Resilient Constrained Reinforcement Learning

Dongsheng Ding **Zhengyan Huan** **Alejandro Ribeiro**
{dongshed, zhhuan, aribeiro}@seas.upenn.edu
University of Pennsylvania

Abstract

We study a class of constrained reinforcement learning (RL) problems in which multiple constraint specifications are not identified before training. It is challenging to identify appropriate constraint specifications due to the undefined trade-off between the reward maximization objective and the constraint satisfaction, which is ubiquitous in constrained decision-making. To tackle this issue, we propose a new constrained RL approach that searches for policy and constraint specifications together. This method features the adaptation of relaxing the constraint according to a relaxation cost introduced in the learning objective. Since this feature mimics how ecological systems adapt to disruptions by altering operation, our approach is termed as resilient constrained RL. Specifically, we provide a set of sufficient conditions that balance the constraint satisfaction and the reward maximization in notion of resilient equilibrium, propose a tractable formulation of resilient constrained policy optimization that takes this equilibrium as an optimal solution, and advocate two resilient constrained policy search algorithms with non-asymptotic convergence guarantees on the optimality gap and constraint satisfaction. Furthermore, we demonstrate the merits and the effectiveness of our approach in computational experiments.

1 INTRODUCTION

Constrained reinforcement learning (RL) is a constrained control problem in which an agent aims to maximize its expected cumulative reward while satisfying a given constraint by interacting with an environment over time. Multiple requirements are of growing interest in constrained RL towards designing an agent to meet more than one constraint, e.g., resource allocation for many users [De Nijs et al., 2021] and safe learning in robotics [Brunke et al., 2022]. Real-world constrained RL often engages practical problems with unwell-specified requirements, e.g., human-satisfaction in human-robot interaction [El-Shamouty et al., 2020] and safety level of robotic agents [Zhang et al., 2020]. Hence, it is challenging to determine reasonable constraint specifications for making trade-off between reward maximization and constraint satisfaction.

Although reward shaping has been widely used to aggregate multiple requirements into a single reward, e.g., [Pérez-D'Arpino et al., 2021], it doesn't guarantee constraint satisfaction for each requirement. A known reason for this single-reward failure is that the solutions generated by standard RL algorithms do not necessarily satisfy required constraints, which is known as "scalarization fallacy" [Szepesvári, 2020, Zahavy et al., 2021, Calvo-Fullana et al., 2023]. Therefore, it is crucial to directly impose the constraints that result from multiple requirements [Roy et al., 2022], which has been studied by a lot of recent efforts, e.g., [Chow et al., 2017, Paternain et al., 2019, Ding et al., 2020]. However, such results are based on known feasible constraints, not applicable in the situations with unknown constraint specifications.

To fill this gap, we aim to automate the constraint specifications during constrained RL training by facilitating the trade-off between reward maximization and constraint satisfaction. The focal RL environment of this paper is the constrained Markov decision process (MDP) that constrains expected cumulative utilities [Altman, 1999], which has been widely-used in many constraint-rich domains, e.g., resource allocation, robotic planning, and financial management; see more in [Garcıa and Fernández, 2015, De Nijs et al., 2021, Gu et al., 2022, Brunke et al., 2022].

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

Contribution. Our contributions are threefold.

- We first introduce nominal constraints that are possibly infeasible, so that they can be relaxed to compromise reward maximization for constraint satisfaction. We provide the sensitivity analysis of the optimal reward value function to the perturbations in constraints. Since the compromise mimics how ecological systems *adapt to disruptions* by changing operating conditions, we term this as *resilient* constrained policy optimization, and broadly as *resilient* constrained RL.

- To specify the levels or thresholds of constraints, we introduce a user-defined cost function that establishes a price for relaxing nominal constraints, and exploit the relative difficulty of relaxing different constraints to define a trade-off solution: resilient equilibrium. We provide a tractable formulation of resilient constrained policy optimization that takes this equilibrium as an optimal solution, and establish its duality theory under less restrictive feasibility assumption.

- To find an optimal pair of policy and constraint specification, we extend two non-resilient policy gradient algorithms for our resilient constrained policy optimization problem, and prove that they converge to a optimal solution with non-asymptotic convergence guarantees on the optimality gap and constraint satisfaction. To the best of our knowledge, for the first time we establish provably resilient constrained policy search algorithms against uncertain constraints. Moreover, we provide computational experiments to show the merits and the effectiveness of our approach.

Related Work. Our problem formulation is based on the constrained MDP framework [Altman, 1999]. Constrained MDPs with well-specified constraints are relatively well-studied in the literature, e.g., model-based algorithms [Ding et al., 2021, Efroni et al., 2020] and policy gradient methods [Ding et al., 2022], under the strict feasibility assumption on constraints; see more related works in this line in [Gu et al., 2022]. However, it is intractable to determine the feasibility of constraints in many scenarios, e.g., budget distribution for many users [Boutilier and Lu, 2016, Vora et al., 2023] and online budget level [Diaz et al., 2023] are unknown for feasibility-checking, and safety constraints in training are different from those for real robotics, which are expensive to model [Kaspar et al., 2020]. Although this issue can be alleviated by some heuristic methods in the references aforementioned, their optimality and constraint satisfaction are not established. We note that the reward and constraint trade-off essentially reduces to the sensitivity of the optimal reward value function, which was studied using the parameter perturbations of a constrained MDP [Altman and Shwartz, 1991, Altman and Gaitsgory, 1993]. Compared with this line of works, in this paper we exploit the sensitivity analysis of the optimal reward value function against the perturbations in constraints to strike a balance between reward maximization and constraint satisfaction. We further develop two constrained policy search algorithms for finding optimal policy and constraint specification simultaneously, with theoretical guarantees.

Our work is also pertinent to recent efforts of augmenting a RL agent with the adaptation to the interference that is potentially catastrophic to system [Yang et al., 2021, Huang et al., 2022]. This capability is often termed as "resilience" that draws the ability of ecological systems to adapt to disrupted environment [Holling, 1973, Holling, 1996]. Resilience to perturbations in agent-environment interaction has been studied in several prior works [Yang et al., 2021, Phan et al., 2021, Gao et al., 2022]; yet the resilience to corrupted constraints on system or performance hasn't been studied. Recently, the adaptation of trained policy to unknown constraint specifications is investigated in control [Chamon et al., 2020], constrained learning [Hounie et al., 2024], and constrained offline RL [Liu et al., 2023, Zhang et al., 2023]. In contrast, we investigate "resilience" for constrained policy optimization and provably convergent constrained policy search algorithms, with a focus on the adaptation of trained policy to unknown constraint specifications.

2 Preliminaries

We consider an infinite-horizon constrained MDP,

$$\text{CMDP}\,(\,S,\,A,\,P,\,r,\,\{u_i\}_{i=1}^m,\,\{b_i\}_{i=1}^m,\,\gamma,\,\rho\,)$$

where S and A are state/action spaces, P is a transition kernel that specifies the probability $P(s'\,|\,s,a)$ from state s to next state s' under action $a\in A$, r, u_i: $S\times A\to[0,1]$ are reward/utility functions, b_i is a constraint threshold for the ith utility, $\gamma\in[0,1)$ is a discount factor, and ρ is an initial distribution. A stochastic policy π: $S\to\Delta(A)$ determines a probability distribution $\Delta(A)$ over the action space A based on the current state, i.e., $a_t\sim\pi(\cdot\,|\,s_t)$ at time t. Let Π be the set of all possible stochastic policies. A policy $\pi\in\Pi$, together with the initial state distribution ρ, induces a distribution over trajectories $\tau=\{(s_t,a_t,r_t,\{u_{i,t}\}_{i=1}^m)\}_{t=0}^\infty$, where $s_0\sim\rho$, $a_t\sim\pi(\cdot\,|\,s_t)$, $r_t=r(s_t,a_t)$, $u_{i,t}=u_i(s_t,a_t)$, and $s_{t+1}\sim P(\cdot\,|\,s_t,a_t)$ for all $t\geq 0$.

Given a policy π, the value functions V_r^π, $V_{u_i}^\pi$: $S\to\mathbb{R}$ associated with the reward r or the utility u_i are given by the expected sums of discounted rewards or utilities received under the policy π, respectively,

$$V_r^\pi(s) \;\; := \;\; \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\,|\,\pi, s_0 = s\right]$$

where $\mathbb{E}$ is expected over the randomness in the trajectory τ induced by π; similarly, we define $V_{u_i}^\pi(s)$ for the utility u_i. The expected values over the initial distribution ρ are given by $V_r^\pi(\rho) = \mathbb{E}_{s\sim\rho}[V_r^\pi(s)]$ and $V_{u_i}^\pi(\rho) = \mathbb{E}_{s\sim\rho}[V_{u_i}^\pi(s)]$. It is useful to introduce the discounted state visitation distribution, $d_{s_0}^\pi(s) = (1-\gamma)\sum_{t=0}^\infty \gamma^t \mathrm{Pr}(s_t = s\,|\,\pi, s_0)$ which adds up discounted probabilities of visiting s in the execution of π starting from s_0. Denote $d_\rho^\pi(s) := \mathbb{E}_{s_0\sim\rho}[d_{s_0}^\pi(s)]$ and thus $d_\rho^\pi(s) \geq (1-\gamma)\rho(s)$ for any ρ and s. Furthermore, for the reward r, we introduce the state-action value function $Q_r^\pi\colon S\times A\to\mathbb{R}$ when the agent begins with a state-action pair (s,a) and follows a policy π, together with its advantage function $A_r^\pi\colon S\times A\to\mathbb{R}$,

$$\begin{aligned} Q_r^\pi(s,a) \;\; &:= \;\; \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\,|\,\pi, s_0 = s, a_0 = a\right] \\ A_r^\pi(s,a) \;\; &:= \;\; Q_r^\pi(s,a) - V_r^\pi(s). \end{aligned}$$

Similarly, we define $Q_{u_i}^\pi$, $A_{u_i}^\pi\colon S\times A\to\mathbb{R}$ for u_i.

The constrained MDP aims to find a policy that maximizes the reward value function $V_r^\pi(\rho)$ while the utility value function $V_{u_i}^\pi(\rho)$ is above some threshold b_i,

$$\begin{array}{ll} \underset{\pi\,\in\,\Pi}{\text{maximize}} & V_r^\pi(\rho) \\ \text{subject to} & V_{u_i}^\pi(\rho) \;\geq\; b_i, \; i = 1,\dots,m \end{array} \tag{1}$$

where b_i is the specified threshold *priori* for the ith utility value function. Since $V_r^\pi(\rho)$, $V_{u_i}^\pi(\rho) \in [0, 1/(1-\gamma)]$, we assume $b_i \in (\,0, 1/(1-\gamma)\,]$. Thus, by taking g_i: $S\times A \to [-1,1]$ with $g_i = u_i - (1-\gamma)b_i$, equivalently we translate the constraint $V_{u_i}^\pi(\rho)\geq b_i$ into $V_{g_i}^\pi(\rho)\geq 0$, which is our focal constraint. Those utility constraints often result from additional requirements on the system operation, e.g., budget or safety constraints [Boutilier and Lu, 2016, Paternain et al., 2022]. We denote the optimal value for Problem (1) by $V^\star$ which takes $V^\star = V_r^\star(\rho)$ at an optimal policy $\pi^\star$ if it is feasible; $V^\star = -\infty$ otherwise.

Although an optimal policy always exists in the unconstrained case, i.e., $m = 0$ in Problem (1), it is not necessarily true in the constrained case because of potentially infeasible constraints. Thus, it is important to specify relevant constraints for Problem (1). In many scenarios, how to specify the constraint thresholds $\{b_i\}_{i=1,\dots,m}$ is not known *priori*. For instance, the threshold b_i means the ith user's negative budget that is often time-varying in resource allocation [Boutilier and Lu, 2016, Vora et al., 2023]; to trade-off many rewards in preference-based RL [Eysenbach et al., 2019, Liang et al., 2022], the threshold b_i is often unknown for preference i; see their details and more examples in Section 2.1. In practice, only a nominal constraint specification is given with *unknown feasibility*, and we have to relax (or tighten) the nominal constraints for guarding the feasibility. With a slight abuse of notation, we use notation $\{b_i\}_{i=1}^m$ to denote the nominal constraint specifications that might be too conservative (or loose).

To study the effect of constraint specifications, we form a variant of constrained MDP with flexible constraints,

$$\begin{array}{lll} V^\star(\xi) \;:=\; & \underset{\pi\,\in\,\Pi}{\text{maximize}} & V_r^\pi(\rho) \\ & \text{subject to} & V_{g_i}^\pi(\rho) \;\geq\; \xi_i, \; i = 1,\dots,m \end{array} \tag{2}$$

where $\xi\in\mathbb{R}^m$ is the *unknown* perturbation that relaxes the constraint when $\xi_i < 0$ (or tightens the ith inequality constraint when $\xi_i > 0$), and $V^\star(\xi)$ is the primal value function: $V^\star(\xi) = V_r^\star(\rho)$ at an optimal policy $\pi^\star(\xi)$ if it is feasible; $V^\star(\xi) = -\infty$ otherwise. Since $V^\star(0) = V^\star$ for $\xi = 0$, Problem (1) is our nominal problem, and Problem (2) is our perturbed problem.

Since $|V_{g_i}^\pi(\rho)| \leq 1/(1-\gamma)$, it is natural to restrict $|\xi_i|\leq 1/(1-\gamma)$ since Problem (2) is infeasible when $\xi_i > 1/(1-\gamma)$ for all $i = 1,\dots,m$, and it is unconstrained when $\xi_i < -1/(1-\gamma)$ for all $i=1,\dots,m$. Denote $\mathbb{R}_\gamma^m := \{\xi\in\mathbb{R}^m\,|\,|\xi_i|\leq 1/(1-\gamma)\}$. Let the domain of $V^\star(\xi)$ be Ξ, which is the set of all ξ for which the constraint set $\{\pi\in\Pi\,|\,V_{g_i}^\pi(\rho)\geq\xi_i, i=1,\dots m\}$ is non-empty, or equivalently, $\Xi := \{\xi\in\mathbb{R}_\gamma^m\,|\,V^\star(\xi) > -\infty\}$.

From the non-convexity of value functions in policy [Agarwal et al., 2021], Problem (2) is non-convex. Nevertheless, we prove that the primal function $V^\star(\xi)$ has several nice properties inherited from the duality analysis. Lemma 1 shows that the primal function $V^\star(\xi)$ enjoys monotonicity and it is concave over the domain Ξ; see Appendix A.1 for proof.

Lemma 1 (Coordinate-Wise Monotonicity and Concavity)**.** *For Problem* (2)*,*

(i) *$V^\star(\xi)$ is monotonically non-increasing with respect to the coordinates of $\xi\in\Xi$, i.e., $V^\star(\xi)\leq V^\star(\xi')$ when $\xi_j > \xi_j'$ for some j and $\xi_i = \xi_i'$ for $i\neq j$;*

(ii) *$V^\star(\xi)$ is a concave function over $\xi\in\Xi$.*

Lemma 1 shows that relaxing the constraints more (or decreasing ξ) may yield a larger optimal value for Problem (2); however this relaxed problem becomes more far away from the nominal problem (1). Similarly, tightening the constraints can decrease the optimal value, even nullifying the constraints.

To stay close to the nominal problem (1) while specifying constraints efficiently, we will exploit the properties of $V^\star(\xi)$ together with a relaxation cost to introduce a trade-off solution for Problem (2) in Section 3. We feature this solution with *resilient* due to its adaptation of primal value function to the changing constraint specifications, and we term the problem of finding optimal policy and constraint specification together as *resilient constrained constrained policy optimization.*

2.1 Examples with Unspecified Constraints

We showcase that constraint specifications are often not *a priori* knowledge for Problem (1).

Resource allocation. In a system of m users sharing a transition kernel [Boutilier and Lu, 2016, Vora et al., 2023], each user i has a reward function r_i and a cost function c_i (or consumption), and the budget $B > 0$ is given. The resource allocation is to decide a budget assignment $\{\bar{c}_i\}_{i=1}^m$ for m users by maximizing the average reward $r = \frac{1}{m}\sum_{i=1}^m r_i$ and restraining the total cost $\sum_{i=1}^m \bar{c}_i \leq B$. Formulation (1) applies when we take $u_i = -c_i$ and $b_i = -\bar{c}_i$, and evaluate their discounted value functions. One application scenario is the robot monitoring problem experimented in Section 5. However, the budget assignment $\{b_i\}_{i=1}^m$ is a decision variable to be determined. Moreover, the budget can be uncertain, e.g., multi-arm bandits with limited resources [Diaz et al., 2023].

Many rewards trade-off. Multiple rewards appear in real RL applications [Shelton, 2000, Liu et al., 2014, Eysenbach et al., 2019, Liang et al., 2022]. For instance, in preference-based RL [Liang et al., 2022], extrinsic rewards $\{r_i\}_{i=1}^m$ are preferences of human feedback while an intrinsic reward r captures the uncertainty in the disagreement among humans. Take $u_i = r_i$ in Problem (1). To encourage exploration under alignment with preferences, Problem (1) aims to maximize the intrinsic reward value function while constraining extrinsic reward value functions above some thresholds $\{b_i\}_{i=1}^m$. However, such thresholds are unknown due to varying human's preferences.

3 RESILIENT CONSTRAINED RL

To principally specify appropriate constraints, we introduce a cost function of relaxing the constraints, and a resilient equilibrium that balances the relaxation and the primal value function in Section 3.1. We formulate a tractable policy optimization based on regularization to find a resilient equilibrium in Section 3.2.

3.1 Resilient Equilibrium

We characterize the change of the primal value function $V^\star(\xi)$ to relaxation ξ via the subgradient and geometric multiplier in nonlinear programming [Bertsekas, 2016]. Let the dual function for Problem (1) be $D(\lambda) := \sup_{\pi\,\in\,\Pi} V^\pi_{r+\lambda^\top g}(\rho)$ and its domain be $\Lambda := \{\lambda \in \mathbb{R}^m_+ \,|\, D(\lambda) > -\infty\}$. A relation between the primal value function and the dual function is, for any $\lambda \geq 0$,

$$D(\lambda) \;=\; \sup_{\xi\,\in\,\Xi}\, \left\{\lambda^\top \xi - (-V^\star(\xi))\right\} \tag{3}$$

which can be derived in Appendix A.2. In other words, $D(\lambda)$ is the conjugate convex function of $-V^\star(\xi)$ for $\xi \in \Xi$ and the domain Λ is a convex set. By the concavity of the primal function $V^\star(\xi)$ in Lemma 1 and Ξ, there exists a subgradient for $V^\star(\xi)$ at any interior point $\xi \in \Xi$. This subgradient naturally connects to the geometric multiplier λ,

$$V^\star(\xi) \;=\; \sup_{\pi\,\in\,\Pi}\, \left\{V^\pi_{r+\lambda^\top g}(\rho) - \lambda^\top \xi\right\}.$$

Thus, we can interpret the subgradient of the negative primal function as a geometric multiplier for Problem (2); see Appendix A.3 for proof.

Lemma 2 (Subgradient and Geometric Multiplier)**.** *In Problem* (2) *with any* $\xi \in \Xi$*, these are equivalent:*

(i) *λ is a subgradient of $-V^\star(\xi)$ at ξ;*

(ii) *λ is a geometric multiplier for Problem* (2)*.*

Having described the sensitivity of the primal function, we next introduce a notion of resilient equilibrium in supervised learning [Hounie et al., 2024] to determine the constraint specifications according to the difficulty in solving the nominal problem (1). The non-increasing property in Lemma 1 is that the primal function $V^\star(\xi)$ would be increased by decreasing ξ coordinate-wise (relaxing the constraints). However, more relaxation ξ yields looser constraints. We introduce a convex cost function $h(\xi)$: $\Xi \to \mathbb{R}$ to measure the relaxation cost. Without loss of generality, we use the case: $\xi = 0$ to match our nominal problem and set the cost to be zero, i.e., $h(0) = 0$. To govern the change in $V^\star(\xi)$, we use the marginal price of relaxing constraint: $\nabla h(\xi)$, to define a resilient equilibrium $\xi^\star$.

Definition 1 (Resilient Equilibrium)**.** *For any cost function h that is continuously differentiable, concave, and non-increasing coordinate-wise, a resilient equilibrium for $V^\star(\xi)$ is a relaxation $\xi^\star \in \Xi$,*

$$\nabla h(\xi^\star) \;\in\; \partial V^\star(\xi^\star).$$

At a resilient equilibrium $\xi^\star$, for some $\epsilon > 0$, relaxing it to $\xi^\star - \epsilon$ would increase the cost by $-\epsilon\nabla h(\xi^\star)$ and the primal function may get larger; similarly, tightening it to $\xi^\star + \epsilon$ would decrease the cost by $\epsilon\nabla h(\xi^\star)$ and the primal function may become smaller. The "resilient" captures how much we can relax (or tighten) the constraints before we observe significant improvement (or degradation) in the primal value function. Thus, a resilient constrained policy optimization problem reduces to solving Problem (2) for some $\xi^\star$ that is a resilient equilibrium for an user-defined cost function h.

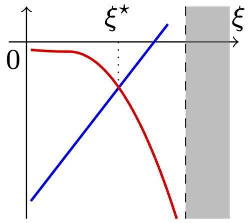


Figure 1: Resilient equilibrium for Problem (2) with $m = 1$ and a quadratic function $h(\xi)$ for $\xi \in \mathbb{R}$. The horizontal axis is the relaxation ξ, and the vertical axis is the (sub)gradient values: $\nabla h(\xi)$ (—) and $\partial V(\xi)$ (—). The shaded area means the infeasibility when ξ is large.

We note that $V^\star(\xi) - h(\xi)$ is a concave function. The existence of resilient equilibrium can be easily obtained in Lemma 3, and we delay its proof to Appendix B.1.

Lemma 3 (Equilibrium Existence)**.** *There exists a resilient equilibrium $\xi^\star \in \Xi$, and it is unique when h is strictly convex.*

Behind the existence, Lemma 4 shows that relaxing the constraint (or decrease ξ) increases the cost sensitivity $|\nabla h(\xi)|$ while the effect on the primal value function is decreased; see Appendix B.2 for proof. Thus, they must cross at a resilient equilibrium as shown in Figure 1.

Lemma 4 (Coordinate-Wise Monotonicity)**.** *Let ξ, $\xi' \in \Xi$ satisfy $\xi_i' < \xi_i$ and $\xi_j' = \xi_j$ for $j \neq i$. Then,*

$$(\nabla h(\xi'))_i \;\leq\; (\nabla h(\xi))_i \;\; \text{and} \;\; (\partial V^\star(\xi))_i \;\leq\; (\partial V^\star(\xi'))_i$$

where $(\cdot)_i$ is the ith entry.

However, the primal value function is unavailable. Theorem 1 gives a sufficient condition on a resilient equilibrium via the geometric multiplier, and we relate it to the duality next; see Appendices B.3–B.4 for proofs.

Theorem 1 (Geometric Multiplier Condition)**.** *For Problem (2) with $\bar{\xi} \in \Xi$, if λ is a geometric multiplier and $\nabla h(\bar{\xi}) + \lambda = 0$, then $\bar{\xi}$ is a resilient equilibrium.*

As a corollary of Theorem 1, we relate the resilient equilibrium to an optimal Lagrange multiplier. Let the standard Lagrangian for Problem (2) for $\xi \in \mathbb{R}^m$ be

$$\begin{aligned} L(\pi, \lambda; \xi) \;\; &= \;\; V_r^\pi(\rho) + \sum_{i=1}^m \lambda_i (V_{g_i}^\pi(\rho) - \xi_i) \\ &:= \;\; V_{r+\lambda^\top g}^\pi(\rho) - \lambda^\top \xi \end{aligned}$$

and the associated dual function be

$$D(\lambda; \xi) \;=\; \max_{\pi \in \Pi} L(\pi, \lambda; \xi) \;\; \text{for any } \lambda \geq 0.$$

The optimal dual function $D^\star(\xi) = \min_{\lambda \geq 0} D(\lambda; \xi)$ is achieved at an optimal Lagrange multiplier $\lambda^\star(\xi)$. By the weak duality, $D^\star(\xi) \geq V^\star(\xi)$ for any $\xi \in \mathbb{R}^m$.

Corollary 1. *Let the strong duality hold for Problem (2) with some $\bar{\xi} \in \Xi$, i.e., $V^\star(\bar{\xi}) = D^\star(\bar{\xi})$. If $\nabla h(\bar{\xi}) + \lambda^\star(\bar{\xi}) = 0$, then $\bar{\xi}$ is a resilient equilibrium.*

The strong duality in Corollary 1 only concerns the relaxed problem (2), which is much weaker than the the strong duality for the nominal problem (e.g., [Paternain et al., 2019, Ding et al., 2020]). We also remark that the strong duality is stronger than the geometric multiplier condition [Bertsekas, 2016], which is more general than the study [Hounie et al., 2024].

3.2 Resilience via Regularization

Although a resilient equilibrium always exists under mild regularity conditions, it is not straightforward to design efficient policy learning algorithms for finding such an equilibrium. To address this issue, we introduce the cost function into Problem (1) as regularization,

$$\begin{array}{ll} \underset{\pi \in \Pi, \, \xi \in \Xi}{\text{maximize}} & V_r^\pi(\rho) - h(\xi) \\ \text{subject to} & V_{g_i}^\pi(\rho) \;\geq\; \xi_i \;\; \text{for} \;\; i = 1, \ldots, m \end{array} \tag{4}$$

where $h(\xi)$ is a regularizer that is monotonically non-increasing coordinate-wise. Relaxing the constraint, i.e., decreasing ξ would increase $h(\xi)$; tightening the constraint is the opposite. Lemma 5 states that Problem (4) provides a resilient equilibrium and an associated resilient policy; see Appendix B.5 for proof.

Lemma 5 (Regularized Solution)**.** *Let $(\bar{\pi}^\star, \bar{\xi}^\star)$ an optimal solution to Problem (4). Then, $\bar{\xi}^\star$ is a resilient equilibrium and $\bar{\pi}^\star$ is an associated resilient policy.*

Problem (4) is a practical extension of constrained policy optimization to jointly optimizing over policy and relaxation. Naturally, we can enable the extension of existing constrained policy search algorithms to being resilient; see two of them in Section 4. Before that, we first show some important properties of Problem (4).

We denote the optimal value for Problem (4) by $V_h^\star := V_r^{\bar{\pi}^\star}(\rho) - h(\bar{\xi}^\star)$ that is evaluated at an optimal solution $(\bar{\pi}^\star, \bar{\xi}^\star)$ if it is feasible; $V_h^\star = -\infty$ otherwise. Let the dual function for Problem (4) be $D_h(\lambda) := \sup_{\pi \in \Pi, \xi \in \Xi} \{V_{r+\lambda^\top g}^\pi(\rho) - h(\xi) - \lambda^\top \xi\}$ and the optimal dual function be $D_h^\star := D_h(\bar{\lambda}^\star)$ that is achieved at an optimal dual variable $\bar{\lambda}^\star$.

Assumption 1 (Strict feasibility)**.** *There exist a pair of $(\bar{\pi}, \bar{\xi}) \in \Pi \times \Xi$ and a constant $c > 0$ such that $V_{g_i}^{\bar{\pi}}(\rho) - \bar{\xi}_i \geq c$ for all $i = 1, \ldots, m$.*

Due to the flexibility of selecting $\bar{\xi}$, Assumption 1 is weaker than the usual Slater condition [Altman, 1999].

Thus, the strong duality and dual boundedness hold for Problem (4); see Appendices B.6–B.7 for proofs.

Theorem 2 (Strong Duality for Regularized Problem). *Let Assumption 1 hold. Then, the strong duality holds for Problem (4), i.e., $V_h^\star = D_h^\star$.*

Corollary 2 (Dual Boundedness). *Let Assumption 1 hold. Then, the optimal dual is bounded, i.e.,*

$$0 \;\leq\; \bar{\lambda}_i^\star \;\leq\; \frac{V_r^\star(\rho) - h(\bar{\xi}^\star) - (V_r^{\bar{\pi}} - h(\bar{\xi}))}{c} \;:=\; C_h.$$

The strict feasibility of $(\bar{\pi}, \bar{\xi})$ and the optimality of $(\bar{\pi}^\star, \bar{\xi}^\star)$ leads to $C_h > 0$. Corollary 2 restricts dual variables in $\Lambda := \{\lambda \in \mathbb{R}_+^m \,|\, \lambda_i \leq C_h, i = 1, \dots, m\}$. Let the standard Lagrangian for Problem (4) be

$$L_h(\pi, \xi; \lambda) \quad := \quad V_{r+\lambda^\top g}^\pi(\rho) - h(\xi) - \lambda^\top \xi.$$

By the strong duality, Problem (4) is equivalent to the following constrained saddle-point problem,

$$\begin{aligned} & \underset{\pi \in \Pi, \, \xi \in \Xi}{\text{maximize}} \; \underset{\lambda \in \Lambda}{\text{minimize}} \; L_h(\pi, \xi; \lambda) \\ & \quad = \; \underset{\lambda \in \Lambda}{\text{minimize}} \; \underset{\pi \in \Pi, \, \xi \in \Xi}{\text{maximize}} \; L_h(\pi, \xi; \lambda). \end{aligned}$$

Let $\Pi^\star \times \Xi^\star \times \Lambda^\star$ be a set of saddle points of $L_h(\pi, \xi; \lambda)$ over $\Pi \times \Xi \times \Lambda$. From Theorem 2, there always exists such a saddle point, i.e., $\Pi^\star \times \Xi^\star \times \Lambda^\star \neq \emptyset$. From the definition of $\Xi^\star$, $|\xi_i| \leq 1/(1-\gamma)$ for any $\xi \in \Xi^\star$. Aided by these nice properties, we next introduce two constrained policy search algorithms to find a near-optimal pair of policy and constraint specification.

