Offline Constrained Multi-Objective Reinforcement Learning via Pessimistic Dual Value Iteration

Runzhe Wu
Shanghai Jiao Tong University
runzhe@sjtu.edu.cn

Yufeng Zhang
Northwestern University
yufengzhang2023@u.northwestern.edu

Zhuoran Yang
Princeton University
zy6@princeton.edu

Zhaoran Wang
Northwestern University
zhaoranwang@gmail.com

Abstract

In constrained multi-objective RL, the goal is to learn a policy that achieves the best performance specified by a multi-objective preference function under a constraint. We focus on the offline setting where the RL agent aims to learn the optimal policy from a given dataset. This scenario is common in real-world applications where interactions with the environment are expensive and the constraint violation is dangerous. For such a setting, we transform the original constrained problem into a primal-dual formulation, which is solved via dual gradient ascent. Moreover, we propose to combine such an approach with pessimism to overcome the uncertainty in offline data, which leads to our Pessimistic Dual Iteration (PEDI). We establish upper bounds on both the suboptimality and constraint violation for the policy learned by PEDI based on an arbitrary dataset, which proves that PEDI is provably sample efficient. We also specialize PEDI to the setting with linear function approximation. To the best of our knowledge, we propose the first provably efficient constrained multi-objective RL algorithm with offline data without any assumption on the coverage of the dataset.

1 Introduction

There has been increased interest in multi-objective RL in recent years. Compared with traditional single-objective RL, the goal of the multi-objective one depends on a preference function, which takes the multiple objectives as input and outputs a scalar. The multi-objective optimization problems are usually constrained, as otherwise, it may cause danger or malfunction in applications. For example, consider a home automation system that helps humans monitor and control home attributes which can be regarded as multi-objectives. Users at different times may value different aspects of its services. Some may think highly of lighting at night, while others may concern the climate. We can formulate the users' preference as a preference function on the multiple objectives. Therefore, the system is dealing with a multi-objective optimization problem. However, the system cannot optimize this problem without any constraints, which might go against human's will, such as generating extreme climate inside a house or performing unacceptably energy-consuming operations.

We formulate the constrained multi-objective Markov decision process (CMOMDP), which is similar to the constrained Markov decision process (CMDP) (Altman, 1999). The difference is that the objectives are multiple and constraints can be nonlinear. We aim to minimize the value of a preference function, which takes multiple objectives as input and outputs a scalar.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Most existing RL methods assume full accessibility of the environment for the agent, which tends to be impractical as the exploration could be expensive (Gottesman et al., 2019) and dangerous (Shalev-Shwartz et al., 2016). Hence, we consider the offline case where the agent has only a historical dataset collected a priori and has no further interactions with the environment. This setting is common and embodied in various scenarios such as healthcare (Chakraborty and Murphy, 2014) and auto-driving (Sun et al., 2020). However, offline RL is less understood theoretically (Levine et al., 2020) than online RL, and how to maximally exploit the dataset remains unclear.

In this paper, we study the offline CMOMDP. The challenges are threefold:

(i) Different from CMDPs, the nonlinearity of the preference function and constraints of CMOMDPs makes the analysis challenging. Moreover, constraints can be neither convex nor concave in the policy, which means the optimization problems are usually non-convex. It brings great difficulty to designing a provably efficient algorithm.

(ii) Since further interaction with the environment is prohibited, the dataset's quality is not guaranteed. The data collected by the experimenter may not sufficiently cover the trajectories induced by the optimal policy. Therefore, the information of the optimal policy could be limited, which probably makes the suboptimality and constraint violation arbitrarily large.

(iii) As the CMOMDP can be considered a generalization of CMDP (see the reduction in Appendix C), any difficulties emerging in the CMDP will arise here too. For example, due to constraints, the Bellman optimality equation may not hold anymore. Therefore, most existing offline RL algorithms based on dynamic programming are inapplicable.

Challenge (i) is inherent in CMOMDPs, and certain requirements on the preference function and constraints are inevitable. Hence, We suppose they satisfy some conditions. For instance, a geometric analog of Slater's condition, which is usually assumed in CMDPs, is imposed. To tackle challenge (ii), existing works have put a great effort. One possible principle is pessimism, which applies penalties to ensure pessimistic estimation. This method is successful in ordinary RL, and we extend it to CMOMDPs. We also note that we only impose a minimal assumption on the offline dataset's compliance in our analysis. For challenge (iii), we apply the convex conjugate and Fenchel's duality to transform the problem into a primal-dual formulation.

In summary, our work answers the following question:

Is it possible to develop a provably efficient offline algorithm for constrained multi-objective reinforcement learning with minimal assumptions on the dataset?

We propose the Pessimistic Dual Iteration (PEDI) algorithm. Theoretical contributions are as follows:

(i) By transforming the original constrained optimization problem of CMOMDPs into a primal-dual formulation via convex conjugate and duality, we develop the algorithm for general CMOMDPs with offline dataset, which iterates in a dual gradient ascent manner. Then, we instantiate the algorithm for linear kernel CMOMDPs, which is a large class of CMOMDPs that includes the tabular case.

(ii) We show in Appendix F the significance of pessimism in offline CMOMDPs. To summarize, we decompose the discrepancy between a value function and the optimal one into spurious correlation, intrinsic uncertainty, and optimization error, among which the spurious correlation is the most difficult to control. However, by maintaining pessimistic estimates of the value functions, PEDI eliminates the spurious correlation.

(iii) We establish theoretical guarantees for PEDI. Specifically, we demonstrate that two metrics, the suboptimality and the constraint violation, can be bounded from above by the optimization error of $\mathcal{O}(1/\sqrt{K})$ and the intrinsic uncertainty, where K is the number of iterations. When specialized to linear kernel CMOMDPs, the upper bound of suboptimality matches the information-theoretic lower bound up to the optimization error and multiplicative factors of constants related to the CMOMDP, which suggests the near-optimality of PEDI.

(iv) We show that the error of PEDI is data-dependent, i.e., it depends on how well the dataset covers the trajectories induced by the optimal policy. When the trajectories are assumed further to be sufficiently covered, the suboptimality and constraint violation are $\widetilde{\mathcal{O}}(1/\sqrt{N})$ where N is the number of trajectories in the dataset.

To the best of our knowledge, we are the first to propose a provably efficient offline algorithm that considers the constrained multi-objective RL without any assumptions on the coverage of the dataset. As a by-product, our method can be viewed as a highly generalized one as it can easily reduce to certain simpler cases such as CMDPs, which is discussed in Appendix C.

