projected-gradient descent method.

Theorem 2. *Under Assumptions 1 and 2, fix any $b \in \mathcal{B}_\sigma(\boldsymbol{\omega})$. Let $(\bar{\pi}_b^{(k)})_{k\geq 0}$ be the exact projected-gradient sequence generated by the inner loop of Algorithm 2 on the product set $Q(b)$, and define*

$$\Delta_b^{(k)} := \bar{V}_{\bar{p}}^{\bar{\pi}_b^{(k)}}(\bar{\boldsymbol{\rho}}) V_p^\pi(\boldsymbol{\rho}) - \min_{q\in Q(b)} V_p^\pi(\boldsymbol{\rho}) V_q^\pi(\boldsymbol{\rho}).$$

Then, for all $k \geq 1$,

$$\Delta_b^{(k)} \leq \frac{128(\gamma|\bar{\mathcal{A}}|+1)|\bar{\mathcal{S}}| r_{\max} |V_p^\pi(\boldsymbol{\rho})|}{(1-\gamma)^5 k} \left\| 1/\bar{\boldsymbol{\rho}} \right\|_\infty^2.$$

In particular, if Algorithm 2 is run with the values of h_ζ and M defined above, then its output u_ζ satisfies

$$u_\zeta \;\geq\; u_{\mathrm{NO}}(\sigma, \boldsymbol{\omega}, p) \;\geq\; u_\zeta - \zeta.$$

The proof of Theorem 2 is provided in Appendix E. The key point is that, after the reduction of Proposition 3, each inner problem is solved over a genuine product set, so the projected-gradient analysis becomes exact and no slack mixing is needed.

Finally, when Algorithm 2 is used inside PTST (Algorithm 1), we replace the unknown transition kernel p with its empirical estimator $\hat{p}_t$, the tolerance ζ with ζ_t, the allocation $\boldsymbol{\omega}$ with $\hat{\boldsymbol{\omega}}(t)$, and the radius σ with the threshold $\beta(t,\delta)/t$.

Remark 3 (Tradeoff of the product-box decomposition)**.** *The exact reduction to product-box budgets transfers the global coupling induced by the KL constraint to an outer optimization over the local budget vector b. For each fixed b, the feasible set becomes a genuine product set, so the inner projected-gradient method admits an exact analysis on each product box, yielding the approximation guarantee of Theorem 2.*

6.2 Optimality of PTST

The following theorem, proved in Appendix F, establishes the asymptotic optimality of the PTST algorithm.

Theorem 3. *Suppose Assumptions 1 and 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.

Proof Sketch of Theorem 3. 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. $\boldsymbol{\omega}$) and such that $\mathbb{P}_p[\mathcal{C}_T(\xi)^c] < \infty$ for large enough T.

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 G) by using Lemma 1. This result is a consequence of a series of theorems in parameterized optimization and real analysis.

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 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. □

7 EXPERIMENTS

In this section, we evaluate the proposed method in several settings. To the best of our knowledge, there is no prior algorithm tailored to policy testing in infinite-horizon discounted MDPs. Consequently, directly applying methods for thresholding bandits or other pure exploration problems is not straightforward. Instead, we adapt a best policy identification method to serve as a baseline: the KLB-TS algorithm of Al Marjani and Proutiere (2021). To align KLB-TS with the policy testing objective, we consider two policies: one identical to π and another, π', satisfying $V_p^{\pi'}(\boldsymbol{\rho}) = 0$. We fix the sampling rule to be uniform over all state–action pairs. The stopping rule of KLB-TS is based on an upper bound obtained by convexifying the original minimax optimization problem. For the empirical study, we use a heuristic variant of PTST whose inner optimization directly projects onto $\bar{\Pi}_\sigma^p$ via SLSQP, rather than the nested product-box solver analyzed in Section 6.1.

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 Tables 1, 2, and 3 in Appendix J for the respective settings. For each setting, we vary δ from 10^{-15} to 10^{-2}. The results are shown in Figure 2. In all three cases, PTST outperforms the

baseline for all values of δ. For further details, please refer to Appendix J.

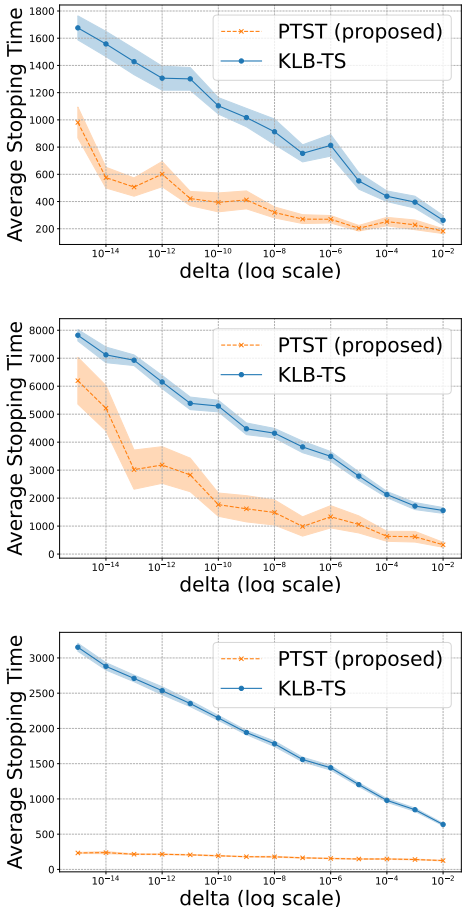

Figure 2: Comparison of average stopping times and δ for the proposed algorithm and KLB-TS. The top, middle, and bottom panels correspond to $|\mathcal{S}| = |\mathcal{A}| = 2, 3, 5$, respectively. Results are averaged over 30 instances. Error bars indicate the standard error.

8 LIMITATIONS AND DISCUSSION

In this paper, we formulated the policy testing problem in discounted, tabular MDPs and characterized its instance-specific complexity under access to a generative model with a static sampling allocation. To the best of our knowledge, this is the first computationally tractable algorithm that achieves instance optimality for pure exploration in MDPs. The key ingredients are a reformulation that converts a non-convex lower-bound problem into a tractable program, an exact decomposition into product-box-constrained subproblems, and projected policy gradient methods in a reversed MDP.

Although our focus is policy testing, we expect the approach to extend to other pure exploration tasks, such as policy evaluation. We outline the required assumptions and corresponding optimization problems in Appendix B.

Our current guarantees are restricted to the static sampling setting. The stopping rule based on the reversed MDP applies directly under static allocations, but extending the full framework, including the sampling design and the optimality guarantees, to adaptive sampling is an interesting direction for future work. Moreover, while Theorem 2 analyzes the nested product-box solver, the empirical study uses the simpler direct-projection heuristic described in Appendix J; providing guarantees for this heuristic is left for future work.

Overall, we hope these findings provide a foundation for developing more efficient algorithms for pure exploration in MDPs.

Acknowledgements

Kaito Ariu is supported by JSPS KAKENHI Grant Number 25K21291. Po-An Wang is supported by NSTC Grant Number 114-2118-M-007 -002 -MY2. A. Proutiere research is supported by Vetenskapprådet, Digital Futures, and the Wallenberg AI, Autonomous Systems and Software program.

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.

Tuynman, A., Degenne, R., and Kaufmann, E. (2024). Finding good policies in average-reward markov decision processes without prior knowledge. *Advances in Neural Information Processing Systems*, 37:109948–109979.

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.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Policy Testing in Markov Decision Processes Supplementary Materials

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 inner projected-gradient analysis (Section 6.1) 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 work with nonlinear KL constraints by decomposing the global feasible set into product-box subproblems, and combine the resulting inner projected-gradient guarantees with an outer budget search and a sensitivity analysis (see Sections 6 and G).

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 general formulations (Russo and Pacchiano, 2025). We discuss extensions to policy evaluation in Section B.