4 RESILIENT CONSTRAINED POLICY LEARNING

We provide two constrained policy gradient algorithms for searching for policy and constraint specification together, in Section 4.1 and Section 4.2, respectively.

4.1 Resilient Policy Gradient Primal-Dual (ResPG-PD) Method

We generalize the policy gradient primal-dual mirror descent [Ding and Jovanović, 2022] for our resilient problem (4) by adding a relaxation update. The resilient policy gradient primal-dual (ResPG-PD) method in Algorithm 1 maintains three sequences for primal and dual variables via Primal update (5a) and Dual update (5b): two primal sequences $(\{\pi_t\}_{t\geq 1}, \{\xi_t\}_{t\geq 1})$ for policy and relaxation, and a dual sequence $\{\lambda_t\}_{t\geq 1}$, where η is the stepsize, π_0 is the uniform distribution over the action space, $\xi_0 = 0$, and $\lambda_0 = 0$. In Primal update (5a), the policy update works as the projected Q-ascent [Bhandari and Russo, 2021, Xiao, 2022] and the relaxation update performs the projected gradient ascent. Dual update (5b) is the standard projected gradient descent. When the relaxation is fixed, i.e., $\xi_t = \xi$, the relaxation update does not impact the dual update, and thus Algorithm 1 reduces to the policy gradient primal-dual method in Euclidean space. By viewing this, we next extend the average-value convergence analysis for our resilient problem by incorporating the additional relaxation update $\{\xi_t\}_{t\geq 1}$ in Theorem 3 and delay its proof to Appendix C.1.

We measure the performance of Algorithm 1 by comparing the sequences $\{\pi_t, \xi_t, \lambda_t\}_{t\geq 1}$ with the optimal solution $(\bar{\pi}^\star, \bar{\xi}^\star)$ in the standard notion of regret,

$$\begin{aligned} R_{\text{opt}} &= \frac{1}{T}\sum_{t=0}^{T-1}(V_r^\star(\rho) - h(\bar{\xi}^\star) - (V_r^{\pi_t}(\rho) - h(\xi_t))) \\ R_{\text{vio}} &= \sum_{i=1}^{m}\left[\frac{1}{T}\sum_{t=0}^{T-1}(\xi_{i,t} - V_{g_i}^{\pi_t}(\rho))\right]_+ \end{aligned}$$

where R_{opt} is the average of the sub-optimal gaps and R_{vio} is the sum of the averaged constraint violations.

Theorem 3 (Regret-Type Performance). *Let Assumption 1 hold. Suppose $\Lambda = [0, 2C_h]$, and $h(\xi)$ has Lipschitz continuous gradient with parameter L_h over $\xi \in \Xi$. If $\eta = 1/\sqrt{T}$ for Algorithm 1, then,*

$$\begin{aligned} R_{\text{opt}} &\leq \frac{m(7 + (L_h + 1)^2)}{\sqrt{T}} \\ R_{\text{vio}} &\leq \frac{(8 + (L_h + 1)^2)m/C_h + mC_h}{(1-\gamma)^2\sqrt{T}}. \end{aligned}$$

Theorem 3 states that the average sub-optimal gaps and constraint violations of the primal-dual iterates of Algorithm 1 decay to zero with rate $1/\sqrt{T}$. This rate matches the rate of non-resilient algorithms [Ding and Jovanović, 2022] and is independent of MDP's dimension. Due to the regularization and relaxation in Problem (4), our proof handles regularized reward value and relaxed utility value together, generalizing the prior art for a broader class of problems. Since each primal-dual iteration involves projections to a probability simplex and intervals, requiring linear complexity, Algorithm 1 has polynomial computational complexity. Denote $\xi_i' := \frac{1}{T}\sum_{t=0}^{T-1}\xi_{i,t}$. After $T = O(1/\epsilon^2)$ iterations of Algorithm 1, we can select the best policy π' from T steps,

$$\begin{aligned} V_r^\star - h(\bar{\xi}^\star) - (V_r^{\pi'} - h(\xi')) &= O(\epsilon) \\ \left[\xi_i' - V_{g_i}^{\pi'}\right]_+ &= O(\epsilon) \end{aligned}$$

However, safety-critical systems demand training stability of policy iterates, which can't be guaranteed by re-

Algorithm 1 Resilient policy gradient primal-dual (ResPG-PD) method

```
Parameters: $\eta > 0$.
Initialization: Let $\pi_0(a\,|\,s) = 1/A$ for $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $\xi_0 = 0$, and $\lambda_0 = 0$.
for step $t = 0, \ldots, T-1$ do
    Primal-dual update
        $$\pi_{t+1}(\cdot\,|\,s) = \operatorname*{argmax}_{\pi(\cdot\,|\,s)\,\in\,\Pi} \left\{ \sum_a \pi(a\,|\,s) Q^{\pi_t}_{r+\lambda_t^\top g}(s,a) - \frac{1}{2\eta}\left\|\pi(\cdot\,|\,s) - \pi_t(\cdot\,|\,s)\right\|^2 \right\}$$
        $$\xi_{t+1} = \operatorname*{argmax}_{\xi\,\in\,\Xi} \left\{ \xi^\top\left(-\nabla h(\xi_t) - \lambda_t\right) - \frac{1}{2\eta}\left\|\xi - \xi_t\right\|^2 \right\} \qquad (5a)$$
        $$\lambda_{t+1} = \operatorname*{argmin}_{\lambda\,\in\,\Lambda} \left\{ \lambda^\top\left(V_g^{\pi_t}(\rho) - \xi_t\right) + \frac{1}{2\eta}\left\|\lambda - \lambda_t\right\|^2 \right\} \qquad (5b)$$
end for
```

gret performance. This issue was addressed by the state-augmentation method [Calvo-Fullana et al., 2023] and the regularized or optimistic policy gradient methods [Ding et al., 2023] in non-resilient problems. We next address this issue in the resilient context by offering a resilient optimistic policy gradient method.

4.2 Resilient Optimistic Policy Gradient Primal-Dual (ResOPG-PD) Method

We extend Algorithm 1 to an optimistic variant via the optimistic method [Rakhlin and Sridharan, 2013], which is detailed in Algorithm 2 in Appendix C.2. The resilient optimistic policy gradient primal-dual (ResOPG-PD) method maintains two sets of primal-dual sequences $\{\pi_t, \xi_t, \lambda_t\}_{t\,\geq\,1}$ and $\{\widehat{\pi}_t, \widehat{\xi}_t, \widehat{\lambda}_t\}_{t\,\geq\,1}$. The update for $\{\widehat{\pi}_t, \widehat{\xi}_t, \widehat{\lambda}_t\}_{t\,\geq\,1}$ is similar as (5) that can be viewed as a real update, except that their gradients are computed at some intermediate iterates $\{\pi_t, \xi_t, \lambda_t\}_{t\,\geq\,1}$ that serve as predictions, instead of previous iterates. Thus, the real step is optimistic about the predictions, which is used to stabilize the dynamics of gradient-based algorithms [Popov, 1980]. We can also view Algorithm 2 as a resilient version of the optimistic policy gradient primal-dual method [Ding et al., 2023] with the introduction of the relaxation update $\{\xi_t, \widehat{\xi}_t\}_{t\,\geq\,1}$. By accounting for the relaxation update, we establish convergence guarantee on the primal-dual iterates in Theorem 4; see Appendix C.3 for proof.

We first state a few notations. The distribution mismatch coefficient over ρ is $\kappa := \sup_\pi \|d_\rho^\pi/\rho\|_\infty$, where the division is component-wise. Clearly, $\kappa \leq 1/\rho_{\min}$, where $\rho_{\min} := \min_s \rho(s)$. The projection operator $\mathcal{P}_X(\cdot)$ is given by $\mathcal{P}_X(x) := \operatorname{argmin}_{x'\,\in\,X}\|x - x'\|$.

Theorem 4 (Last-Iterate Convergence)**.** *Let Assumption 1 hold. Suppose $\Lambda = [0, 2C_h]$, $\rho_{\min} > 0$, $h(\xi)$ is strongly convex and has Lipschitz continuous gradient with parameter L_h, and the optimal state visitation distribution is unique, i.e., $d_\rho^{\pi^\star} = d_\rho^\pi$ for $\pi \in \Pi^\star$. If we set stepsize $\eta \leq \eta_{\max}$, where $\eta_{\max}$ is given in Appendix C.3 for Algorithm 2 in Appendix C.2, then for any t,*

$$\frac{1}{2(1-\gamma)}\sum_s d_\rho^{\pi^\star}(s)\left\|\mathcal{P}_{\Pi^\star}(\widehat{\pi}_t(\cdot\,|\,s)) - \widehat{\pi}_t(\cdot\,|\,s)\right\|^2 + \frac{1}{2}\left\|\mathcal{P}_{\Xi^\star}(\widehat{\xi}_t) - \widehat{\xi}_t\right\|^2 + \frac{1}{2}\left\|\mathcal{P}_{\Lambda^\star}(\widehat{\lambda}_t) - \widehat{\lambda}_t\right\|^2 = O\left(\frac{1}{t}\right)$$

where $O(\cdot)$ hides a problem-dependent constant $C_{\rho,\gamma,\sigma}$ in Lemma 11 in Appendix C.3.

Theorem 4 states that the primal-dual iterates of Algorithm 2 converge to a set of $(\bar{\pi}^\star, \bar{\xi}^\star, \bar{\lambda}^\star)$ in a sublinear rate. Due to the introduction of regularization and relaxation, our proof distinguishes itself from the problem-dependent linear rate [Ding et al., 2023], e.g., a new quadratic term enters into the lower bound in Lemma 11. An immediate implication of Theorem 4 is that the primal iterates $(\widehat{\pi}_t, \widehat{\xi}_t)$ are ϵ-near optimal after $O(1/\epsilon^2)$ iterations; see Appendix C.4 for proof.

Corollary 3 (Near-Optimal Policy and Relaxation)**.** *Let assumptions in Theorem 4 hold. For a desired level of accuracy $\epsilon > 0$, if the stepsize η is provided by Theorem 4, then for $t = \Omega(1/\epsilon^2)$,*

$$\begin{aligned} V_r^\star(\rho) - h(\bar{\xi}^\star) - (V_r^{\widehat{\pi}_t}(\rho) - h(\widehat{\xi}_t)) &= O(\epsilon) \\ \left\|\widehat{\xi}_t - V_g^{\widehat{\pi}_t}(\rho)\right\| &= O(\epsilon) \end{aligned}$$

where $\Omega(\cdot)$ hides some problem-dependent constant.

Corollary 3 states that the last primal iterate $(\widehat{\pi}_t, \widehat{\xi}_t)$ is ϵ-near optimal after $\Omega(1/\epsilon^2)$ iterations. This iteration complexity is similar as the one for Algorithm 1, as well as the computational complexity. However, policy convergence in Corollary 3 is stated per iterate, which is stronger than the one for the best iterate.

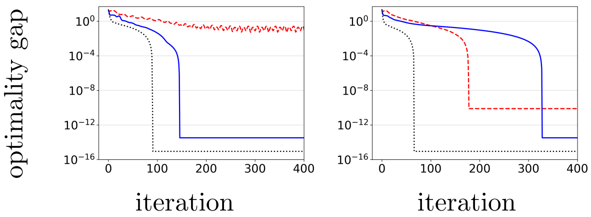


Figure 2: Policy optimality gaps of ResPG-PD (Algorithm 1, left) and ResOPG-PD (Algorithm 2, right), with three cost functions $h(\xi) = \alpha\xi^2$ for $\alpha = 0.03$ (- - -) $\alpha = 0.2$ (——), $\alpha = 1$ (⋯⋯), and stepsize $\eta = 0.2$.

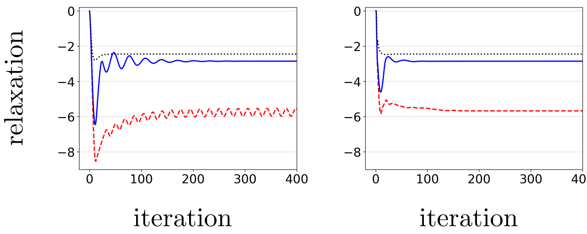


Figure 3: Relaxation of ResPG-PD (Algorithm 1, left) and ResOPG-PD (Algorithm 2, right), with three cost functions $h(\xi) = \alpha\xi^2$ for $\alpha = 0.03$ (- - -), $\alpha = 0.2$ (——), $\alpha = 1$ (⋯⋯) and stepsize $\eta = 0.2$.

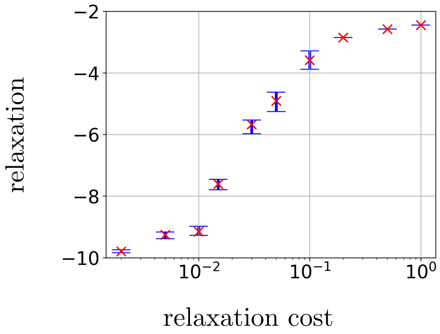


Figure 4: Constraint specifications under different relaxation costs for Algorithm 1 (ResPG-PD, ⌶) and Algorithm 2 (ResOPG-PD, ×). The relaxation cost function is $h(\xi) = \alpha\xi^2$. The horizontal axis is the value of α and the vertical axis is the relaxation ξ. The height of ⌶ is the oscillation magnitude of ResPG-PD. We run algorithms for 2000 iterations with stepsize $\eta = 0.2$ and uniform initial distribution ρ.

5 EXPERIMENTS

We show the merits and the effectiveness of our resilient policy search algorithms: ResPG-PD (Algorithm 1) and ResOPG-PD (Algorithm 2) in three experiments.

We first use a randomly-generated constrained MDP with state/action size $(20, 5)$ and calculate its optimal policy and relaxation as a sanity check; see Appendix D for the detail. Figure 2 shows that ResOPG-PD's policy iterates converge for different cost functions and so does ResPG-PD except for a small cost function, which verifies the average and the last-iterate performance in Theorems 3–4; see the convergence of relaxation in Figure 3. This indicates the capability of ResPG-PD and ResOPG-PD to find an optimal policy associated with a proper constraint specification. Figure 4 shows that increasing relaxing cost leads to less relaxation, providing relaxation in accord with the relaxing cost.

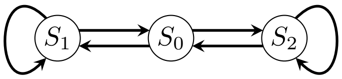


Figure 5: Robot monitoring of three locations.

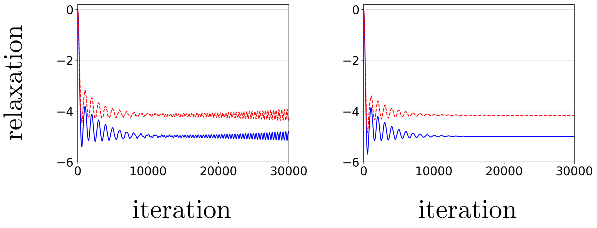


Figure 6: Relaxations (ξ_1: ——, ξ_2: - - -) of ResPG-PD (Algorithm 1, left) and ResOPG-PD (Algorithm 2, right), with a cost functions $h(\xi) = \alpha \left\|\xi\right\|^2$ for $\alpha = 0.1$, and stepsize $\eta = 0.005$.

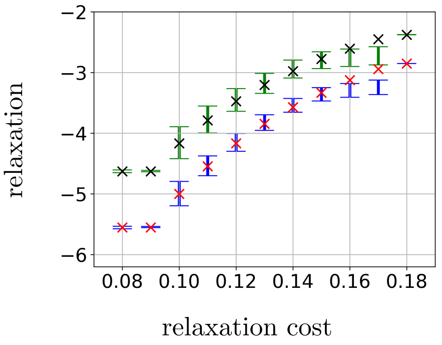


Figure 7: Constraint specifications under different relaxation costs for Algorithm 1 (ResPG-PD, ξ_1: ⌶, ξ_2: ⌶) and Algorithm 2 (ResOPG-PD, ξ_1: ×, ξ_2: ×). The relaxation cost function is $h(\xi) = \alpha \left\|\xi\right\|^2$. The horizontal axis is the value of α and the vertical axes are relaxations ξ_1 and ξ_2. The height of ⌶ is the oscillation magnitude of ResPG-PD. We run algorithms for 100000 iterations with stepsize $\eta = 0.005$ and uniform initial distribution ρ.

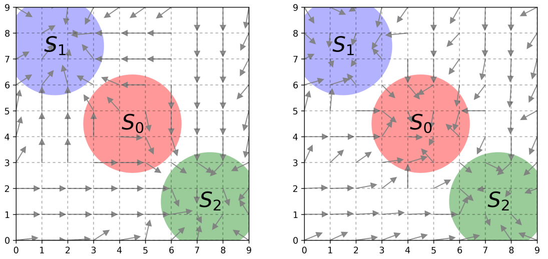
Second, we consider a monitoring problem in Figure 5 in which an agent needs to spend as much time as possible at S_0 and must stay S_1 and S_2 with some time [Calvo-Fullana et al., 2023]; see Appendix D for the detail. Due to the unknown feasibility of the time for S_1 or S_2, it warrants a resilient approach to relax (or tighten) either constraints. We choose the initial constraints to be *infeasible*. Figure 6 shows that both algorithms can adapt two relaxations to the difficulty of constraints: one is relaxed more than the other, and Figure 7 shows two relaxation curves for two algorithms against different relaxation cost functions.

Third, we generalize the previous monitoring problem to a larger state/action space in Figure 8, where in a given time a robot has to stay in blue/green areas for a minimum amount of time while maximizing the time in red area in Figure 8; see Appendix D for the detail. We also choose *infeasible* initial constraints. Figure 8(a) shows that applying a non-resilient method to this infeasible problem yields a policy that does not monitor S_0, where the non-resilient method is ResOPG-PD without relaxation update [Ding et al., 2023]. ResOPG-PD's policy in Figure 8(b) balances three areas. Figure 9 shows that ResOPG-PD gets higher reward value than the non-resilient method by modifying the constraints. However, the non-resilient method does not balance the reward and the constraints. Figure 10 shows two relaxation curves for ResPG-PD and ResOPG-PD against different relaxation cost functions.

6 CONCLUDING REMARKS

To specify the constraint specifications, we have presented an approach by making trade-off between the marginal decrease in the optimal reward value function that results from relaxation and the marginal increase in relaxation cost. We show the existence of such a resilient equilibrium under some mild regularity conditions, and provide a tractable constrained policy optimization that takes this equilibrium as an optimal solution. We provide two constrained policy search algorithms to search for such a resilient equilibrium with convergence guarantees on the optimality gap and constraint violation, generating nearly optimal policy and constraint specification. A series of computational experiments have demonstrated that our resilient constrained policy search methods effectively sustain the trade-off between the reward maximization and the constraint satisfaction even if the problem is infeasible.

For future work, our approach readily enhances existing constrained RL algorithms with resilience in different learning settings. It is also of our interest to study tighter convergence analysis, function approximation, and sample-based algorithms. For application, it is important to investigate the resilient policy learning in other practical constrained RL problems.



(a) non-resilient policy (b) resilient policy

Figure 8: Robot monitoring of three areas. Arrows mean the moving directions of a policy generated by (a) a non-resilient algorithm: Algorithm 2 (ResOPG-PD) without relaxation updates; (b) a resilient algorithm: Algorithm 2 (ResOPG-PD), with a cost functions $h(\xi) = \alpha \|\xi\|^2$ for $\alpha = 0.08$, and stepsize $\eta = 0.05$.

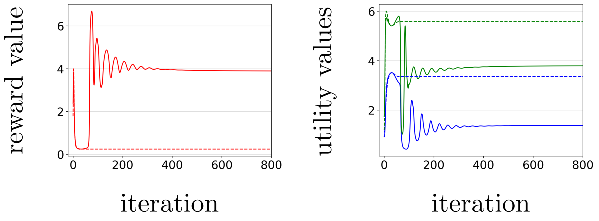


Figure 9: Convergence performance of a non-resilient method ($V_r^\pi(\rho)$: - - - , $V_{g_1}^\pi(\rho)$: - - - , $V_{g_2}^\pi(\rho)$: - - -) and Algorithm 2 (ResOPG-PD, $V_r^\pi(\rho)$: ——, $V_{g_1}^\pi(\rho)$: ——, $V_{g_2}^\pi(\rho)$: ——), with a cost functions $h(\xi) = \alpha \|\xi\|^2$ for $\alpha = 0.08$, and stepsize $\eta = 0.05$.

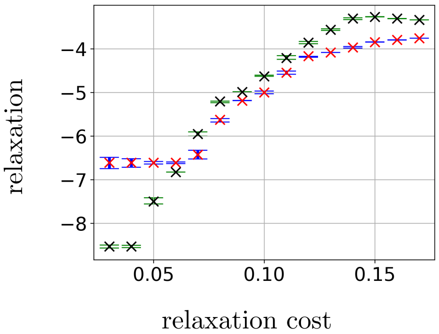


Figure 10: Constraint specifications under different relaxation costs for Algorithm 1 (ResPG-PD, ξ_1: ⌶, ξ_2: ⌶) and Algorithm 2 (ResOPG-PD, ξ_1: ×, ξ_2: ×). The relaxation cost function is $h(\xi) = \alpha \|\xi\|^2$. The horizontal axis is the value of α and the vertical axes are relaxations ξ_1 and ξ_2. The height of ⌶ is the oscillation magnitude of ResPG-PD. We run ResPG-PD for 5000 iterations with stepsize $\eta = 0.01$, and ResOPG-PD for 2000 iterations with stepsize $\eta = 0.05$. The initial distribution ρ is uniform.

Acknowledgements

We thank the anonymous reviewers for their valuable comments and suggestions. D. Ding and A. Ribeiro were supported by THEORINET Simons-NSF MoDL, and DCIST CRA.

References

[Agarwal et al., 2021] Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506.

[Altman, 1999] Altman, E. (1999). *Constrained Markov decision processes*, volume 7. CRC Press.

[Altman and Gaitsgory, 1993] Altman, E. and Gaitsgory, V. A. (1993). Stability and singular perturbations in constrained Markov decision problems. *IEEE Transactions on Automatic Control*, 38(6):971–975.

[Altman and Shwartz, 1991] Altman, E. and Shwartz, A. (1991). Sensitivity of constrained Markov decision processes. *Annals of Operations Research*, 32(1):1–22.

[Bertsekas, 2016] Bertsekas, D. (2016). *Nonlinear Programming*, volume 4. Athena Scientific.

[Bhandari and Russo, 2021] Bhandari, J. and Russo, D. (2021). On the linear convergence of policy gradient methods for finite MDPs. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 2386–2394.

[Boutilier and Lu, 2016] Boutilier, C. and Lu, T. (2016). Budget allocation using weakly coupled, constrained Markov decision processes. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 52–61.

[Brunke et al., 2022] Brunke, L., Greeff, M., Hall, A. W., Yuan, Z., Zhou, S., Panerati, J., and Schoellig, A. P. (2022). Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:411–444.

[Calvo-Fullana et al., 2023] Calvo-Fullana, M., Paternain, S., Chamon, L. F., and Ribeiro, A. (2023). State augmented constrained reinforcement learning: Overcoming the limitations of learning with rewards. *IEEE Transactions on Automatic Control.*

[Chamon et al., 2020] Chamon, L. F., Amice, A., Paternain, S., and Ribeiro, A. (2020). Resilient control: Compromising to adapt. In *Proceedings of the 59th IEEE Conference on Decision and Control*, pages 5703–5710.

[Chow et al., 2017] Chow, Y., Ghavamzadeh, M., Janson, L., and Pavone, M. (2017). Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120.

[De Nijs et al., 2021] De Nijs, F., Walraven, E., De Weerdt, M., and Spaan, M. (2021). Constrained multiagent Markov decision processes: A taxonomy of problems and algorithms. *Journal of Artificial Intelligence Research*, 70:955–1001.

[Diaz et al., 2023] Diaz, P. R., Killian, J. A., Xu, L., Suggala, A. S., Taneja, A., and Tambe, M. (2023). Flexible budgets in restless bandits: a primal-dual algorithm for efficient budget allocation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12103–12111.

[Ding and Jovanović, 2022] Ding, D. and Jovanović, M. R. (2022). Policy gradient primal-dual mirror descent for constrained MDPs with large state spaces. In *Proceedings of the IEEE 61st Conference on Decision and Control*, pages 4892–4897.

[Ding et al., 2023] Ding, D., Wei, C.-Y., Zhang, K., and Ribeiro, A. (2023). Last-iterate convergent policy gradient primal-dual methods for constrained MDPs. In *Proceedings of the Advances in Neural Information Processing Systems.*

[Ding et al., 2021] Ding, D., Wei, X., Yang, Z., Wang, Z., and Jovanovic, M. (2021). Provably efficient safe exploration via primal-dual policy optimization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 3304–3312.

[Ding et al., 2020] Ding, D., Zhang, K., Basar, T., and Jovanovic, M. (2020). Natural policy gradient primal-dual method for constrained Markov decision processes. In *Advances in Neural Information Processing Systems*, pages 8378–8390.

[Ding et al., 2022] Ding, D., Zhang, K., Duan, J., Başar, T., and Jovanović, M. R. (2022). Convergence and sample complexity of natural policy gradient primal-dual methods for constrained MDPs. *arXiv preprint arXiv:2206.02346.*

[Efroni et al., 2020] Efroni, Y., Mannor, S., and Pirotta, M. (2020). Exploration-exploitation in constrained MDPs. *arXiv preprint arXiv:2003.02189.*

[El-Shamouty et al., 2020] El-Shamouty, M., Wu, X., Yang, S., Albus, M., and Huber, M. F. (2020). Towards safe human-robot collaboration using deep reinforcement learning. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 4899–4905.

[Eysenbach et al., 2019] Eysenbach, B., Tyo, J., Gu, S., Brain, G., Salakhutdinov, R., Lipton, Z., and Levine, S. (2019). Reinforcement learning with unknown reward functions. In *Task-Agnostic Reinforcement Learning Workshop at ICLR 2019*.

[Gao et al., 2022] Gao, W., Deng, C., Jiang, Y., and Jiang, Z.-P. (2022). Resilient reinforcement learning and robust output regulation under denial-of-service attacks. *Automatica*, 142:110366.

[Garcıa and Fernández, 2015] Garcıa, J. and Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480.

[Gu et al., 2022] Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., Yang, Y., and Knoll, A. (2022). A review of safe reinforcement learning: Methods, theory and applications. *arXiv preprint arXiv:2205.10330*.

[Holling, 1973] Holling, C. S. (1973). Resilience and stability of ecological systems. *Annual Review of Ecology and Systematics*, 4(1):1–23.

[Holling, 1996] Holling, C. S. (1996). Engineering resilience versus ecological resilience. *Engineering within Ecological Constraints*, 31(1996):32.

[Hounie et al., 2024] Hounie, I., Ribeiro, A., and Chamon, L. F. (2024). Resilient constrained learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36.

[Huang et al., 2022] Huang, Y., Huang, L., and Zhu, Q. (2022). Reinforcement learning for feedback-enabled cyber resilience. *Annual Reviews in Control*, 53:273–295.

[Kaspar et al., 2020] Kaspar, M., Osorio, J. D. M., and Bock, J. (2020). Sim2real transfer for reinforcement learning without dynamics randomization. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4383–4388.

[Liang et al., 2022] Liang, X., Shu, K., Lee, K., and Abbeel, P. (2022). Reward uncertainty for exploration in preference-based reinforcement learning. *arXiv preprint arXiv:2205.12401*.

[Liu et al., 2014] Liu, C., Xu, X., and Hu, D. (2014). Multiobjective reinforcement learning: A comprehensive overview. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(3):385–398.

[Liu et al., 2023] Liu, Z., Guo, Z., Yao, Y., Cen, Z., Yu, W., Zhang, T., and Zhao, D. (2023). Constrained decision transformer for offline safe reinforcement learning. In *Proceedings of the International Conference on Machine Learning*.

[Paternain et al., 2022] Paternain, S., Calvo-Fullana, M., Chamon, L. F., and Ribeiro, A. (2022). Safe policies for reinforcement learning via primal-dual methods. *IEEE Transactions on Automatic Control*.

[Paternain et al., 2019] Paternain, S., Chamon, L., Calvo-Fullana, M., and Ribeiro, A. (2019). Constrained reinforcement learning has zero duality gap. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 32.

[Pérez-D'Arpino et al., 2021] Pérez-D'Arpino, C., Liu, C., Goebel, P., Martín-Martín, R., and Savarese, S. (2021). Robot navigation in constrained pedestrian environments using reinforcement learning. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1140–1146.

[Phan et al., 2021] Phan, T., Belzner, L., Gabor, T., Sedlmeier, A., Ritz, F., and Linnhoff-Popien, C. (2021). Resilient multi-agent reinforcement learning with adversarial value decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11308–11316.

[Popov, 1980] Popov, L. D. (1980). A modification of the Arrow-Hurwicz method for search of saddle points. *Mathematical Notes of the Academy of Sciences of the USSR*, 28:845–848.

[Rakhlin and Sridharan, 2013] Rakhlin, A. and Sridharan, K. (2013). Online learning with predictable sequences. In *Proceedings of the Conference on Learning Theory*, pages 993–1019.

[Recht and Wright, 2019] Recht, B. and Wright, S. J. (2019). *Optimization for Modern Data Analysis*. Preprint available at http://eecs.berkeley.edu/~brecht/opt4mlbook.