1.1 Related works

Constrained RL. Our work is closely related to CMDPs (Altman, 1999). Several recent works (Efroni et al., 2020; Ding et al., 2021; Qiu et al., 2020; Brantley et al., 2020; Tessler et al., 2018) have studied the MDP with linear constraints and others (Bhatnagar and Lakshmanan, 2012; Chow et al., 2017; Paternain et al., 2019) have considered RL with more complex constraints. Most of them have provided algorithms with both regret and total constraint violation guarantees. However, while these approaches are successful in CMDPs, they cannot be applied to CMOMDPs where objectives are multiple and constraints can be nonlinear. The difference between their methods and ours are (i) they usually assume the Slater's condition, while we consider its geometric analog, and (ii) most works utilize the Lagrangian multiplier to handle constraints, while in our method, we obtain a primal-dual formulation via conjugate.

Multi-objective RL. Existing studies on multi-objective RL can be divided into two groups: the first maintains a set of Pareto-optimal policies (Barrett and Narayanan, 2008; Castelletti et al., 2011, 2012; Wang and Sebag, 2013), and the second collapses multi-objective vector into a scalar and then apply a standard RL method (Barrett and Narayanan, 2008; Natarajan and Tadepalli, 2005; Van Moffaert et al., 2013; Cheung, 2019). However, these methods are available in unconstrained multi-objective RL but unsuitable for constrained cases. The difficulty is that the constraints invalidate the Bellman optimality principle, which is the cornerstone of these works.

Offline RL. Recent successful offline RL (Lange et al., 2012; Levine et al., 2020) methods (Fujimoto et al., 2019; Laroche et al., 2019; Jaques et al., 2019; Wu et al., 2019; Kumar et al., 2019, 2020; Agarwal et al., 2020; Yu et al., 2020; Kidambi et al., 2020; Siegel et al., 2020; Nair et al., 2020; Liu et al., 2020; Wang et al., 2020a,b; Tosatto et al., 2017; Farahmand et al., 2016, 2010) fall into two categories (Buckman et al., 2020): (i) uncertainty-aware pessimistic algorithms and (ii) proximal pessimistic algorithms. (i) gives a lower estimate for states and actions less covered by the dataset, while (ii) avoids visiting these less visited states and actions by regularization. However, existing methods are largely based on approximate dynamic programming and do not take constraints into account, thus failing to apply. Our proposed method generalizes the ordinary pessimistic algorithm so that it is compatible with multiple objectives and constraints.

In the analysis, we only assume the compliance of the dataset. In comparison with existing works, which assume the sufficient coverage of the dataset such as the finite concentrability coefficients (Antos et al., 2008; Farahmand et al., 2010; Scherrer et al., 2015; Le et al., 2019; Chen and Jiang, 2019; Fu et al., 2020) and uniformly lower bounded densities of visitation measures (Yin et al., 2021), we require no assumptions as they often fail to hold in practice. As Wang et al. (2020a) have studied the influence of insufficient coverage of the dataset, their conclusion is also reflected in our main results. Moreover, we do not impose any constraints on the affinity of the behavior policy and the learned policy (Liu et al., 2020). In addition, our general algorithm uses convex conjugate, a technique used similarly in Yu et al. (2021); Miryoosefi and Jin (2021) but their method cannot apply to offline situations and Yu et al. (2021) do not address the function approximation setting.

1.2 Notations

For simplicity, we write $\langle f, g\rangle_{\mathcal{X}} = \int_{\mathcal{X}} f(x)g(x)\,\mathrm{d}x$ as the inner product on space $\mathcal{X}$ for functions $f, g : \mathcal{X} \to \mathbb{R}$. We also define the norms $\|f\|_{2,\mathcal{X}} = \sqrt{\langle f, f\rangle_{\mathcal{X}}}$ and $\|f\|_{\infty,\mathcal{X}} = \sup_{x\in\mathcal{X}} |f(x)|$. We denote by $[n]$ the set of positive integers not greater than n, i.e., $[n] = \{i \in \mathbb{Z} | 1 \le i \le n\}$. For two vectors $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^D$, we write $\boldsymbol{x} \ge \boldsymbol{x}'$ (or $\boldsymbol{x} \le \boldsymbol{x}'$) if $\boldsymbol{x}$ is not smaller (or larger) than $\boldsymbol{x}'$ in an elementwise manner. We set $|\boldsymbol{x}| = (|x_1|, |x_2|, \ldots, |x_D|)^\top$ and $\boldsymbol{x}_+ = (\max\{x_1, 0\}, \max\{x_2, 0\}, \ldots, \max\{x_D, 0\})^\top$. We denote by $\mathcal{B}^D$ the D-dimensional unit Euclidean ball, i.e., $\mathcal{B}^D = \{\boldsymbol{x} \in \mathbb{R}^D : \|\boldsymbol{x}\|_2 \le 1\}$. The set of probability distributions on a set $\mathcal{X}$ is denoted by $\Delta(\mathcal{X})$. We denote by $\widetilde{O}(\cdot)$ the big O notation ignoring logarithmic factors.

2 Preliminaries

2.1 Constrained multi-objective Markov decision process (CMOMDP)

We consider an episodic CMOMDP given by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, \boldsymbol{c})$ where $\mathcal{S}$ is a state space, $\mathcal{A}$ is an action space, $H \in \mathbb{N}_+$ is the horizon, $\mathcal{P} = \{\mathcal{P}_h\}_{h=1}^H$ is a collection of transition kernels $\mathcal{P}_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, and $\boldsymbol{c} = \{\boldsymbol{c}_h\}_{h=1}^H$ is a collection of cost functions $\boldsymbol{c}_h : \mathcal{S} \times \mathcal{A} \to [0,1]^D$. Note that by $\mathcal{P}$ we mean both the probability density function and the probability mass function, as they can be unified (see Appendix A for details). We suppose a fixed initial state $\underline{s} \in \mathcal{S}$ for simplicity. However, a stochastic initial state poses no extra difficulty to our analysis. In each episode, given a policy $\pi = \{\pi_h\}_{h=1}^H$ where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$, the agent interacts with the environment as follows. At state s_h, the agent takes action $a_h \in \mathcal{A}$ according to $\pi_h(\cdot \mid s_h)$, and then receives a stochastic cost $\boldsymbol{c}_h \in [0,1]^D$, for which we assume $\mathbb{E}[\boldsymbol{c}_h | s_h, a_h] = \boldsymbol{c}_h(s,a)$. The system then transits to the next state s_{h+1} according to $\mathcal{P}_h(\cdot \mid s_h, a_h)$. The episode terminates at state s_{H+1} where no action is taken and the cost is zero.