B Extensions Toward Other Pure Exploration Tasks

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 3 for policy evaluation. Notice that if Assumption 3 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 3. *$\rho_s > 0$ 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 6 and 5.2, solving the stopping condition boils down to identify whether the minimal value of the following two optimization problems is larger than $\beta(t,\delta)/t$.

$$\inf_q \sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p,q) \quad \text{s.t. } V_q^\pi(\boldsymbol{\rho}) - V_p^\pi(\boldsymbol{\rho}) + \varepsilon < 0, \tag{13}$$

and

$$\inf_q \sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p,q) \quad \text{s.t. } V_p^\pi(\boldsymbol{\rho}) - V_q^\pi(\boldsymbol{\rho}) + \varepsilon < 0. \tag{14}$$

Under Assumption 3, 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 (14) 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 4 holds with the corresponding substitution in Lemma 5, then one can implement a projected policy gradient method to approximate the value of its dual problem, as described in Section 6.1.

$$\min_q V_p^\pi(\boldsymbol{\rho}) - V_q^\pi(\boldsymbol{\rho}) + \varepsilon \quad \text{s.t.} \sum_{sa} \omega_{sa} \mathrm{KL}_{sa}(p, q) \leq \frac{\beta(t, \delta)}{t}. \tag{15}$$

Lemma 5. *When $\{q \in \mathcal{P} : V_p^\pi(\boldsymbol{\rho}) - V_q^\pi(\boldsymbol{\rho}) + \varepsilon < 0\} \neq \emptyset$, Assumption 4 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. □

Best Policy Identification. Here we discuss the extension of our methods to best policy identification. As shown in Al Marjani and Proutiere (2021), $\mathrm{Alt}(p) := \bigcup_{s \in \mathcal{S}} \bigcup_{a \neq \pi(s)} \{q \in \mathcal{P} : Q_q^\pi(s, a) > V_q^\pi(s)\}$. The nonconvex optimization we are interested in is

$$\inf_{q \in \mathrm{Alt}(p)} \sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p, q) = \min_{s \in \mathcal{S}} \min_{a \neq \pi(s)} \inf_{q : Q_q^\pi(s,a) > V_q^\pi(s)} \sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p, q).$$

For a fixed pair (s, a) such that $s \in \mathcal{S}, a \neq \pi(s)$, we define (NC-u) as:

$$\inf_{q \in \mathcal{P}} \sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p, q) \quad \text{s.t. } V_q^\pi(s) - Q_q^\pi(s, a) < u.$$

As described in Section 5.2, (NC-u) can be transformed into (NO-σ) written below.

$$\min_{q \in \mathcal{P}} V_q^\pi(s) - Q_q^\pi(s, a) \quad \text{s.t. } \sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p, q) \leq \sigma.$$

Since $V_q^\pi(s) = Q_q^\pi(s, \pi(s))$, the objective function of (NO-σ) becomes $Q_q^\pi(s, \pi(s)) - Q_q^\pi(s, a)$.

Leveraging the construction of the reversed MDP in Section 5.3, we obtain

$$Q_q^\pi(s, \pi(s)) - Q_q^\pi(s, a) = \bar{V}_{\bar{p}}^\pi(\bar{s}) - \bar{V}_{\bar{p}}^\pi(\bar{s}'),$$

where $\bar{s} = (s, \pi(s))$ and $\bar{s}' = (s, a)$. The objective function is now the difference between the value functions of two states. Optimizing such objective functions may not be possible using the standard policy gradient method (as in our paper). However, we expect that this issue could be resolved by employing a more sophisticated algorithm design and analysis. This would then yield an instance-optimal algorithm for the best policy identification setting.

Toward the Forward Model. Extending our results to the forward (online, single-trajectory) interaction model—in which the next state depends on the current state and action—remains an important direction for future work, likely building on techniques from Al Marjani et al. (2021). Such extensions may require assuming that all state–action pairs with positive probability under the policy are visited. Most algorithms in this literature assume that the transition kernel induces a communicating MDP, i.e., every state is reachable from every other state under some deterministic policy (Al Marjani et al., 2021; Russo and Vannella, 2024; Taupin et al., 2023). However, this assumption may be unrealistic in practice, particularly when the presence of transient states is unknown. Recent work by Tuynman et al. (2024) weakens this requirement by assuming that the MDPs are weakly communicating, which is still stronger than Assumption 1.

While this extension is important, it is largely orthogonal to our main contribution, which establishes instance-specific optimality in the generative-model setting with computational efficiency. We therefore view our results as a step toward addressing more general and realistic settings.

C Instance-Specific Sample Complexity Lower Bound–Proof of Theorem 1

Proof of Theorem 1. Consider two cases, (i) $\inf_{q\in \mathrm{Alt}(p)} \sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p,q) = \infty$; (ii) $\inf_{q\in \mathrm{Alt}(p)} \sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p,q) < \infty$. Observe that our theorem holds directly in case (i), one only needs to focus on the case (ii). By our Lemma 6, one can derive $\tilde{q} \in \mathrm{Alt}(p)$ such that $\mathrm{KL}_{sa}(p,\tilde{q}) < \infty$ for all $s \in \mathcal{S}, a \in \mathcal{A}$

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,

$$\sum_{s,a} \mathbb{E}_p[N_{sa}(\tau)] \mathrm{KL}_{sa}(p,q) \geq \mathrm{kl}(\mathbb{P}_p[E], \mathbb{P}_q[E]), \tag{16}$$

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 $\mathrm{kl}(\mathbb{P}_p[E], \mathbb{P}_q[E]) \geq \mathrm{kl}(\delta, 1-\delta)$. Thanks to the assumption on sampling rule, for any ε, one can find $c_\varepsilon > 0$ such that $\mathbb{E}_p[N_{sa}(t)] \leq t(\omega_{sa}+\varepsilon) + c_\varepsilon$, $\forall s \in \mathcal{S}, a \in \mathcal{A}$. As (16) holds for any $q \in \mathrm{Alt}(p)$,

$$\begin{aligned}
\mathrm{kl}(\delta, 1-\delta) &\leq \inf_{q\in \mathrm{Alt}(p)} \mathbb{E}_p[\tau] \sum_{s,a} \frac{\mathbb{E}_p[N_{sa}(\tau)]}{\mathbb{E}_p[\tau]} \mathrm{KL}_{sa}(p,q) \\
&\leq \mathbb{E}_p[\tau] \left(\inf_{q\in \mathrm{Alt}(p)} \sum_{s,a} (\omega_{sa} + \varepsilon + \frac{c_\varepsilon}{\mathbb{E}_p[\tau]}) \mathrm{KL}_{sa}(p,q) \right) \qquad (17) \\
&\leq \mathbb{E}_p[\tau] \left(\sum_{s,a} (\omega_{sa} + \varepsilon + c_\varepsilon) \mathrm{KL}_{sa}(p,\tilde{q}) \right), \qquad (18)
\end{aligned}$$

where the last inequality follows as $\mathbb{E}_p[\tau] \geq 1$ and $\tilde{q} \in \mathrm{Alt}(p)$. Since $\left(\sum_{s,a} (\omega_{sa} + \varepsilon + c_\varepsilon) \mathrm{KL}_{sa}(p,\tilde{q}) \right)$ is finite, we conclude $\mathbb{E}_p[\tau] \to \infty$ as $\mathrm{kl}(\delta, 1-\delta) \to \infty$ if $\delta \to 0$ from (18). Rearranging (17) yields that

$$\left(\inf_{q\in \mathrm{Alt}(p)} \sum_{s,a} (\omega_{sa} + \varepsilon + \frac{c_\varepsilon}{\mathbb{E}_p[\tau]}) \mathrm{KL}_{sa}(p,q) \right)^{-1} \leq \frac{\mathbb{E}_p[\tau]}{\mathrm{kl}(\delta, 1-\delta)} \tag{19}$$

Using the fact that $\mathrm{kl}(\delta, 1-\delta) \approx \log(1/\delta)$ and $\mathbb{E}_p[\tau] \to \infty$ when δ goes to zero, one can conclude the theorem by taking the limit inferior on both sides of (19) as ε is taken arbitrarily. □

Lemma 6. *If* $\inf_{q\in \mathrm{Alt}(p)} \sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p,q) < \infty$, *there is* $\tilde{q} \in \mathrm{Alt}(p)$ *such that* $\mathrm{KL}_{sa}(p,\tilde{q}) < \infty$, $\forall s \in \mathcal{S}, a \in \mathcal{A}$.

Proof. As $\inf_{q\in \mathrm{Alt}(p)} \sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p,q) < \infty$, there is $q' \in \mathrm{Alt}(p)$ (or equivalently $V_p^\pi(\boldsymbol{\rho}) V_q^\pi(\boldsymbol{\rho}) < 0$) such that $\sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p,q') < \infty$. It is done if $\omega_{sa} > 0$ for all s, a as one can take $\tilde{q} = q'$.
Suppose there are pairs (s,a) such that $\omega_{sa} = 0$, $\mathrm{KL}_{sa}(p,q') = \infty$. Observe that $\mathrm{KL}_{sa}(p,q') = \infty$ if and only if $\exists s' \in \mathcal{S}$ such that $q'(s' \mid s,a) = 0$ but $p(s' \mid s,a) > 0$. Since $V_p^\pi(\boldsymbol{\rho}) V_q^\pi(\boldsymbol{\rho})$ is a continuous mapping on q by Lemma 2, one can hence obtain $\tilde{q} \in \mathrm{Alt}(p)$ which satisfies $\tilde{q}(s' \mid s,a) > 0$ for all s', s, a by slightly perturbing q'. Thus, one has $\mathrm{KL}_{sa}(p,\tilde{q}) < \infty$, $\forall s \in \mathcal{S}, a \in \mathcal{A}$. □

D 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

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

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 $\sigma_{\mathrm{NC}}(u)$ and $u_{\mathrm{NO}}(\sigma)$ denote the value of (NC-u) and that of (NO-σ), respectively. As one can see, in Section 5.2, we make the substitutions $\mathcal{X}=\mathcal{P}$, $x=q$, $h(q)=\sum_{sa}\omega_{sa}\mathrm{KL}_{sa}(p,q)$, and $g(q)=V_p^\pi(\boldsymbol{\rho})V_q^\pi(\boldsymbol{\rho})$. As stated in Proposition 4, the desired bijection holds under the mild assumptions in Assumption 4. In Appendix D.1, we verify that these assumptions hold for the aforementioned substitution.