[Roy et al., 2022] Roy, J., Girgis, R., Romoff, J., Bacon, P.-L., and Pal, C. J. (2022). Direct behavior specification via constrained reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 18828–18843.

[Shelton, 2000] Shelton, C. (2000). Balancing multiple sources of reward in reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 13.

[Szepesvári, 2020] Szepesvári, C. (2020). Constrained MDPs and the reward hypothesis.

http://readingsml.blogspot.com/2020/03/constrained-mdps-and-reward-hypothesis.html. Access on October 11, 2023.

[Vora et al., 2023] Vora, M., Thangeda, P., Grussing, M. N., and Ornik, M. (2023). Welfare maximization algorithm for solving budget-constrained multi-component POMDPs. *IEEE Control Systems Letters*.

[Wei et al., 2020] Wei, C.-Y., Lee, C.-W., Zhang, M., and Luo, H. (2020). Linear last-iterate convergence in constrained saddle-point optimization. In *Proceedings of the International Conference on Learning Representations*.

[Xiao, 2022] Xiao, L. (2022). On the convergence rates of policy gradient methods. *The Journal of Machine Learning Research*, 23(1):12887–12922.

[Yang et al., 2021] Yang, C.-H. H., Danny, I., Hung, T., Ouyang, Y., and Chen, P.-Y. (2021). Causal inference Q-network: Toward resilient reinforcement learning. In *Self-Supervision for Reinforcement Learning Workshop-ICLR 2021*.

[Zahavy et al., 2021] Zahavy, T., O'Donoghue, B., Desjardins, G., and Singh, S. (2021). Reward is enough for convex MDPs. *Proceedings of the Advances in Neural Information Processing Systems*, pages 25746–25759.

[Zhang et al., 2020] Zhang, J., Cheung, B., Finn, C., Levine, S., and Jayaraman, D. (2020). Cautious adaptation for reinforcement learning in safety-critical settings. In *Proceedings of the International Conference on Machine Learning*, pages 11055–11065.

[Zhang et al., 2023] Zhang, Q., Zhang, L., Shen, L., Xu, H., Wang, B., Yuan, B., Chang, Y., and Wang, X. (2023). Saformer: A conditional sequence modeling approach to offline safe reinforcement learning. In *International Conference on Learning Representations 2023 Workshop on Scene Representations for Autonomous Driving*.

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. [No]
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 that 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]

Resilient Constrained Reinforcement Learning: Supplementary Materials

A Proofs in Section 2

We state proofs for claims made in Section 2.

A.1 Proof of Lemma 1

By the definition, $V^\star(\xi)$ is real-valued for $\xi \in \Xi$. The monotonic property of $V^\star(\xi)$ over $\xi \in \Xi$ is straightforward from Problem (2). We next prove the concavity.

We first check the convexity of the domain Ξ. For any ξ, $\xi' \in \Xi$, there exist π, $\pi' \in \Pi$ such that $V_{g_i}^{\pi}(\rho) \geq \xi_i$ and $V_{g_i}^{\pi'}(\rho) \geq \xi_i'$ for $i = 1, \ldots, m$. Let the occupancy measures associated with π, π' be q, q', respectively. Thus, $\langle g_i, q \rangle \geq \xi_i$ and $\langle g_i, q' \rangle \geq \xi_i$ for $i = 1, \ldots, m$, which implies $\langle g_i, \alpha q + (1-\alpha) q' \rangle \geq \alpha \xi_i + (1-\alpha)\xi_i'$, i.e., the policy induced by $\alpha q + (1-\alpha) q'$ meets the constraint. Therefore, $\alpha \xi + (1-\alpha)\xi' \in \Xi$.

We next show the convexity of $V^\star(\xi)$ over $\xi \in \Xi$. We re-formulate Problem (2) in terms of occupancy measure,

$$\begin{array}{ll} \underset{q \,\in\, \mathcal{Q}}{\text{maximize}} & \langle r, q \rangle \\ \text{subject to} & \langle u_i, q \rangle \;\geq\; \xi_i \;\; \text{for all} \;\; i = 1, \ldots, m \end{array} \tag{6}$$

where q is the occupancy measure that lives in a polytope $\mathcal{Q}$ specified by Bellman flow equations [Altman, 1999]. Instead of policy π, we work with the occupancy measure q in Problem (6). The primal function $V^\star(\xi)$, $\xi \in \Xi$ does not change, because of the one-to-one correspondence between π and q. The rest is straightforward from convex analysis. We define the function,

$$F(q, \xi) \;=\; \begin{cases} \langle r, q \rangle & \text{if} \;\; \langle g_i, q \rangle \geq \xi_i \;\; \text{for} \; i = 1, \ldots, m; \\ -\infty & \text{otherwise.} \end{cases}$$

and its domain,

$$\operatorname{dom}(F) \;=\; \left\{ (q, \xi) \,|\, q \in \mathcal{Q}, \xi \in \mathbb{R}^m_\gamma, \langle g_i, q \rangle \geq \xi_i \;\; \text{for} \; i = 1, \ldots, m \right\}.$$

Because of linearity, $\operatorname{dom}(F)$ is a convex set, and $F(q, \xi)$ is a concave function in its domain. We notice that $V^\star(\xi) = \sup_{q \,\in\, \mathcal{Q}} F(q, \xi)$ and the convex domain Ξ. Therefore, $V^\star(\xi)$ is a concave function.

A.2 Proof of Equation (3)

$$\begin{aligned} D(\lambda) \;&=\; \sup_{\pi \,\in\, \Pi} \; V^{\pi}_{r + \lambda^\top g}(\rho) \\ &=\; \sup_{\{(\xi, \pi) \,|\, \pi \,\in\, \Pi, \, \xi \,\in\, \Xi, V^{\pi}_{g_i}(\rho) \,\geq\, \xi_i, \, i = 1, \ldots \, m\}} \; V^{\pi}_{r + \lambda^\top g}(\rho) \\ &=\; \sup_{\{(\xi, \pi) \,|\, \pi \,\in\, \Pi, \, \xi \,\in\, \Xi, V^{\pi}_{g_i}(\rho) \,\geq\, \xi_i, \, i = 1, \ldots \, m\}} \; \left\{ V^{\pi}_{r}(\rho) + \lambda^\top \xi \right\} \\ &=\; \sup_{\xi \,\in\, \Xi} \; \sup_{\{\pi \,\in\, \Pi, \, V^{\pi}_{g_i}(\rho) \,\geq\, \xi_i, \, i = 1, \ldots \, m\}} \; \left\{ V^{\pi}_{r}(\rho) + \lambda^\top \xi \right\} \\ &=\; \sup_{\xi \,\in\, \Xi} \; \left\{ \lambda^\top \xi - (-V^\star(\xi)) \right\}. \end{aligned}$$

A.3 Proof of Lemma 2

There are two directions.

(ii) $\Longrightarrow$ (i): From the geometric multiplier λ for Problem (2),

$$\begin{aligned} V^\star(\xi) &= \sup_{\pi\,\in\,\Pi} \left\{ V^\pi_{r+\lambda^\top g}(\rho) - \lambda^\top \xi \right\} \\ &= \sup_{\xi'\,\in\,\Xi} \left\{ \lambda^\top \xi' - (-V^\star(\xi')) \right\} - \lambda^\top \xi \end{aligned}$$

where the second equality is due to (3). Therefore,

$$-V^\star(\xi') \;\geq\; -V^\star(\xi) + \lambda^\top(\xi' - \xi) \text{ for all } \xi' \in \mathbb{R}^m \tag{7}$$

which shows that λ is a subgradient of $-V^\star(\xi)$ at $\xi \in \Xi$.

(i) $\Longrightarrow$ (ii): Assume that (7) holds for some λ. Since $V^\star(\xi)$ is monotonically non-increasing with respect to the coordinates of ξ, thus $\lambda \geq 0$. Otherwise, $V^\star(\xi') + \lambda^\top(\xi' - \xi)$ would be unbounded below which makes (7) invalid. From (7), we have

$$V^\star(\xi) \;\geq\; \sup_{\xi'\,\in\,\Xi} \left\{ V^\star(\xi') + \lambda^\top(\xi' - \xi) \right\} \;=\; D(\lambda) - \lambda^\top \xi$$

where the second equality is due to (3). We notice that $D(\lambda) - \lambda^\top \xi$ is the dual function for Problem (2) and the weak duality $V^\star(\xi) \leq D(\lambda) - \lambda^\top \xi$. Therefore, λ is a geometric multiplier for Problem (2).

B Proofs in Section 3

We state proofs for claims made in Section 3.

B.1 Proof of Lemma 3

To show the existence, it is equivalent to show existence of subgradients for the function $\mathbf{0} \in \partial(-V^\star(\bar{\xi}) + h(\bar{\xi}))$ for some $\bar{\xi} \in \Xi$. We introduce the set $\bar{\Xi}$,

$$\bar{\Xi} \;:=\; \underset{\xi\,\in\,\Xi}{\operatorname{argmin}} \;\{ -V^\star(\xi) \,+\, h(\xi) \}\,. \tag{8}$$

We notice that Ξ is an effective domain for $-V^\star(\xi) + h(\xi)$. Because of the concavity of $V^\star(\xi)$ in Lemma 1 and the convexity of $h(\xi)$, $-V^\star(\xi) + h(\xi)$ is a convex function on Ξ. Thus, the set $\bar{\Xi}$ is nonempty and $\mathbf{0} \in \partial(-V^\star(\xi') + h(\xi'))$ for any $\xi' \in \bar{\Xi}$ according to the first-order optimality [Recht and Wright, 2019, Theorem 8.2]. Therefore, the existence is proved by simply taking a resilient equilibrium $\xi^\star = \bar{\xi} \in \bar{\Xi}$.

From the further hypothesis on h, the function $-V^\star(\xi) + h(\xi)$ is strictly convex. Thus, the minimizer in Problem (8) is unique or $\bar{\Xi}$ is a singleton. Therefore, the uniqueness holds.

B.2 Proof of Lemma 4

By the concavity of $V^\star$ in Lemma 1, if $p \in \partial V^\star(\xi)$ and $p' \in \partial V^\star(\xi')$ for ξ, $\xi' \in \Xi$, then,

$$\begin{aligned} V^\star(\xi') &\leq V^\star(\xi) + \langle p, \xi' - \xi \rangle \\ V^\star(\xi) &\leq V^\star(\xi') + \langle p', \xi - \xi' \rangle. \end{aligned}$$

Thus,

$$\langle p' - p, \xi - \xi' \rangle \;\geq\; 0$$

which implies the second inequality. Similarly, the convexity of h yields

$$\langle \xi - \xi', \nabla h(\xi') - \nabla h(\xi) \rangle \;\leq\; 0$$

which implies the first inequality.

B.3 Proof of Theorem 1

By the geometric multiplier $\lambda \geq 0$,

$$\begin{array}{rcl} V^\star(\bar\xi) & = & \sup\limits_{\pi\,\in\,\Pi} \left\{ V_{r+\lambda^\top g}^{\pi}(\rho) - \lambda^\top \bar\xi \right\} \\ & \geq & V_{r+\lambda^\top g}^{\pi^\star(\xi)}(\rho) - \lambda^\top \bar\xi \\ & = & V_r^{\pi^\star(\xi)}(\rho) + \lambda^\top (V_g^{\pi^\star(\xi)}(\rho) - \bar\xi) \\ & \geq & V^\star(\xi) + \lambda^\top(\xi - \bar\xi) \end{array}$$

where we set $\pi = \pi^\star(\xi)$ in the first inequality, and the second inequality is due to that $V^\star(\xi) := V_r^{\pi^\star(\xi)}(\rho)$ and $V_{g_i}^{\pi^\star(\xi)}(\rho) \geq \xi_i$ for all $i = 1, \ldots, m$.

Therefore, $-V^\star(\xi) \geq -V^\star(\bar\xi) + \lambda^\top(\xi - \bar\xi)$ for all $\xi \in \Xi$, i.e., λ is a subgradient of $-V^\star(\xi)$ at $\bar\xi$. By the assumption, $\nabla h(\bar\xi)$ is a subgradient of $V^\star(\bar\xi)$, which proves that $\bar\xi$ is a resilient equilibrium.

B.4 Proof of Corollary 1

It is straightforward to verify that $\lambda^\star(\bar\xi)$ is a geometric multiplier,

$$V^\star(\bar\xi) \; = \; D^\star(\bar\xi) \; = \; D(\lambda^\star(\bar\xi); \bar\xi) \; = \; \sup_{\pi\,\in\,\Pi} \left\{ V_{r+(\lambda^\star(\bar\xi))^\top g}^{\pi}(\rho) - (\lambda^\star(\bar\xi))^\top \bar\xi \right\}.$$

By Theorem 1 and $\nabla h(\bar\xi) + \lambda^\star(\bar\xi) = 0$, $\bar\xi$ is a resilient equilibrium.

B.5 Proof of Lemma 5

It is equivalent to show that

$$\bar\xi^\star \;\in\; \operatorname*{argmax}_{\xi\,\in\,\Xi} \; \{V^\star(\xi) - h(\xi)\}$$

because of the concavity of $V^\star(\xi) - h(\xi)$ over $\xi \in \Xi$, and that the first-order optimality condition $\mathbf{0} \in \partial\left(-V^\star(\bar\xi^\star) + h(\bar\xi^\star)\right)$ equals to the resilient equilibrium's condition.

By the optimality of $(\bar\pi^\star, \bar\xi^\star)$,

$$V_r^{\bar\pi^\star}(\rho) - h(\bar\xi^\star) \;\geq\; V_r^{\pi^\star(\xi)}(\rho) - h(\xi) \;=\; V^\star(\xi) - h(\xi)$$

where the right hand side of inequality particularly uses a pair $(\pi^\star(\xi), \xi)$ in which $\pi^\star(\xi)$ is an optimal policy of Problem (2) for some fixed $\xi \in \Xi$, and the equality is clear from $V^\star(\xi) := V_r^{\pi^\star(\xi)}(\rho)$. For the left hand side of the inequality, application of $\pi^\star(\bar\xi^\star)$ leads to $V^\star(\bar\xi^\star) := V_r^{\pi^\star(\bar\xi^\star)}(\rho) \geq V_r^{\bar\pi^\star}(\rho)$ and thereby,

$$V^\star(\bar\xi^\star) - h(\bar\xi^\star) \;\geq\; V^\star(\xi) - h(\xi) \;\text{ for all } \xi \in \Xi.$$

Therefore, $\mathbf{0}$ is a subgradient of $-V^\star(\xi) + h(\xi)$ at $\bar\xi^\star \in \Xi$.

B.6 Proof of Theorem 2

By the weak duality, $V_h^\star \leq D_h^\star$. The rest is to show that $V_h^\star \geq D_h^\star$. To proceed, we first make a few observations. It is easy to show the convexity of the set $\mathcal{Z}$,

$$\mathcal{Z} \;:=\; \left\{ z \in \mathbb{R}^{m+1} \,|\, \exists(\pi, \xi) \in \Pi \times \Xi, V_r^\pi(\rho) - h(\xi) \geq z_0 \text{ and } V_{g_i}^\pi(\rho) - \xi_i \geq z_i \text{ for all } i = 1, \ldots, m \right\}.$$

By the optimality of $(\bar\pi^\star, \bar\xi^\star)$, $V_h^\star = V_r^{\bar\pi^\star}(\rho) - h(\bar\xi^\star)$ and $V_{g_i}^{\bar\pi^\star}(\rho) - \bar\xi_i^\star \geq 0$ for all $i = 1, \ldots, m$, i.e., $(V_h^\star, \mathbf{0}) \in \mathcal{Z}$. Hence, $\mathcal{Z}$ is non-empty. In fact, $\mathcal{Z}$ is a convex set. Assume z, $z' \in \mathcal{Z}$ and $\alpha \in [0, 1]$. Thus, there exist π, $\pi' \in \Pi$ and ξ, $\xi' \in \Xi$ such that

$$V_r^\pi(\rho) - h(\xi) \;\geq\; z_0 \;\text{ and }\; V_r^{\pi'}(\rho) - h(\xi') \;\geq\; z_0'$$

$$V_{g_i}^\pi(\rho) - \xi_i \;\geq\; z_i \;\text{ and }\; V_{g_i}^{\pi'}(\rho) - \xi_i' \;\geq\; z_i' \;\text{ for all } i = 1, \ldots, m.$$

Let the occupancy measures associated with π, π' be q, q', respectively. Thus, $\langle r, q\rangle - h(\xi) \geq z_0$ and $\langle r, q'\rangle - h(\xi') \geq z_0'$, which implies $\langle r, \alpha q + (1-\gamma) q'\rangle - h(\alpha\xi + (1-\alpha)\xi') \geq \alpha z_0 + (1-\alpha) z_0'$. Meanwhile, $\langle g_i, q\rangle - \xi_i \geq z_i$ and $\langle g_i, q'\rangle - \xi_i \geq z_i'$ for $i = 1, \ldots, m$, which implies $\langle g_i, \alpha q + (1-\alpha) q'\rangle - (\alpha \xi_i + (1-\alpha)\xi_i') \geq \alpha z_i + (1-\alpha) z_i'$, i.e., the policy induced by $\alpha q + (1-\alpha) q'$ meets the constraint in $\mathcal{Z}$. Therefore, $\alpha z + (1-\alpha) z' \in \mathcal{Z}$.

To show that $V_h^\star \geq D_h^\star$, it is sufficient to prove that there exists $\lambda \in \Lambda$ such that

$$V_h^\star \;\geq\; D_h(\lambda) \;:=\; \sup_{\pi \,\in\, \Pi} \left\{ V_{r+\lambda^\top g}^{\pi}(\rho) - h(\xi) - \lambda^\top \xi \right\}. \tag{9}$$

We notice that $(V_h^\star, \mathbf{0}) \in \partial Z$, where ∂Z is the boundary set of $\mathcal{Z}$. If not, then $(V_h^\star, \mathbf{0}) \in \operatorname{int}(\mathcal{Z})$, i.e., there exists a small ball around $(V_h^\star, \mathbf{0})$ inside $\mathcal{Z}$, which contradicts the optimality of $V_h^\star$. Since $\mathcal{Z}$ is a convex set and $(V_h^\star, \mathbf{0}) \in \partial Z$, by the supporting hyperplane theorem, there exists $\widehat{\lambda} := (\widehat{\lambda}_0, \widehat{\lambda}_1, \ldots, \widehat{\lambda}_m) \in \mathbb{R}^{m+1}$ such that

$$\begin{bmatrix} V_h^\star & \mathbf{0}^\top \end{bmatrix} \widehat{\lambda} \;\geq\; z^\top \widehat{\lambda} \;\text{ for all } z \in \mathcal{Z}. \tag{10}$$

We can show that $\widehat{\lambda} \geq 0$ and $\widehat{\lambda}_0 > 0$ by contradiction. Assume $\widehat{\lambda}_i < 0$ for some i. Since $\mathcal{Z}$ is unbounded from below, we can always select a very negative z_i that fails (10). Hence, $\widehat{\lambda}_i \geq 0$ for all $i = 0, 1, \ldots, m$. On the other hand, assume $\widehat{\lambda}_0 = 0$. Thus, (10) reduces to $0 \geq z^\top \widehat{\lambda}$ for all $z \in \mathcal{Z}$. Non-negativity of $\widehat{\lambda}$ implies that there exists i from 1 to m such that $z_i \leq 0$, which contradicts the strict feasibility that demands a positive $z \in \mathcal{Z}$. Therefore, we can denote $\lambda^\dagger := \widehat{\lambda}/\widehat{\lambda}_0$ and

$$(\pi^\dagger, \xi^\dagger) \;:=\; \operatorname*{argmax}_{\pi \,\in\, \Pi,\, \xi \,\in\, \Xi} \left\{ V_{r+(\lambda^\dagger)^\top g}^{\pi}(\rho) - h(\xi) - (\lambda^\dagger)^\top \xi \right\}.$$

We notice that $(V_r^{\pi^\dagger}(\rho) - h(\xi^\dagger), V_g^{\pi^\dagger}(\rho) - \xi^\dagger) \in \mathcal{Z}$, and $\lambda_0^\dagger = 1$. By the dual function and (10),

$$D_h(\lambda^\dagger) \;=\; \left\langle (V_r^{\pi^\dagger}(\rho) - h(\xi^\dagger), V_g^{\pi^\dagger}(\rho) - \xi^\dagger), \lambda^\dagger \right\rangle \;\leq\; V_h^\star$$

which proves the existence (9).

B.7 Proof of Corollary 2

We denote the level set of the dual function by $\Lambda_a := \{\lambda \in \mathbb{R}_+^m \,|\, D_h(\lambda) \leq a\}$ for $a \in \mathbb{R}$. For any $\lambda \in \Lambda_a$,

$$a \;\geq\; D_h(\lambda) \;=\; V_r^{\bar\pi}(\rho) - h(\bar\xi) + \lambda^\top (V_g^{\bar\pi}(\rho) - \bar\xi) \;\geq\; V_r^{\bar\pi}(\rho) - h(\bar\xi) + c\, \lambda^\top \mathbf{1}$$

where $(\bar\pi, \bar\xi)$ is a Slater point in Assumption 1. Taking $a = D_h^\star$ or $V_r^\star(\rho) - h(\bar\xi^\star)$ leads to $\Lambda_a = \Lambda^\star$ and

$$\sum_{i\,=\,1}^{m} \lambda_i \;\leq\; \frac{V_r^\star(\rho) - h(\bar\xi^\star) - (V_r^{\bar\pi} - h(\bar\xi))}{c} \;\text{ for all } \lambda \in \Lambda_a$$

which implies the bound on $\bar\lambda_i^\star$ for $i = 1, \ldots, m$.

C Proofs in Section 4

We state proofs for claims made in Section 4.

C.1 Proof of Theorem 3

Lemma 6. *In Algorithm 1, for any $\pi \in \Delta(A)$ and $\xi \in \Xi$*

$$\begin{aligned} & \eta \langle Q_{r+\lambda_t^\top g}^{\pi_t}(s, \cdot), (\pi - \pi_{t+1})(\cdot \,|\, s)\rangle + \frac{1}{2}\left\|\pi_{t+1}(\cdot\,|\,s) - \pi_t(\cdot\,|\,s)\right\|^2 \\ \leq\;& \frac{1}{2}\left\|\pi(\cdot\,|\,s) - \pi_t(\cdot\,|\,s)\right\|^2 - \frac{1}{2}\left\|\pi(\cdot\,|\,s) - \pi_{t+1}(\cdot\,|\,s)\right\|^2 \end{aligned}$$

and

$$\begin{aligned} & \eta \langle -\nabla h(\xi_t) - \lambda_t, \xi - \xi_{t+1}\rangle + \frac{1}{2}\left\|\xi_{t+1} - \xi_t\right\|^2 \\ \leq\;& \frac{1}{2}\left\|\xi - \xi_t\right\|^2 - \frac{1}{2}\left\|\xi - \xi_{t+1}\right\|^2. \end{aligned}$$

Proof. By the optimality of π_{t+1},

$$\langle \eta Q_{r+\lambda_t^\top g}^{\pi_t}(s,\cdot) - (\pi_{t+1}-\pi_t)(\cdot\,|\,s), (\pi - \pi_{t+1})(\cdot\,|\,s)\rangle \;\leq\; 0 \;\text{ for any } \pi.$$

Direct application of the equality $\frac{1}{2}\left\|\pi(\cdot\,|\,s) - \pi_t(\cdot\,|\,s)\right\|^2 = \frac{1}{2}\left\|\pi_{t+1}(\cdot\,|\,s) - \pi_t(\cdot\,|\,s)\right\|^2 + \langle \pi_{t+1}(\cdot\,|\,s) - \pi_t(\cdot\,|\,s), \pi(\cdot\,|\,s) - \pi_{t+1}(\cdot\,|\,s)\rangle + \frac{1}{2}\left\|\pi(\cdot\,|\,s) - \pi_{t+1}(\cdot\,|\,s)\right\|^2$ leads to the first inequality. Similarly, the optimality of ξ_{t+1} shows that

$$\langle \eta(-h(\xi_t) - \lambda_t) - (\xi_{t+1} - \xi_t), \xi - \xi_{t+1}\rangle \;\leq\; 0.$$

In combination of the equality $\frac{1}{2}\left\|\xi - \xi_t\right\|^2 = \frac{1}{2}\left\|\xi_{t+1} - \xi_t\right\|^2 + \langle \xi_{t+1} - \xi_t, \xi - \xi_{t+1}\rangle + \frac{1}{2}\left\|\xi - \xi_{t+1}\right\|^2$, we conclude the second inequality. □

Lemma 7. *In Algorithm 1, for any s and t,*

$$V_r^{\pi_{t+1}}(s) - V_r^{\pi_t}(s) + \lambda_t^\top(V_g^{\pi_{t+1}}(s) - V_g^{\pi_t}(s)) \quad\geq\quad \frac{1}{\eta(1-\gamma)}\mathbb{E}_{s'\,\sim\, d_s^{\pi_{t+1}}}\left[\left\|\pi_{t+1}(\cdot\,|\,s') - \pi_t(\cdot\,|\,s')\right\|^2\right].$$

Proof. By the performance difference lemma,

$$\begin{aligned}
& V_r^{\pi_{t+1}}(s) - V_r^{\pi_t}(s) + \lambda_t^\top(V_g^{\pi_{t+1}}(s) - V_g^{\pi_t}(s)) \\
=\;\; & \frac{1}{1-\gamma}\mathbb{E}_{s'\,\sim\, d_s^{\pi_{t+1}}}\left[\langle Q_{r+\lambda_t^\top g}(s',\cdot), (\pi_{t+1} - \pi_t)(\cdot\,|\,s')\rangle\right].
\end{aligned}$$

Application of the first inequality in Lemma 6 with $\pi = \pi_t$ leads to our desired inequality. □

Lemma 8. *In Algorithm 1, for any $T > 0$,*

$$\begin{aligned}
& \frac{1}{T}\sum_{t\,=\,0}^{T-1}(V_r^\star(\rho) - h(\bar\xi^\star) - (V_r^{\pi_t}(\rho) - h(\xi_t))) + \frac{1}{T}\sum_{t\,=\,0}^{T-1}\lambda_t^\top(V_g^\star(\rho) - \bar\xi^\star - (V_g^{\pi_t}(\rho) - \xi_t)) \\
\leq\;\; & \frac{1}{(1-\gamma)^2 T} + \frac{1}{\eta(1-\gamma)T} + \frac{4\eta m}{(1-\gamma)^2} + \frac{\eta(L_h+1)^2 m}{(1-\gamma)^2}.
\end{aligned}$$

Proof. By the performance difference lemma,

$$\begin{aligned}
& V_r^\star(s) - h(\bar\xi^\star) - (V_r^{\pi_t}(s) - h(\xi_t)) + \lambda_t^\top(V_g^\star(s) - \bar\xi^\star - (V_g^{\pi_t}(s) - \xi_t)) \\
=\;\; & \frac{1}{1-\gamma}\mathbb{E}_{s'\,\sim\, d_s^\star}\left[\langle Q_{r+\lambda_t^\top g}^{\pi_t}(s',\cdot), (\bar\pi^\star - \pi_t)(\cdot\,|\,s')\right] - (h(\bar\xi^\star) - h(\xi_t)) + \lambda_t^\top(-\bar\xi^\star + \xi_t) \\
=\;\; & \frac{1}{1-\gamma}\mathbb{E}_{s'\,\sim\, d_s^\star}\left[\langle Q_{r+\lambda_t^\top g}^{\pi_t}(s',\cdot), (\bar\pi^\star - \pi_{t+1})(\cdot\,|\,s')\right] \\
& + \frac{1}{1-\gamma}\mathbb{E}_{s'\,\sim\, d_s^\star}\left[\langle Q_{r+\lambda_t^\top g}^{\pi_t}(s',\cdot), (\pi_{t+1} - \bar\pi^\star)(\cdot\,|\,s')\right] \\
& - (h(\bar\xi^\star) - h(\xi_t)) + \lambda_t^\top(-\bar\xi^\star + \xi_t) \\
\leq\;\; & \frac{1}{2\eta(1-\gamma)}\mathbb{E}_{s'\,\sim\, d_s^\star}\left[\left\|\bar\pi^\star(\cdot\,|\,s) - \pi_t(\cdot\,|\,s)\right\|^2 - \left\|\bar\pi^\star(\cdot\,|\,s) - \pi_{t+1}(\cdot\,|\,s)\right\|^2\right] \\
& + \frac{1}{1-\gamma}\mathbb{E}_{s'\,\sim\, d_s^\star}\left[\langle Q_{r+\lambda_t^\top g}^{\pi_t}(s',\cdot), (\pi_{t+1} - \bar\pi^\star)(\cdot\,|\,s')\right] \\
& - (h(\bar\xi^\star) - h(\xi_t)) + \lambda_t^\top(-\bar\xi^\star + \xi_t)
\end{aligned} \tag{11}$$

where the inequality is due to the first inequality in Lemma 6 with $\pi = \bar\pi^\star$. To further bound the inequality above, we first notice that $\langle Q_{r+\lambda_t^\top g}^{\pi_t}(s,\cdot), (\pi_{t+1} - \pi_t)(\cdot\,|\,s)\rangle \geq 0$ for any s, when we set $\pi = \pi_t$ in the first inequality in Lemma 6. Thus,

$$\begin{aligned}
& \mathbb{E}_{s'\,\sim\, d_\rho^\star}\left[\langle Q_{r+\lambda_t^\top g}^{\pi_t}(s',\cdot), (\pi_{t+1} - \pi_t)(\cdot\,|\,s')\rangle\right] \\
=\;\; & \sum_{s'}\frac{d_\rho^\star(s')}{d_{d_\rho^\star}^{\pi_{t+1}}(s')} d_{d_\rho^\star}^{\pi_{t+1}}(s')\langle Q_{r+\lambda_t^\top g}^{\pi_t}(s',\cdot), (\pi_{t+1} - \pi_t)(\cdot\,|\,s')\rangle \\
\leq\;\; & \frac{1}{1-\gamma}\sum_{s'} d_{d_\rho^\star}^{\pi_{t+1}}(s')\langle Q_{r+\lambda_t^\top g}^{\pi_t}(s',\cdot), (\pi_{t+1} - \pi_t)(\cdot\,|\,s')\rangle \\
=\;\; & (V_r^{\pi_{t+1}}(d_\rho^\star) - V_r^{\pi_t}(d_\rho^\star)) + \lambda_t^\top(V_g^{\pi_{t+1}}(d_\rho^\star) - V_g^{\pi_t}(d_\rho^\star))
\end{aligned}$$

where the inequality is due to that $d^{\pi_{t+1}}_{d^\star_\rho} \geq (1-\gamma) d^\star_\rho$. Hence, we further bound (11) as

$$\begin{aligned}
& V_r^\star(\rho) - h(\bar{\xi}^\star) - (V_r^{\pi_t}(\rho) - h(\xi_t)) + \lambda_t^\top (V_g^\star(\rho) - \bar{\xi}^\star - (V_g^{\pi_t}(\rho) - \xi_t)) \\
\leq \;\; & \frac{1}{2\eta(1-\gamma)} \mathbb{E}_{s' \sim d^\star_\rho} \left[\left\| \bar{\pi}^\star(\cdot \,|\, s) - \pi_t(\cdot \,|\, s) \right\|^2 - \left\| \bar{\pi}^\star(\cdot \,|\, s) - \pi_{t+1}(\cdot \,|\, s) \right\|^2 \right] \\
& + \frac{1}{1-\gamma} \left((V_r^{\pi_{t+1}}(d^\star_\rho) - V_r^{\pi_t}(d^\star_\rho)) + \lambda_t^\top (V_g^{\pi_{t+1}}(d^\star_\rho) - V_g^{\pi_t}(d^\star_\rho)) \right) \\
& - (h(\bar{\xi}^\star) - h(\xi_t)) + \lambda_t^\top (-\bar{\xi}^\star + \xi_t).
\end{aligned}$$