Given a policy π, the state-value function $\boldsymbol{V}_h^\pi : \mathcal{S} \to [0,H]^D$ and the action-value function $\boldsymbol{Q}_h^\pi : \mathcal{S} \times \mathcal{A} \to [0,H]^D$ at the h-th step are defined as $\boldsymbol{V}_h^\pi(s) = \mathbb{E}_\pi[\sum_{i=h}^H \boldsymbol{c}_i(s_i,a_i) \mid s_h = s]$ and $\boldsymbol{Q}_h^\pi(s,a) = \mathbb{E}_\pi[\sum_{i=h}^H \boldsymbol{c}_i(s_i,a_i) \mid s_h = s, a_h = a]$ where the expectation $\mathbb{E}_\pi$ is taken with respect to $a_i \sim \pi_i(\cdot \mid s_i)$ and $s_{i+1} \sim \mathcal{P}(\cdot \mid s_i, a_i)$ for $i \in [h:H]$. Note that those value functions are D-dimensional vector-valued, and we use bold letters to represent vectors. We denote their i-th scalar components by $V_h^{i,\pi}$ and $Q_h^{i,\pi}$ respectively. For notational simplicity, we write: $\mathbb{P}_h[\boldsymbol{V}](s,a) = \mathbb{E}_{s' \sim \mathcal{P}_h(\cdot \mid s,a)}[\boldsymbol{V}(s')]$, $\mathbb{B}_h[\boldsymbol{V}](s,a) = \boldsymbol{c}_h(s,a) + \mathbb{P}_h[\boldsymbol{V}](s,a)$, and $\mathbb{D}_\pi[\boldsymbol{Q}](s) = \mathbb{E}_{a \sim \pi(\cdot \mid s)}[\boldsymbol{Q}(s,a)]$. Thereby, the Bellman equation for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$ can be written as $\boldsymbol{Q}_h^\pi(s,a) = \mathbb{B}_h[\boldsymbol{V}_h^\pi](s,a)$, and $\boldsymbol{V}_h^\pi(s) = \mathbb{D}_\pi[\boldsymbol{Q}_h^\pi](s)$. Our goal is to minimize the value of a preference function subject to some constraints. Formally, we define $g : \mathbb{R}^D \to \mathbb{R}$ as the preference function and assume that g is 1-Lipschitz and convex and that $g(\boldsymbol{x}) \geq g(\boldsymbol{x}')$ holds as long as $\boldsymbol{x} \geq \boldsymbol{x}'$. For example, $g(\boldsymbol{x})$ can be the summation over $\{x_i\}_{i=1}^n$ or the maximum among $\{x_i\}_{i=1}^n$. We aim to find the optimal policy π^* of the following constrained optimization problem,

$$\min_{\pi \in \Delta(\mathcal{A} \mid \mathcal{S}, H)} g\big(\boldsymbol{V}_1^\pi(\underline{s})\big) \quad \text{s.t.} \quad \boldsymbol{V}_1^\pi(\underline{s}) \in \mathcal{W}^*, \tag{1}$$

where $\mathcal{W}^* \subseteq [0,H]^D$ is a target set. We use $\boldsymbol{V}_1^*(\underline{s}) = \boldsymbol{V}_1^{\pi^*}(\underline{s})$ to denote the state-value function of the optimal policy of (1) for simplicity.

Solving (1) is impossible when we have no geometric requirements on the target set $\mathcal{W}^*$. Corresponding to the Slater's condition[1] commonly imposed in CMDPs, we establish its geometric analogue for CMOMDPs, as studied in Yu et al. (2021). Let $\mathcal{V}$ be the set of achievable values, i.e., $\mathcal{V} = \{\boldsymbol{V}_1^\pi(\underline{s}) : \text{any policy } \pi\}$, and $\mathcal{W} = \mathcal{V} \cap \mathcal{W}^*$ be the achievable values within the target set. We suppose $\mathcal{W}$ is nonempty. We denote by $\partial\mathcal{W}^*$ and $\partial\mathcal{V}$ the boundaries of $\mathcal{W}^*$ and $\mathcal{V}$, respectively. With a little abuse of notations, we set $\partial\mathcal{W} = \partial\mathcal{W}^* \cap \partial\mathcal{V}$ as the intersection of boundaries. If $\partial\mathcal{W}$ is nonempty, then for each $\boldsymbol{W} \in \partial\mathcal{W}$, we define the angle between support vectors[2] at $\boldsymbol{W}$ as

$$\gamma(\boldsymbol{W}) = \min\big\{\angle(\boldsymbol{a},\boldsymbol{b}) \mid \boldsymbol{a},\boldsymbol{b} \text{ are support vectors of } \mathcal{W}^* \text{ and } \mathcal{V} \text{ at } \boldsymbol{W}, \text{ respectively}\big\}$$

where $\angle(\boldsymbol{a},\boldsymbol{b})$ denotes the angle between $\boldsymbol{a}$ and $\boldsymbol{b}$. We impose geometric requirements on the target set $\mathcal{W}^*$ as stated below.

Assumption 1. *We assume the target set $\mathcal{W}^*$ is closed and convex, and there exists an upper bound $\gamma_{\max} \in [\frac{\pi}{2}, \pi)$ such that $\max_{\boldsymbol{W} \in \partial\mathcal{W}} \gamma(\boldsymbol{W}) < \gamma_{\max}$. Moreover, we assume $\mathcal{W}^*$ is a lower set, i.e., for any $\boldsymbol{W} \in \mathcal{W}^*$, it holds that $\boldsymbol{W}' \in \mathcal{W}^*$ for any $\boldsymbol{0} \leq \boldsymbol{W}' \leq \boldsymbol{W}$.*

We write $\rho = 2/\sin(\gamma_{\max})$ for notational simplicity. Explanations and justifications of Assumption 1 are provided in Appendix B where we show that this assumption is reasonable and closely related to Slater's condition. However, the optimization problem (1) is still hard even with these assumptions. Instead, we solve it by considering the suboptimality and constraint violation as performance metrics, which are defined as

$$\mathrm{SubOpt}(\pi) = g\big(\boldsymbol{V}_1^\pi(\underline{s})\big) - g\big(\boldsymbol{V}_1^*(\underline{s})\big), \quad \mathrm{Violation}(\pi) = \mathrm{dist}\big(\boldsymbol{V}_1^\pi(\underline{s}), \mathcal{W}^*\big), \tag{2}$$

[1] For CMDPs with constraints in the form of $\boldsymbol{V}^\pi \leq \boldsymbol{b}$, the Slater's condition assumes the existence of a policy π such that $\boldsymbol{V}^\pi < \boldsymbol{b}$.

[2] The support vectors at a point are normals to the support hyperplanes at this point.

where we define $\operatorname{dist}(\boldsymbol{V}_1^\pi(\underline{s}), \mathcal{W}^*) = \min_{\boldsymbol{W}\in\mathcal{W}^*} \|\boldsymbol{V}_1^\pi(\underline{s}) - \boldsymbol{W}\|_2$.

To show that the formulation of CMOMDPs is reasonable, we state the relationship between CMOMDPs and CMDPs in the following proposition, which suggests that the CMOMDP is a generalization of the CMDP. The proof is deterred to Appendix C.1.

Proposition 1. *The CMDP is a special case of the CMOMDP, and the Slater's condition assumed in CMDPs corresponds to the existence of $\gamma_{\max}$ in Assumption 1 in CMOMDPs.*

2.2 Offline learning

We consider the offline setting where we only have access to a dataset $\mathcal{D} = \{(s_h^\tau, a_h^\tau, \boldsymbol{c}_h^\tau)\}_{h,\tau=1}^{H,N}$ with N trajectories collected a priori by an experimentor. We only make the following assumption on the data collecting process.