Assumption 4. *The following conditions hold.*

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

(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 4, there exists a bijection between problems (NC-u) and (NO-σ).

Proposition 4. *Under Assumption 4, we have*

$$\text{for all } \sigma\ge 0 \text{ such that } u_{\mathrm{NO}}(\sigma) > \min_{x\in\mathcal{X}} g(x), \quad \sigma_{\mathrm{NC}}(u_{\mathrm{NO}}(\sigma)) = \sigma, \tag{20}$$

$$\text{for all } u\in(\min_{x\in\mathcal{X}} g(x), u_{\mathrm{NO}}(0)] \quad u_{\mathrm{NO}}(\sigma_{\mathrm{NC}}(u)) = u. \tag{21}$$

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

Proof of Proposition 4. We first show (20). Let $\sigma\ge 0$ and $u = u_{\mathrm{NO}}(\sigma) > \min_{x\in\mathcal{X}} g(x)$. By Lemma 7 and Lemma 8, we get $\sigma_{\mathrm{NC}}(u)\le\sigma$ and $\sigma_{\mathrm{NC}}(u)\ge\sigma$ respectively, which directly yield (20).
For proving (21), we consider $u\in(\min_{x\in\mathcal{X}} g(x), u_{\mathrm{NO}}(0)]$. Using intermediate value theorem and the continuity of $u_{\mathrm{NO}}(\cdot)$ proved in Lemma 22 given in Appendix H.3, we deduce that there is $\sigma\in[0,\infty)$ such that $u_{\mathrm{NO}}(\sigma)=u$. As a consequence,

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

where the second equation is the application of (20).

□

Lemma 7. *Under Assumption 4, whenever* $\sigma\ge 0, u\in(\min_{x\in\mathcal{X}} g(x),\infty)$*, and* $u_{\mathrm{NO}}(\sigma)\le u$*, then* $\sigma_{\mathrm{NC}}(u)\le\sigma$ *holds.*

Proof of Lemma 7. As $u_{\mathrm{NO}}(\sigma)\le u$, $\exists x'\in\mathcal{X}$, such that $h(x')\le\sigma$, and $g(x')\le u$. If $g(x')<u$ (case i), then x' is a feasible point for (NC-u), and therefore $\sigma_{\mathrm{NC}}(u)\le h(x')\le\sigma$. 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 4-(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 4-(d), $\lim_{n\to\infty} h(x_n) = h(x')\le\sigma$, which yields $\sigma_{\mathrm{NC}}(u)\le\sigma$. □

Lemma 8. *Whenever* $\sigma\ge 0, u\in\mathbb{R}$ *and* $u_{\mathrm{NO}}(\sigma)\ge u$*,* $\sigma_{\mathrm{NC}}(u)\ge\sigma$ *holds.*

Proof of Lemma 8. Suppose in contrast, $\sigma_{\mathrm{NC}}(u)<\sigma$. There exist $x\in\mathcal{X}$ such that $g(x)<u$, $h(x)<\sigma$. Hence $u_{\mathrm{NO}}(\sigma)<u$, which contradicts that $u_{\mathrm{NO}}(x)\ge u$. □

D.1 Proof of Proposition 2

Proof of Proposition 2. Thanks to Proposition 4, the proof is completed by verifying that the following substitutions satisfy Assumption 4 when $p \in \mathcal{P}_{\mathrm{Test}}$.

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

(a) Let $\underline{x} = p$, one has that $h(\underline{x}) = \sum_{sa} w_{sa} \mathrm{KL}_{sa}(p, p) = 0$.
(b) Due to Assumption 1 and inequalities (1), there exists $q \in \mathcal{P}$ such that $\mathrm{sign}(V_q^\pi(\boldsymbol{\rho})) \neq \mathrm{sign}(V_p^\pi(\boldsymbol{\rho}))$, 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 5.3. 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).

□

E Convergence Analysis of Nested Projected Gradient Descent–Proof of Theorem 2

Notation. Throughout this appendix, we work with the sign-normalized reversed-MDP objective

$$f(\bar{\pi}) := \mathrm{sign}(V_p^\pi(\boldsymbol{\rho}))\, \bar{V}_{\bar{p}}^{\bar{\pi}}(\bar{\boldsymbol{\rho}}),$$

where $\bar{p}$ is the reversed-MDP transition kernel induced by the fixed target policy π as defined in Section 5.3. Recall from Section 6.1 that

$$Q(b) = \{\bar{\pi} \in \mathcal{P} : \mathrm{KL}_{sa}(p, \bar{\pi}) \leq b_{sa},\ \forall (s,a) \in \mathcal{S} \times \mathcal{A}\}.$$

For any $b \in \mathcal{B}_\sigma(\boldsymbol{\omega})$, define

$$f_b^\star := \min_{\bar{\pi} \in Q(b)} f(\bar{\pi}).$$

Roadmap. This appendix proves Theorem 2. We first analyze the inner projected-gradient method on a fixed $b \in \mathcal{B}_\sigma(\boldsymbol{\omega})$ and establish its convergence rate (Lemma 11). The proof follows the same general strategy as (Xiao, 2022): Lemma 10 provides the weak gradient-domination property that turns control of the gradient mapping into control of value suboptimality, while Lemma 4 gives the smoothness estimate needed for the projected-gradient descent argument. We then combine this inner-loop guarantee with the discretization of the outer search over $\mathcal{B}_\sigma(\boldsymbol{\omega})$; here Lemma 12 controls the error induced by the budget mesh. Putting these ingredients together yields the proof of Theorem 2.

For readability, we defer two auxiliary components to separate subsections. Appendix E.1 recalls the general projected-gradient convergence template used in the proof, and Appendix E.2 collects the product-box reduction argument underlying the fixed-b analysis.

To derive the weak-gradient domination, we need an analogue of variational gradient domination (Lemma 4 in (Agarwal et al., 2021)) in a product-set.

Lemma 9 (Product-set variational gradient domination)**.** *Fix $b \in \mathcal{B}_\sigma(\boldsymbol{\omega})$ and let $\bar{\pi}_b^\star \in \arg\min_{\bar{\pi} \in Q(b)} f(\bar{\pi})$. Then, for every $\bar{\pi} \in Q(b)$,*

$$f(\bar{\pi}) - f_b^\star \leq \left\| \frac{d_{\bar{p},\bar{\boldsymbol{\rho}}}^{\bar{\pi}_b^\star}}{d_{\bar{p},\bar{\boldsymbol{\rho}}}^{\bar{\pi}}} \right\|_\infty \max_{\bar{\pi}' \in Q(b)} \langle \nabla_{\bar{\pi}} f(\bar{\pi}),\, \bar{\pi} - \bar{\pi}' \rangle .$$

In particular,

$$f(\bar{\pi}) - f_b^\star \leq \frac{1}{1-\gamma} \left\| 1/\bar{\boldsymbol{\rho}} \right\|_\infty \max_{\bar{\pi}' \in Q(b)} \langle \nabla_{\bar{\pi}} f(\bar{\pi}),\, \bar{\pi} - \bar{\pi}' \rangle .$$

Proof. Fix $\bar{\pi} \in Q(b)$ and let $\bar{\pi}_b^\star \in \arg\min_{\bar{\pi}' \in Q(b)} f(\bar{\pi}')$. Write the reversed-MDP advantage as

$$A_{\bar{p}}^{\bar{\pi}}((s,a),s') := \bar{Q}_{\bar{p}}^{\bar{\pi}}((s,a),s') - \bar{V}_{\bar{p}}^{\bar{\pi}}(s,a).$$

For each (s,a), define

$$g_b^{\bar{\pi}}(s,a) := \max_{\substack{q(\cdot|s,a)\in\Delta(\mathcal{S}):\\ \mathrm{KL}_{sa}(p,q)\le b_{sa}}} \sum_{s'\in\mathcal{S}} q(s')\big(-A_{\bar{p}}^{\bar{\pi}}((s,a),s')\big).$$

Since $\bar{\pi}_b^\star(\cdot \mid s,a)$ satisfies $\mathrm{KL}_{sa}(p,\bar{\pi}_b^\star) \le b_{sa}$, we have

$$\sum_{s'} \bar{\pi}_b^\star(s' \mid s,a)\big(-A_{\bar{p}}^{\bar{\pi}}((s,a),s')\big) \le g_b^{\bar{\pi}}(s,a).$$