By the convexity of h, $h(\bar{\xi}^\star) \geq h(\xi_t) + \langle \nabla h(\xi_t), \bar{\xi}^\star - \xi_t \rangle$. Thus,

$$\begin{aligned}
& -(h(\bar{\xi}^\star) - h(\xi_t)) + \lambda_t^\top (-\bar{\xi}^\star + \xi_t) \\
\leq \;\; & \langle -\nabla h(\xi_t) - \lambda_t, \bar{\xi}^\star - \xi_t \rangle \\
= \;\; & \langle -\nabla h(\xi_t) - \lambda_t, \bar{\xi}^\star - \xi_{t+1} \rangle + \langle -\nabla h(\xi_t) - \lambda_t, \xi_{t+1} - \xi_t \rangle \\
\leq \;\; & \frac{1}{2\eta} \left\| \bar{\xi}^\star - \xi_t \right\|^2 - \frac{1}{2\eta} \left\| \bar{\xi}^\star - \xi_{t+1} \right\|^2 + \langle -\nabla h(\xi_t) - \lambda_t, \xi_{t+1} - \xi_t \rangle
\end{aligned}$$

where the last inequality is due to the second inequality in Lemma 6 with $\xi = \bar{\xi}^\star$. Therefore, (11) reduces to

$$\begin{aligned}
& V_r^\star(\rho) - h(\bar{\xi}^\star) - (V_r^{\pi_t}(\rho) - h(\xi_t)) + \lambda_t^\top (V_g^\star(\rho) - \bar{\xi}^\star - (V_g^{\pi_t}(\rho) - \xi_t)) \\
\leq \;\; & \frac{1}{2\eta(1-\gamma)} \mathbb{E}_{s' \sim d^\star_\rho} \left[\left\| \bar{\pi}^\star(\cdot \,|\, s) - \pi_t(\cdot \,|\, s) \right\|^2 - \left\| \bar{\pi}^\star(\cdot \,|\, s) - \pi_{t+1}(\cdot \,|\, s) \right\|^2 \right] \\
& + \frac{1}{1-\gamma} \left((V_r^{\pi_{t+1}}(d^\star_\rho) - V_r^{\pi_t}(d^\star_\rho)) + \lambda_t^\top (V_g^{\pi_{t+1}}(d^\star_\rho) - V_g^{\pi_t}(d^\star_\rho)) \right) \\
& + \frac{1}{2\eta} \left\| \bar{\xi}^\star - \xi_t \right\|^2 - \frac{1}{2\eta} \left\| \bar{\xi}^\star - \xi_{t+1} \right\|^2 + \langle -\nabla h(\xi_t) - \lambda_t, \xi_{t+1} - \xi_t \rangle.
\end{aligned}$$

Summing up the inequality above from $t = 0$ to $t = T-1$ and dividing it by T yield,

$$\begin{aligned}
& \frac{1}{T} \sum_{t=0}^{T-1} \left(V_r^\star(\rho) - h(\bar{\xi}^\star) - (V_r^{\pi_t}(\rho) - h(\xi_t)) \right) + \frac{1}{T} \sum_{t=0}^{T-1} \lambda_t^\top (V_g^\star(\rho) - \bar{\xi}^\star - (V_g^{\pi_t}(\rho) - \xi_t)) \\
\leq \;\; & \frac{1}{2\eta(1-\gamma)T} \mathbb{E}_{s' \sim d^\star_\rho} \left[\left\| \bar{\pi}^\star(\cdot \,|\, s) - \pi_0(\cdot \,|\, s) \right\|^2 - \left\| \bar{\pi}^\star(\cdot \,|\, s) - \pi_T(\cdot \,|\, s) \right\|^2 \right] \\
& + \frac{1}{(1-\gamma)T} (V_r^{\pi_T}(d^\star_\rho) - V_r^{\pi_0}(d^\star_\rho)) + \frac{1}{1-\gamma} \frac{1}{T} \sum_{t=0}^{T-1} \lambda_t^\top (V_g^{\pi_{t+1}}(d^\star_\rho) - V_g^{\pi_t}(d^\star_\rho)) \\
& + \frac{1}{2\eta T} \left\| \bar{\xi}^\star - \xi_0 \right\|^2 - \frac{1}{2\eta T} \left\| \bar{\xi}^\star - \xi_T \right\|^2 + \frac{1}{T} \sum_{t=0}^{T-1} \langle -\nabla h(\xi_t) - \lambda_t, \xi_{t+1} - \xi_t \rangle.
\end{aligned} \tag{12}$$

We notice that $\lambda_0 = 0$, $\lambda_T = \sum_{t=0}^{T-1} (\lambda_{t+1} - \lambda_t)$, and $|V_{g_i}^{\pi_t}(\rho) - \xi_{i,t}| \leq \frac{2}{1-\gamma}$ for $i = 1, \ldots, m$. From the λ-update in (5), we have $|\lambda_{i,t} - \lambda_{i,t+1}| \leq \frac{2\eta}{1-\gamma}$ and $|\lambda_{i,T}| \leq \frac{2\eta T}{1-\gamma}$. Thus,

$$\begin{aligned}
& \sum_{t=0}^{T-1} \lambda_t^\top (V_g^{\pi_{t+1}}(d^\star_\rho) - V_g^{\pi_t}(d^\star_\rho)) \\
= \;\; & \sum_{t=0}^{T-1} (\lambda_{t+1}^\top V_g^{\pi_{t+1}}(d^\star_\rho) - \lambda_t^\top V_g^{\pi_t}(d^\star_\rho)) + \sum_{t=0}^{T-1} (\lambda_t - \lambda_{t+1})^\top V_g^{\pi_{t+1}}(d^\star_\rho) \\
\leq \;\; & \lambda_T^\top V_g^{\pi_T}(d^\star_\rho) + \sum_{t=0}^{T-1} \sum_{i=1}^{m} |\lambda_{i,t} - \lambda_{i,t+1}| V_{g_i}^{\pi_{t+1}}(d^\star_\rho) \\
\leq \;\; & \frac{4\eta m T}{(1-\gamma)^2}.
\end{aligned}$$

Meanwhile, from the ξ-update in (5), we have $|\xi_{i,t+1} - \xi_{i,t}| \leq \eta | -\nabla h(\xi_t) - \lambda_t|_i \leq \frac{\eta(L_h+1)}{1-\gamma}$. Thus,

$$\sum_{t=0}^{T-1} \langle -\nabla h(\xi_t) - \lambda_t, \xi_{t+1} - \xi_t \rangle \leq \frac{\eta (L_h+1)^2 m T}{(1-\gamma)^2}.$$

Finally, we combine these inequalities above with (12) to get our desired inequality. □

Proof. To show the first inequality, we notice that $\lambda_0 = 0$, $\lambda_{i,T}^2 = \sum_{t=0}^{T-1}(\lambda_{i,t+1}^2 - \lambda_{i,t}^2)$, and $|V_{g_i}^{\pi_t}(\rho) - \xi_{i,t}| \leq \frac{2}{1-\gamma}$ for $i = 1, \ldots, m$. From the λ-update in (5), we have $|\lambda_{i,t} - \lambda_{i,t+1}| \leq \frac{2\eta}{1-\gamma}$ and $|\lambda_{i,T}| \leq \frac{2\eta T}{1-\gamma}$. Thus,

$$
\begin{aligned}
\lambda_{i,T}^2 &= \sum_{t=0}^{T-1} (\lambda_{i,t+1})^2 - (\lambda_{i,t})^2 \\
&= \sum_{t=0}^{T-1} -2\eta\lambda_{i,t}(V_{g_i}^{\pi_t}(\rho) - \xi_{i,t}) + \eta^2 (V_{g_i}^{\pi_t}(\rho) - \xi_{i,t})^2 \\
&\leq 2\eta \sum_{t=0}^{T-1} \lambda_{i,t}((V_{g_i}^{\star}(\rho) - \bar{\xi}_i^{\star}) - (V_{g_i}^{\pi_t}(\rho) - \xi_{i,t})) + \eta^2 (V_{g_i}^{\pi_t}(\rho) - \xi_{i,t})^2 \\
&\leq 2\eta \sum_{t=0}^{T-1} \lambda_{i,t}((V_{g_i}^{\star}(\rho) - \bar{\xi}_i^{\star}) - (V_{g_i}^{\pi_t}(\rho) - \xi_{i,t})) + \frac{4\eta^2 T}{(1-\gamma)^2}
\end{aligned}
$$

where the first inequality is due to the feasibility $V_{g_i}^{\star}(\rho) \geq \bar{\xi}_i^{\star}$ for $i = 1, \ldots, m$. Thus,

$$
-\frac{1}{T} \sum_{t=0}^{T-1} \lambda_t^\top ((V_g^{\star}(\rho) - \bar{\xi}^{\star}) - (V_g^{\pi_t}(\rho) - \xi_t)) \;\leq\; \frac{2\eta m}{(1-\gamma)^2}
$$

which can be added to the inequality in Lemma 8 from both sides,

$$
\begin{aligned}
& \frac{1}{T} \sum_{t=0}^{T-1} (V_r^{\star}(\rho) - h(\bar{\xi}^{\star}) - (V_r^{\pi_t}(\rho) - h(\xi_t))) \\
\leq\;& \frac{1}{(1-\gamma)^2 T} + \frac{1}{\eta(1-\gamma)T} + \frac{4\eta m}{(1-\gamma)^2} + \frac{\eta(L_h+1)^2 m}{(1-\gamma)^2} + \frac{2\eta m}{(1-\gamma)^2}.
\end{aligned}
$$

Hence, we obtain the first inequality by taking $\eta = \frac{1}{\sqrt{T}}$.

We next prove the second inequality. From the λ-update in (5), for any $\lambda \in \Lambda := \{\lambda \in \mathbb{R}^m \,|\, 0 \leq \lambda_i \leq C_h, i = 1, \ldots, m\}$,

$$
(\lambda_{i,t+1} - \lambda_i)^2 \;\leq\; (\lambda_{i,t} - \lambda_i)^2 - 2\eta(V_{g_i}^{\pi_t}(\rho) - \xi_{i,t})(\lambda_{i,t} - \lambda_i) + \eta^2 (V_{g_i}^{\pi_t}(\rho) - \xi_{i,t})^2
$$

which combines with $|V_{g_i}^{\pi_t}(\rho) - \xi_{i,t}| \leq \frac{2}{1-\gamma}$ to give us,

$$
\begin{aligned}
& \frac{1}{T} \sum_{t=0}^{T-1} \sum_{i=1}^{m} (V_{g_i}^{\pi_t}(\rho) - \xi_{i,t})(\lambda_{i,t} - \lambda_i) \\
\leq\;& \frac{1}{2\eta T} \sum_{t=0}^{T-1} \sum_{i=1}^{m} \left((\lambda_{i,t} - \lambda_i)^2 - (\lambda_{i,t+1} - \lambda_i)^2 \right) + \frac{2\eta m}{\eta(1-\gamma)^2} \\
\leq\;& \frac{1}{2\eta T} \sum_{i=1}^{m} \lambda_i^2 + \frac{2\eta m}{(1-\gamma)^2}.
\end{aligned}
\tag{13}
$$

Notice that $V_g^{\star}(\rho) \geq \bar{\xi}^{\star}$. If we add (13) to the inequality in Lemma 8, then

$$
\begin{aligned}
& \frac{1}{T} \sum_{t=0}^{T-1} (V_r^{\star}(\rho) - h(\bar{\xi}^{\star}) - (V_r^{\pi_t}(\rho) - h(\xi_t))) + \frac{1}{T} \sum_{t=0}^{T-1} \lambda^\top (-(V_g^{\pi_t}(\rho) - \xi_t)) \\
\leq\;& \frac{1}{(1-\gamma)^2 T} + \frac{1}{\eta(1-\gamma)T} + \frac{4\eta m}{(1-\gamma)^2} + \frac{\eta(L_h+1)^2 m}{(1-\gamma)^2} + \frac{1}{2\eta T} \|\lambda\|^2 + \frac{2\eta m}{(1-\gamma)^2}
\end{aligned}
$$

where the RHS can be upper bounded by, if we take $\eta = \frac{1}{\sqrt{T}}$,

$$
\frac{2 + (6 + (L_h+1)^2)m}{(1-\gamma)^2 \sqrt{T}} + \frac{1}{2\sqrt{T}} \|\lambda\|^2 .
$$

We next apply a useful property of constrained convex optimization. We notice that $V_r^{\pi_t}(\rho)$ and $V_g^{\pi_t}$ are linear in the occupancy measure induced by π_t. By the convexity of the occupancy measure set, $\frac{1}{T}\sum_{t=0}^{T-1} V_r^{\pi_t}(\rho)$ and $\frac{1}{T}\sum_{t=0}^{T-1} V_g^{\pi_t}(\rho)$ are linear in an occupancy measure induced by some policy π' and we denote them as $V_r^{\pi'}(\rho)$ and $V_g^{\pi'}(\rho)$. Since h is convex, there exists ξ' such that $\frac{1}{T}\sum_{t=0}^{T-1} h(\xi_t) \geq h(\xi')$. If we choose $\lambda_i = 2C_h$ if $V_{g_i}^{\pi_t}(\rho) \leq \xi_i$ and $\lambda_i = 0$ otherwise, then

$$\begin{aligned} & V_r^{\star}(\rho) - h(\bar{\xi}^{\star}) - (V_r^{\pi'}(\rho) - h(\xi')) + 2C_h \sum_{i=1}^{m} \left[\frac{1}{T} \sum_{t=0}^{T-1} \xi_{i,t} - V_{g_i}^{\pi'}(\rho) \right]_+ \\ \leq\ & \frac{2 + (6 + (L_h+1)^2)m}{(1-\gamma)^2\sqrt{T}} + \frac{2mC_h^2}{\sqrt{T}}. \end{aligned}$$

Due to $2C_h \geq 2\lambda^\star$ and the strong duality, application of Lemma 12 leads to

$$\sum_{i=1}^{m} \left[\frac{1}{T} \sum_{t=0}^{T-1} \xi_{i,t} - V_{g_i}^{\pi'}(\rho) \right]_+ \leq \frac{2 + (6 + (L_h+1)^2)m}{(1-\gamma)^2 C_h \sqrt{T}} + \frac{2mC_h}{\sqrt{T}}$$

which shows the second inequality by replacing $V_{g_i}^{\pi'}(\rho)$ by $\frac{1}{T}\sum_{t=0}^{T-1} V_{g_i}^{\pi_t}(\rho)$. □

C.2 Resilient Optimistic Policy Gradient Primal-Dual Method

Algorithm 2 Resilient optimistic policy gradient primal-dual (ResOPG-PD) method

```
Parameters: $\eta > 0$.
Initialization: Let $\pi_0(a\,|\,s) = \widehat{\pi}_0(a\,|\,s) = 1/A$ for $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $\xi_0 = \widehat{\xi}_0 = 0$, and $\lambda_0 = \widehat{\lambda}_0 = 0$.
for step $t = 1, \ldots, T$ do
    Primal-dual update
        $\pi_t(\cdot\,|\,s) = \underset{\pi(\cdot\,|\,s)\,\in\,\Pi}{\operatorname{argmax}} \left\{ \sum_a \pi(a\,|\,s) Q_{r+\lambda_{t-1}^\top g}^{\pi_{t-1}}(s,a) - \frac{1}{2\eta} \left\| \pi(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s) \right\|^2 \right\}$
        $\xi_t = \underset{\xi\,\in\,\Xi}{\operatorname{argmax}} \left\{ \xi^\top \left( -\nabla h(\xi_{t-1}) - \lambda_{t-1} \right) - \frac{1}{2\eta} \left\| \xi - \widehat{\xi}_t \right\|^2 \right\}$
        $\widehat{\pi}_{t+1}(\cdot\,|\,s) = \underset{\pi(\cdot\,|\,s)\,\in\,\Pi}{\operatorname{argmax}} \left\{ \sum_a \pi(a\,|\,s) Q_{r+\lambda_t^\top g}^{\pi_t}(s,a) - \frac{1}{2\eta} \left\| \pi(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s) \right\|^2 \right\}$
        $\widehat{\xi}_{t+1} = \underset{\xi\,\in\,\Xi}{\operatorname{argmax}} \left\{ \xi^\top \left( -\nabla h(\xi_t) - \lambda_t \right) - \frac{1}{2\eta} \left\| \xi - \widehat{\xi}_t \right\|^2 \right\}$        (14a)

        $\lambda_t = \underset{\lambda\,\in\,\Lambda}{\operatorname{argmax}} \left\{ \lambda^\top \left( V_g^{\pi_{t-1}}(\rho) - \xi_{t-1} \right) + \frac{1}{2\eta} \left\| \lambda - \widehat{\lambda}_t \right\|^2 \right\}$
        $\widehat{\lambda}_{t+1} = \underset{\lambda\,\in\,\Lambda}{\operatorname{argmax}} \left\{ \lambda^\top \left( V_g^{\pi_t}(\rho) - \xi_t \right) + \frac{1}{2\eta} \left\| \lambda - \widehat{\lambda}_t \right\|^2 \right\}$        (14b)
end for
```

C.3 Proof of Theorem 4

We define

$$\begin{aligned} \Theta_{t+1} \;:=\; & \frac{1}{2(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \mathcal{P}_{\Pi^\star}(\widehat{\pi}_{t+1}(\cdot\,|\,s)) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\|^2 + \frac{1}{2} \left\| \mathcal{P}_{\Xi^\star}(\widehat{\xi}_{t+1}) - \widehat{\xi}_{t+1} \right\|^2 + \frac{1}{2} \left\| \mathcal{P}_{\Lambda^\star}(\widehat{\lambda}_{t+1}) - \widehat{\lambda}_{t+1} \right\|^2 \\ & + \frac{1}{4(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t+1}(\cdot\,|\,s) - \pi_t(\cdot\,|\,s) \right\|^2 + \frac{1}{4} \left\| \widehat{\xi}_{t+1} - \xi_t \right\|^2 + \frac{1}{4} \left\| \widehat{\lambda}_{t+1} - \lambda_t \right\|^2 \end{aligned}$$

$$
\begin{aligned}
\zeta_t \;\; := \;\; & \frac{1}{2(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\|\widehat{\pi}_{t+1}(\cdot\,|\,s) - \pi_t(\cdot\,|\,s)\right\|^2 + \frac{1}{2}\left\|\widehat{\xi}_{t+1} - \xi_t\right\|^2 + \frac{1}{2}\left\|\widehat{\lambda}_{t+1} - \lambda_t\right\|^2 \\
& + \frac{1}{2(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\|\pi_t(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\right\|^2 + \frac{1}{2}\left\|\xi_t - \widehat{\xi}_t\right\|^2 + \frac{1}{2}\left\|\lambda_t - \widehat{\lambda}_t\right\|^2
\end{aligned}
$$

and

$$
\iota \; := \; \max\left(\frac{|A|}{(1-\gamma)^2} + 1, L_h^2 + 1, 2|A|^2 \frac{\gamma(1+m(C_h)^2)\kappa_\rho}{(1-\gamma)^4 \rho_{\min}} + \frac{|A| m \kappa_\rho^2}{(1-\gamma)^5} \right)
$$

$$
\eta_{\max} \; := \; \min\left(\frac{1}{4\sqrt{|A|}}, \frac{1}{2(L_h+1)}, \frac{1}{5\sqrt{m|A|}\kappa_\rho}, \frac{\rho_{\min}}{4\gamma\sqrt{m}C_h|A|}, \frac{1}{2\sqrt{2\iota}}, \frac{4\max(\frac{\kappa_\rho}{1-\gamma}, 1)}{\sqrt{\Theta_1 C_{\rho,\gamma,\sigma}(1-\gamma)}} \right).
$$

Lemma 9. *Let assumptions in Theorem 4 hold. In Algorithm 2, for (14) with* $\eta \leq \frac{1}{2\sqrt{2\iota}}$,

$$
\Theta_{t+1} \; \leq \; \Theta_t - \frac{1}{2}\zeta_t.
$$

Proof. For any $(\pi^\star, \xi^\star, \lambda^\star) \in \Pi^\star \times \Xi^\star \times \Lambda^\star$,

$$
\begin{aligned}
& V_{r+\lambda_t^\top g}^{\pi^\star}(\rho) - h(\xi^\star) - \lambda_t^\top \xi^\star - \left(V_{r+(\lambda^\star)^\top g}^{\pi_t}(\rho) - h(\xi_t) - (\lambda^\star)^\top \xi_t \right) \\
= \;\; & \underbrace{\left(V_{r+\lambda_t^\top g}^{\pi^\star}(\rho) - V_{r+\lambda_t^\top g}^{\pi_t}(\rho) \right)}_{(a)} + \underbrace{\left(-h(\xi^\star) - \lambda_t^\top \xi^\star + h(\xi_t) + \lambda_t^\top \xi_t \right)}_{(b)} \\
& + \underbrace{\left(V_{r+\lambda_t^\top g}^{\pi_t}(\rho) - V_{r+(\lambda^\star)^\top g}^{\pi_t}(\rho) - \lambda_t^\top \xi_t + (\lambda^\star)^\top \xi_t \right)}_{(c)}.
\end{aligned}
\tag{15}
$$

We next analyze three terms (a), (b), and (c), separately.

For (a), we have

$$
\begin{aligned}
& V_{r+\lambda_t^\top g}^{\pi^\star}(\rho) - V_{r+\lambda_t^\top g}^{\pi_t}(\rho) \\
= \;\; & \frac{1}{1-\gamma} \sum_{s,a} d_\rho^{\pi^\star}(s) (\pi^\star(a\,|\,s) - \pi_t(a\,|\,s)) Q_{r+\lambda_t^\top g}^{\pi_t}(s,a) \\
= \;\; & \frac{1}{1-\gamma} \sum_{s,a} d_\rho^{\pi^\star}(s) (\pi^\star(a\,|\,s) - \widehat{\pi}_{t+1}(a\,|\,s)) Q_{r+\lambda_t^\top g}^{\pi_t}(s,a) + \frac{1}{1-\gamma} \sum_{s,a} d_\rho^{\pi^\star}(s) (\widehat{\pi}_{t+1}(a\,|\,s) - \pi_t(a\,|\,s)) Q_{r+\lambda_{t-1}^\top g}^{\pi_{t-1}}(s,a) \\
& + \frac{1}{1-\gamma} \sum_{s,a} d_\rho^{\pi^\star}(s) (\widehat{\pi}_{t+1}(a\,|\,s) - \pi_t(a\,|\,s)) \big(Q_{r+\lambda_t^\top g}^{\pi_t}(s,a) - Q_{r+\lambda_{t-1}^\top g}^{\pi_{t-1}}(s,a) \big) \\
\leq \;\; & \frac{1}{2\eta(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left(\|\pi^\star(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 - \|\pi^\star(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s)\|^2 - \|\widehat{\pi}_{t+1}(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 \right) \\
& + \frac{1}{2\eta(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left(\|\widehat{\pi}_{t+1}(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 - \|\widehat{\pi}_{t+1}(\cdot\,|\,s) - \pi_t(\cdot\,|\,s)\|^2 - \|\pi_t(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 \right) \\
& + \frac{1}{1-\gamma} \sum_{s,a} d_\rho^{\pi^\star}(s) (\widehat{\pi}_{t+1}(a\,|\,s) - \pi_t(a\,|\,s)) \big(Q_{r+\lambda_t^\top g}^{\pi_t}(s,a) - Q_{r+\lambda_{t-1}^\top g}^{\pi_{t-1}}(s,a) \big) \\
\leq \;\; & \frac{1}{2\eta(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left(\|\pi^\star(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 - \|\pi^\star(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s)\|^2 - \|\widehat{\pi}_{t+1}(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 \right) \\
& + \frac{1}{2\eta(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left(\|\widehat{\pi}_{t+1}(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 - \|\widehat{\pi}_{t+1}(\cdot\,|\,s) - \pi_t(\cdot\,|\,s)\|^2 - \|\pi_t(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 \right) \\
& + \frac{4\eta|A|}{(1-\gamma)^3} \left(\left\|\lambda_t - \widehat{\lambda}_t\right\|^2 + \left\|\widehat{\lambda}_t - \lambda_{t-1}\right\|^2 \right) \\
& + 8\eta|A|^2 \frac{\gamma(1+m(C_h)^2)\kappa_\rho}{(1-\gamma)^5\rho_{\min}} \sum_s d_\rho^{\pi^\star}(s) \left(\|\pi_t(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 + \|\widehat{\pi}_t(\cdot\,|\,s) - \pi_{t-1}(\cdot\,|\,s)\|^2 \right)
\end{aligned}
$$

where the first inequality is due to the optimality of $\widehat{\pi}_{t+1}$ and π_t that results from Lemma 13, and the second inequality is due to that

$$
\begin{aligned}
& \sum_{s,a} d_\rho^{\pi^\star}(s)(\widehat{\pi}_{t+1}(a\,|\,s)-\pi_t(a\,|\,s))\big(Q^{\pi_t}_{r+\lambda_t^\top g}(s,a)-Q^{\pi_{t-1}}_{r+\lambda_{t-1}^\top g}(s,a)\big) \\
\leq\;\; & \eta\sum_s d_\rho^{\pi^\star}(s)\left\|Q^{\pi_t}_{r+\lambda_t^\top g}(s,\cdot)-Q^{\pi_{t-1}}_{r+\lambda_{t-1}^\top g}(s,\cdot)\right\|^2 \\
\leq\;\; & 2\eta\sum_s d_\rho^{\pi^\star}(s)\left\|(\lambda_t-\lambda_{t-1})^\top Q_g^{\pi_t}(s,\cdot)\right\|^2+2\eta\sum_s d_\rho^{\pi^\star}(s)\left\|Q^{\pi_t}_{r+\lambda_{t-1}^\top g}(s,\cdot)-Q^{\pi_{t-1}}_{r+\lambda_{t-1}^\top g}(s,\cdot)\right\|^2 \\
\leq\;\; & \frac{2\eta|A|}{(1-\gamma)^2}\left\|\lambda_t-\lambda_{t-1}\right\|^2+2\eta|A|\sum_s d_\rho^{\pi^\star}(s)\left\|Q^{\pi_t}_{r+\lambda_{t-1}^\top g}(s,\cdot)-Q^{\pi_{t-1}}_{r+\lambda_{t-1}^\top g}(s,\cdot)\right\|_\infty^2 \\
\leq\;\; & \frac{2\eta|A|}{(1-\gamma)^2}\left\|\lambda_t-\lambda_{t-1}\right\|^2+4\eta|A|\sum_s d_\rho^{\pi^\star}(s)\left\|Q_r^{\pi_t}(s,\cdot)-Q_r^{\pi_{t-1}}(s,\cdot)\right\|_\infty^2 \\
& +4\eta|A|\sum_s d_\rho^{\pi^\star}(s)\left\|\lambda_{t-1}\right\|^2 m\max_i\left\|Q_{g_i}^{\pi_t}(s,\cdot)-Q_{g_i}^{\pi_{t-1}}(s,\cdot)\right\|_\infty^2 \\
\leq\;\; & \frac{2\eta|A|}{(1-\gamma)^2}\left\|\lambda_t-\lambda_{t-1}\right\|^2+4\eta|A|\frac{\gamma(1+m(C_h)^2)}{(1-\gamma)^2}\max_s\left\|\pi_t(\cdot\,|\,s)-\pi_{t-1}(\cdot\,|\,s)\right\|_1^2 \\
\leq\;\; & \frac{4\eta|A|}{(1-\gamma)^2}\left(\left\|\lambda_t-\widehat{\lambda}_t\right\|^2+\left\|\widehat{\lambda}_t-\lambda_{t-1}\right\|^2\right) \\
& +8\eta|A|^2\frac{\gamma(1+m(C_h)^2)\kappa_\rho}{(1-\gamma)^4\rho_{\min}}\sum_s d_\rho^{\pi^\star}(s)\left(\left\|\pi_t(\cdot\,|\,s)-\widehat{\pi}_t(\cdot\,|\,s)\right\|^2+\left\|\widehat{\pi}_t(\cdot\,|\,s)-\pi_{t-1}(\cdot\,|\,s)\right\|^2\right)
\end{aligned}
$$

where the first inequality is due to Lemma 14, we use the inequality $\|x+y\|^2\leq 2\|x\|^2+2\|y\|^2$ in the second inequality, the third inequality is due to $\|x\|\leq\sqrt{d}\,\|x\|_\infty$ for any $x\in\mathbb{R}^d$, and the fourth inequality is due to the inequalities $\|x+y\|^2\leq 2\|x\|^2+2\|y\|^2$, $\|x\|\leq\sqrt{m}\,\|x\|_\infty$ for any $x\in\mathbb{R}^m$, and $\|\sum_i x_i\|_\infty\leq\sum_i\|x_i\|_\infty$, application of Lemma 15 leads to the fifth inequality together with $\|\lambda_{t-1}\|\leq C_h$, and the last inequality is due to $\|x\|_1\leq\sqrt{d}\,\|x\|$ for any $x\in\mathbb{R}^d$, $\|x+y\|^2\leq 2\|x\|^2+2\|y\|^2$, and the property of κ_ρ,

$$
\frac{\kappa_\rho}{1-\gamma}d_\rho^{\pi^\star}(s)\;\geq\;d_\rho^\pi(s)\;\geq\;(1-\gamma)\rho_{\min}.
$$