Assumption 2 (Compliance of dataset)**.** *We assume that the dataset $\mathcal{D}$ is compliant with the CMOMDP $\mathcal{M}$, i.e., for any $\boldsymbol{c} \in [0,1]^D, s' \in \mathcal{S}, h \in [H]$, and $\tau \in [N]$,*

$$\Pr_{\mathcal{D}}\big(s_{h+1}^\tau = s', \boldsymbol{c}_h^\tau = \boldsymbol{c} \,\big|\, \mathcal{F}_h^\tau\big) = \Pr\big(s_{h+1} = s', \boldsymbol{c}_h = \boldsymbol{c} \,\big|\, s_h = s_h^\tau, a_h = a_h^\tau\big), \tag{3}$$

where $\mathcal{F}_h^\tau$ is the σ-algebra generated via $\mathcal{F}_h^\tau = \sigma(\{(s_i^n, a_i^n, c_i^n) \,|\, (n-1)H+i \le (\tau-1)H+h\})$. The probability on the left-hand side of (3) *is with respect to the joint distribution of the data collecting process, and that of the right-hand side is with respect to the underlying CMOMDP.*

Assumption 2 is adapted from Jin et al. (2020b). It implies that the dataset $\mathcal{D}$ is generated from the CMOMDP and possesses the Markov property. Such an assumption holds when the experimenter collects the dataset by interacting with the CMOMDP. In particular, we do not require that the data collecting process well explores the state-action space. In other words, we do not impose any assumption on the coverage of the dataset.

3 Pessimistic Dual Iteration (PEDI)

In this section, we first motivate our method to solve CMOMDPs in Section 3.1. Then, we develop our algorithm, Pessimistic Dual Iteration (PEDI), for general CMOMDPs in Section 3.2 and introduce an instantiation for linear kernel CMOMDPs in Section 3.3.

3.1 Primal-dual formulation

To solve problems like (1), recent works usually apply Lagrangian multipliers to handle the constraints. However, we seek to use another dual method. Specifically, we consider the following problem:

$$\mathsf{p}^* = \min_\pi \big(\operatorname{SubOpt}(\pi) + \nu \operatorname{Violation}(\pi)\big), \tag{4}$$

where ν is a scaling constant that serves as a trade-off between suboptimality and constraint violation. In Appendix D, we show that when $\nu > 1$, (1) and (4) share the solution, and thus they are equivalent. We note that the formulation of (4) is different from existing works on CMDPs where ν usually plays the role of the Lagrangian multiplier. Intuitively, a large ν lies more emphasis on satisfying the constraints, while a small ν focus on reducing the suboptimality.

However, p^* is hard to solve in (4), since the relationship between the objective and the policy π is complex. Therefore, we substitute the policy π with the state-value function in the following way,

$$\begin{aligned}\mathsf{p}^* &= \min_\pi \Big(g\big(\boldsymbol{V}_1^\pi(\underline{s})\big) - g\big(\boldsymbol{V}_1^*(\underline{s})\big) + \nu \operatorname{dist}\big(\boldsymbol{V}_1^\pi(\underline{s}), \mathcal{W}^*\big)\Big) \\ &= \min_{\boldsymbol{V}\in\mathcal{V}} \big(g(\boldsymbol{V}) - g\big(\boldsymbol{V}_1^*(\underline{s})\big) + \nu \operatorname{dist}(\boldsymbol{V}, \mathcal{W}^*)\big)\end{aligned} \tag{5}$$

Nevertheless, solving this problem is still hard since the objective is nonlinear in $\boldsymbol{V}$ and thus standard RL cannot apply. Therefore, we utilize the convex conjugate to "linearize" the objective. This technique is widely used in related works (Miryoosefi and Jin, 2021; Yu et al., 2021). Specifically, by convex conjugate, we have

$$g(\boldsymbol{V}) = \max_{\boldsymbol{\beta}\in\mathcal{B}^D} \big(\boldsymbol{\beta}^\top \boldsymbol{V} - g^*(\boldsymbol{\beta})\big), \quad \operatorname{dist}(\boldsymbol{V}, \mathcal{W}^*) = \max_{\boldsymbol{\alpha}\in\mathcal{B}^D} (\boldsymbol{\alpha}^\top \boldsymbol{V} - \max_{\boldsymbol{x}\in\mathcal{W}^*} \boldsymbol{\alpha}^\top \boldsymbol{x}) \tag{6}$$

for any $\boldsymbol{V} \in \mathcal{V}$. The effective domains, $\mathcal{B}^D$, of both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are obtained by the 1-Lipschitzness and Corollary 13.3.3 in Rockafellar (1970). To see why the second equality of (6) holds, we notice that

$$\operatorname{dist}(\boldsymbol{V}, \mathcal{W}^*) = \max_{\boldsymbol{\alpha} \in \mathcal{B}^D} \Big(\boldsymbol{\alpha}^\top \boldsymbol{V} - \underbrace{\max_{\boldsymbol{x} \in \mathbb{R}^D} \big(\boldsymbol{\alpha}^\top \boldsymbol{x} - \operatorname{dist}(\boldsymbol{x}, \mathcal{W}^*) \big)}_{(*)} \Big),$$

and $(*)$ attains the maximum when $\boldsymbol{x} \in \mathcal{W}^*$ since $\boldsymbol{\alpha} \in \mathcal{B}^D$. We call $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ the dual variables throughout this paper. By plugging (6) into (5), we have

$$\mathsf{p}^* = \min_{\boldsymbol{V} \in \mathcal{V}} \max_{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{B}^D} \mathcal{L}(\boldsymbol{V}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{V} - g^*(\boldsymbol{\beta}) - g\big(\boldsymbol{V}_1^*(\underline{s})\big) + \nu \boldsymbol{\alpha}^\top \boldsymbol{V} - \nu \max_{\boldsymbol{x} \in \mathcal{W}^*} \boldsymbol{\alpha}^\top \boldsymbol{x}. \tag{7}$$

Its dual problem can be written as $\mathsf{d}^* = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{B}^D} \mathsf{D}(\boldsymbol{\alpha}, \boldsymbol{\beta})$, where $\mathsf{D}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is the dual function defined as $\mathsf{D}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\boldsymbol{V} \in \mathcal{V}} \mathcal{L}(\boldsymbol{V}; \boldsymbol{\alpha}, \boldsymbol{\beta})$. We notice that $\mathcal{L}(\boldsymbol{V}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ is convex in $\boldsymbol{V} \in \mathcal{V}$ and concave in $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{B}^D$, which implies the strong duality, $\mathsf{p}^* = \mathsf{d}^*$. Hence, dual methods can apply.