By the performance-difference lemma on the reversed MDP (see Lemma 2),

$$f(\bar{\pi}) - f_b^\star = \frac{1}{1-\gamma} \sum_{s,a} d_{\bar{p},\bar{\boldsymbol{\rho}}}^{\bar{\pi}_b^\star}(s,a) \sum_{s'} \bar{\pi}_b^\star(s' \mid s,a)\big(-A_{\bar{p}}^{\bar{\pi}}((s,a),s')\big)$$

$$\le \frac{1}{1-\gamma} \sum_{s,a} d_{\bar{p},\bar{\boldsymbol{\rho}}}^{\bar{\pi}_b^\star}(s,a) g_b^{\bar{\pi}}(s,a).$$

Multiplying and dividing by $d_{\bar{p},\bar{\boldsymbol{\rho}}}^{\bar{\pi}}(s,a)$ yields

$$f(\bar{\pi}) - f_b^\star \le \frac{1}{1-\gamma} \left\| \frac{d_{\bar{p},\bar{\boldsymbol{\rho}}}^{\bar{\pi}_b^\star}}{d_{\bar{p},\bar{\boldsymbol{\rho}}}^{\bar{\pi}}} \right\|_\infty \sum_{s,a} d_{\bar{p},\bar{\boldsymbol{\rho}}}^{\bar{\pi}}(s,a) g_b^{\bar{\pi}}(s,a).$$

For each (s,a), choose

$$q_{sa}^\star \in \arg\max_{\substack{q(\cdot|s,a)\in\Delta(\mathcal{S}):\\ \mathrm{KL}_{sa}(p,q)\le b_{sa}}} \sum_{s'} q(s')(-A_{\bar{p}}^{\bar{\pi}}((s,a),s')),$$

and define $\hat{\pi} \in Q(b)$ by $\hat{\pi}(\cdot \mid s,a) = q_{sa}^\star(\cdot)$. Because $Q(b)$ is a direct product, $\hat{\pi} \in Q(b)$. Using $\sum_{s'} \bar{\pi}(s' \mid s,a) A_{\bar{p}}^{\bar{\pi}}((s,a),s') = 0$ for all (s,a) and the directional policy-gradient identity from Lemma 3, we obtain

$$\sum_{s,a} d_{\bar{p},\bar{\boldsymbol{\rho}}}^{\bar{\pi}}(s,a) g_b^{\bar{\pi}}(s,a) = (1-\gamma)\,\langle \nabla_{\bar{\pi}} f(\bar{\pi}),\, \bar{\pi} - \hat{\pi} \rangle$$

$$\le (1-\gamma) \max_{\bar{\pi}' \in Q(b)} \langle \nabla_{\bar{\pi}} f(\bar{\pi}),\, \bar{\pi} - \bar{\pi}' \rangle .$$

Substituting this bound yields the first claim. The second one follows from $d_{\bar{p},\bar{\boldsymbol{\rho}}}^{\bar{\pi}}(s,a) \ge (1-\gamma)\bar{\rho}(s,a)$ for all (s,a). □

Lemma 10 (Weak gradient-mapping domination on a fixed product box)**.** *Fix $b \in \mathcal{B}_\sigma(\boldsymbol{\omega})$ and assume that f is L-smooth on $Q(b)$. Define*

$$T_{L,b}(\bar{\pi}) := \mathbf{proj}_{Q(b)}\left(\bar{\pi} - \frac{1}{L}\nabla_{\bar{\pi}} f(\bar{\pi})\right), \qquad G_{L,b}(\bar{\pi}) := L\big(\bar{\pi} - T_{L,b}(\bar{\pi})\big).$$

Then, for every $\bar{\pi} \in Q(b)$,

$$f(T_{L,b}(\bar{\pi})) - f_b^\star \le \frac{2\sqrt{2|\bar{\mathcal{S}}|}}{1-\gamma} \left\|1/\bar{\boldsymbol{\rho}}\right\|_\infty \|G_{L,b}(\bar{\pi})\|_2.$$

Proof. Fix $\bar{\pi} \in Q(b)$ and let $T := T_{L,b}(\bar{\pi})$. Applying Lemma 9 at $T \in Q(b)$ gives

$$f(T) - f_b^\star \le \frac{1}{1-\gamma} \left\|1/\bar{\boldsymbol{\rho}}\right\|_\infty \max_{\bar{\pi}' \in Q(b)} \langle \nabla_{\bar{\pi}} f(T),\, T - \bar{\pi}' \rangle .$$

Since f is L-smooth and T is the exact projected-gradient step of $\bar{\pi}$ on $Q(b)$, Theorem 1 of Nesterov (2013) yields

$$\langle \nabla_{\bar{\pi}} f(T),\, T - \bar{\pi}' \rangle \le 2\|G_{L,b}(\bar{\pi})\|_2 \, \|T - \bar{\pi}'\|_2, \qquad \forall \bar{\pi}' \in Q(b).$$

Now $Q(b) \subseteq \mathcal{P} = \Delta(\bar{\mathcal{A}})^{\bar{\mathcal{S}}}$, and the Euclidean diameter of $\mathcal{P}$ is at most $\sqrt{2|\bar{\mathcal{S}}|}$. Hence

$$\max_{\bar{\pi}' \in Q(b)} \|T - \bar{\pi}'\|_2 \leq \sqrt{2|\bar{\mathcal{S}}|},$$

which implies the claim. □

Lemma 11 (Inner projected-gradient rate on a fixed product box)**.** *Fix $b \in \mathcal{B}_\sigma(\boldsymbol{\omega})$ and let $(\bar{\pi}_b^{(k)})_{k\geq 0}$ be the exact projected-gradient sequence on $Q(b)$:*

$$\bar{\pi}_b^{(k+1)} = \mathbf{proj}_{Q(b)}\left(\bar{\pi}_b^{(k)} - \frac{1}{L}\nabla_{\bar{\pi}} f(\bar{\pi}_b^{(k)})\right).$$

Assume that f is L-smooth on $Q(b)$. Then, for every $k \geq 1$,

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

Proof. By Lemma 10, the gradient-mapping domination condition holds on $Q(b)$ with

$$\omega_b := \frac{(1-\gamma)^2}{16|\bar{\mathcal{S}}|\,\|1/\bar{\boldsymbol{\rho}}\|_\infty^2}.$$

Applying Theorem 4 with this value of ω_b yields the claim. □

Lemma 12 (Explicit grid discretization for product-box budgets)**.** *Define*

$$\varphi(b) := \min_{q \in Q(b)} V_p^\pi(\boldsymbol{\rho}) V_q^\pi(\boldsymbol{\rho}).$$

Let $b^\star \in \arg\min_{b\in\mathcal{B}_\sigma(\boldsymbol{\omega})} \varphi(b)$ and let $h > 0$. Define $\tilde{b} \in \mathcal{B}_\sigma(\boldsymbol{\omega})$ by coordinatewise flooring:

$$\tilde{b}_{sa} := h\,\lfloor b^\star_{sa}/h \rfloor, \qquad (s,a) \in \mathcal{S}\times\mathcal{A}.$$

Then

$$\varphi(\tilde{b}) - \varphi(b^\star) \leq \frac{\sqrt{2}\,|\mathcal{S}||\mathcal{A}|\,|V_p^\pi(\boldsymbol{\rho})|\,r_{\max}}{(1-\gamma)^2}\sqrt{h}.$$

In particular, if

$$h_\zeta := \frac{\zeta^2(1-\gamma)^4}{18\,|\mathcal{S}|^2|\mathcal{A}|^2\,|V_p^\pi(\boldsymbol{\rho})|^2\,r_{\max}^2},$$

then

$$\varphi(\tilde{b}) \leq \varphi(b^\star) + \zeta/3.$$

Proof. Choose $q^\star \in \Pi_\sigma^p$ such that

$$V_p^\pi(\boldsymbol{\rho}) V_{q^\star}^\pi(\boldsymbol{\rho}) = u_{\mathrm{NO}}(\sigma, \boldsymbol{\omega}, p),$$

which exists because Π_σ^p is compact and the objective is continuous. Set

$$b^\star_{sa} := \mathrm{KL}_{sa}(p, q^\star), \qquad (s,a) \in \mathcal{S}\times\mathcal{A}.$$

Then $b^\star \in \mathcal{B}_\sigma(\boldsymbol{\omega})$, $q^\star \in Q(b^\star)$, and Proposition 3 gives

$$\varphi(b^\star) = u_{\mathrm{NO}}(\sigma, \boldsymbol{\omega}, p) = V_p^\pi(\boldsymbol{\rho}) V_{q^\star}^\pi(\boldsymbol{\rho}).$$

For each (s,a) with $b^\star_{sa} > 0$, define

$$\alpha_{sa} := 1 - \tilde{b}_{sa}/b^\star_{sa}, \qquad \tilde{q}(\cdot \mid s,a) := (1-\alpha_{sa})q^\star(\cdot \mid s,a) + \alpha_{sa}p(\cdot \mid s,a).$$

If $b^\star_{sa} = 0$, then $q^\star(\cdot \mid s,a) = p(\cdot \mid s,a)$ and we set $\alpha_{sa} := 0$, so again $\tilde{q}(\cdot \mid s,a) = q^\star(\cdot \mid s,a)$. By convexity of $q \mapsto \mathrm{KL}_{sa}(p,q)$ in its second argument,

$$\mathrm{KL}_{sa}(p,\tilde{q}) \le (1-\alpha_{sa})\mathrm{KL}_{sa}(p,q^\star) \le (1-\alpha_{sa})b^\star_{sa} = \tilde{b}_{sa},$$

so $\tilde{q} \in Q(\tilde{b})$.