For (b), we have

$$
\begin{aligned}
& -h(\xi^\star)-\lambda_t^\top\xi^\star+h(\xi_t)+\lambda_t^\top\xi_t \\
\leq\;\; & (-\nabla h(\xi_t)-\lambda_t)^\top(\xi^\star-\xi_t) \\
=\;\; & (-\nabla h(\xi_t)-\lambda_t)^\top(\xi^\star-\widehat{\xi}_{t+1})+(-\nabla h(\xi_{t-1})-\lambda_{t-1})^\top(\widehat{\xi}_{t+1}-\xi_t) \\
& +(-\nabla h(\xi_t)-\lambda_t+\nabla h(\xi_{t-1})+\lambda_{t-1})^\top(\widehat{\xi}_{t+1}-\xi_t) \\
\leq\;\; & \frac{1}{2\eta}\left(\left\|\xi^\star-\widehat{\xi}_t\right\|^2-\left\|\xi^\star-\widehat{\xi}_{t+1}\right\|^2-\left\|\widehat{\xi}_{t+1}-\widehat{\xi}_t\right\|^2\right)+\frac{1}{2\eta}\left(\left\|\widehat{\xi}_{t+1}-\widehat{\xi}_t\right\|^2-\left\|\widehat{\xi}_{t+1}-\xi_t\right\|^2-\left\|\xi_t-\widehat{\xi}_t\right\|^2\right) \\
& +(-\nabla h(\xi_t)-\lambda_t+\nabla h(\xi_{t-1})+\lambda_{t-1})^\top(\widehat{\xi}_{t+1}-\xi_t) \\
\leq\;\; & \frac{1}{2\eta}\left(\left\|\xi^\star-\widehat{\xi}_t\right\|^2-\left\|\xi^\star-\widehat{\xi}_{t+1}\right\|^2-\left\|\widehat{\xi}_{t+1}-\widehat{\xi}_t\right\|^2\right)+\frac{1}{2\eta}\left(\left\|\widehat{\xi}_{t+1}-\widehat{\xi}_t\right\|^2-\left\|\widehat{\xi}_{t+1}-\xi_t\right\|^2-\left\|\xi_t-\widehat{\xi}_t\right\|^2\right) \\
& +4\eta\left(\left\|\lambda_t-\widehat{\lambda}_t\right\|^2+\left\|\widehat{\lambda}_t-\lambda_{t-1}\right\|^2\right)+4\eta L_h^2\left(\left\|\xi_t-\widehat{\xi}_t\right\|^2+\left\|\widehat{\xi}_t-\xi_{t-1}\right\|^2\right)
\end{aligned}
$$

where the first inequality is due to the convexity of h: $h(\xi^\star)\geq h(\xi_t)+\langle\nabla h(\xi_t),\xi^\star-\xi_t\rangle$, the second inequality is due to the optimality of $\widehat{\xi}_{t+1}$ and ξ_t that results from Lemma 13, and the last inequality is due to that

$$
\begin{aligned}
& (-\nabla h(\xi_t)-\lambda_t+\nabla h(\xi_{t-1})+\lambda_{t-1})^\top(\widehat{\xi}_{t+1}-\xi_t) \\
\leq\;\; & \eta\left\|-\nabla h(\xi_t)-\lambda_t+\nabla h(\xi_{t-1})+\lambda_{t-1}\right\|^2 \\
\leq\;\; & 2\eta\left\|\lambda_t-\lambda_{t-1}\right\|^2+2\eta\left\|-\nabla h(\xi_t)+\nabla h(\xi_{t-1})\right\|^2 \\
\leq\;\; & 4\eta\left(\left\|\lambda_t-\widehat{\lambda}_t\right\|^2+\left\|\widehat{\lambda}_t-\lambda_{t-1}\right\|^2\right)+4\eta L_h^2\left(\left\|\xi_t-\widehat{\xi}_t\right\|^2+\left\|\widehat{\xi}_t-\xi_{t-1}\right\|^2\right)
\end{aligned}
$$

where the first inequality is due to Lemma 14, the second inequality is due to $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$, the third inequality is due to the Lipschitz continuous gradient of h: $\|\nabla h(\xi) - \nabla h(\xi')\| \leq L_h \|\xi - \xi'\|$ for any ξ, $\xi' \in \Xi$, and $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$.

For (c), we have

$$
\begin{aligned}
& V_{r+\lambda_t^\top g}^{\pi_t}(\rho) - V_{r+(\lambda^\star)^\top g}^{\pi_t}(\rho) - \lambda_t^\top \xi_t + (\lambda^\star)^\top \xi_t \\
= \;\; & (\lambda_t - \lambda^\star)^\top \left(V_g^{\pi_t}(\rho) - \xi_t\right) \\
= \;\; & (\lambda_t - \widehat{\lambda}_{t+1})^\top \left(V_g^{\pi_{t-1}}(\rho) - \xi_{t-1}\right) + (\lambda_t - \widehat{\lambda}_{t+1})^\top \left(V_g^{\pi_t}(\rho) - \xi_t - V_g^{\pi_{t-1}}(\rho) + \xi_{t-1}\right) \\
& + (\widehat{\lambda}_{t+1} - \lambda^\star)^\top \left(V_g^{\pi_t}(\rho) - \xi_t\right) \\
\leq \;\; & \frac{1}{2\eta}\left(\left\|\widehat{\lambda}_{t+1} - \widehat{\lambda}_t\right\|^2 - \left\|\widehat{\lambda}_{t+1} - \lambda_t\right\|^2 - \left\|\lambda_t - \widehat{\lambda}_t\right\|^2\right) \\
& + (\lambda_t - \widehat{\lambda}_{t+1})^\top \left(V_g^{\pi_t}(\rho) - \xi_t - V_g^{\pi_{t-1}}(\rho) + \xi_{t-1}\right) \\
& + \frac{1}{2\eta}\left(\left\|\lambda^\star - \widehat{\lambda}_t\right\|^2 - \left\|\lambda^\star - \widehat{\lambda}_{t+1}\right\|^2 - \left\|\widehat{\lambda}_{t+1} - \widehat{\lambda}_t\right\|^2\right) \\
\leq \;\; & \frac{1}{2\eta}\left(\left\|\widehat{\lambda}_{t+1} - \widehat{\lambda}_t\right\|^2 - \left\|\widehat{\lambda}_{t+1} - \lambda_t\right\|^2 - \left\|\lambda_t - \widehat{\lambda}_t\right\|^2\right) + \frac{1}{2\eta}\left(\left\|\lambda^\star - \widehat{\lambda}_t\right\|^2 - \left\|\lambda^\star - \widehat{\lambda}_{t+1}\right\|^2 - \left\|\widehat{\lambda}_{t+1} - \widehat{\lambda}_t\right\|^2\right) \\
& + \frac{4\eta|A|m\kappa_\rho^2}{(1-\gamma)^6}\sum_s d_\rho^{\pi^\star}(s)\left(\|\pi_t(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 + \|\widehat{\pi}_t(\cdot\,|\,s) - \pi_{t-1}(\cdot\,|\,s)\|^2\right) \\
& + 4\eta\left(\left\|\xi_t - \widehat{\xi}_t\right\|^2 + \left\|\widehat{\xi}_t - \xi_{t-1}\right\|^2\right)
\end{aligned}
$$

where the first inequality is due to the optimality of λ_t and $\widehat{\lambda}_{t+1}$, and the the second inequality is due to that

$$
\begin{aligned}
& (\lambda_t - \widehat{\lambda}_{t+1})^\top \left(V_g^{\pi_t}(\rho) - \xi_t - V_g^{\pi_{t-1}}(\rho) + \xi_{t-1}\right) \\
\leq \;\; & \eta\left\|V_g^{\pi_t}(\rho) - \xi_t - V_g^{\pi_{t-1}}(\rho) + \xi_{t-1}\right\|^2 \\
\leq \;\; & 2\eta m\left(\frac{\kappa_\rho}{(1-\gamma)^3}\sum_s d_\rho^{\pi^\star}(s)\,\|\pi_t(\cdot\,|\,s) - \pi_{t-1}(\cdot\,|\,s)\|_1\right)^2 + 2\eta\,\|\xi_t - \xi_{t-1}\|^2 \\
\leq \;\; & \frac{2\eta m\kappa_\rho^2}{(1-\gamma)^6}\sum_s\left(\sqrt{d_\rho^{\pi^\star}(s)}\,\|\pi_t(\cdot\,|\,s) - \pi_{t-1}(\cdot\,|\,s)\|_1\right)^2 + 2\eta\,\|\xi_t - \xi_{t-1}\|^2 \\
\leq \;\; & \frac{2\eta|A|m\kappa_\rho^2}{(1-\gamma)^6}\sum_s d_\rho^{\pi^\star}(s)\,\|\pi_t(\cdot\,|\,s) - \pi_{t-1}(\cdot\,|\,s)\|^2 + 2\eta\,\|\xi_t - \xi_{t-1}\|^2 \\
\leq \;\; & \frac{4\eta|A|m\kappa_\rho^2}{(1-\gamma)^6}\sum_s d_\rho^{\pi^\star}(s)\left(\|\pi_t(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 + \|\widehat{\pi}_t(\cdot\,|\,s) - \pi_{t-1}(\cdot\,|\,s)\|^2\right) \\
& + 4\eta\left(\left\|\xi_t - \widehat{\xi}_t\right\|^2 + \left\|\widehat{\xi}_t - \xi_{t-1}\right\|^2\right)
\end{aligned}
$$

where the first inequality is due to Lemma 14, the second inequality is due to $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$, $\|x\| \leq \|x\|_1$, and Lemma 15, the third inequality is due to Cauchy-Schwarz inequality, and the last inequality is due to $\|x\|_1 \leq \sqrt{d}\,\|x\|$ for any $x \in \mathbb{R}^d$ and $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$.

By applying the above upper bounds on (a), (b), and (c) to (15), it is ready to have

$$
\begin{aligned}
& V_{r+\lambda_t^\top g}^{\pi^\star}(\rho) - h(\xi^\star) - \lambda_t^\top \xi^\star - \left(V_{r+(\lambda^\star)^\top g}^{\pi_t}(\rho) - h(\xi_t) - (\lambda^\star)^\top \xi_t \right) \\
\leq \quad & \frac{1}{2\eta(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left(\|\pi^\star(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 - \|\pi^\star(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s)\|^2 - \|\widehat{\pi}_{t+1}(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 \right) \\
& + \frac{1}{2\eta(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left(\|\widehat{\pi}_{t+1}(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 - \|\widehat{\pi}_{t+1}(\cdot\,|\,s) - \pi_t(\cdot\,|\,s)\|^2 - \|\pi_t(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 \right) \\
& + \frac{4\eta|A|}{(1-\gamma)^3} \left(\left\|\lambda_t - \widehat{\lambda}_t\right\|^2 + \left\|\widehat{\lambda}_t - \lambda_{t-1}\right\|^2 \right) \\
& + 8\eta|A|^2 \frac{\gamma(1+m(C_h)^2)\kappa_\rho}{(1-\gamma)^5 \rho_{\min}} \sum_s d_\rho^{\pi^\star}(s) \left(\|\pi_t(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 + \|\widehat{\pi}_t(\cdot\,|\,s) - \pi_{t-1}(\cdot\,|\,s)\|^2 \right) \\
& + \frac{1}{2\eta} \left(\left\|\xi^\star - \widehat{\xi}_t\right\|^2 - \left\|\xi^\star - \widehat{\xi}_{t+1}\right\|^2 - \left\|\widehat{\xi}_{t+1} - \widehat{\xi}_t\right\|^2 \right) + \frac{1}{2\eta} \left(\left\|\widehat{\xi}_{t+1} - \widehat{\xi}_t\right\|^2 - \left\|\widehat{\xi}_{t+1} - \xi_t\right\|^2 - \left\|\xi_t - \widehat{\xi}_t\right\|^2 \right) \\
& + 4\eta \left(\left\|\lambda_t - \widehat{\lambda}_t\right\|^2 + \left\|\widehat{\lambda}_t - \lambda_{t-1}\right\|^2 \right) + 4\eta L_h^2 \left(\left\|\xi_t - \widehat{\xi}_t\right\|^2 + \left\|\widehat{\xi}_t - \xi_{t-1}\right\|^2 \right) \\
& + \frac{1}{2\eta} \left(\left\|\widehat{\lambda}_{t+1} - \widehat{\lambda}_t\right\|^2 - \left\|\widehat{\lambda}_{t+1} - \lambda_t\right\|^2 - \left\|\lambda_t - \widehat{\lambda}_t\right\|^2 \right) + \frac{1}{2\eta} \left(\left\|\lambda^\star - \widehat{\lambda}_t\right\|^2 - \left\|\lambda^\star - \widehat{\lambda}_{t+1}\right\|^2 - \left\|\widehat{\lambda}_{t+1} - \widehat{\lambda}_t\right\|^2 \right) \\
& + \frac{4\eta|A|m\kappa_\rho^2}{(1-\gamma)^6} \sum_s d_\rho^{\pi^\star}(s) \left(\|\pi_t(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 + \|\widehat{\pi}_t(\cdot\,|\,s) - \pi_{t-1}(\cdot\,|\,s)\|^2 \right) \\
& + 4\eta \left(\left\|\xi_t - \widehat{\xi}_t\right\|^2 + \left\|\widehat{\xi}_t - \xi_{t-1}\right\|^2 \right).
\end{aligned}
$$

We notice that $V_{r+\lambda_t^\top g}^{\pi^\star}(\rho) - h(\xi^\star) - \lambda_t^\top \xi^\star - \left(V_{r+(\lambda^\star)^\top g}^{\pi_t}(\rho) - h(\xi_t) - (\lambda^\star)^\top \xi_t \right) \geq 0$. Thus,

$$
\begin{aligned}
& \frac{1}{2(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \, \|\pi^\star(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s)\|^2 + \frac{1}{2} \left\|\xi^\star - \widehat{\xi}_{t+1}\right\|^2 + \frac{1}{2} \left\|\lambda^\star - \widehat{\lambda}_{t+1}\right\|^2 \\
\leq \quad & \frac{1}{2(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \, \|\pi^\star(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 + \frac{1}{2} \left\|\xi^\star - \widehat{\xi}_t\right\|^2 + \frac{1}{2} \left\|\lambda^\star - \widehat{\lambda}_t\right\|^2 \\
& - \frac{1}{2(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \, \|\widehat{\pi}_{t+1}(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 - \frac{1}{2} \left\|\widehat{\xi}_{t+1} - \widehat{\xi}_t\right\|^2 - \frac{1}{2} \left\|\widehat{\lambda}_{t+1} - \widehat{\lambda}_t\right\|^2 \\
& + \frac{1}{2(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left(\|\widehat{\pi}_{t+1}(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 - \|\widehat{\pi}_{t+1}(\cdot\,|\,s) - \pi_t(\cdot\,|\,s)\|^2 - \|\pi_t(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 \right) \\
& + \frac{1}{2} \left(\left\|\widehat{\xi}_{t+1} - \widehat{\xi}_t\right\|^2 - \left\|\widehat{\xi}_{t+1} - \xi_t\right\|^2 - \left\|\xi_t - \widehat{\xi}_t\right\|^2 \right) \\
& + \frac{1}{2} \left(\left\|\widehat{\lambda}_{t+1} - \widehat{\lambda}_t\right\|^2 - \left\|\widehat{\lambda}_{t+1} - \lambda_t\right\|^2 - \left\|\lambda_t - \widehat{\lambda}_t\right\|^2 \right) \\
& + 4\eta^2 \iota \frac{1}{1-\gamma} \sum_s d_\rho^{\pi^\star}(s) \left(\|\pi_t(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\|^2 + \|\widehat{\pi}_t(\cdot\,|\,s) - \pi_{t-1}(\cdot\,|\,s)\|^2 \right) \\
& + 4\eta^2 \iota \left(\left\|\xi_t - \widehat{\xi}_t\right\|^2 + \left\|\widehat{\xi}_t - \xi_{t-1}\right\|^2 \right) \\
& + 4\eta^2 \iota \left(\left\|\lambda_t - \widehat{\lambda}_t\right\|^2 + \left\|\widehat{\lambda}_t - \lambda_{t-1}\right\|^2 \right)
\end{aligned}
$$

where we use the defintion of ι. After some re-combination, we have

$$
\begin{aligned}
& \frac{1}{2(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \pi^\star(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\|^2 + \frac{1}{2} \left\| \xi^\star - \widehat{\xi}_{t+1} \right\|^2 + \frac{1}{2} \left\| \lambda^\star - \widehat{\lambda}_{t+1} \right\|^2 \\
\leq \;\; & \frac{1}{2(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \pi^\star(\cdot\,|\,s) - \widehat{\pi}_{t}(\cdot\,|\,s) \right\|^2 + \frac{1}{2} \left\| \xi^\star - \widehat{\xi}_{t} \right\|^2 + \frac{1}{2} \left\| \lambda^\star - \widehat{\lambda}_{t} \right\|^2 \\
& - \frac{1}{2(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t+1}(\cdot\,|\,s) - \pi_{t}(\cdot\,|\,s) \right\|^2 - \frac{1}{2} \left\| \widehat{\xi}_{t+1} - \xi_t \right\|^2 - \frac{1}{2} \left\| \widehat{\lambda}_{t+1} - \lambda_t \right\|^2 \\
& - \left(\frac{1}{2} - 4\eta^2 \iota \right) \left(\frac{1}{1-\gamma} \sum_s d_\rho^{\pi^\star}(s) \left\| \pi_{t}(\cdot\,|\,s) - \widehat{\pi}_{t}(\cdot\,|\,s) \right\|^2 + \left\| \xi_t - \widehat{\xi}_t \right\|^2 + \left\| \lambda_t - \widehat{\lambda}_t \right\|^2 \right) \\
& + 4\eta^2 \iota \frac{1}{1-\gamma} \sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t}(\cdot\,|\,s) - \pi_{t-1}(\cdot\,|\,s) \right\|^2 + 4\eta^2\iota \left\| \widehat{\xi}_t - \xi_{t-1} \right\|^2 + 4\eta^2\iota \left\| \widehat{\lambda}_t - \lambda_{t-1} \right\|^2.
\end{aligned}
$$

By taking

$$
\pi^\star(\cdot\,|\,s) \; = \; \mathcal{P}_{\Pi^\star}(\widehat{\pi}_t(\cdot\,|\,s)), \; \xi^\star \; = \; \mathcal{P}_{\Xi^\star}(\widehat{\xi}_t), \;\; \text{and} \;\; \lambda^\star \; = \; \mathcal{P}_{\Lambda^\star}(\widehat{\lambda}_t)
$$

and using the non-expansivenss of projection operators $\mathcal{P}_{\Pi^\star}$, $\mathcal{P}_{\Xi^\star}$, and $\mathcal{P}_{\Lambda^\star}$, we obtain the following inequality,

$$
\begin{aligned}
& \frac{1}{2(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \mathcal{P}_{\Pi^\star}(\widehat{\pi}_{t+1}(\cdot\,|\,s)) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\|^2 + \frac{1}{2} \left\| \mathcal{P}_{\Xi^\star}(\widehat{\xi}_{t+1}) - \widehat{\xi}_{t+1} \right\|^2 + \frac{1}{2} \left\| \mathcal{P}_{\Lambda^\star}(\widehat{\lambda}_{t+1}) - \widehat{\lambda}_{t+1} \right\|^2 \\
\leq \;\; & \frac{1}{2(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \mathcal{P}_{\Pi^\star}(\widehat{\pi}_{t}(\cdot\,|\,s)) - \widehat{\pi}_{t}(\cdot\,|\,s) \right\|^2 + \frac{1}{2} \left\| \mathcal{P}_{\Xi^\star}(\widehat{\xi}_{t}) - \widehat{\xi}_{t} \right\|^2 + \frac{1}{2} \left\| \mathcal{P}_{\Lambda^\star}(\widehat{\lambda}_{t}) - \widehat{\lambda}_{t} \right\|^2 \\
& - \frac{1}{2(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t+1}(\cdot\,|\,s) - \pi_{t}(\cdot\,|\,s) \right\|^2 - \frac{1}{2} \left\| \widehat{\xi}_{t+1} - \xi_t \right\|^2 - \frac{1}{2} \left\| \widehat{\lambda}_{t+1} - \lambda_t \right\|^2 \\
& - \left(\frac{1}{2} - 4\eta^2 \iota \right) \left(\frac{1}{1-\gamma} \sum_s d_\rho^{\pi^\star}(s) \left\| \pi_{t}(\cdot\,|\,s) - \widehat{\pi}_{t}(\cdot\,|\,s) \right\|^2 + \left\| \xi_t - \widehat{\xi}_t \right\|^2 + \left\| \lambda_t - \widehat{\lambda}_t \right\|^2 \right) \\
& + 4\eta^2 \iota \frac{1}{1-\gamma} \sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t}(\cdot\,|\,s) - \pi_{t-1}(\cdot\,|\,s) \right\|^2 + 4\eta^2\iota \left\| \widehat{\xi}_t - \xi_{t-1} \right\|^2 + 4\eta^2\iota \left\| \widehat{\lambda}_t - \lambda_{t-1} \right\|^2.
\end{aligned}
$$

If we choose $\eta > 0$ such that $\frac{1}{2} - 4\eta^2\iota \geq \frac{1}{4}$ and do some re-arrangement, we have

$$
\begin{aligned}
& \frac{1}{2(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \mathcal{P}_{\Pi^\star}(\widehat{\pi}_{t+1}(\cdot\,|\,s)) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\|^2 + \frac{1}{2} \left\| \mathcal{P}_{\Xi^\star}(\widehat{\xi}_{t+1}) - \widehat{\xi}_{t+1} \right\|^2 + \frac{1}{2} \left\| \mathcal{P}_{\Lambda^\star}(\widehat{\lambda}_{t+1}) - \widehat{\lambda}_{t+1} \right\|^2 \\
& + \frac{1}{4(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t+1}(\cdot\,|\,s) - \pi_{t}(\cdot\,|\,s) \right\|^2 + \frac{1}{4} \left\| \widehat{\xi}_{t+1} - \xi_t \right\|^2 + \frac{1}{4} \left\| \widehat{\lambda}_{t+1} - \lambda_t \right\|^2 \\
\leq \;\; & \frac{1}{2(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \mathcal{P}_{\Pi^\star}(\widehat{\pi}_{t}(\cdot\,|\,s)) - \widehat{\pi}_{t}(\cdot\,|\,s) \right\|^2 + \frac{1}{2} \left\| \mathcal{P}_{\Xi^\star}(\widehat{\xi}_{t}) - \widehat{\xi}_{t} \right\|^2 + \frac{1}{2} \left\| \mathcal{P}_{\Lambda^\star}(\widehat{\lambda}_{t}) - \widehat{\lambda}_{t} \right\|^2 \\
& - \frac{1}{4(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t+1}(\cdot\,|\,s) - \pi_{t}(\cdot\,|\,s) \right\|^2 - \frac{1}{4} \left\| \widehat{\xi}_{t+1} - \xi_t \right\|^2 - \frac{1}{4} \left\| \widehat{\lambda}_{t+1} - \lambda_t \right\|^2 \\
& - \frac{1}{4(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \pi_{t}(\cdot\,|\,s) - \widehat{\pi}_{t}(\cdot\,|\,s) \right\|^2 - \frac{1}{4} \left\| \xi_t - \widehat{\xi}_t \right\|^2 - \frac{1}{4} \left\| \lambda_t - \widehat{\lambda}_t \right\|^2 \\
& + \frac{1}{4(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t}(\cdot\,|\,s) - \pi_{t-1}(\cdot\,|\,s) \right\|^2 + \frac{1}{4} \left\| \widehat{\xi}_t - \xi_{t-1} \right\|^2 + \frac{1}{4} \left\| \widehat{\lambda}_t - \lambda_{t-1} \right\|^2.
\end{aligned}
$$

Finally, our desired inequality is obtained by using notation Θ and ζ. □

Lemma 10. *In Algorithm 2, for* (14) *with* $\eta \le \min\left(\frac{1}{4\sqrt{|A|}}, \frac{1}{2(L_h+1)}, \frac{1}{5\sqrt{m|A|}\kappa_\rho}, \frac{\rho_{\min}}{4\gamma\sqrt{m}C_h|A|}\right)$,

$$\begin{aligned} & \frac{1}{1-\gamma}\sum_s d_\rho^{\pi^\star}(s)\left\|\widehat{\pi}_{t+1}(\cdot\,|\,s) - \pi_t(\cdot\,|\,s)\right\|^2 + \left\|\widehat{\xi}_{t+1} - \xi_t\right\|^2 + \left\|\widehat{\lambda}_{t+1} - \lambda_t\right\|^2 \\ & + \frac{1}{1-\gamma}\sum_s d_\rho^{\pi^\star}(s)\left\|\pi_t(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\right\|^2 + \left\|\xi_t - \widehat{\xi}_t\right\|^2 + \left\|\lambda_t - \widehat{\lambda}_t\right\|^2 \\ \geq \quad & \frac{\eta^2}{9\max\left(\frac{\kappa_\rho}{1-\gamma}, 1\right)^2} \frac{\left[\left(V^{\pi}_{r+\widehat{\lambda}_{t+1}^\top g}(\rho) - h(\xi) - \widehat{\lambda}_{t+1}^\top \xi\right) - \left(V^{\widehat{\pi}_{t+1}}_{r+\lambda^\top g}(\rho) - h(\widehat{\xi}_{t+1}) - \lambda^\top \widehat{\xi}_{t+1}\right)\right]_+^2}{\left(\max_s \left\|\pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s))\right\| + \left\|\xi - \widehat{\xi}_{t+1}\right\| + \left\|\lambda - \widehat{\lambda}_{t+1}\right\|\right)^2} \end{aligned}$$

for any $(\pi, \xi, \lambda) \neq (\widehat{\pi}_{t+1}, \widehat{\xi}_{t+1}, \widehat{\lambda}_{t+1})$.