When $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are fixed, solving $\mathsf{D}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is a planning problem. When the model is known, it suffices to conduct value iteration on the model. However, what we have is a dataset collected a priori in offline RL. Therefore, we develop the offline planning algorithm that is a pessimistic variant of the value iteration (Jin et al., 2020b), which applies penalties to ensure pessimism. We explain why pessimism is crucial to offline CMOMDP in Appendix F. We now describe the high-level intuition of our approach. Consider a meta-algorithm that constructs the empirical transition kernel $\widehat{\mathcal{P}}$ and the empirical cost function $\widehat{\boldsymbol{c}}$ based on the dataset $\mathcal{D}$. As defined below, we introduce the uncertainty quantifier that bounds the estimation error from above with a confidence parameter $\xi \in (0, 1)$.

Definition 1 (ξ-uncertainty quantifier)**.** *We call $\{(\Gamma_h^{\mathcal{P}}, \Gamma_h^{\boldsymbol{c}})\}_{h=1}^H$ a ξ-uncertainty quantifier if the following event*

$$\begin{aligned} \mathcal{E} = \Big\{ & \big|\widehat{\mathcal{P}}_h(s'|s,a) - \mathcal{P}_h(s'|s,a)\big| \le \Gamma_h^{\mathcal{P}}(s,a,s'),\ \big|\widehat{c}_h^i(s,a) - c_h^i(s,a)\big| \le \Gamma_h^{c^i}(s,a) \\ & \text{for all } (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}, h \in [H], i \in [D] \Big\} \end{aligned} \tag{8}$$

holds with probability at least $1 - \xi$. Here we write $\Gamma_h^{\boldsymbol{c}} = (\Gamma_h^{c^1}, \Gamma_h^{c^2}, \ldots, \Gamma_h^{c^D})$.

Then, we can construct the empirical counterparts of $\mathbb{P}_h$ and $\mathbb{B}_h$ by $\widehat{\mathbb{P}}_h[\boldsymbol{V}](s,a) = \langle \boldsymbol{V}(\cdot), \widehat{\mathcal{P}}_h(\cdot \,|\, s,a) \rangle_{\mathcal{S}}$ and $\widehat{\mathbb{B}}_h[\boldsymbol{V}](s,a) = \widehat{\boldsymbol{c}}_h(s,a) + \widehat{\mathbb{P}}_h[\boldsymbol{V}](s,a)$. Note that for any function $V : \mathcal{S} \to [0, H]$, under the event $\mathcal{E}$, it holds for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and any $h \in [H]$ that

$$\big|(\widehat{\mathbb{P}}_h - \mathbb{P}_h)[V](s,a)\big| = \left| \int_{\mathcal{S}} \big(\widehat{\mathcal{P}}_h(s'|s,a) - \mathcal{P}_h(s'|s,a)\big) V(s') \,\mathrm{d}s' \right| \le H \int_{\mathcal{S}} \Gamma_h^{\mathcal{P}}(s,a,s') \,\mathrm{d}s'.$$

Therefore, we define $\Gamma_h(s,a) = H \int_{\mathcal{S}} \Gamma_h^{\mathcal{P}}(s,a,s') \,\mathrm{d}s'$, which quantifies the uncertainty arising from the transition kernel. We use the ξ-uncertainty quantifier as the penalty function for pessimistic planning, which leads to a conservative estimation of the value function. We formally describe our planning method in Algorithm 1, where $\boldsymbol{\theta}$ denotes a projection vector to be specified later. By adding the pessimistic penalty $\Gamma_h \cdot \mathbf{1} + \Gamma_h^{\boldsymbol{c}}$ to the estimated value, we ensure a conservative estimation. We also truncate $\boldsymbol{Q}_h^k$ and $\boldsymbol{V}_h^k$ in $[0, H - h + 1]^D$ to improve the estimation.

It remains to solve $\max_{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{B}^D} \mathsf{D}(\boldsymbol{\alpha}, \boldsymbol{\beta})$. We utilize the projected subgradient method which produces a sequence of dual variables $\{\boldsymbol{\alpha}^k\}_{k=1}^K$ and $\{\boldsymbol{\beta}^k\}_{k=1}^K$ that approximately solves the dual problem. A detailed description of projected subgradient method is provided in Appendix H.6.2. For a brief description, we update the dual variables via

$$\begin{aligned} \boldsymbol{\alpha}^{k+1} &\leftarrow \Pi_{\mathcal{B}^D}\big\{\boldsymbol{\alpha}^k + \eta^k (\boldsymbol{V}^k - \operatorname*{arg\,max}_{\boldsymbol{x} \in \mathcal{W}^*} (\boldsymbol{\alpha}^k)^\top \boldsymbol{x})\big\}, \\ \boldsymbol{\beta}^{k+1} &\leftarrow \Pi_{\mathcal{B}^D}\Big\{\boldsymbol{\beta}^k + \eta^k \big(\boldsymbol{V}^k - \partial g^*(\boldsymbol{\beta}^k)\big)\Big\}. \end{aligned} \tag{9}$$

where $\boldsymbol{V}^k \in \arg\min_{\boldsymbol{V} \in \mathcal{V}} \mathcal{L}(\boldsymbol{V}; \boldsymbol{\alpha}^k, \boldsymbol{\beta}^k)$ and η^k is the step length at the k-th iteration.

As claimed in Proposition 1, the CMDP is a special case of the CMOMDP. Moreover, we show in Appendix C.2 that when reducing to the CMDP, the objective (7) reduces to the Lagrangian formulation of the CMDP problem. Hence, our proposed method has good generalization ability.

Algorithm 1 Pessimistic planning.

function PESSPLANNING($\boldsymbol{\theta}, \mathcal{D}, \{(\Gamma_h^{\mathcal{P}}, \Gamma_h^{\mathbf{c}})\}_{h=1}^H$)
  for step $h = H, H-1, \ldots, 1$ **do**
    $\overline{\boldsymbol{Q}}_h(\cdot,\cdot) \leftarrow \widehat{\boldsymbol{c}}_h(\cdot,\cdot) + \widehat{\mathbb{P}}_h[\boldsymbol{V}_{h+1}](\cdot,\cdot) + \Gamma_h \cdot \mathbf{1} + \Gamma_h^{\mathbf{c}}$
    $\boldsymbol{Q}_h(\cdot,\cdot) \leftarrow \min\{\overline{\boldsymbol{Q}}_h(\cdot,\cdot), (H-h+1)\cdot\mathbf{1}\}_+$
    $\pi_h(\cdot) \leftarrow \arg\min_{a\in\mathcal{A}} \boldsymbol{Q}_h(\cdot,\cdot)\cdot\boldsymbol{\theta}$
    $\boldsymbol{V}_h(\cdot) \leftarrow \mathbb{D}_{\pi_h}[\boldsymbol{Q}_h](\cdot)$
  end for
  return $\pi = \{\pi_h\}_{h=1}^H, \boldsymbol{Q} = \{\boldsymbol{Q}_h\}_{h=1}^H, \boldsymbol{V} = \{\boldsymbol{V}_h\}_{h=1}^H$
end function