For each block, Pinsker's inequality and the identity $\alpha_{sa}b^\star_{sa} = b^\star_{sa} - \tilde{b}_{sa} \le h$ give

$$\|\tilde{q}(\cdot \mid s,a) - q^\star(\cdot \mid s,a)\|_1 = \alpha_{sa}\|p(\cdot \mid s,a) - q^\star(\cdot \mid s,a)\|_1 \le \alpha_{sa}\sqrt{2\,\mathrm{KL}_{sa}(p,q^\star)} \le \sqrt{2h}.$$

Summing over (s,a) yields

$$\|\tilde{q} - q^\star\|_1 \le |\mathcal{S}||\mathcal{A}|\sqrt{2h}.$$

Now let

$$F(q) := V^\pi_p(\boldsymbol{\rho})V^\pi_q(\boldsymbol{\rho}).$$

By Lemma 3,

$$\frac{\partial F(q)}{\partial q(s' \mid s,a)} = \frac{V^\pi_p(\boldsymbol{\rho})}{1-\gamma} d^\pi_{q,\boldsymbol{\rho}}(s,a)\big(r(s,a) + \gamma V^\pi_q(s')\big).$$

Since $d^\pi_{q,\boldsymbol{\rho}}(s,a) \le 1$ and $|V^\pi_q(s')| \le r_{\max}/(1-\gamma)$, we obtain

$$\left|\frac{\partial F(q)}{\partial q(s' \mid s,a)}\right| \le \frac{|V^\pi_p(\boldsymbol{\rho})| r_{\max}}{(1-\gamma)^2}, \qquad \forall q \in \mathcal{P},\ \forall s,s',a.$$

Therefore,

$$|F(\tilde{q}) - F(q^\star)| \le \frac{|V^\pi_p(\boldsymbol{\rho})| r_{\max}}{(1-\gamma)^2}\|\tilde{q} - q^\star\|_1 \le \frac{\sqrt{2}\,|\mathcal{S}||\mathcal{A}|\,|V^\pi_p(\boldsymbol{\rho})|\,r_{\max}}{(1-\gamma)^2}\sqrt{h}.$$

Because $\tilde{q} \in Q(\tilde{b})$,

$$\varphi(\tilde{b}) - \varphi(b^\star) \le F(\tilde{q}) - F(q^\star) \le |F(\tilde{q}) - F(q^\star)|.$$

This proves the first claim. The second follows by substituting the displayed choice of h_ζ. □

Proof of Theorem 2. If $V^\pi_p(\boldsymbol{\rho}) = 0$, then $u_{\mathrm{NO}}(\sigma,\boldsymbol{\omega},p) = 0$ for every σ, and Algorithm 2 returns 0. Hence we may assume $V^\pi_p(\boldsymbol{\rho}) \ne 0$. Define the proof-only shorthand

$$\varphi(b) := \min_{q \in Q(b)} V^\pi_p(\boldsymbol{\rho})V^\pi_q(\boldsymbol{\rho}).$$

$$\Delta^{(k)}_b := \bar{V}^{\bar{\pi}^{(k)}_b}_{\bar{p}}(\bar{\boldsymbol{\rho}})V^\pi_p(\boldsymbol{\rho}) - \varphi(b).$$

Smoothness of the objective. Recall that

$$f(\bar{\pi}) := \mathrm{sign}(V^\pi_p(\boldsymbol{\rho}))\,\bar{V}^{\bar{\pi}}_{\bar{p}}(\bar{\boldsymbol{\rho}}).$$

For each $\bar{s} = (s,a) \in \bar{\mathcal{S}}$, we have $\bar{V}^{\bar{\pi}}_{\bar{p}}(\bar{s}) = Q^\pi_q(s,a)$ under the identification $q = \bar{\pi}$, and

$$\bar{V}^{\bar{\pi}}_{\bar{p}}(\bar{\boldsymbol{\rho}}) = \sum_{\bar{s} \in \bar{\mathcal{S}}} \bar{\rho}(\bar{s})\bar{V}^{\bar{\pi}}_{\bar{p}}(\bar{s}).$$

Hence Lemma 4 implies that, for any $\bar{\pi}, \tilde{\bar{\pi}} \in \mathcal{P}$,

$$\begin{aligned}\|\nabla_{\bar{\pi}}\bar{V}^{\bar{\pi}}_{\bar{p}}(\bar{\boldsymbol{\rho}}) - \nabla_{\bar{\pi}}\bar{V}^{\tilde{\bar{\pi}}}_{\bar{p}}(\bar{\boldsymbol{\rho}})\|_2 &\le \sum_{\bar{s} \in \bar{\mathcal{S}}} \bar{\rho}(\bar{s})\|\nabla_{\bar{\pi}}\bar{V}^{\bar{\pi}}_{\bar{p}}(\bar{s}) - \nabla_{\bar{\pi}}\bar{V}^{\tilde{\bar{\pi}}}_{\bar{p}}(\bar{s})\|_2 \\ &\le \frac{2\gamma|\bar{\mathcal{A}}|r_{\max}}{(1-\gamma)^3}\|\bar{\pi} - \tilde{\bar{\pi}}\|_2.\end{aligned}$$

Thus $\bar{V}_{\bar{p}}^{\bar{\pi}}(\bar{\boldsymbol{\rho}})$ is L_0-smooth on $\mathcal{P}$ with

$$L_0 := \frac{2\gamma|\bar{\mathcal{A}}|r_{\max}}{(1-\gamma)^3}.$$

Multiplication by $\operatorname{sign}(V_p^{\pi}(\boldsymbol{\rho}))$ does not change the smoothness constant, so f is also L_0-smooth. In particular, any larger constant is admissible in the projected-gradient analysis. For the sequel, we use the slightly looser choice

$$L := \frac{2(\gamma|\bar{\mathcal{A}}|+1)r_{\max}}{(1-\gamma)^3},$$

where the extra $+1$ is an inessential slack term used only to simplify the presentation.

Fixed-budget inner rate. Since $Q(b) \subseteq \mathcal{P}$, the objective f is L-smooth on $Q(b)$. Applying Lemma 11 with this choice of L and then multiplying the resulting inequality by $|V_p^{\pi}(\boldsymbol{\rho})|$, we obtain, for every $b \in \mathcal{B}_{\sigma}(\boldsymbol{\omega})$ and every $k \geq 1$,

$$\Delta_b^{(k)} \leq \max\left\{ \frac{128(\gamma|\bar{\mathcal{A}}|+1)|\bar{\mathcal{S}}|r_{\max}|V_p^{\pi}(\boldsymbol{\rho})|}{(1-\gamma)^5 k} \left\|1/\bar{\boldsymbol{\rho}}\right\|_{\infty}^2, \frac{\Delta_b^{(0)}}{2^{k/2}} \right\}.$$

Since $\bar{\pi}_b^{(0)} = p$ and $|\bar{V}_{\bar{p}}^{\bar{\pi}}(\bar{\boldsymbol{\rho}})| \leq r_{\max}/(1-\gamma)$ for every $\bar{\pi}$, we have

$$\Delta_b^{(0)} \leq \frac{2|V_p^{\pi}(\boldsymbol{\rho})|r_{\max}}{1-\gamma}.$$

Since $2^{-k/2} \leq 2/k$ for all $k \geq 1$ and

$$\frac{128(\gamma|\bar{\mathcal{A}}|+1)|\bar{\mathcal{S}}|r_{\max}|V_p^{\pi}(\boldsymbol{\rho})|}{(1-\gamma)^5} \left\|1/\bar{\boldsymbol{\rho}}\right\|_{\infty}^2 \geq \frac{4|V_p^{\pi}(\boldsymbol{\rho})|r_{\max}}{1-\gamma} \geq 2\Delta_b^{(0)},$$

where we used $|\bar{\mathcal{S}}| \geq 1$, $\|1/\bar{\boldsymbol{\rho}}\|_{\infty} \geq 1$, $\gamma|\bar{\mathcal{A}}|+1 \geq 1$, and $(1-\gamma)^4 \leq 1$, we obtain

$$\Delta_b^{(k)} \leq \frac{128(\gamma|\bar{\mathcal{A}}|+1)|\bar{\mathcal{S}}|r_{\max}|V_p^{\pi}(\boldsymbol{\rho})|}{(1-\gamma)^5 k} \left\|1/\bar{\boldsymbol{\rho}}\right\|_{\infty}^2.$$