Proof. By the optimality of $\widehat{\pi}_{t+1}$,

$$\left\langle Q^{\pi_t}_{r+\lambda_t^\top g}(s,\cdot) - \frac{1}{\eta}\left(\widehat{\pi}_{t+1}(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\right), \widehat{\pi}_{t+1}(\cdot\,|\,s) - \pi(\cdot\,|\,s)\right\rangle \;\geq\; 0 \;\text{ for all } \pi \in \Pi.$$

Hence,

$$\begin{aligned} & \langle \widehat{\pi}_{t+1}(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s), \pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s)\rangle \\ \geq \quad & \eta\left\langle Q^{\pi_t}_{r+\lambda_t^\top g}(s,\cdot), \pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s)\right\rangle \\ = \quad & \eta\left\langle Q^{\pi_{t+1}}_{r+\widehat{\lambda}_{t+1}^\top g}(s,\cdot), \pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s)\right\rangle + \eta\left\langle Q^{\pi_t}_{r+\lambda_t^\top g}(s,\cdot) - Q^{\widehat{\pi}_{t+1}}_{r+\lambda_t^\top g}(s,\cdot), \pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s)\right\rangle \\ & + \eta\left\langle Q^{\widehat{\pi}_{t+1}}_{r+\lambda_t^\top g}(s,\cdot) - Q^{\widehat{\pi}_{t+1}}_{r+\widehat{\lambda}_{t+1}^\top g}(s,\cdot), \pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s)\right\rangle. \end{aligned} \tag{16}$$

Similarly, the optimality of π_{t+1},

$$\left\langle Q^{\pi_t}_{r+\lambda_t^\top g}(s,\cdot) - \frac{1}{\eta}\left(\pi_{t+1}(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s)\right), \pi_{t+1}(\cdot\,|\,s) - \pi(\cdot\,|\,s)\right\rangle \;\geq\; 0 \;\text{ for all } \pi \in \Pi$$

implies that

$$\begin{aligned} & \langle \pi_{t+1}(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s), \pi(\cdot\,|\,s) - \pi_{t+1}(\cdot\,|\,s)\rangle \\ \geq \quad & \eta\left\langle Q^{\pi_t}_{r+\lambda_t^\top g}(s,\cdot), \pi(\cdot\,|\,s) - \pi_{t+1}(\cdot\,|\,s)\right\rangle \\ = \quad & \eta\left\langle Q^{\pi_{t+1}}_{r+\lambda_{t+1}^\top g}(s,\cdot), \pi(\cdot\,|\,s) - \pi_{t+1}(\cdot\,|\,s)\right\rangle + \eta\left\langle Q^{\pi_t}_{r+\lambda_t^\top g}(s,\cdot) - Q^{\pi_{t+1}}_{r+\lambda_t^\top g}(s,\cdot), \pi(\cdot\,|\,s) - \pi_{t+1}(\cdot\,|\,s)\right\rangle \\ & + \eta\left\langle Q^{\pi_{t+1}}_{r+\lambda_t^\top g}(s,\cdot) - Q^{\pi_{t+1}}_{r+\lambda_{t+1}^\top g}(s,\cdot), \pi(\cdot\,|\,s) - \pi_{t+1}(\cdot\,|\,s)\right\rangle. \end{aligned}$$

By the optimality of $\widehat{\xi}_{t+1}$,

$$\left\langle -\nabla h(\xi_t) - \lambda_t - \frac{1}{\eta}\left(\widehat{\xi}_{t+1} - \widehat{\xi}_t\right), \widehat{\xi}_{t+1} - \xi\right\rangle \;\geq\; 0 \;\text{ for all } \xi \in \Xi.$$

Hence,

$$\begin{aligned} & \left\langle \widehat{\xi}_{t+1} - \widehat{\xi}_t, \xi - \widehat{\xi}_{t+1}\right\rangle \\ \geq \quad & \eta\left\langle -\nabla h(\xi_t) - \lambda_t, \xi - \widehat{\xi}_{t+1}\right\rangle \\ = \quad & \eta\left\langle -\nabla h(\widehat{\xi}_{t+1}) - \widehat{\lambda}_{t+1}, \xi - \widehat{\xi}_{t+1}\right\rangle + \eta\left\langle -\nabla h(\xi_t) - \lambda_t + \nabla h(\widehat{\xi}_{t+1}) + \lambda_t, \xi - \widehat{\xi}_{t+1}\right\rangle \\ & + \eta\left\langle -\nabla h(\widehat{\xi}_{t+1}) - \lambda_t + \nabla h(\widehat{\xi}_{t+1}) + \widehat{\lambda}_{t+1}, \xi - \widehat{\xi}_{t+1}\right\rangle. \end{aligned} \tag{17}$$

Similarly, the optimality of ξ_{t+1},

$$\left\langle -\nabla h(\xi_t) - \lambda_t - \frac{1}{\eta}\left(\xi_{t+1} - \widehat{\xi}_{t+1}\right), \xi_{t+1} - \xi \right\rangle \;\geq\; 0 \;\text{ for all } \xi \in \Xi$$

implies that

$$\begin{aligned}
& \left\langle \xi_{t+1} - \widehat{\xi}_{t+1}, \xi - \xi_{t+1} \right\rangle \\
\geq \;\; & \eta \left\langle -\nabla h(\xi_t) - \lambda_t, \xi - \xi_{t+1} \right\rangle \\
= \;\; & \eta \left\langle -\nabla h(\xi_{t+1}) - \lambda_{t+1}, \xi - \xi_{t+1} \right\rangle + \eta \left\langle -\nabla h(\xi_t) - \lambda_t + \nabla h(\xi_{t+1}) + \lambda_t, \xi - \xi_{t+1} \right\rangle \\
& + \eta \left\langle -\nabla h(\xi_{t+1}) - \lambda_t + \nabla h(\xi_{t+1}) + \lambda_{t+1}, \xi - \xi_{t+1} \right\rangle.
\end{aligned}$$

By the optimality of $\widehat{\lambda}_{t+1}$,

$$\left\langle V_g^{\pi_t}(\rho) - \xi_t + \frac{1}{\eta}\left(\widehat{\lambda}_{t+1} - \widehat{\lambda}_t\right), \widehat{\lambda}_{t+1} - \lambda \right\rangle \;\leq\; 0 \;\text{ for all } \lambda \in \Lambda.$$

Hence,

$$\begin{aligned}
& \left\langle \widehat{\lambda}_{t+1} - \widehat{\lambda}_t, \lambda - \widehat{\lambda}_{t+1} \right\rangle \\
\geq \;\; & \eta \left\langle V_g^{\pi_t}(\rho) - \xi_t, \widehat{\lambda}_{t+1} - \lambda \right\rangle \\
= \;\; & \eta \left\langle V_g^{\widehat{\pi}_{t+1}}(\rho) - \widehat{\xi}_{t+1}, \widehat{\lambda}_{t+1} - \lambda \right\rangle + \eta \left\langle V_g^{\pi_t}(\rho) - \xi_t - V_g^{\widehat{\pi}_{t+1}}(\rho) + \widehat{\xi}_{t+1}, \widehat{\lambda}_{t+1} - \lambda \right\rangle.
\end{aligned} \tag{18}$$

Similarly, the optimality of λ_{t+1},

$$\left\langle V_g^{\pi_t}(\rho) - \xi_t + \frac{1}{\eta}\left(\lambda_{t+1} - \widehat{\lambda}_{t+1}\right), \lambda_{t+1} - \lambda \right\rangle \;\leq\; 0 \;\text{ for all } \lambda \in \Lambda$$

implies that

$$\begin{aligned}
& \left\langle \lambda_{t+1} - \widehat{\lambda}_{t+1}, \lambda - \lambda_{t+1} \right\rangle \\
\geq \;\; & \eta \left\langle V_g^{\pi_t}(\rho) - \xi_t, \lambda_{t+1} - \lambda \right\rangle \\
= \;\; & \eta \left\langle V_g^{\pi_{t+1}}(\rho) - \xi_{t+1}, \lambda_{t+1} - \lambda \right\rangle + \eta \left\langle V_g^{\pi_t}(\rho) - \xi_t - V_g^{\pi_{t+1}}(\rho) + \xi_{t+1}, \lambda_{t+1} - \lambda \right\rangle.
\end{aligned}$$

Summing up the inequalities (16), (17), and (18) from both sides, with some state distribution d_ρ^π, yields,

$$
\begin{aligned}
& \frac{1}{1-\gamma} \sum_s d_\rho^\pi(s) \left\langle \widehat{\pi}_{t+1}(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s), \pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\rangle \\
& + \left\langle \widehat{\xi}_{t+1} - \widehat{\xi}_t, \xi - \widehat{\xi}_{t+1} \right\rangle + \left\langle \widehat{\lambda}_{t+1} - \widehat{\lambda}_t, \lambda - \widehat{\lambda}_{t+1} \right\rangle \\
\geq \quad & \frac{\eta}{1-\gamma} \sum_s d_\rho^\pi(s) \left\langle Q_{r+\widehat{\lambda}_{t+1}^\top g}^{\pi_{t+1}}(s,\cdot), \pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\rangle \\
& + \frac{\eta}{1-\gamma} \sum_s d_\rho^\pi(s) \left\langle Q_{r+\lambda_t^\top g}^{\pi_t}(s,\cdot) - Q_{r+\lambda_t^\top g}^{\widehat{\pi}_{t+1}}(s,\cdot), \pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\rangle \\
& + \frac{\eta}{1-\gamma} \sum_s d_\rho^\pi(s) \left\langle Q_{r+\lambda_t^\top g}^{\widehat{\pi}_{t+1}}(s,\cdot) - Q_{r+\widehat{\lambda}_{t+1}^\top g}^{\widehat{\pi}_{t+1}}(s,\cdot), \pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\rangle \\
& + \eta \left\langle -\nabla h(\widehat{\xi}_{t+1}) - \widehat{\lambda}_{t+1}, \xi - \widehat{\xi}_{t+1} \right\rangle + \eta \left\langle -\nabla h(\xi_t) - \lambda_t + \nabla h(\widehat{\xi}_{t+1}) + \lambda_t, \xi - \widehat{\xi}_{t+1} \right\rangle \\
& + \eta \left\langle -\nabla h(\widehat{\xi}_{t+1}) - \lambda_t + \nabla h(\widehat{\xi}_{t+1}) + \widehat{\lambda}_{t+1}, \xi - \widehat{\xi}_{t+1} \right\rangle \\
& + \eta \left\langle V_g^{\widehat{\pi}_{t+1}}(\rho) - \widehat{\xi}_{t+1}, \widehat{\lambda}_{t+1} - \lambda \right\rangle + \eta \left\langle V_g^{\pi_t}(\rho) - \xi_t - V_g^{\widehat{\pi}_{t+1}}(\rho) + \widehat{\xi}_{t+1}, \widehat{\lambda}_{t+1} - \lambda \right\rangle \\
\geq \quad & \eta \left(V_{r+\widehat{\lambda}_{t+1}^\top g}^{\pi}(\rho) - V_{r+\lambda^\top g}^{\widehat{\pi}_{t+1}}(\rho) \right) + \eta \left(h(\widehat{\xi}_{t+1}) - h(\xi) \right) + \eta \left\langle -\widehat{\lambda}_{t+1}, \xi - \widehat{\xi}_{t+1} \right\rangle + \eta \left\langle -\widehat{\xi}_{t+1}, \widehat{\lambda}_{t+1} - \lambda \right\rangle \\
& + \frac{\eta}{1-\gamma} \sum_s d_\rho^\pi(s) \left\langle Q_{r+\lambda_t^\top g}^{\pi_t}(s,\cdot) - Q_{r+\lambda_t^\top g}^{\widehat{\pi}_{t+1}}(s,\cdot), \pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\rangle \\
& + \frac{\eta}{1-\gamma} \sum_s d_\rho^\pi(s) \left\langle Q_{r+\lambda_t^\top g}^{\widehat{\pi}_{t+1}}(s,\cdot) - Q_{r+\widehat{\lambda}_{t+1}^\top g}^{\widehat{\pi}_{t+1}}(s,\cdot), \pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\rangle \\
& + \eta \left\langle -\nabla h(\xi_t) - \lambda_t + \nabla h(\widehat{\xi}_{t+1}) + \lambda_t, \xi - \widehat{\xi}_{t+1} \right\rangle \\
& + \eta \left\langle -\nabla h(\widehat{\xi}_{t+1}) - \lambda_t + \nabla h(\widehat{\xi}_{t+1}) + \widehat{\lambda}_{t+1}, \xi - \widehat{\xi}_{t+1} \right\rangle \\
& + \eta \left\langle V_g^{\pi_t}(\rho) - \xi_t - V_g^{\widehat{\pi}_{t+1}}(\rho) + \widehat{\xi}_{t+1}, \widehat{\lambda}_{t+1} - \lambda \right\rangle \\
\geq \quad & \eta \left(V_{r+\widehat{\lambda}_{t+1}^\top g}^{\pi}(\rho) - V_{r+\lambda^\top g}^{\widehat{\pi}_{t+1}}(\rho) \right) + \eta \left(h(\widehat{\xi}_{t+1}) - h(\xi) \right) + \eta \left\langle -\widehat{\lambda}_{t+1}, \xi - \widehat{\xi}_{t+1} \right\rangle + \eta \left\langle -\widehat{\xi}_{t+1}, \widehat{\lambda}_{t+1} - \lambda \right\rangle \\
& - \frac{\eta}{1-\gamma} \sum_s d_\rho^\pi(s) \left\| Q_{r+\lambda_t^\top g}^{\pi_t}(s,\cdot) - Q_{r+\lambda_t^\top g}^{\widehat{\pi}_{t+1}}(s,\cdot) \right\|_\infty \left\| \pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\|_1 \\
& - \frac{\eta}{1-\gamma} \sum_s d_\rho^\pi(s) \left\| Q_{r+\lambda_t^\top g}^{\widehat{\pi}_{t+1}}(s,\cdot) - Q_{r+\widehat{\lambda}_{t+1}^\top g}^{\widehat{\pi}_{t+1}}(s,\cdot) \right\|_\infty \left\| \pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\|_1 \\
& -\eta \left\| \nabla h(\xi_t) - \nabla h(\widehat{\xi}_{t+1}) \right\| \left\| \xi - \widehat{\xi}_{t+1} \right\| - \eta \left\| \lambda_t - \widehat{\lambda}_{t+1} \right\| \left\| \xi - \widehat{\xi}_{t+1} \right\| \\
& - \eta \left\| V_g^{\pi_t}(\rho) - V_g^{\widehat{\pi}_{t+1}}(\rho) \right\| \left\| \widehat{\lambda}_{t+1} - \lambda \right\| - \eta \left\| \xi_t - \widehat{\xi}_{t+1} \right\| \left\| \widehat{\lambda}_{t+1} - \lambda \right\| \\
\geq \quad & \eta \left(V_{r+\widehat{\lambda}_{t+1}^\top g}^{\pi}(\rho) - V_{r+\lambda^\top g}^{\widehat{\pi}_{t+1}}(\rho) \right) + \eta \left(h(\widehat{\xi}_{t+1}) - h(\xi) \right) + \eta \left\langle -\widehat{\lambda}_{t+1}, \xi - \widehat{\xi}_{t+1} \right\rangle + \eta \left\langle -\widehat{\xi}_{t+1}, \widehat{\lambda}_{t+1} - \lambda \right\rangle \\
& - \frac{\gamma \eta \sqrt{m} C_h}{(1-\gamma)^4} \max_s \left\| \pi_t(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\|_1 \sum_s d_\rho^\pi(s) \left\| \pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\|_1 \\
& - \frac{\eta}{(1-\gamma)^2} \left\| \lambda_t - \widehat{\lambda}_{t+1} \right\| \sum_s d_\rho^\pi(s) \left\| \pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\|_1 \\
& -\eta L_h \left\| \xi_t - \widehat{\xi}_{t+1} \right\| \left\| \xi - \widehat{\xi}_{t+1} \right\| - \eta \left\| \lambda_t - \widehat{\lambda}_{t+1} \right\| \left\| \xi - \widehat{\xi}_{t+1} \right\| \\
& - \frac{\eta \sqrt{m} \kappa_\rho}{(1-\gamma)^3} \left\| \widehat{\lambda}_{t+1} - \lambda \right\| \sum_s d_\rho^{\pi^\star}(s) \left\| \pi_t(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\|_1 - \eta \left\| \xi_t - \widehat{\xi}_{t+1} \right\| \left\| \widehat{\lambda}_{t+1} - \lambda \right\|
\end{aligned}
$$

where the second inequality is due to the performance difference lemma and the convexity of $h(\xi)$,

$$
h(\xi) \;\geq\; h(\widehat{\xi}_{t+1}) + \left\langle \nabla h(\widehat{\xi}_{t+1}), \xi - \widehat{\xi}_{t+1} \right\rangle
$$

and the the last two inequalities is due to Cauchy–Schwarz inequality, and the inequalities in Lemma 15 and the smoothness of h,

$$
\left\| Q_{r+\lambda_t^\top g}^{\pi_t}(s,\cdot) - Q_{r+\lambda_t^\top g}^{\widehat{\pi}_{t+1}}(s,\cdot) \right\|_\infty \;\leq\; \frac{\gamma \sqrt{m} C_h}{(1-\gamma)^3} \max_s \; \left\| \pi_t(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\|_1
$$

$$\left\| Q^{\widehat{\pi}_{t+1}}_{r+\lambda_t^\top g}(s,\cdot) - Q^{\widehat{\pi}_{t+1}}_{r+\widehat{\lambda}_{t+1}^\top g}(s,\cdot) \right\|_\infty \;\leq\; \frac{1}{1-\gamma} \left\| \lambda_t - \widehat{\lambda}_{t+1} \right\|$$

$$\left\| V_g^{\pi_t}(\rho) - V_g^{\widehat{\pi}_{t+1}}(\rho) \right\| \;\leq\; \frac{\sqrt{m}\kappa_\rho}{(1-\gamma)^3} \sum_s d_\rho^{\pi^\star}(s) \left\| \pi_t(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\|_1$$

$$\left\| \nabla h(\xi_t) - \nabla h(\widehat{\xi}_{t+1}) \right\| \;\leq\; L_h \left\| \xi_t - \widehat{\xi}_{t+1} \right\|.$$

We further notice that

$$\begin{aligned}
& \frac{1}{1-\gamma} \sum_s d_\rho^\pi(s) \left\langle \widehat{\pi}_{t+1}(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s), \pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\rangle \\
& + \left\langle \widehat{\xi}_{t+1} - \widehat{\xi}_t, \xi - \widehat{\xi}_{t+1} \right\rangle + \left\langle \widehat{\lambda}_{t+1} - \widehat{\lambda}_t, \lambda - \widehat{\lambda}_{t+1} \right\rangle \\
\leq \quad & \frac{1}{1-\gamma} \max_s \left\| \pi(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s) \right\| \sum_s d_\rho^\pi(s) \left\| \widehat{\pi}_{t+1}(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s) \right\| \\
& + \left\| \widehat{\xi}_{t+1} - \widehat{\xi}_t \right\| \left\| \xi - \widehat{\xi}_{t+1} \right\| + \left\| \widehat{\lambda}_{t+1} - \widehat{\lambda}_t \right\| \left\| \lambda - \widehat{\lambda}_{t+1} \right\|.
\end{aligned}$$

Application of the inequality $ac + bd \leq (a+b)(c+d)$ for $a \geq 0$, $b \geq 0$, $c \geq 0$, and $d \geq 0$ and $d_\rho^\pi(s) \leq \frac{\kappa_\rho}{1-\gamma} d_\rho^{\pi^\star}(s)$ leads to

$$\begin{aligned}
& \eta \left(V^\pi_{r+\widehat{\lambda}_{t+1}^\top g}(\rho) - h(\xi) - \widehat{\lambda}_{t+1}^\top \xi \right) - \eta \left(V^{\widehat{\pi}_{t+1}}_{r+\lambda^\top g}(\rho) - h(\widehat{\xi}_{t+1}) - \lambda^\top \widehat{\xi}_{t+1} \right) \\
\leq \quad & \left(\max_s \left\| \pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s)) \right\| + \left\| \xi - \widehat{\xi}_{t+1} \right\| + \left\| \lambda - \widehat{\lambda}_{t+1} \right\| \right) \\
& \times \Bigg(\frac{1}{1-\gamma} \sum_s d_\rho^\pi(s) \left\| \widehat{\pi}_{t+1}(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s) \right\| + \left\| \widehat{\xi}_{t+1} - \widehat{\xi}_t \right\| + \left\| \widehat{\lambda}_{t+1} - \widehat{\lambda}_t \right\| \\
& \qquad + \frac{\gamma\eta\sqrt{m}C_h|A|}{(1-\gamma)^4} \max_s \left\| \pi_t(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\| \\
& \qquad + \frac{2\eta\sqrt{|A|}}{(1-\gamma)^2} \left\| \lambda_t - \widehat{\lambda}_{t+1} \right\| + \eta(L_h+1) \left\| \xi_t - \widehat{\xi}_{t+1} \right\| \\
& \qquad + \frac{\eta\sqrt{m|A|}\kappa_\rho}{(1-\gamma)^3} \sum_s d_\rho^{\pi^\star}(s) \left\| \pi_t(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\| \Bigg) \\
\leq \quad & \left(\max_s \left\| \pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s)) \right\| + \left\| \xi - \widehat{\xi}_{t+1} \right\| + \left\| \lambda - \widehat{\lambda}_{t+1} \right\| \right) \\
& \underbrace{\begin{aligned}
& \times \Bigg(\frac{\kappa_\rho}{(1-\gamma)^2} \sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t+1}(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s) \right\| \\
& \qquad + \left(\frac{\eta\sqrt{m|A|}\kappa_\rho}{(1-\gamma)^2} + \frac{\gamma\eta\sqrt{m}C_h|A|}{(1-\gamma)^4\rho_{\min}} \right) \frac{1}{1-\gamma} \sum_s d_\rho^{\pi^\star}(s) \left\| \pi_t(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s) \right\| \\
& \qquad + \frac{2\eta\sqrt{|A|}}{(1-\gamma)^2} \left\| \lambda_t - \widehat{\lambda}_{t+1} \right\| + \left\| \widehat{\lambda}_{t+1} - \widehat{\lambda}_t \right\| + \eta(L_h+1) \left\| \xi_t - \widehat{\xi}_{t+1} \right\| + \left\| \widehat{\xi}_{t+1} - \widehat{\xi}_t \right\| \Bigg)
\end{aligned}}_{:= \,\mathrm{Diff}}.
\end{aligned}$$

If we take $\eta > 0$ such that

$$\frac{2\eta\sqrt{|A|}}{(1-\gamma)^2} \;\leq\; \frac{1}{2},\; \eta(L_h+1) \;\leq\; \frac{1}{2},\; \frac{\eta\sqrt{m|A|}\kappa_\rho}{(1-\gamma)^2} \;\leq\; \frac{1}{4},\; \text{ and } \frac{\gamma\eta\sqrt{m}C_h|A|}{(1-\gamma)^4\rho_{\min}} \;\leq\; \frac{1}{4}$$

then,

$$\begin{aligned}
\text{Diff}^2 \;\; &\leq \;\; \max\left(\frac{\kappa_\rho}{1-\gamma}, 1\right)^2 \left(\frac{1}{1-\gamma}\sum_s d_\rho^{\pi^\star}(s)\left(\left\|\widehat{\pi}_{t+1}(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\right\| + \frac{1}{2}\left\|\pi_t(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s)\right\|\right)\right. \\
&\qquad \left. + \left\|\widehat{\lambda}_{t+1} - \widehat{\lambda}_t\right\| + \frac{1}{2}\left\|\lambda_t - \widehat{\lambda}_{t+1}\right\| + \left\|\widehat{\xi}_{t+1} - \widehat{\xi}_t\right\| + \frac{1}{2}\left\|\xi_t - \widehat{\xi}_{t+1}\right\|\right)^2 \\
&\leq \;\; \max\left(\frac{\kappa_\rho}{1-\gamma}, 1\right)^2 \left(\frac{1}{1-\gamma}\sum_s d_\rho^{\pi^\star}(s)\left(\left\|\pi_t(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\right\| + \frac{3}{2}\left\|\pi_t(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s)\right\|\right)\right. \\
&\qquad \left. + \left\|\lambda_t - \widehat{\lambda}_t\right\| + \frac{3}{2}\left\|\lambda_t - \widehat{\lambda}_{t+1}\right\| + \left\|\xi_t - \widehat{\xi}_t\right\| + \frac{3}{2}\left\|\xi_t - \widehat{\xi}_{t+1}\right\|\right)^2 \\
&\leq \;\; 9\max\left(\frac{\kappa_\rho}{1-\gamma}, 1\right)^2 \left(\frac{1}{1-\gamma}\sum_s d_\rho^{\pi^\star}(s)\left(\left\|\pi_t(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\right\|^2 + \left\|\pi_t(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s)\right\|^2\right)\right. \\
&\qquad \left. + \left\|\lambda_t - \widehat{\lambda}_t\right\|^2 + \left\|\lambda_t - \widehat{\lambda}_{t+1}\right\|^2 + \left\|\xi_t - \widehat{\xi}_t\right\|^2 + \left\|\xi_t - \widehat{\xi}_{t+1}\right\|^2\right)
\end{aligned}$$

where the second inequality is due to triangle inequality and the last inequality is due to relaxing the multiplier, $(x+y)^2 \leq 2x^2 + 2y^2$, and Jensen's inequality. □

Lemma 11. *In Algorithm 2,*

$$\begin{aligned}
&\sup_{\pi \in \Pi, \xi \in \Xi, \lambda \in \Lambda} \frac{\left(V^{\pi}_{r+\widehat{\lambda}_{t+1}^\top g}(\rho) - h(\xi) - \widehat{\lambda}_{t+1}^\top \xi\right) - \left(V^{\widehat{\pi}_{t+1}}_{r+\lambda^\top g}(\rho) - h(\widehat{\xi}_{t+1}) - \lambda^\top \widehat{\xi}_{t+1}\right)}{\max_s \left\|\pi(\cdot\,|\,s) - \widehat{\pi}_{t+1}(\cdot\,|\,s))\right\| + \left\|\xi - \widehat{\xi}_{t+1}\right\| + \left\|\lambda - \widehat{\lambda}_{t+1}\right\|} \\
&\geq \;\; C_{\rho,\gamma,\sigma}\left(\sum_s d_\rho^{\pi^\star}(s)\left\|\widehat{\pi}_{t+1}(\cdot\,|\,s) - \mathcal{P}_{\Pi^\star}(\widehat{\pi}_{t+1}(\cdot\,|\,s))\right\| + \left\|\widehat{\xi}_{t+1} - \mathcal{P}_{\Lambda^\star}(\widehat{\xi}_{t+1})\right\|^2 + \left\|\widehat{\lambda}_{t+1} - \mathcal{P}_{\Lambda^\star}(\widehat{\lambda}_{t+1})\right\|\right)
\end{aligned}$$

where $C_{\rho,\gamma,\sigma} > 0$ is some problem-dependent constant.

Proof. Using the saddle-point property of $(\pi^\star, \xi^\star)$, we first show that for all $\widehat{\pi}_{t+1}$ and $\widehat{\xi}_{t+1}$,

$$\begin{aligned}
\min_{\lambda \in \Lambda}\left(V^{\widehat{\pi}_{t+1}}_{r+\lambda^\top g}(\rho) - h(\widehat{\xi}_{t+1}) - \lambda^\top \widehat{\xi}_{t+1}\right) \;\; \leq \;\; & \frac{1}{2}\min_{\lambda \in \Lambda}\left(V^{\pi^\star}_{r+\lambda^\top g}(\rho) - h(\widehat{\xi}_{t+1}) - \lambda^\top \widehat{\xi}_{t+1}\right) \\
& + \frac{1}{2}\min_{\lambda \in \Lambda}\left(V^{\widehat{\pi}_{t+1}}_{r+\lambda^\top g}(\rho) - h(\xi^\star) - \lambda^\top \xi^\star\right).
\end{aligned} \tag{19}$$

This can be proved using the linear program formulation of MDP, we can express the value function in terms of the occupancy measure, i.e., $V^\pi_{r+\lambda^\top g}(\rho) = \langle q^\pi, r + \lambda^\top g\rangle$, where q^π is the occupancy measure that lives in a polytope $\mathcal{Q}$. Let the envelope function be $F(q^\pi, \xi) := \min_{\lambda \in \Lambda}(\langle q^\pi, r + \lambda^\top g\rangle - h(\xi) - \lambda^\top \xi)$. We notice that the point-wise minimization is over a family of affine functions $\langle q^\pi, r + \lambda^\top g\rangle - \lambda^\top \xi$ in terms of (q^π, ξ), and $h(\xi)$ is strongly convex. Thus, $F(q^\pi, \xi)$ is a concave function,

$$F(q^\pi, \xi) \;\leq\; F(q^\pi, \xi^\star) + \partial_\xi F(q^\pi, \xi^\star)^\top(\xi - \xi^\star) \;=\; F(q^\pi, \xi^\star) + \partial_\xi F(q^{\pi^\star}, \xi^\star)^\top(\xi - \xi^\star)$$

$$F(q^\pi, \xi) \;\leq\; F(q^{\pi^\star}, \xi) + \partial_q F(q^{\pi^\star}, \xi)^\top(q^\pi - q^{\pi^\star}) \;=\; F(q^{\pi^\star}, \xi) + \partial_q F(q^{\pi^\star}, \xi^\star)^\top(q^\pi - q^{\pi^\star})$$

where $q^{\pi^\star}$ corresponds to $\pi^\star$ in the one-to-one way, and we notice that $\partial_\xi F(q^\pi, \xi^\star) = \partial_\xi F(q^{\pi^\star}, \xi^\star)$ and $\partial_q F(q^{\pi^\star}, \xi^\star) = \partial_q F(q^{\pi^\star}, \xi)$ because of the decoupled structure of q^π and ξ. From the saddle-point property of $(q^{\pi^\star}, \xi^\star)$, it also reaches the maximum of $F(q^\pi, \xi)$. Thus, by the optimality of $(q^{\pi^\star}, \xi^\star)$,

$$F(q^\pi, \xi) \;\leq\; \frac{1}{2}\left(F(q^\pi, \xi^\star) + F(q^{\pi^\star}, \xi)\right) \quad \text{for all } q^\pi \in \mathcal{Q} \text{ and } \xi \in \Xi$$

which proves (19) by taking $\pi = \widehat{\pi}_{t+1}$ and $\xi = \widehat{\xi}_{t+1}$.