3.2 General CMOMDPs

Based on the above analysis, we now develop the Pessimistic Dual Iteration (PEDI) (Algorithm 2) for general CMOMDPs, which is motivated by the dual gradient method in solving the minimax optimization problem. Specifically, PEDI solves $\mathsf{D}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ via PESSPLANNING (Algorithm 1) and updates dual variables via (9), alternately. The output policy $\widehat{\pi}$ is a mixed policy executed by randomly selecting a policy π^k from $\{\pi^k\}_{k=1}^K$ with equal probability beforehand and then exclusively following π^k thereafter. Hence, it holds that, for the mixed policy $\widehat{\pi}$, $\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s}) = \frac{1}{K}\sum_{k=1}^K \boldsymbol{V}_1^{\pi^k}(\underline{s})$. Note that the mixed policy may not be a Markov policy (Altman, 1999). However, it would be useful to extend the definition of a policy $\pi = \{\pi_h\}_{h=1}^H$ so as to allow π_h to depend not only on h but also on some initial randomizing mechanism. That is, we may have a set of policies $\mathcal{U}$ and a distribution $q \in \Delta(\mathcal{U})$. Then we define the mixed policy of $\mathcal{U}$, $\pi_{\mathcal{U}}$, as a policy to be executed by first using q to choose some policy $\pi \in \mathcal{U}$ and then proceeding with only that policy. The optima of (1) over the mixed policy is still attained by the Markov policy π^*. Thus, $\mathrm{SubOpt}(\widehat{\pi})$ remains a reasonable performance metric of our algorithm.

Algorithm 2 Pessimistic Dual Iteration (PEDI).

Input: offline dataset $\mathcal{D}$, step length $\{\eta^k\}_{k=1}^K$, scaling parameter ν
Output: the mixed policy $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$ of $\{\pi^k\}_{k=1}^K$

randomly initialize dual variables $\boldsymbol{\alpha}^1, \boldsymbol{\beta}^1 \in \mathcal{B}^D$, and set $\boldsymbol{\theta}^1 = \nu\boldsymbol{\alpha}^1 + \boldsymbol{\beta}^1$
construct $\{\widehat{\boldsymbol{c}}_h\}_{h=1}^H$ and $\{\widehat{\mathcal{P}}_h\}_{h=1}^H$ based on $\mathcal{D}$
construct the ξ-uncertainty quantifier $\{(\Gamma_h^{\mathcal{P}}, \Gamma_h^{\mathbf{c}})\}_{h=1}^H$ based on $\mathcal{D}$
for iteration $k = 1, 2, \ldots, K$ **do**
  $\pi^k, \boldsymbol{Q}^k, \boldsymbol{V}^k \leftarrow$ PESSPLANNING$\Big(\boldsymbol{\theta}^k, \mathcal{D}, \{(\Gamma_h^{\mathcal{P}}, \Gamma_h^{\mathbf{c}})\}_{h=1}^H\Big)$
  update $\boldsymbol{\alpha}^k, \boldsymbol{\beta}^k$ to $\boldsymbol{\alpha}^{k+1}, \boldsymbol{\beta}^{k+1}$ via (9)
  $\boldsymbol{\theta}^{k+1} \leftarrow \nu\boldsymbol{\alpha}^{k+1} + \boldsymbol{\beta}^{k+1}$
end for

3.3 Linear kernel CMOMDPs

In this section, we study PEDI on settings with linear function approximation such as linear kernel CMOMDPs. To that end, we first introduce the linear kernel CMOMDP, which is a generalization of the linear kernel MDP (Yang and Wang, 2019; Jin et al., 2020a; Cai et al., 2020). Then, we propose an instantiation of Algorithm 2 for linear kernel CMOMDPs.

Definition 2 (Linear kernel CMOMDP)**.** *The CMOMDP $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, \mathbf{c})$ is a linear kernel CMOMDP with a known kernel feature map $\psi : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}^{d_1}$ and a known value feature map $\varphi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^{d_2}$, if for any $h \in [H]$ there exist unknown vectors $\theta_h \in \mathbb{R}^{d_1}$ and $\theta_h^{c^i} \in \mathbb{R}^{d_2}$ for any $i \in [D]$ such that for any $(s, a, s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$,*

$$\mathcal{P}_h(s' \mid s, a) = \psi(s, a, s')^\top \theta_h, \quad c_h^i(s, a) = \varphi(s, a)^\top \theta_h^{c^i}.$$

Let $d = \max\{d_1, d_2\}$. We assume there exists a constant $R > 0$ such that

$$R^{-2} \max_{s' \in \mathcal{S}} \left(y^\top \psi(s,a,s')\right)^2 \le \int_{\mathcal{S}} \left(y^\top \psi(s,a,s')\right)^2 \mathrm{d}s' \le d$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $y : \|y\|_2 \le 1$. Moreover, we assume $\|\theta_h\|_2 \le \sqrt{d}$ and $\|\theta_h^{c^i}\|_2 \le \sqrt{d}$ for $i \in [d]$.

The existence of constant R naturally holds when $\psi(s,a,\cdot)$ is upper bounded and Lipschitz continuous for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Note that the tabular CMOMDP is a special case of the linear kernel CMOMDP with $R = 1$ (see Appendix G for a detailed discussion).

An instantiation of PEDI for linear kernel CMOMDPs is deferred to Appendix E where we propose a method to estimate the empirical transition kernel $\widehat{\mathcal{P}}$ and cost function $\widehat{c}$ and thereby construct a ξ-uncertainty quantifier.

4 Theoretical results

In this section, we first show upper bounds for suboptimality and constraint violation for PEDI on general CMOMDPs in Section 4.1. In Section 4.2, we show that PEDI achieves the minimax optimality up to the optimization error and multiplicative factors of constants related to the CMOMDP when specified to linear kernel CMOMDPs. Moreover, when the dataset has sufficient coverage over the trajectories induced by the optimal policy, the intrinsic uncertainty is guaranteed to be $\widetilde{\mathcal{O}}(1/\sqrt{N})$ where N is the number of trajectories in the dataset.

4.1 Results of general CMOMDPs

By using pessimistic planning (Algorithm 1), we ensure a pessimistic policy $\widehat{\pi}$. We formally state this notion in the following lemma.