Hence the choice of M in Algorithm 2 guarantees that, for every $b \in \mathcal{G}_{h_\zeta}$,

$$u_b \geq \varphi(b) \geq u_b - \zeta/3. \tag{$\star$}$$

Outer discretization. Let $b^{\star} \in \arg\min_{b \in \mathcal{B}_{\sigma}(\boldsymbol{\omega})} \varphi(b)$. Define $\tilde{b} \in \mathcal{G}_{h_\zeta}$ by coordinatewise flooring:

$$\tilde{b}_{sa} := h_\zeta \left\lfloor b^{\star}_{sa}/h_\zeta \right\rfloor, \qquad (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Then $\tilde{b} \in \mathcal{G}_{h_\zeta}$, and Lemma 12 together with the choice of h_ζ in Algorithm 2 gives

$$\varphi(\tilde{b}) \leq \varphi(b^{\star}) + \zeta/3 = u_{\mathrm{NO}}(\sigma, \boldsymbol{\omega}, p) + \zeta/3.$$

Conclusion. Using $(\star)$ at $b = \tilde{b}$,

$$u_\zeta \leq u_{\tilde{b}} \leq \varphi(\tilde{b}) + \zeta/3 \leq u_{\mathrm{NO}}(\sigma, \boldsymbol{\omega}, p) + \frac{2\zeta}{3}.$$

Equivalently,

$$u_{\mathrm{NO}}(\sigma, \boldsymbol{\omega}, p) \geq u_\zeta - \frac{2\zeta}{3} \geq u_\zeta - \zeta.$$

On the other hand, from $(\star)$,

$$u_\zeta = \min_{b \in \mathcal{G}_{h_\zeta}} u_b \geq \min_{b \in \mathcal{G}_{h_\zeta}} \varphi(b) \geq \min_{b \in \mathcal{B}_{\sigma}(\boldsymbol{\omega})} \varphi(b) = u_{\mathrm{NO}}(\sigma, \boldsymbol{\omega}, p),$$

where the last equality is Proposition 3. This proves

$$u_\zeta \geq u_{\mathrm{NO}}(\sigma, \boldsymbol{\omega}, p) \geq u_\zeta - \zeta.$$

□

E.1 Convergence rate of projected gradient descent

Definition 2 (gradient-mapping domination (Xiao, 2022)). *Consider an L-smooth function f on a compact convex set Q. We say that f satisfies the gradient-mapping domination condition if there exists $\omega > 0$ such that*

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

where $f^\star = \min_{x\in Q} f(x)$ and

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

Theorem 4 (Xiao (2022)). *Consider the minimization of an L-smooth function f over a compact convex set Q. Suppose that f satisfies the gradient-mapping domination condition with constant $\omega > 0$. Then the projected-gradient sequence*

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

satisfies, for all $k \geq 1$,

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

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 $s \geq 0$,

$$f(x^{(s)}) - f(x^{(s+1)}) \geq \frac{\omega}{L}(f(x^{(s+1)}) - f^\star)^2. \tag{22}$$

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

$$\delta_k \leq \frac{L}{\omega rck}.$$

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 (22), we further obtain $\delta_{s+1} \leq \delta_s$ for each $s \in \{0, \dots, k-1\}$. 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. □

E.2 Proof of Proposition 3

Proof of Proposition 3. If $q \in \Pi^p_\sigma$, define $b(q) \in \mathbb{R}_+^{|\mathcal{S}||\mathcal{A}|}$ by

$$b(q)_{sa} := \mathrm{KL}_{sa}(p,q), \qquad (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Then $b(q) \in \mathcal{B}_\sigma(\boldsymbol{\omega})$ and $q \in Q(b(q))$. Hence

$$u_{\mathrm{NO}}(\sigma, \boldsymbol{\omega}, p) = \min_{q \in \Pi^p_\sigma} V_p^\pi(\boldsymbol{\rho}) V_q^\pi(\boldsymbol{\rho}) \geq \min_{b \in \mathcal{B}_\sigma(\boldsymbol{\omega})} \min_{q \in Q(b)} V_p^\pi(\boldsymbol{\rho}) V_q^\pi(\boldsymbol{\rho}).$$

Conversely, if $b \in \mathcal{B}_\sigma(\boldsymbol{\omega})$ and $q \in Q(b)$, then

$$\sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p,q) \leq \sum_{s,a} \omega_{sa} b_{sa} \leq \sigma,$$

so $q \in \Pi^p_\sigma$. Therefore,

$$\min_{b \in \mathcal{B}_\sigma(\boldsymbol{\omega})} \min_{q \in Q(b)} V_p^\pi(\boldsymbol{\rho}) V_q^\pi(\boldsymbol{\rho}) \geq u_{\mathrm{NO}}(\sigma, \boldsymbol{\omega}, p).$$

This proves the equality. □

F Upper Bound on the Sample Complexity–Proof of Theorem 3

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

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

$$u_{\mathrm{NO}}(\beta(\tau,\delta)/\tau, \hat{\boldsymbol{\omega}}(\tau), \hat{p}_\tau) \geq u_{\zeta_\tau} - \zeta_\tau \geq 0. \tag{23}$$

As Assumption 1 implies that $\min_q V_q^\pi(\boldsymbol{\rho}) V_{\hat{p}_\tau}^\pi(\boldsymbol{\rho}) < 0$, inequality (23) yields that

$$u_{\mathrm{NO}}(\beta(\tau,\delta)/\tau, \hat{\boldsymbol{\omega}}(\tau), \hat{p}_\tau) > \min_q V_q^\pi(\boldsymbol{\rho}) V_{\hat{p}_\tau}^\pi(\boldsymbol{\rho}). \tag{24}$$

Observing that $\sigma_{\mathrm{NC}}(\cdot, \hat{\boldsymbol{\omega}}(\tau), \hat{p}_\tau)$ is a decreasing function, (23) also yields that

$$\sigma_{\mathrm{NC}}(0, \hat{\boldsymbol{\omega}}(\tau), \hat{p}_\tau) \geq \sigma_{\mathrm{NC}}(u_{\mathrm{NO}}(\beta(\tau,\delta)/\tau, \hat{\boldsymbol{\omega}}(\tau), \hat{p}_\tau), \hat{\boldsymbol{\omega}}(\tau), \hat{p}_\tau) = \beta(\tau,\delta)/\tau, \tag{25}$$

where the last equality follows from Proposition 2 with $\sigma = \beta(\tau,\delta)/\tau \geq 0$ and condition (24). One can notice that (25) is equivalent to (6). Hence, if $\mathrm{Ans}(\hat{p}_\tau) \neq \mathrm{Ans}(p)$ (in other words, if $p \in \mathrm{Alt}(\hat{p}_\tau)$), then

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

By Proposition 1 in Jonsson et al. (2020) (see (7)), we deduce that $\mathbb{P}_p[\mathrm{Ans}(\hat{p}_\tau) \neq \mathrm{Ans}(p)] \leq \delta$.

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

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

and $\varepsilon \in (0, F(\boldsymbol{\omega}, p)/2)$. As shown in Lemma 20 (Appendix H.1), F is a continuous function on $\Sigma \times \mathcal{P}^+_{\mathrm{Test}}$; thus, there exists $\xi_1 \in (0,1)$ such that

$$|F(\boldsymbol{\omega}, p) - F(\boldsymbol{\omega}', p')| < \varepsilon \quad \text{if } \max\{\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_1, \|p' - p\|_1\} \le \xi_1. \tag{26}$$

Moreover, an application of Theorem 5 in Appendix G with $u = 0$ and $p = p$ implies that there exist $\xi_2 \in (0,1)$ and $c > 0$ such that $u_{\mathrm{NO}}(\cdot, \hat{\boldsymbol{\omega}}(t), \hat{p}_t)$ 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'

$$\mathcal{C}_T(\xi) = \bigcap_{\sqrt{T} \le t \le T} \left\{ \max\left\{ \|\hat{\boldsymbol{\omega}}(t) - \boldsymbol{\omega}\|_1, \|\hat{p}_t - p\|_1 \right\} \le \xi \right\}. \tag{27}$$

By Proposition 5 (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 $T \ge T_2(\xi)$,

$$\beta(T,\delta) + \frac{\zeta_T T}{c} \le \log(1/\delta) + (F(\boldsymbol{\omega}, p) - \varepsilon)\xi T. \tag{28}$$

Finally, we define

$$T_3(\xi, \varepsilon, \delta) = \frac{(F(\boldsymbol{\omega}, p) - \varepsilon)^{-1} \log(1/\delta)}{1 - \xi}. \tag{29}$$