Denote $V_h^\star := V^\star_{r+(\lambda^\star)^\top g}(\rho) - h(\xi^\star) - (\lambda^\star)^\top \xi^\star$ and

$$D_{\max} \; := \max_{\pi,\pi' \,\in\, \Pi,\, \xi,\xi' \,\in\, \Xi,\, \lambda,\lambda' \,\in\, \Lambda} \left(\max_s \left\| \pi(\cdot \,|\, s) - \pi'(\cdot \,|\, s) \right\| + \left\| \xi - \xi' \right\| + \left\| \lambda - \lambda' \right\| \right).$$

Thus,

$$\begin{aligned}
& \sup_{\pi \,\in\, \Pi, \xi \,\in\, \Xi, \lambda \,\in\, \Lambda} \frac{\left(V^{\pi}_{r+\widehat{\lambda}_{t+1}^\top g}(\rho) - h(\xi) - \widehat{\lambda}_{t+1}^\top \xi \right) - \left(V^{\widehat{\pi}_{t+1}}_{r+\lambda^\top g}(\rho) - h(\widehat{\xi}_{t+1}) - \lambda^\top \widehat{\xi}_{t+1} \right)}{\max_s \left\| \pi(\cdot \,|\, s) - \widehat{\pi}_{t+1}(\cdot \,|\, s)) \right\| + \left\| \xi - \widehat{\xi}_{t+1} \right\| + \left\| \lambda - \widehat{\lambda}_{t+1} \right\|} \\
\geq \;\; & \frac{1}{D_{\max}} \sup_{\pi \,\in\, \Pi, \xi \,\in\, \Xi, \lambda \,\in\, \Lambda} \left(\left(V^{\pi}_{r+\widehat{\lambda}_{t+1}^\top g}(\rho) - h(\xi) - \widehat{\lambda}_{t+1}^\top \xi \right) - \left(V^{\widehat{\pi}_{t+1}}_{r+\lambda^\top g}(\rho) - h(\widehat{\xi}_{t+1}) - \lambda^\top \widehat{\xi}_{t+1} \right) \right) \\
= \;\; & \frac{1}{2D_{\max}} \left(\sup_{\pi \,\in\, \Pi} \left(V^{\pi}_{r+\widehat{\lambda}_{t+1}^\top g}(\rho) - h(\xi^\star) - \widehat{\lambda}_{t+1}^\top \xi^\star \right) - \inf_{\lambda \,\in\, \Lambda} \left(V^{\widehat{\pi}_{t+1}}_{r+\lambda^\top g}(\rho) - h(\widehat{\xi}_{t+1}) - \lambda^\top \widehat{\xi}_{t+1} \right) \right) \\
& + \frac{1}{2D_{\max}} \left(\sup_{\xi \,\in\, \Xi} \left(V^{\pi^\star}_{r+\widehat{\lambda}_{t+1}^\top g}(\rho) - h(\xi) - \widehat{\lambda}_{t+1}^\top \xi \right) - \inf_{\lambda \,\in\, \Lambda} \left(V^{\widehat{\pi}_{t+1}}_{r+\lambda^\top g}(\rho) - h(\widehat{\xi}_{t+1}) - \lambda^\top \widehat{\xi}_{t+1} \right) \right) \\
\geq \;\; & \frac{1}{2D_{\max}} \left(\sup_{\pi \,\in\, \Pi} \left(V^{\pi}_{r+\widehat{\lambda}_{t+1}^\top g}(\rho) - h(\xi^\star) - \widehat{\lambda}_{t+1}^\top \xi^\star \right) - \inf_{\lambda \,\in\, \Lambda} \left(V^{\widehat{\pi}_{t+1}}_{r+\lambda^\top g}(\rho) - h(\xi^\star) - \lambda^\top \xi^\star \right) \right) \\
& + \frac{1}{2D_{\max}} \left(\sup_{\xi \,\in\, \Xi} \left(V^{\pi^\star}_{r+\widehat{\lambda}_{t+1}^\top g}(\rho) - h(\xi) - \widehat{\lambda}_{t+1}^\top \xi \right) - \inf_{\lambda \,\in\, \Lambda} \left(V^{\pi^\star}_{r+\lambda^\top g}(\rho) - h(\widehat{\xi}_{t+1}) - \lambda^\top \widehat{\xi}_{t+1} \right) \right)
\end{aligned}$$

where the first inequality is due to the domain's diameter, and the second inequality is due to (19).

Denote $V_h^\star := V^{\pi^\star}_{r+(\lambda^\star)^\top g}(\rho) - h(\xi^\star) - (\lambda^\star)^\top \xi^\star$. If we can prove that there exist constants $c_1 > 0$, $c_2 > 0$, and $c_3 > 0$ such that

$$\max_{\pi \,\in\, \Pi} \left(V^{\pi}_{r+\widehat{\lambda}_{t+1}^\top g}(\rho) - h(\xi^\star) - \widehat{\lambda}_{t+1}^\top \xi^\star \right) - V_h^\star \;\geq\; c_1 \left\| \widehat{\lambda}_{t+1} - \mathcal{P}_{\Lambda^\star}(\widehat{\lambda}_{t+1}) \right\| \tag{20a}$$

$$V_h^\star - \min_{\lambda \,\in\, \Lambda} \left(V^{\widehat{\pi}_{t+1}}_{r+\lambda^\top g}(\rho) - h(\xi^\star) - \lambda^\top \xi^\star \right) \;\geq\; c_2 \sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t+1}(\cdot \,|\, s) - \mathcal{P}_{\Pi^\star}(\widehat{\pi}_{t+1}(\cdot \,|\, s)) \right\| \tag{20b}$$

$$\sup_{\xi \,\in\, \Xi} \left(V^{\pi^\star}_{r+\widehat{\lambda}_{t+1}^\top g}(\rho) - h(\xi) - \widehat{\lambda}_{t+1}^\top \xi \right) - \inf_{\lambda \,\in\, \Lambda} \left(V^{\pi^\star}_{r+\lambda^\top g}(\rho) - h(\widehat{\xi}_{t+1}) - \lambda^\top \widehat{\xi}_{t+1} \right) \;\geq\; c_3 \left\| \widehat{\xi}_{t+1} - \mathcal{P}_{\Lambda^\star}(\widehat{\xi}_{t+1}) \right\|^2 \tag{20c}$$

then,

$$\begin{aligned}
& \sup_{\pi \,\in\, \Pi, \xi \,\in\, \Xi, \lambda \,\in\, \Lambda} \frac{\left(V^{\pi}_{r+\widehat{\lambda}_{t+1}^\top g}(\rho) - h(\xi) - \widehat{\lambda}_{t+1}^\top \xi \right) - \left(V^{\widehat{\pi}_{t+1}}_{r+\lambda^\top g}(\rho) - h(\widehat{\xi}_{t+1}) - \lambda^\top \widehat{\xi}_{t+1} \right)}{\max_s \left\| \pi(\cdot \,|\, s) - \widehat{\pi}_{t+1}(\cdot \,|\, s)) \right\| + \left\| \xi - \widehat{\xi}_{t+1} \right\| + \left\| \lambda - \widehat{\lambda}_{t+1} \right\|} \\
\geq \;\; & \frac{\min(c_1, c_2, c_3)}{2D_{\max}} \left(\sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t+1}(\cdot \,|\, s) - \mathcal{P}_{\Pi^\star}(\widehat{\pi}_{t+1}(\cdot \,|\, s)) \right\| + \left\| \widehat{\xi}_{t+1} - \mathcal{P}_{\Lambda^\star}(\widehat{\xi}_{t+1}) \right\|^2 + \left\| \widehat{\lambda}_{t+1} - \mathcal{P}_{\Lambda^\star}(\widehat{\lambda}_{t+1}) \right\| \right)
\end{aligned}$$

which proves the desired inequality by taking $C_{\rho,\gamma,\sigma} = \frac{\min(c_1,c_2,c_3)}{2D_{\max}}$.

To complete the proof, we first show (20a) and (20b) by resorting the following partial bilinear saddle-point problem,

$$\mathop{\text{maximize}}_{q^\pi \,\in\, \mathcal{Q}} \; \mathop{\text{minimize}}_{\lambda \,\in\, \Lambda} \; \langle q^\pi, r + \lambda^\top g \rangle - h(\xi^\star) - \lambda^\top \xi^\star \;=\; \mathop{\text{minimize}}_{\lambda \,\in\, \Lambda} \; \mathop{\text{maximize}}_{q^\pi \,\in\, \mathcal{Q}} \; \langle q^\pi, r + \lambda^\top g \rangle - h(\xi^\star) - \lambda^\top \xi^\star.$$

Hence, we can apply [Ding et al., 2023, Lemma 10] to obtain (20a) and (20b) directly. To see (20c), we consider the following partial saddle-point problem,

$$\mathop{\text{maximize}}_{\xi \,\in\, \Xi} \; \mathop{\text{minimize}}_{\lambda \,\in\, \Lambda} \; \langle q^{\pi^\star}, r + \lambda^\top g \rangle - h(\xi) - \lambda^\top \xi \;=\; \mathop{\text{minimize}}_{\lambda \,\in\, \Lambda} \; \mathop{\text{maximize}}_{\xi \,\in\, \Xi} \; \langle q^{\pi^\star}, r + \lambda^\top g \rangle - h(\xi) - \lambda^\top \xi$$

which has a strongly concave and linear saddle-point objective function.

We notice that $V_h^\star = \langle q^{\pi^\star}, r + (\lambda^\star)^\top g\rangle - h(\xi^\star) - (\lambda^\star)^\top \xi^\star$. It is straightforward to verify that

$$\begin{aligned}
&\max_{\xi\,\in\,\Xi}\left(\langle q^{\pi^\star}, r+\lambda^\top g\rangle - h(\xi) - \lambda^\top\xi\right) - \min_{\lambda\,\in\,\Lambda}\left(\langle q^{\pi^\star}, r+\lambda^\top g\rangle - h(\xi) - \lambda^\top\xi\right)\\
\geq\;\; & \left(\langle q^{\pi^\star}, r+\lambda^\top g\rangle - h(\xi^\star) - \lambda^\top\xi^\star\right) - V_h^\star\\
&+V_h^\star - \left(\langle q^{\pi^\star}, r+(\lambda^\star)^\top g\rangle - h(\xi) - (\lambda^\star)^\top\xi\right)\\
\geq\;\; & \frac{\sigma}{2}\left\|\xi-\xi^\star\right\|^2
\end{aligned}$$

where we fix $\xi = \xi^\star$ and $\lambda = \lambda^\star$ for the first inequality and the second inequality is from the strong convexity and the optimality of $(\xi^\star, \lambda^\star)$,

$$\langle q^{\pi^\star}, r+(\lambda^\star)^\top g\rangle - h(\xi) - (\lambda^\star)^\top\xi \;\leq\; V_h^\star + (\nabla h(\xi^\star) - \lambda^\star)^\top(\xi-\xi^\star) - \frac{\sigma}{2}\left\|\xi-\xi^\star\right\|^2$$

$$\langle q^{\pi^\star}, r+\lambda^\top g\rangle - h(\xi^\star) - \lambda^\top\xi^\star \;\geq\; V_h^\star + (\langle q^{\pi^\star}, g\rangle - \xi^\star)^\top(\lambda-\lambda^\star).$$

Finally, we notice that replacing $\xi^\star$ by $\mathcal{P}_{\Xi^\star}(\xi)$ does not alter the argument above, which proves (20c) by taking $c_3 = \frac{\sigma}{2}$. □

Proof. By the non-increasing sequence in Lemma 9 and the definition of ζ_t, we have

$$\frac{1}{2(1-\gamma)}\sum_s d_\rho^{\pi^\star}(s)\left\|\widehat{\pi}_{t+1}(\cdot\,|\,s) - \pi_t(\cdot\,|\,s)\right\|^2 + \frac{1}{2}\left\|\widehat{\xi}_{t+1}-\xi_t\right\|^2 + \frac{1}{2}\left\|\widehat{\lambda}_{t+1}-\lambda_t\right\|^2 \;\leq\; \zeta_t \;\leq\; 2\Theta_t \;\leq\; 2\Theta_1.$$

Meanwhile, by the definition of ζ_t and Lemma 10, we have

$$
\begin{aligned}
\zeta_t \;\; & \geq \;\; \frac{1}{4(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t+1}(\cdot \,|\, s) - \pi_t(\cdot \,|\, s) \right\|^2 + \frac{1}{4(1-\gamma)} \left\| \widehat{\xi}_{t+1} - \xi_t \right\|^2 + \frac{1}{4} \left\| \widehat{\lambda}_{t+1} - \lambda_t \right\|^2 \\
& \quad + \frac{1}{4(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t+1}(\cdot \,|\, s) - \pi_t(\cdot \,|\, s) \right\|^2 + \frac{1}{4} \left\| \widehat{\xi}_{t+1} - \xi_t \right\|^2 + \frac{1}{4} \left\| \widehat{\lambda}_{t+1} - \lambda_t \right\|^2 \\
& \quad + \frac{1}{4(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \pi_t(\cdot \,|\, s) - \widehat{\pi}_t(\cdot \,|\, s) \right\|^2 + \frac{1}{4} \left\| \xi_t - \widehat{\xi}_t \right\|^2 + \frac{1}{4} \left\| \lambda_t - \widehat{\lambda}_t \right\|^2 \\
& \geq \;\; \frac{1}{4(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t+1}(\cdot \,|\, s) - \pi_t(\cdot \,|\, s) \right\|^2 + \frac{1}{4} \left\| \widehat{\xi}_{t+1} - \xi_t \right\|^2 + \frac{1}{4} \left\| \widehat{\lambda}_{t+1} - \lambda_t \right\|^2 \\
& \quad + \frac{\eta^2}{32 \max\left(\frac{\kappa_\rho}{1-\gamma}, 1\right)^2} \frac{\left[\left(V^\pi_{r+\widehat{\lambda}_{t+1}^\top g}(\rho) - h(\xi) - \widehat{\lambda}_{t+1}^\top \xi \right) - \left(V^{\widehat{\pi}_{t+1}}_{r+\lambda^\top g}(\rho) - h(\widehat{\xi}_{t+1}) - \lambda^\top \widehat{\xi}_{t+1} \right) \right]_+^2}{\left(\max_s \left\| \pi(\cdot \,|\, s) - \widehat{\pi}_{t+1}(\cdot \,|\, s)) \right\| + \left\| \xi - \widehat{\xi}_{t+1} \right\| + \left\| \lambda - \widehat{\lambda}_{t+1} \right\| \right)^2} \\
& \geq \;\; \frac{1}{4(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t+1}(\cdot \,|\, s) - \pi_t(\cdot \,|\, s) \right\|^2 + \frac{1}{4} \left\| \widehat{\xi}_{t+1} - \xi_t \right\|^2 + \frac{1}{4} \left\| \widehat{\lambda}_{t+1} - \lambda_t \right\|^2 \\
& \quad + \frac{\eta^2 C_{\rho,\gamma,\sigma}}{32 \max\left(\frac{\kappa_\rho}{1-\gamma}, 1\right)^2} \left(\sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t+1}(\cdot \,|\, s) - \mathcal{P}_{\Pi^\star}(\widehat{\pi}_{t+1}(\cdot \,|\, s)) \right\|^2 \right. \\
& \qquad\qquad\qquad\qquad \left. + \left\| \widehat{\xi}_{t+1} - \mathcal{P}_{\Lambda^\star}(\widehat{\xi}_{t+1}) \right\|^4 + \left\| \widehat{\lambda}_{t+1} - \mathcal{P}_{\Lambda^\star}(\widehat{\lambda}_{t+1}) \right\|^2 \right) \\
& \gtrsim \;\; \frac{1}{4(1-\gamma)} \sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t+1}(\cdot \,|\, s) - \pi_t(\cdot \,|\, s) \right\|^4 + \frac{1}{4} \left\| \widehat{\xi}_{t+1} - \xi_t \right\|^4 + \frac{1}{4} \left\| \widehat{\lambda}_{t+1} - \lambda_t \right\|^4 \\
& \quad + \frac{\eta^2 C_{\rho,\gamma,\sigma}}{32 \max\left(\frac{\kappa_\rho}{1-\gamma}, 1\right)^2} \left(\sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t+1}(\cdot \,|\, s) - \mathcal{P}_{\Pi^\star}(\widehat{\pi}_{t+1}(\cdot \,|\, s)) \right\|^4 \right. \\
& \qquad\qquad\qquad\qquad \left. + \left\| \widehat{\xi}_{t+1} - \mathcal{P}_{\Lambda^\star}(\widehat{\xi}_{t+1}) \right\|^4 + \left\| \widehat{\lambda}_{t+1} - \mathcal{P}_{\Lambda^\star}(\widehat{\lambda}_{t+1}) \right\|^4 \right) \\
& \geq \;\; \min\left(\frac{1}{4}, \frac{\eta^2 C_{\rho,\gamma,\sigma}(1-\gamma)}{32 \max\left(\frac{\kappa_\rho}{1-\gamma}, 1\right)^2} \right) \left(\frac{1}{1-\gamma} \sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t+1}(\cdot \,|\, s) - \pi_t(\cdot \,|\, s) \right\|^2 + \left\| \widehat{\xi}_{t+1} - \xi_t \right\|^2 + \left\| \widehat{\lambda}_{t+1} - \lambda_t \right\|^2 \right. \\
& \quad \left. + \frac{1}{1-\gamma} \sum_s d_\rho^{\pi^\star}(s) \left\| \widehat{\pi}_{t+1}(\cdot \,|\, s) - \mathcal{P}_{\Pi^\star}(\widehat{\pi}_{t+1}(\cdot \,|\, s)) \right\|^2 + \left\| \widehat{\xi}_{t+1} - \mathcal{P}_{\Lambda^\star}(\widehat{\xi}_{t+1}) \right\|^2 + \left\| \widehat{\lambda}_{t+1} - \mathcal{P}_{\Lambda^\star}(\widehat{\lambda}_{t+1}) \right\|^2 \right)^2 \\
& \geq \;\; \min\left(\frac{1}{4}, \frac{\eta^2 C_{\rho,\gamma,\sigma}(1-\gamma)}{32 \max\left(\frac{\kappa_\rho}{1-\gamma}, 1\right)^2} \right) \Theta_{t+1}^2
\end{aligned}
$$

where $\gtrsim$ means $\geq$ up to some normalization constants for (ξ, λ)-relevant terms that can be normalized to one due to the boundedness.

Denote $C_\eta := \min\left(\frac{1}{4}, \frac{\eta^2 C_{\rho,\gamma,\sigma}(1-\gamma)}{32 \max\left(\frac{\kappa_\rho}{1-\gamma}, 1\right)^2} \right)$. Thus,

$$
\Theta_{t+1} \;\leq\; \Theta_t - C_\eta \Theta_{t+1}^2.
$$

By Lemma 16, we have

$$
\Theta_t \;=\; O\left(\frac{1}{t} \right)
$$

where the stepsize η satisfies

$$
\eta \;\leq\; \min\left(\frac{1}{4\sqrt{|A|}}, \frac{1}{2(L_h+1)}, \frac{1}{5\sqrt{m|A|}\kappa_\rho}, \frac{\rho_{\min}}{4\gamma\sqrt{m} C_h |A|}, \frac{1}{2\sqrt{2}\iota}, \frac{4\max(\frac{\kappa_\rho}{1-\gamma}, 1)}{\sqrt{\Theta_1 C_{\rho,\gamma,\sigma}(1-\gamma)}} \right) \;:=\; \eta_{\max}.
$$

□

C.4 Proof of Corollary 3

According to Theorem 4, if $t = \Omega(1/\epsilon)$, then,

$$\begin{aligned}
\frac{1}{2}\sum_s d_\rho^{\pi^\star}(s)\left\|\mathcal{P}_{\Pi^\star}(\widehat{\pi}_t(\cdot\,|\,s)) - \widehat{\pi}_t(\cdot\,|\,s)\right\|^2 &= O(\epsilon) \\
\frac{1}{2}\left\|\mathcal{P}_{\Xi^\star}(\widehat{\xi}_t) - \widehat{\xi}_t\right\|^2 &= O(\epsilon) \\
\frac{1}{2}\left\|\mathcal{P}_{\Lambda^\star}(\widehat{\lambda}_t) - \widehat{\lambda}_t\right\|^2 &= O(\epsilon).
\end{aligned}$$

Let $\widehat{\pi}_t^\star(\cdot\,|\,s) := \mathcal{P}_{\Pi^\star}(\widehat{\pi}_t(\cdot\,|\,s))$, $\widehat{\xi}_t^\star := \mathcal{P}_{\Xi^\star}(\widehat{\xi}_t)$, and $\widehat{\lambda}_t^\star := \mathcal{P}_{\Lambda^\star}(\widehat{\lambda}_t)$. Because of the interchangeability of saddle points, $(\widehat{\pi}_t^\star, \widehat{\xi}_t^\star, \widehat{\lambda}_t^\star)$ is a saddle point in $\Pi^\star \times \Xi^\star \times \Lambda^\star$.

First, we have

$$\begin{aligned}
V_r^\star(\rho) - h(\bar{\xi}^\star) - (V_r^{\widehat{\pi}_t}(\rho) - h(\widehat{\xi}_t)) &= \frac{1}{1-\gamma}\sum_{s,a} d_\rho^{\widehat{\pi}_{t+1}}(s)(\widehat{\pi}_t^\star(a\,|\,s) - \widehat{\pi}_t(a\,|\,s))Q_r^{\widehat{\pi}_t}(s,a) - h(\bar{\xi}^\star) + h(\widehat{\xi}_t) \\
&\leq \frac{1}{(1-\gamma)^2}\sum_s d_\rho^{\widehat{\pi}_{t+1}}(s)\left\|\widehat{\pi}_t^\star(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\right\|_1 - h(\bar{\xi}^\star) + h(\widehat{\xi}_t) \\
&\leq \frac{\sqrt{|A|}}{(1-\gamma)^2}\sum_s d_\rho^{\widehat{\pi}_{t+1}}(s)\left\|\widehat{\pi}_t^\star(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\right\| - h(\bar{\xi}^\star) + h(\widehat{\xi}_t) \\
&\leq \frac{\sqrt{|A|}}{(1-\gamma)^2}\sqrt{\sum_s d_\rho^{\widehat{\pi}_{t+1}}(s)\left\|\widehat{\pi}_t^\star(\cdot\,|\,s) - \widehat{\pi}_t(\cdot\,|\,s)\right\|^2} - h(\bar{\xi}^\star) + h(\widehat{\xi}_t)
\end{aligned}$$

where we use Cauchy–Schwarz inequality in the first and third inequalities, and the second inequality is due to that $\|x\|_1 \leq \sqrt{d}\,\|x\|_2$ for $x \in \mathbb{R}^d$. We note that h is continuous. Thus,

$$V_r^\star(\rho) - h(\bar{\xi}^\star) - (V_r^{\widehat{\pi}_t}(\rho) - h(\widehat{\xi}_t)) \;=\; O(\sqrt{\epsilon})$$

where $V_r^\star(\rho) = V_r^{\widehat{\pi}_t^\star}(\rho)$, and $|h(\bar{\xi}^\star) - h(\widehat{\xi}_t)| \leq \sqrt{\epsilon}$ for small ϵ.

Second, we have

$$\begin{aligned}
\left\|\widehat{\xi}_t - V_g^{\widehat{\pi}_t}(\rho)\right\| &\leq \left\|V_g^{\widehat{\pi}_t^\star}(\rho) - \widehat{\xi}_t^\star + \widehat{\xi}_t - V_g^{\widehat{\pi}_t}(\rho)\right\| \\
&= \left\|\widehat{\xi}_t - \widehat{\xi}_t^\star\right\| + \left\|V_g^{\widehat{\pi}_t^\star}(\rho) - V_g^{\widehat{\pi}_t}(\rho)\right\| \\
&\leq O(\sqrt{\epsilon})
\end{aligned}$$

where $|V_{g_i}^{\widehat{\pi}_t^\star}(\rho) - V_{g_i}^{\widehat{\pi}_t}(\rho)| = O(\sqrt{\epsilon})$ is similar as we did for the reward value function.

Finally, we replace $\sqrt{\epsilon}$ by ϵ and combine big O notation to complete the proof.

D Experiment Setup and Additional Results

We provide details of our experiments and additional results. We conduct all experiments on Google Colab in Jupyter Notebook.

D.1 Resilient Policy Search

In this experiment, we consider a tabular constrained MDP with a randomly generated transition kernel, a discount factor $\gamma = 0.9$, uniform rewards $r \in [0,1]$ and utilities $g \in [-1,1]$, and a uniform initial state distribution ρ. The relaxation cost function is $h(\xi) = \alpha\xi^2$, where parameter α balances the closeness to original constraints and the reward maximization objective. The initial constraint is $V_g^\pi(\rho) \geq c$, where c is some constant that often makes the problem infeasible. For comparison, we solve a linear program in occupancy-measure space to find the maximal utility value, which is the minimum value of c to make this problem infeasible. Then, we solve a quadratic program to find the maximal value $V_r^{\pi^\star}(\rho) - h(\xi)$ at an optimal policy $\pi^\star$. Throughout all experiments, the random seed is fixed, and the minimal c to make the problem infeasible is 5.56.

As an example, we take $c = 8$, presenting an infeasible constraint. We report the reward and utility value convergence of our two methods: Algorithm 1 (ResPG-PD) and Algorithm 2 (ResOPG-PD) in Figures 11–12, which show similar convergence behavior as the relaxation in Figure 3. We note, for large α, ResPG-PD behaves similar as ResOPG-PD, which is mainly due to the strongly convex cost function h. However, when α is small, ResPG-PD starts to oscillate while ResOPG-PD still converges, which demonstrates the advantage of the last-iterate convergence in Theorem 4.

We then report the reward, utility, and relaxation gaps by comparing the iterates generated by the two methods with the optimal values from the quadratic program in Figures 13–15. For three choices of α, reward, utility, and relaxation of ResOPG-PD converge to the optimal ones, reaching low platforms due to the accuracy of quadratic program. ResPG-PD behaves similarly for large α: 0.2 or 1, however, oscillates when small α: 0.03.

D.2 Resilient Monitoring: Small State Space

We consider a partial monitoring problem with three locations S_0, S_1, and S_2 in Figure 5. Each location represents a state of an agent. In each state, the choice of action determines the next state, $s_{t+1} = a_t$. In state S_0, possible actions are $\{S_1, S_2\}$. In state S_i with $i \neq 0$, possible actions are $\{S_0, S_i\}$. We define the reward functions as

$$r_i(s_t, a_t) \;=\; b_i \mathbb{I}(s_t = S_i) \;\text{ for }\; i = 0, 1, 2$$

where b_i is the reward for a single time step in a state S_i. To formulate a constrained MDP, we introduce: (i) two constraints $V_i^\pi \geq c_i$ for $i = 1, 2$, where the value functions V_i^π are associated with the corresponding rewards $r_i(s_t, a_t)$; (ii) an objective V_0^π is the value function associated with the reward $r_0(s_t, a_t)$. We initialize the agent using a uniform initial distribution ρ.

As an example, we choose $b_0 = b_1 = 1, b_2 = 1.2$, a discount factor $\gamma = 0.9$ and initial constraints $V_{g_1}^\pi(\rho) = V_1^\pi(\rho) \geq 7$ and $V_{g_2}^\pi(\rho) = V_2^\pi(\rho) \geq 9$, which are infeasible to satisfy. The relaxation controls the trade-off between the closeness to the original constraints and the gain of rewards. One extreme case is that the agent always moves to S_0 if not S_0. In this case, the reward reaches the maximum value and the relaxation is the maximum. Another extreme case is that the agent always stays in its original state S_i if not in S_0. In this case, the agent spends as much time as possible in S_1, S_2 to gain utility, which makes relaxed constraints close to the original constraints and makes the reward value smallest. The relaxation cost function is $h(\xi) = \alpha \|\xi\|^2$ in which we set $\alpha = 0.1$. We observe that our two methods can relax the two initial constraints to make this problem feasible, but keep away from the extreme cases.

We recall Figure 7 that both methods can successfully relax the two constraints, which are initially infeasible, to be feasible for different cost functions. To show the reward and utility convergence performance of our two methods: Algorithm 1 (ResPG-PD) and Algorithm 2 (ResOPG-PD), we report Figures 16–17 for $\alpha = 0.1$. We show the reward, utility and relaxation optimality gaps of the two methods in Figures 18–20. As we expect, ResPG-PD often behave oscillating during training while ResOPG-PD can overcome oscillation, yielding a nearly-optimal policy in the last iterate.

D.3 Resilient Monitoring: Large State Space

We generalize the problem in Section D.2 to a robot monitoring problem in a 10×10 grid as shown in Figure 8, where each grid point is a state. In each state, four possible actions are: left, right, up, and down. The choice of the action and current state determines next state. The next state is the current state moving towards the action selected for one unit. If the next state falls outside the grid, the robot remains in the current state. Three shaded circles S_0, S_1, S_2 in the grid represent three areas to be monitored. We define the reward functions as

$$r_i(s_t, a_t) \;=\; b_i \mathbb{I}(s_t \in S_i) \;\text{ for }\; i = 0, 1, 2$$

where b_i is the reward for a single time step in an area S_i. We also have: (i) two constraints $V_i^\pi \geq c_i$ for $i = 1, 2$, where the value function V_i^π is associated with the reward $r_i(s_t, a_t)$; (ii) an objective V_0^π is the value function associated with the reward $r_0(s_t, a_t)$. The initial distribution ρ is uniform.

We use the same problem parameter setting as Section D.2, i.e., $b_0 = b_1 = 1, b_2 = 1.2$, a discount factor $\gamma = 0.9$ and initial constraints $V_{g_1}^\pi(\rho) = V_1^\pi(\rho) \geq 7$ and $V_{g_2}^\pi(\rho) = V_2^\pi(\rho) \geq 9$, which are infeasible to satisfy. The relaxation cost function is $h(\xi) = \alpha \|\xi\|^2$, where $\alpha = 0.08$.

As previously shown in Figure 10, both methods can relax the infeasible constraints. We then report reward and utility convergence performance for the two methods in Figure 21–22 and report the reward, utility and relaxation optimality gaps of two methods in Figures 23–25. The convergence behavior of two methods in a large state/action space is similar as the described in Section D.2.