Lemma 1 (Pessimism)**.** *Suppose that Assumptions 1 and 2 hold. By Algorithm 2, under event $\mathcal{E}$ defined in* (8)*, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$, it holds that $\boldsymbol{V}_h^k(s) \ge \boldsymbol{V}_h^{\pi^k}(s)$, $\boldsymbol{Q}_h^k(s,a) \ge \boldsymbol{Q}_h^{\pi^k}(s,a)$, and $\operatorname{dist}\left(\boldsymbol{V}_1^k(\underline{s}), \mathcal{W}^*\right) \ge \operatorname{dist}\left(\boldsymbol{V}_1^{\pi^k}(\underline{s}), \mathcal{W}^*\right)$. Furthermore, it holds that*

$$g\big(\widehat{\boldsymbol{V}}_1(\underline{s})\big) \ge g\big(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s})\big) \quad \text{and} \quad \operatorname{dist}\left(\widehat{\boldsymbol{V}}_1(\underline{s}), \mathcal{W}^*\right) \ge \operatorname{dist}\left(\boldsymbol{V}_1^*(\underline{s}), \mathcal{W}^*\right),$$

where $\widehat{\boldsymbol{V}}_1(\underline{s}) = \frac{1}{K}\sum_{k=1}^K \boldsymbol{V}_1^k(\underline{s})$ and $\widehat{\pi}$ is the output of Algorithm 2.

Proof of Lemma 1. See Appendix H.1 for a detailed proof. □

We are now ready to establish the following theorem, which characterizes the suboptimality and constraint violation of PEDI for general offline CMOMDPs.

Theorem 1 (Suboptimality and constraint violation for general CMOMDPs)**.** *Suppose that Assumptions 1 and 2 hold. Under event $\mathcal{E}$ defined in* (8)*, for the output policy $\widehat{\pi}$ of Algorithm 2, when we set $\nu = \rho$ and $\eta^k = 2G^{-1}\sqrt{D/k}$ (or $2G^{-1}\sqrt{D/K}$ if K is predefined), where $G = 2(1+\nu)H\sqrt{D}$, it holds that*

$$\mathrm{SubOpt}(\widehat{\pi}) \le \epsilon_K + \mathrm{IntUncert}_{\mathcal{D}}^{\pi^*}, \quad \mathrm{Violation}(\widehat{\pi}) \le \frac{2}{\rho}(\epsilon_K + \mathrm{IntUncert}_{\mathcal{D}}^{\pi^*}), \tag{10}$$

where we let C denote an absolute constant and define

$$\epsilon_K = C(1+\rho)\sqrt{\frac{DH^2}{K}}, \quad \mathrm{IntUncert}_{\mathcal{D}}^{\pi^*} = 2(1+\rho)\sqrt{D}\sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\Gamma_h(s_h,a_h) + \|\Gamma_h^{\boldsymbol{c}}(s_h,a_h)\|_\infty \mid s_1 = \underline{s}\right].$$

Proof of Theorem 1. See Appendix H.2 for a detailed proof. Here we provide a proof sketch, which consists of two steps.

First step: upper bounding $\mathrm{SubOpt}(\hat{\pi}) + \nu\,\mathrm{Violation}(\hat{\pi})$, which is our objective (4). Here $\hat{\pi}$ is the output of PEDI (Algorithm 2). By pessimism, we have $\mathrm{SubOpt}(\hat{\pi}) + \nu\,\mathrm{Violation}(\hat{\pi}) \le$

$g(\widehat{\boldsymbol{V}}_1(\underline{s})) - g\left(\boldsymbol{V}_1^*(\underline{s})\right) + \nu \operatorname{dist}(\widehat{\boldsymbol{V}}_1(\underline{s}), \mathcal{W}^*)$ where $\widehat{\boldsymbol{V}}_1(\underline{s}) = \frac{1}{K}\sum_{k=1}^K \boldsymbol{V}_1^k(\underline{s})$. We do linearization via convex conjugate (6) and then get

$$\begin{aligned}\operatorname{SubOpt}(\hat{\pi}) + \nu \operatorname{Violation}(\hat{\pi}) \leq & K^{-1} \max_{\|\boldsymbol{\beta}\| \leq 1} \left\{ \boldsymbol{\beta} \cdot \sum_{k=1}^K \boldsymbol{V}_1^k(\underline{s}) - \sum_{k=1}^K g^*(\boldsymbol{\beta}) \right\} - g\left(\boldsymbol{V}_1^*(\underline{s})\right) \\ & + K^{-1}\nu \max_{\|\boldsymbol{\alpha}\| \leq 1} \left\{ \boldsymbol{\alpha} \cdot \sum_{k=1}^K \boldsymbol{V}_1^k(\underline{s}) - \sum_{k=1}^K \max_{\boldsymbol{x} \in \mathcal{W}^*} \boldsymbol{\alpha} \cdot \boldsymbol{x} \right\}.\end{aligned}$$

Recall that PEDI uses the projected subgradient method, which approximates the max operator above by a sequence of variables at the expense of a sublinear approximation error, which leads to

$$\begin{aligned}\operatorname{SubOpt}(\hat{\pi}) + \nu \operatorname{Violation}(\hat{\pi}) \leq & K^{-1} \sum_{k=1}^K \left\{ \left(\boldsymbol{\beta}^k\right)^\top \boldsymbol{V}_1^k(\underline{s}) - g^*\left(\boldsymbol{\beta}^k\right) \right\} - g\left(\boldsymbol{V}_1^*(\underline{s})\right) \\ & + K^{-1}\nu \sum_{k=1}^K \left\{ \left(\boldsymbol{\alpha}^k\right)^\top \boldsymbol{V}_1^k(\underline{s}) - \max_{\boldsymbol{x} \in \mathcal{W}^*} \left(\boldsymbol{\alpha}^k\right)^\top \boldsymbol{x} \right\} \\ & + C(1+\nu)\sqrt{DH^2/K},\end{aligned}$$

where C is a constant. We rearrange the right-hand side above with some tricks and then obtain

$$\operatorname{SubOpt}(\hat{\pi}) + \nu \operatorname{Violation}(\hat{\pi}) \leq K^{-1} \sum_{k=1}^K \left[\left(\boldsymbol{\theta}^k\right)^\top \left(\boldsymbol{V}_1^k(\underline{s}) - \boldsymbol{V}_1^*(\underline{s}) \right) \right] + C(1+\nu)\sqrt{DH^2/K}.$$

The last term in the right hand side above is exactly ϵ_K. Intuitively, the term $(\boldsymbol{\theta}^k)^\top(\boldsymbol{V}_1^k(\underline{s}) - \boldsymbol{V}_1^*(\underline{s}))$ indicates the projected difference of state-value functions along $\boldsymbol{\theta}$, which motivates the PESSPLANNING (Algorithm 1). We apply results from this pessimistic approach and then conclude

$$\operatorname{SubOpt}(\hat{\pi}) + \nu \operatorname{Violation}(\hat{\pi}) = g\left(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s})\right) - g\left(\boldsymbol{V}_1^*(\underline{s})\right) + \nu \operatorname{dist}\left(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s}), \mathcal{W}^*\right) \leq \epsilon_K + \operatorname{IntUncert}_{\mathcal{D}}^{\pi^*}.$$

Second step: upper bounding $\operatorname{SubOpt}(\hat{\pi})$ and $\operatorname{Violation}(\hat{\pi})$ individually. This is done by two facts: (i) $\operatorname{dist}(\cdot,\cdot) \geq 0$, and (ii) $g(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s})) - g\left(\boldsymbol{V}_1^*(\underline{s})\right) \geq -\nu \operatorname{dist}(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s}), \boldsymbol{V}_1^*(\underline{s}))/2$. We plug them into the resulting inequality in the first step and then complete the proof. □

As shown in Theorem 1, our pessimistic algorithm bounds the suboptimality and constraint violation from above by the sum of the optimization error $\epsilon_K \in \mathcal{O}(1/\sqrt{K})$ and the intrinsic uncertainty $\operatorname{IntUncert}_{\mathcal{D}}^{\pi^*}$. The optimization error ϵ_K vanishes as K goes into infinity. Later, we will show that the intrinsic uncertainty, $\operatorname{IntUncert}_{\mathcal{D}}^{\pi^*}$, is impossible to eliminate through an analysis of PEDI on linear kernel CMOMDPs in Section 4.2.