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

$$\begin{aligned} \frac{\zeta_T T}{c} + \beta(T,\delta) &\le \log(1/\delta) + (F(\boldsymbol{\omega}, p) - \varepsilon)\xi T \\ &\le (F(\boldsymbol{\omega}, p) - \varepsilon) T \\ &\le F(\hat{\boldsymbol{\omega}}(T), \hat{p}_T) T, \end{aligned} \tag{30}$$

where the first inequality is (28); the second one follows from (29); the last one is a consequence of (27) and $T \ge T_1(\xi)$. Recall that $F(\hat{\boldsymbol{\omega}}(T), \hat{p}_T)$ is exactly $\sigma_{\mathrm{NC}}(0, \hat{\boldsymbol{\omega}}(T), \hat{p}_T)T$. Applying Theorem 5 with $u = 0$, $\hat{\boldsymbol{\omega}} = \hat{\boldsymbol{\omega}}(T)$, $\hat{p} = \hat{p}_T$, $\sigma_1 = \beta(T,\delta)/T$, $\sigma_2 = \sigma_{\mathrm{NC}}(0, \hat{\boldsymbol{\omega}}(T), \hat{p}_T)$, (30) implies that

$$\begin{aligned} \zeta_T &\le u_{\mathrm{NO}}(\beta(T,\delta)/T, \hat{\boldsymbol{\omega}}(T), \hat{p}_T) - u_{\mathrm{NO}}(\sigma_{\mathrm{NC}}(0, \hat{\boldsymbol{\omega}}(T), \hat{p}_T), \hat{\boldsymbol{\omega}}(T), \hat{p}_T) \\ &= u_{\mathrm{NO}}(\beta(T,\delta)/T, \hat{\boldsymbol{\omega}}(T), \hat{p}_T), \end{aligned} \tag{31}$$

where the last equality follows from Proposition 2. By Theorem 2, we have $u_{\zeta_T} \ge u_{\mathrm{NO}}(\beta(T,\delta)/T, \hat{\boldsymbol{\omega}}(T), \hat{p}_T)$, hence (31) implies $u_{\zeta_T} - \zeta_T \ge 0$. Therefore, $\tau \le T$. 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

$$\begin{aligned} \mathbb{E}_p[\tau] &\le \max\{T_1(\xi), T_2(\xi), T_3(\xi,\varepsilon,\delta)\} + \sum_{T = \max\{T_1(\xi), T_2(\xi), T_3(\xi,\varepsilon,\delta)\}+1}^{\infty} \mathbb{P}_p[\tau > T] \\ &\le \max\{T_1(\xi), T_2(\xi), T_3(\xi,\varepsilon,\delta)\} + \sum_{T = \lceil T_1(\xi) \rceil}^{\infty} \mathbb{P}_p[\mathcal{C}_T(\xi)^c]. \end{aligned} \tag{32}$$

From Proposition 5, $\sum_{T=\lceil T_1(\xi)\rceil}^{\infty} \mathbb{P}_p[\mathcal{C}_T(\xi)^c] \le 8\,|\mathcal{S}|^4\,|\mathcal{A}|^3\,/\xi^2 \min_{s,a} w_{sa}$. As a consequence of (32),

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

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

□

Proposition 5. *Under Assumption 2, in Algorithm 1, for any $\xi \in (0,1), \boldsymbol{\omega} \in \Sigma$, there exists $T_1(\xi) > 0$ such that*

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

where $\mathcal{C}_T(\xi)$ is introduced in (27).

Proof. Due Assumption 2, $\mathcal{W} := \{(s,a) : \omega_{sa} > 0\} = \mathcal{S} \times \mathcal{A}$. By Lemma 14, $\left\| \hat{\boldsymbol{\omega}}(t) - \boldsymbol{\omega} \right\|_1 \leq \sum_{s,a} |\mathcal{S}| \left| \mathcal{A} \right| / t \leq \left| \mathcal{S} \right|^2 \left| \mathcal{A} \right|^2 / t$. We then derive that when $T \geq \left| \mathcal{S} \right|^4 \left| \mathcal{A} \right|^4 / \xi^2$ and $t \geq \sqrt{T}$, one can deduce that $\left\| \hat{\boldsymbol{\omega}}(t) - \boldsymbol{\omega} \right\|_1 \leq \xi$. As for the estimate on p, we apply Lemma 14 again to have that for each (s,a),

$$N_{sa}(t) \geq t \min_{s,a} \omega_{sa} - \left| \mathcal{S} \right| \left| \mathcal{A} \right| \geq t \min_{s,a} \omega_{sa} / 2 \tag{33}$$

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

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

where the last inequality follows from Lemma 13 and (33). 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

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

The proof is completed by applying Lemma 15 with $A = \frac{t \xi^2 \min_{s,a} w_{sa}}{2 |\mathcal{S}|^3 |\mathcal{A}|^2}$, $\alpha = 1/2$, $\beta = 1$. □

Lemma 13 (Proposition 1 in Weissman et al. (2003))**.** *Suppose one has samples the state-action pair (s,a) for $n \geq 1$ times, then the empirical estimate on $p(\cdot \mid s,a)$, $\hat{p}_n(\cdot \mid s,a)$ satisfies that*

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

Lemma 14. *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 $t \geq |\mathcal{W}|$, one has

$$t \omega_{sa} - |\mathcal{W}| \leq N_{sa}(t) \leq t \omega_{sa} + 1. \tag{34}$$

Proof. If $(s,a) \notin \mathcal{W}$, $N_{sa}(t) = 0$ for all $t \in \mathbb{N}$, (34) holds directly.
For a fixed $(s,a) \in \mathcal{W}$, we prove the upper bound in (34) by induction. When $t = |\mathcal{W}|$, $N_{sa}(t) = 1 \leq t \omega_{sa} + 1$. Now suppose $N_{sa}(t-1) \leq (t-1)\omega_{sa} + 1$, and consider two following cases, (i) $A_t \neq (s,a)$; (ii) $A_t = (s,a)$.
When (i) $A_t \neq (s,a)$, using the inductive hypothesis yields that

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

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

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

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

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

Thus the upper bound in (34) is obtained by multiplying $t\omega_{sa}$ on the the both sides of the above inequality.

We now prove the lower bound in (34). 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 (34).

□

Lemma 15 (Lemma 5 in Wang et al. (2021))**.** *Let $\alpha, \beta \in (0,1)$ and $A > 0$.*

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

G 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 $\bar{\sigma} = \max_{\|\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}\|_1 \leq \xi} \max_{\|\hat{p} - p\|_1 \leq \xi} \{\sigma_{\mathrm{NC}}(u, \hat{\boldsymbol{\omega}}, \hat{p})\}$, then*

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

Proof. One can assume $\hat{p}$ is full-supported. Otherwise, due to the continuity of $u_{\mathrm{NO}}(\sigma, \hat{\boldsymbol{\omega}}, \cdot)$ with respect to its third argument (the kernel), as shown in Lemma 21 (Appendix H.2), for an arbitrary $\varepsilon > 0$, one can always find a full-supported kernel $\tilde{p}$ sufficiently close to $\hat{p}$ such that

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

Because $u_{\mathrm{NO}}(\sigma, \hat{\boldsymbol{\omega}}, \hat{p})$ is a decreasing function of σ, an application of Monotone difference lemma (16) implies that $u_{\mathrm{NO}}(\sigma, \hat{\boldsymbol{\omega}}, \hat{p})$ as a function of σ is differentiable almost everywhere. Let $\sigma \in [\sigma_1, \sigma_2]$ be a point at which $u_{\mathrm{NO}}(\sigma, \hat{\boldsymbol{\omega}}, \hat{p})$ 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 $\sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(\hat{p}, q) - \sigma \leq 0$, $-q(s' \mid s, a) \leq 0$, $\sum_{s' \in \mathcal{S}} q(s' \mid s, a) - 1 = 0$ respectively. An application of Envelope Theorem (Theorem 7) yields that

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

By Lemma 17, we know

$$\eta_\sigma = \frac{\gamma V_{\hat{p}}^\pi(\boldsymbol{\rho}) d_{q_\sigma, \boldsymbol{\rho}}(s, a) \left(V_{q_\sigma}^\pi(s_\sigma^M) - V_{q_\sigma}^\pi(s_\sigma^m) \right)}{\hat{\omega}_{sa}(1-\gamma) \left(\frac{\hat{p}(s_\sigma^M \mid s,a)}{q_\sigma(s_\sigma^M \mid s,a)} - \frac{\hat{p}(s_\sigma^m \mid s,a)}{q_\sigma(s_\sigma^m \mid s,a)} \right)}, \quad \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{35}$$

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

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

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

$$\min_{q\in\mathcal{P}} \max_{s,s'} V_{\hat{p}}^\pi(\boldsymbol{\rho}) V_q^\pi(s) - V_{\hat{p}}^\pi(\boldsymbol{\rho}) V_q^\pi(s') \geq c_1, \quad \forall \left\|\hat{p} - p\right\|_1 < \xi. \tag{36}$$