E Some Useful Lemmas

Lemma 12. *Let Assumption 1 hold for Problem (4). For any $C \geq 2\bar{\lambda}^\star$, if there exists a policy $\pi \in \Pi$, $\xi \in \Xi$, and $\delta > 0$ such that $V_r^\star(\rho) - h(\bar{\xi}^\star) - (V_r^\pi(\rho) - h(\xi)) + C\sum_{i=1}^m [\xi_i - V_{g_i}^\pi(\rho)]_+ \leq \delta$, then $\sum_{i=1}^m [\xi_i - V_{g_i}^\pi(\rho)]_+ \leq 2\delta/C$, where $[x]_+ = \max(x,0)$.*

Proof. For Problem (4), we introduce its value function as

$$v(\tau) \;=\; \underset{\pi\,\in\,\Pi,\xi\,\in\,\Xi}{\text{maximize}} \;\left\{\, V_r^\pi(\rho) - h(\xi) \,|\, V_{g_i}^\pi(\rho) \geq \xi_i + \tau_i, i = 1,\ldots,m \,\right\}$$

We note that, $v(\tau)$ is a concave function, which is similar as Lemma 1, and $v(0) = V_h^\star = V_r^\star(\rho) - h(\bar{\xi}^\star)$. The rest of proof is straightforward from [Ding et al., 2022, Lemma 4].

□

For any convex differentiable function ψ: $X \to \mathbb{R}$, the Bregman divergence of x, $x' \in X$ is given by $D_\psi(x',x) := \psi(x') - \psi(x) - \langle \nabla\psi(x), x'-x\rangle$. When ψ is σ-strongly convex, $D_\psi(x',x) \geq \frac{\sigma}{2}\left\|x-x'\right\|^2$ for any x, $x' \in X$. Specifically, when $\psi(x) = \frac{1}{2}\left\|x\right\|^2$, $D_\psi(x',x) = \frac{1}{2}\left\|x'-x\right\|^2$.

Lemma 13. *Let X be a convex set. If $x' = \operatorname{argmin}_{\bar{x}\,\in\,X}\langle\bar{x},g\rangle + D_\psi(\bar{x},x)$, then for any $x^\star \in X$,*

$$\langle x'-x^\star, g\rangle \;\leq\; D_\psi(x^\star,x) - D_\psi(x^\star,x') - D_\psi(x',x).$$

Proof. See [Wei et al., 2020, Lemma 10]. □

Lemma 14. *Assume that $D_\psi(x,x') \geq \frac{1}{2}\left\|x-x'\right\|_p^2$ for some ψ and $p \geq 1$. If*

$$x_1 \;=\; \underset{\bar{x}\,\in\,X}{\operatorname{argmin}}\;\langle\bar{x},g_1\rangle + D_\psi(\bar{x},x) \;\;\text{and}\;\; x_2 \;=\; \underset{\bar{x}\,\in\,X}{\operatorname{argmin}}\;\langle\bar{x},g_2\rangle + D_\psi(\bar{x},x)$$

then,

$$\left\|x_1 - x_2\right\|_p \;\leq\; \left\|g_1 - g_2\right\|_q$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

Proof. See [Wei et al., 2020, Lemma 11]. □

Lemma 15. *For any two policies π and π', we have*

$$\begin{aligned}\left\|Q_r^\pi(\cdot,\cdot) - Q_r^{\pi'}(\cdot,\cdot)\right\|_\infty \;&\leq\; \frac{\gamma}{(1-\gamma)^2}\max_s\;\left\|\pi(\cdot\,|\,s) - \pi'(\cdot\,|\,s)\right\|_1\\ \left|V_r^\pi(\rho) - V_r^{\pi'}(\rho)\right| \;&\leq\; \frac{\kappa_\rho}{(1-\gamma)^3}\sum_s d_\rho^{\pi^\star}(s)\left\|\pi(\cdot\,|\,s) - \pi'(\cdot\,|\,s)\right\|_1.\end{aligned}$$

Proof. See [Ding et al., 2023, Lemma 11]. □

Lemma 16. *Let a non-negative sequence $\{B_t\}_{t\,\geq\,1}$ satisfy that for any $p > 0$ and $q > 0$,*

$$B_{t+1} \;\leq\; B_t - qB_{t+1}^{p+1} \;\;\text{and}\;\; q(1+p)B_1^p \;\leq\; 1.$$

Then, $B_t \leq C\,t^{-1/p}$, where $C := \max(B_1, (1/(pq))^{1/p})$.

Proof. See [Wei et al., 2020, Lemma 12]. □

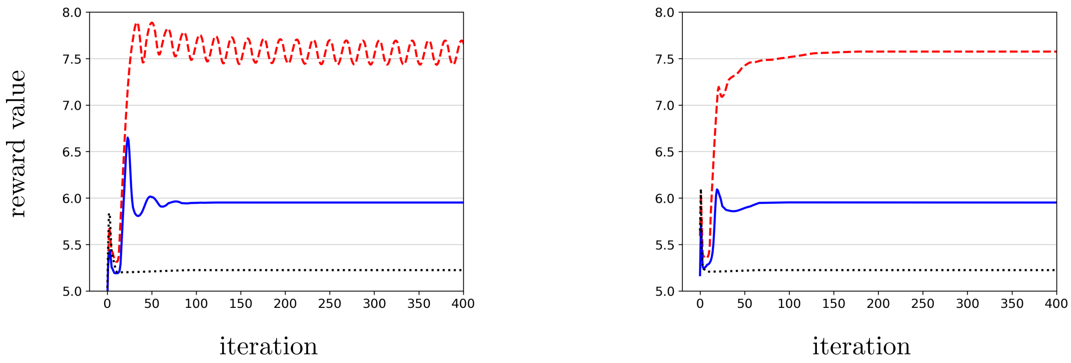


Figure 11: Reward value convergence of ResPG-PD (Algorithm 1, left) and ResOPG-PD (Algorithm 2, right), with different cost functions $h(\xi) = \alpha\xi^2$, where $\alpha = 0.03$ (- - -), $\alpha = 0.2$ (——), $\alpha = 1$ (⋯⋯), and stepsize $\eta = 0.2$ in the policy search problem of Section D.1.

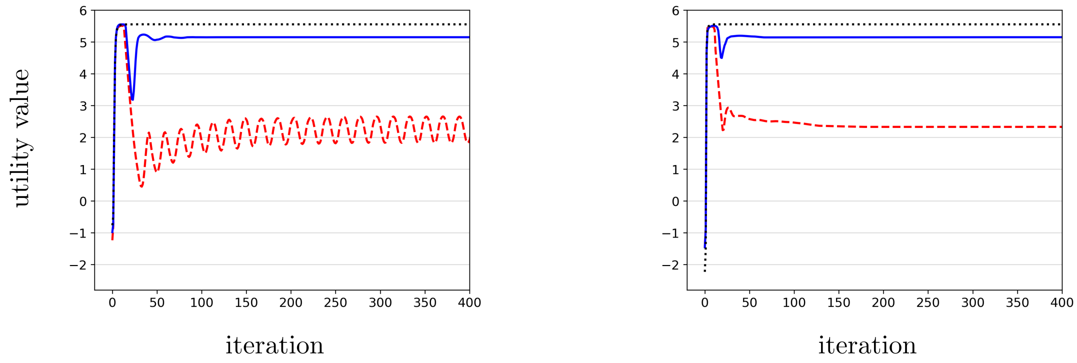


Figure 12: Utility value convergence of ResPG-PD (Algorithm 1, left) and ResOPG-PD (Algorithm 2, right), with different cost functions $h(\xi) = \alpha\xi^2$, where $\alpha = 0.03$ (- - -), $\alpha = 0.2$ (——), $\alpha = 1$ (⋯⋯), and stepsize $\eta = 0.2$ in the policy search problem of Section D.1.

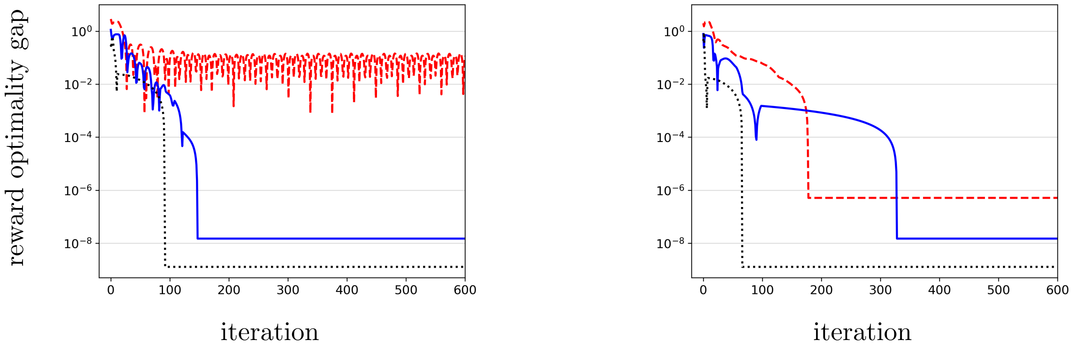


Figure 13: Reward optimality gap of ResPG-PD (Algorithm 1, left) and ResOPG-PD (Algorithm 2, right), with different cost functions $h(\xi) = \alpha\xi^2$, where $\alpha = 0.03$ (- - -), $\alpha = 0.2$ (——), $\alpha = 1$ (⋯⋯), and stepsize $\eta = 0.2$ in the policy search problem of Section D.1.

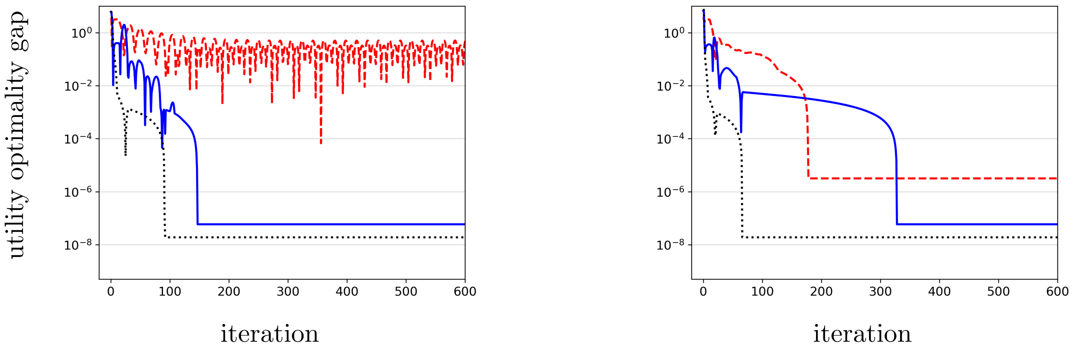


Figure 14: Utility optimality gap of ResPG-PD (Algorithm 1, left) and ResOPG-PD (Algorithm 2, right), with different cost functions $h(\xi) = \alpha\xi^2$, where $\alpha = 0.03$ (- - -), $\alpha = 0.2$ (——), $\alpha = 1$ (······), and stepsize $\eta = 0.2$ in the policy search problem of Section D.1.

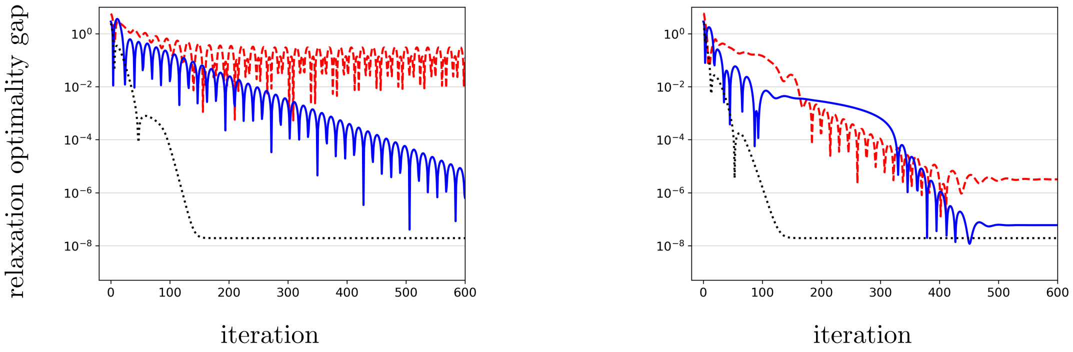


Figure 15: Relaxation optimality gap of ResPG-PD (Algorithm 1, left) and ResOPG-PD (Algorithm 2, right), with different cost functions $h(\xi) = \alpha\xi^2$, where $\alpha = 0.03$ (- - -), $\alpha = 0.2$ (——), $\alpha = 1$ (······), and stepsize $\eta = 0.2$ in the policy search problem of Section D.1.

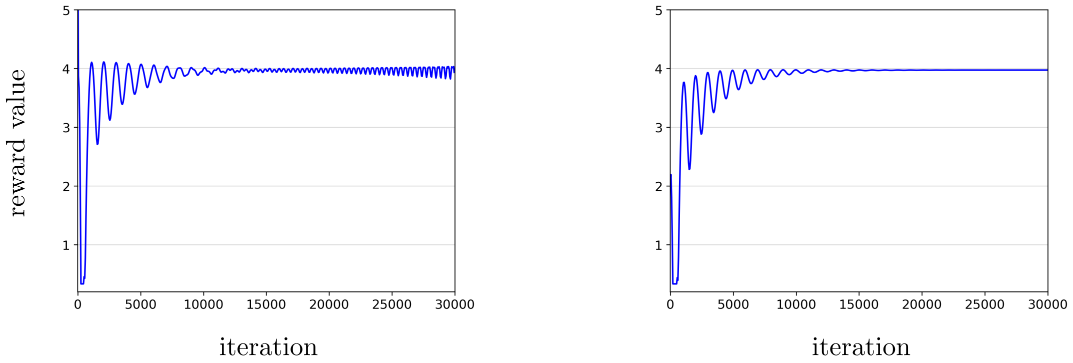


Figure 16: Reward value convergence of ResPG-PD (Algorithm 1, left) and ResOPG-PD (Algorithm 2, right), with a cost functions $h(\xi) = \alpha \left\| \xi \right\|^2$, where $\alpha = 0.1$, and stepsize $\eta = 0.005$ in the monitoring problem of Section D.2.

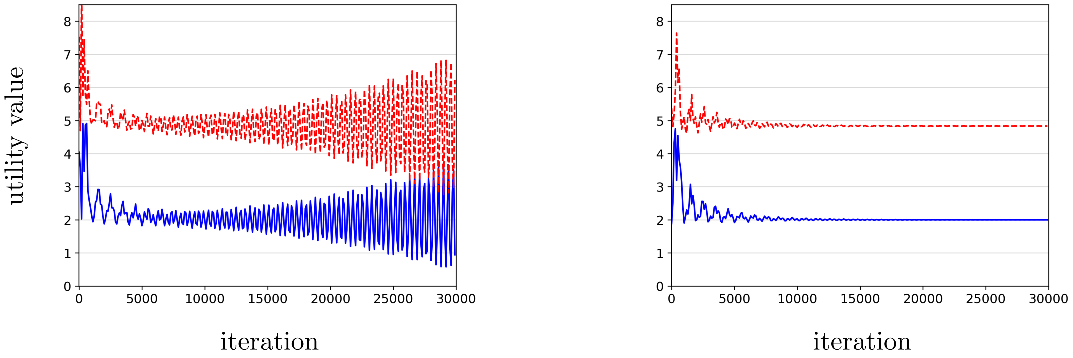


Figure 17: Utility value convergence ($V_{g_1}^{\pi}(\rho)$: ——, $V_{g_2}^{\pi}(\rho)$: - - -) of ResPG-PD (Algorithm 1, left) and ResOPG-PD (Algorithm 2, right), with a cost functions $h(\xi) = \alpha \left\| \xi \right\|^2$ for $\alpha = 0.1$, and stepsize $\eta = 0.005$ in the monitoring problem of Section D.2.

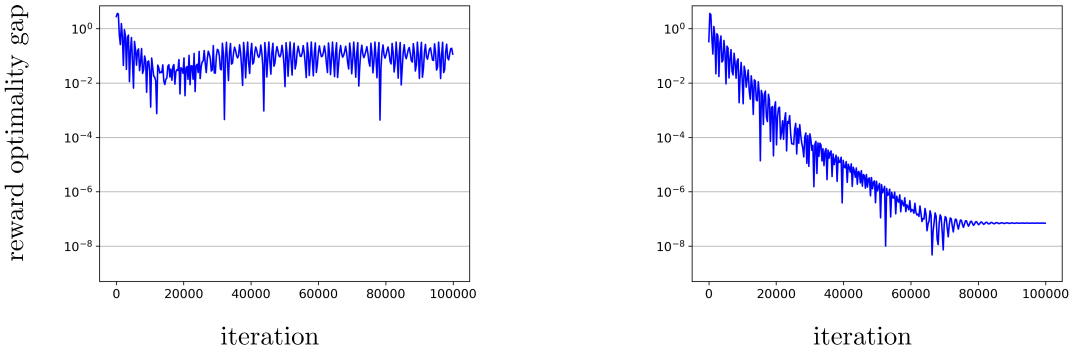


Figure 18: Reward optimality gap of ResPG-PD (Algorithm 1, left) and ResOPG-PD (Algorithm 2, right), with a cost functions $h(\xi) = \alpha \|\xi\|^2$, where $\alpha = 0.1$ and stepsize $\eta = 0.005$ in the monitoring problem of Section D.2.

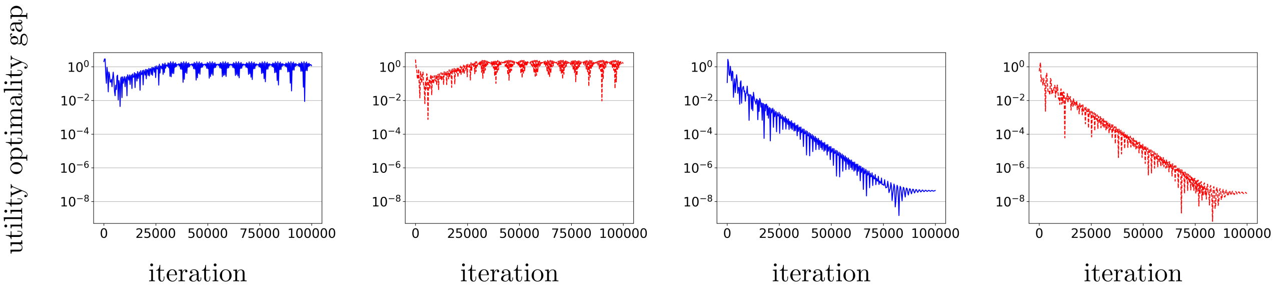


Figure 19: Utility optimality gap ($V_{g_1}^{\pi}(\rho)$: ——, $V_{g_2}^{\pi}(\rho)$: - - -) of ResPG-PD (Algorithm 1, two figures on the left) and ResOPG-PD (Algorithm 2, two figures on the right), with a cost functions $h(\xi) = \alpha \|\xi\|^2$, where $\alpha = 0.1$ and stepsize $\eta = 0.005$ in the monitoring problem of Section D.2.

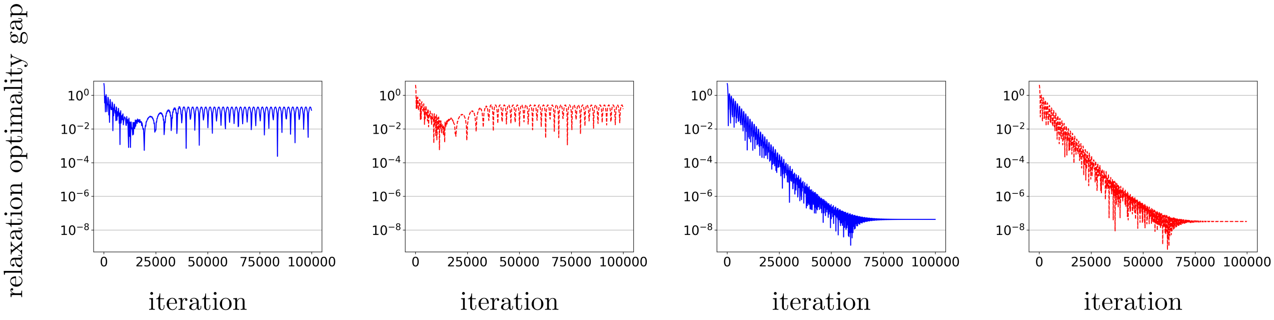


Figure 20: Relaxation optimality gap (ξ_1: ——, ξ_2: - - -) of ResPG-PD (Algorithm 1, two figures on the left) and ResOPG-PD (Algorithm 2, two figures on the right), with a cost functions $h(\xi) = \alpha \|\xi\|^2$, where $\alpha = 0.1$ and stepsize $\eta = 0.005$ in the monitoring problem of Section D.2.

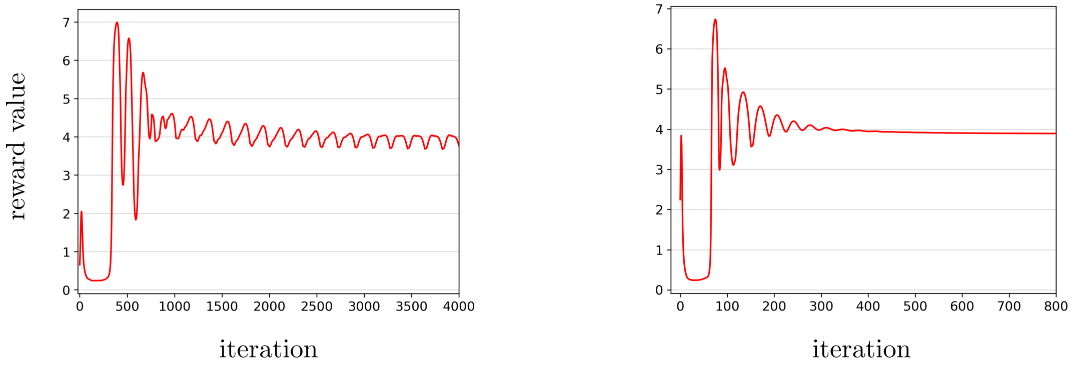


Figure 21: Reward value convergence of ResPG-PD (Algorithm 1, left) and ResOPG-PD (Algorithm 2, right), with a cost functions $h(\xi) = \alpha \left\|\xi\right\|^2$, where $\alpha = 0.08$, in the monitoring problem of Section D.3. The stepsize for ResPG-PD is $\eta = 0.01$ and the stepsize for ResOPG-PD is $\eta = 0.05$.

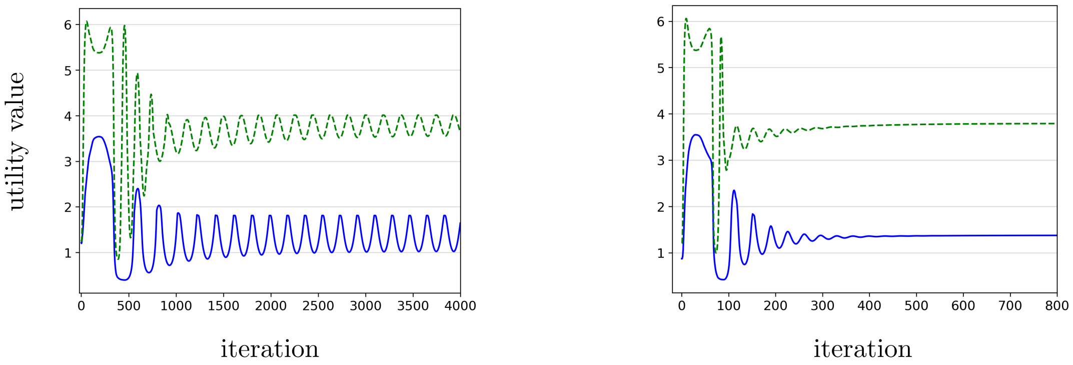


Figure 22: Utility value convergence ($V_{g_1}^{\pi}(\rho)$: ——, $V_{g_2}^{\pi}(\rho)$: - - -) of ResPG-PD (Algorithm 1, left) and ResOPG-PD (Algorithm 2, right), with a cost functions $h(\xi) = \alpha \left\|\xi\right\|^2$ for $\alpha = 0.08$, in the monitoring problem of Section D.3. The stepsize for ResPG-PD is $\eta = 0.01$ and the stepsize for ResOPG-PD is $\eta = 0.05$.

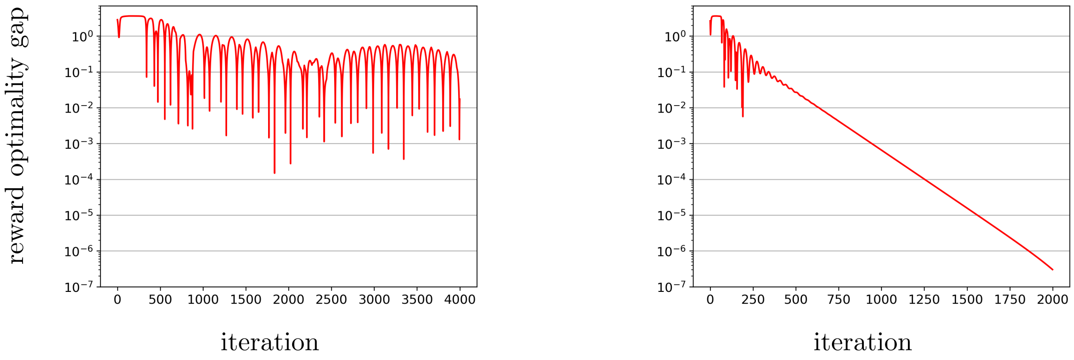


Figure 23: Reward optimality gap of ResPG-PD (Algorithm 1, left) and ResOPG-PD (Algorithm 2, right), with a cost functions $h(\xi) = \alpha \|\xi\|^2$, where $\alpha = 0.08$ in the monitoring problem of Section D.3. The stepsize for ResPG-PD is $\eta = 0.01$ and the stepsize for ResOPG-PD is $\eta = 0.05$.

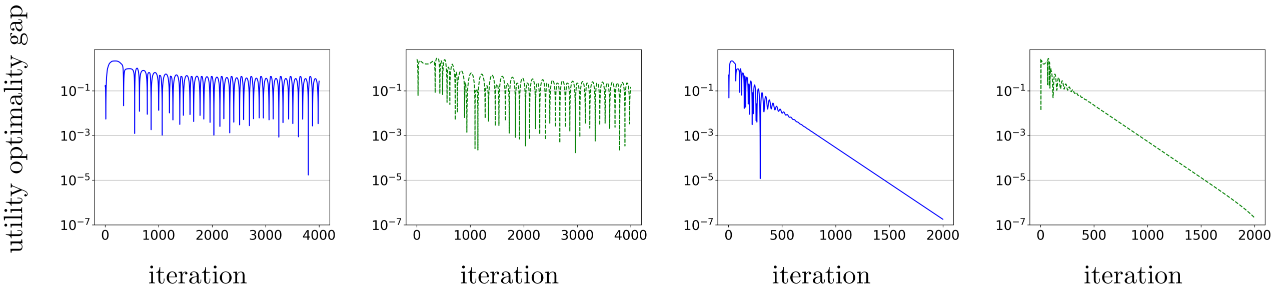


Figure 24: Utility optimality gap ($V_{g_1}^{\pi}(\rho)$: ——, $V_{g_2}^{\pi}(\rho)$: - - -) of ResPG-PD (Algorithm 1, two figures on the left) and ResOPG-PD (Algorithm 2, two figures on the right), with a cost functions $h(\xi) = \alpha \|\xi\|^2$, where $\alpha = 0.08$ in the monitoring problem of Section D.3. The stepsize for ResPG-PD is $\eta = 0.01$ and the stepsize for ResOPG-PD is $\eta = 0.05$.

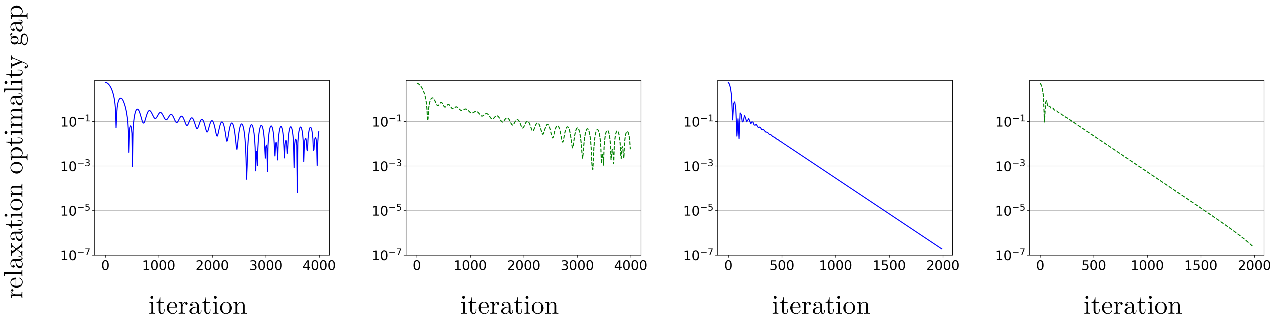


Figure 25: Relaxation optimality gap (ξ_1: ——, ξ_2: - - -) of ResPG-PD (Algorithm 1, two figures on the left) and ResOPG-PD (Algorithm 2, two figures on the right), with a cost functions $h(\xi) = \alpha \|\xi\|^2$, where $\alpha = 0.08$ in the monitoring problem of Section D.3. The stepsize for ResPG-PD is $\eta = 0.01$ and the stepsize for ResOPG-PD is $\eta = 0.05$.