Further observe that

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

which is upper bounded by some $c_2 > 0$ for any $\left\|\hat{p} - p\right\|_1 \leq \xi$. 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. □

G.1 Technical Lemmas

Lemma 16 (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*

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

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*

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

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 $x^*(p) \in \mathrm{Int}(X)$ 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

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

G.2 The Value of the Lagrangian Multiplier

Lemma 17. *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 $\sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p, q) \le \sigma$ is*

$$\eta_\sigma = \frac{\gamma V_p^\pi(\boldsymbol{\rho}) d_{q_\sigma,\boldsymbol{\rho}}(s,a) \left(V_{q_\sigma}^\pi(s_\sigma^M) - V_{q_\sigma}^\pi(s_\sigma^m)\right)}{\omega_{sa}(1-\gamma)\left(\frac{p(s_\sigma^M|s,a)}{q_\sigma(s_\sigma^M|s,a)} - \frac{p(s_\sigma^m|s,a)}{q_\sigma(s_\sigma^m|s,a)}\right)} \quad \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{37}$$

where $s_\sigma^M \in \arg\max_{s\in\mathcal{S}} V_p^\pi(\boldsymbol{\rho}) V_{q_\sigma}^\pi(s)$ and $s_\sigma^m \in \arg\min_{s\in\mathcal{S}} V_p^\pi(\boldsymbol{\rho}) V_{q_\sigma}^\pi(s)$.

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

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

where $\eta \ge 0$, $\boldsymbol{\lambda} \in \mathbb{R}_+^{|\mathcal{S}|^2|\mathcal{A}|}$ 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 $\sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p, q_\sigma) \le \sigma$). That is, $q_\sigma(s'|s,a) > 0, \forall s, s', a$. Further using Corollary 1 in Appendix G.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

$$\begin{aligned} &\frac{\partial}{\partial q(s'|s,a)} \sum_{s,a} \omega_{sa}\mathrm{KL}_{sa}(p,q_\sigma) - \sigma = -\frac{\omega_{sa} p(s'|s,a)}{q_\sigma(s'|s,a)} \quad \forall s', s, a, \\ &\text{and } \frac{\partial}{\partial q(s'|s,a)} \sum_{s'\in\mathcal{S}} q_\sigma(s'|s,a) - 1 = 1, \quad \forall s', s, a. \end{aligned}$$

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 $\eta_\sigma \ge 0, \boldsymbol{\lambda} \in \mathbb{R}_+^{|\mathcal{S}|^2|\mathcal{A}|}$ and $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ such that

$$\begin{aligned} &\frac{\partial}{\partial q(s'|s,a)} V_p^\pi(\boldsymbol{\rho}) V_{q_\sigma}^\pi(\boldsymbol{\rho}) - \frac{\eta\omega_{sa} p(s'|s,a)}{q_\sigma(s'|s,a)} + \mu_{sa} - \lambda_{s'sa} \quad = 0, \quad \forall s', s, a, && \text{(Stationarity)} \\ &\text{and } \eta\left(\sum_{s,a} \omega_{sa}\mathrm{KL}_{sa}(p,q_\sigma) - \sigma\right) = 0,\ \lambda_{s'sa}(-q_\sigma(s'|s,a)) \quad = 0, \quad \forall s', s, a, && \text{(Complementary slackness)} \end{aligned}$$

As q_σ is full-supported, we derive $\lambda_{s'sa} = 0, \forall s', s, a$, from Complementary slackness. From (12) in Lemma 3, Stationarity can be rewritten as:

$$\frac{V_p^\pi(\boldsymbol{\rho})}{1-\gamma} d_{q_\sigma,\boldsymbol{\rho}}(s,a)\left(r(s,a) + \gamma V_{q_\sigma}^\pi(s')\right) - \frac{\eta\omega_{sa} p(s' \mid s,a)}{q_\sigma(s' \mid s,a)} = -\mu_{sa}, \forall s', s, a. \tag{38}$$

By taking difference of the equations (38) with $s' = s_\sigma^M$ and $s' = s_\sigma^m$, we obtain

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

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

G.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 18.

Lemma 18. *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 $\mathcal{P}_\sigma$*[2].

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 19, we derive that $\forall (s', a') \in \mathcal{S} \times \mathcal{A}$, $\exists \alpha_{s'a'} \in \mathbb{R}$ such that for each $s'' \in \mathcal{S}$,

$$\alpha_{s'a'} = \frac{\partial V_{q_o}^\pi(\boldsymbol{\rho}) V_p^\pi(\boldsymbol{\rho})}{\partial q(s''|s', a')} = \frac{V_p^\pi(\boldsymbol{\rho})}{1-\gamma} \sum_{s,a} \rho(s) \pi(a|s) d_{q_o,s,a}^\pi(s', a') (r(s, a) + \gamma V_{q_o}^\pi(s'')), \tag{39}$$

where the last equation stems directly from Lemma 3. Let $\alpha := \sum_{s',a'} \alpha_{s'a'} / V_p^\pi(\boldsymbol{\rho})$ and sum (39) 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*

$$\sum_{sa} \omega_{sa} \mathrm{KL}_{sa}(p, q_\sigma) = \sigma \tag{40}$$

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 18, q_σ is on the boundary, we deduce (40). □

Lemma 19. *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 □

H 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:

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

[2]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).

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.

H.1 Proof of Lemma 20

Here we divide $\mathcal{P}_{\text{Test}}$ into two disjoint sets, $\mathcal{P}_{\text{Test}} = \mathcal{P}^+_{\text{Test}} \cup \mathcal{P}^-_{\text{Test}}$, where $\mathcal{P}^+_{\text{Test}} := \{p \in \mathcal{P}_{\text{Test}} : \text{Ans}(p) = +\}$, $\mathcal{P}^-_{\text{Test}} := \{p \in \mathcal{P}_{\text{Test}} : \text{Ans}(p) = -\}$.

Lemma 20. *A function $F : \Sigma \times \mathcal{P}^+_{\text{Test}} \to \mathbb{R}$ defined as $F(\boldsymbol{\omega}, q) := \inf_{q \in \text{Alt}(p)} \sum_{s,a} \omega_{sa} \text{KL}_{sa}(p, q)$ 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. □

H.2 Proof of Lemma 21

Lemma 21. *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:

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

As the objective function, ψ, is a continuous function on $\mathbb{R}_+ \times \mathcal{P}$, it suffices to show the corresponding $\Phi(\sigma, p) = \{q \in \mathcal{P} : \sum_{s,a} \omega_{sa} \text{KL}_{sa}(p, q) \le \sigma\}$ 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 $\sum_{s,a} \omega_{sa} \text{KL}_{sa}(p_n, q_n) - \sigma_n \le 0$, and $q_n \xrightarrow{n\to\infty} q$. By continuity of KL-divergence, one has

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

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 $n \ge N$. 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 $\sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p_n, q) = 0$, $\alpha_n = 1$. When $\sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p_n, q) \neq 0$ Due to the joint convexity, one has

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

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

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

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 $n \to \infty$.

□

H.3 Proof of Lemma 22

Lemma 22. *Under Assumption 4 and the notion in Appendix D, $u_{\mathrm{NO}}(\cdot)$ 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 4-(c), $\Phi(\sigma)$ is always a compact set. Additionally, $h(\underline{x}) \leq 0$ from Assumption 4-(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 4-(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 4-(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.

□

I PROOF OF THE REMAINING LEMMA AND PROPOSITION

I.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 $\max_s r^\pi(s) > \min_s r^\pi(s)$. □

I.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 $\sum_{sa} \omega_{sa} \mathrm{KL}_{sa}(p, q^\star) = 0$ 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, $\omega_{sa} > 0$ by Assumption 2, 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 $p \in \mathcal{P}_{\mathrm{Test}}$. □

J 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 $\zeta_t = \frac{5}{t^{3/2}}$. In the experiments, the inner optimization is implemented heuristically by taking gradient steps in the reversed MDP and numerically projecting each iterate onto $\bar{\Pi}_\sigma^p$ using SLSQP. We fixed $L = 400.0$ and capped the maximum value of M at 20.

As $|\mathcal{S}|$ and $|\mathcal{A}|$ increased, PTST tended to statistically perform better than the baseline. However, the computational cost grew with problem size; in particular, projection onto $\bar{\Pi}_\sigma^p$ dominated the runtime. An important future direction is to develop projection-free variants of our method (e.g., conditional gradient updates in the reversed MDP) that avoid Euclidean projections onto a convex set and further reduce computational overhead.

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 ($\mathcal{S} = \mathcal{A} = \{0,1\}$).

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 ($\mathcal{S} = \mathcal{A} = \{0,1,2\}$).

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 ($\mathcal{S} = \mathcal{A} = \{0,1,2,3,4\}$).

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