4.2 Results of linear kernel CMOMDPs

The following theorem characterizes the suboptimality and constraint violation of PEDI for linear kernel CMOMDPs. Recall that we instantiate PEDI to linear kernel CMOMDPs in Appendix E.

Theorem 2 (Suboptimality and constraint violation for linear kernel CMOMDPs)**.** *We set $\lambda = 1$ and $\kappa = CR\sqrt{d\log(dN) + \log(DH/\xi)}$ where $C > 0$ is an absolute constant. Then, under Assumptions 1 and 2, $\{\Gamma_h^{\mathcal{P}}, \Gamma_h^r\}_{h=1}^H$ defined in* (24) *(Appendix E) is a ξ-uncertainty quantifier. Hence, under the same settings of ν and η^k in Theorem 1,* (10) *holds with probability at least $1-\xi$ for the output $\widehat{\pi}$, and the intrinsic uncertainty is specified to*

$$\operatorname{IntUncert}_{\mathcal{D}}^{\pi^*} = 2(1+\rho)\sqrt{D}\sum_{h=1}^H \mathbb{E}_{\pi^*}\left[H \int_{\mathcal{S}} \kappa \cdot \|\psi(s_h, a_h, s')\|_{\Lambda_h^{-1}} \,\mathrm{d}s' + \kappa \cdot \|\varphi(s_h, a_h)\|_{\Lambda_{\varphi,h}^{-1}} \middle| s_1 = \underline{s} \right].$$

Proof of Theorem 2. See Appendix H.3 for a detailed proof. □

In what follows, we show that when the dataset sufficiently covers the trajectories of the optimal policy, the intrinsic uncertainty is $\mathcal{O}(1/\sqrt{N})$ where N is the number of trajectories in the dataset.

Corollary 1 (Dataset with sufficient coverage)**.** *Suppose that the τ-th trajectory $\{(s_h^\tau, a_h^\tau, \boldsymbol{c}_h^\tau)\}_{h=1}^H$ of the dataset $\mathcal{D}$ is generated by a behavior policy $\pi^{\mathrm{b},\tau}$ for $\tau \in [N]$. Meanwhile, we assume that there exists a constant $\varsigma > 0$ such that $\mu_h^*(s,a)/\mu_h^{\mathrm{b},\tau}(s,a) \le \varsigma$ for any $h \in [H], \tau \in [N]$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$. Here $\{\mu_h^*\}_{h=1}^H$ and $\{\mu_h^{\mathrm{b},\tau}\}_{h=1}^H$ are the visitation measures of π^* and $\pi^{\mathrm{b},\tau}$. Then, we have with probability at least $1-\xi$ that*

$$\mathrm{IntUncert}_{\mathcal{D}}^{\pi^*} \in \widetilde{\mathcal{O}}\big(C\varsigma\kappa(1+\rho)H^2\sqrt{d|\mathcal{S}|/N}\big).$$

Proof of Corollary 1. See Appendix H.5 for a detailed proof. □

Here $|\mathcal{S}|$ denotes a certain measure on $\mathcal{S}$. For instance, when $\mathcal{S} = [0,1]$ and we take the Lebesgue measure, we have $|\mathcal{S}| = 1$. Note that Corollary 1 holds under a weaker condition than the uniform coverage (Wang et al., 2020a; Rashidinejad et al., 2021), where the condition is $\max_\pi \mu^\pi/\mu^{\mathrm{b}} \le C$ for a certain constant C. However, our algorithm is still able to achieve the intrinsic uncertainty of $\widetilde{O}(\sqrt{1/N})$. In particular, our condition, $\mu^*/\mu^{\mathrm{b}} \le \varsigma$, can hold when we have expert demonstration in the dataset $\mathcal{D}$ (Buckman et al., 2020).

In Theorem 2, we notice that when the number of iteration K goes into infinity, the optimization error ϵ_K vanishes, and the suboptimality and constraint violation are bounded from above by the intrinsic uncertainty $\mathrm{IntUncert}_{\mathcal{D}}^{\pi^*}$ only. However, this is the best effort that we can put. The following theorem indicates that the term $\mathrm{IntUncert}_{\mathcal{D}}^{\pi^*}$ is impossible to eliminate since it arises from the information-theoretic lower bound.

Theorem 3 (Information-theoretic lower bound)**.** *For the output $\widehat{\pi}$ of any algorithm of offline CMOMDPs, there exists a linear kernel CMOMDP $\mathcal{M}$ with initial state $\underline{s} \in \mathcal{S}$ and a dataset $\mathcal{D}$ compliant with $\mathcal{M}$, such that*

$$\mathbb{E}_{\mathcal{D}}\left[\frac{\mathrm{SubOpt}(\widehat{\pi})}{\sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\|\varphi(s_h,a_h)\|_{\Lambda_{\varphi,h}^{-1}} + |\mathcal{S}|^{-1}\int_{\mathcal{S}}\|\psi(s_h,a_h,s')\|_{\Lambda_h^{-1}}\,\mathrm{d}s' \,\middle|\, s_1 = \underline{s}\right]}\right] \ge c,$$

where the expectation is taken with respect to the data collecting process, and $c > 0$ is an absolute constant.

Proof of Theorem 3. See Appendix H.4 for a detailed proof. □

By Theorem 2 and Theorem 3, we find that the upper bound of suboptimality of Algorithm 2 for linear kernel CMOMDP matches the information-theoretic lower bound up to the optimization error $\mathcal{O}(1/\sqrt{K})$ and some multiplicative factors of horizons, dimensions, and geometric properties of the target set, which means that PEDI is near-optimal for linear kernel CMOMDP.

5 Experimental Results

We also conduct numerical experiments to verify our theory. Please see Appendix I for details. The results show that the proposed algorithm is not only provably efficient but also applicable.

6 Conclusion

In this paper, we propose a general algorithm, Pessimistic Dual Iteration (PEDI), for constrained multi-objective reinforcement learning with offline data. We show our algorithm delivers a data-dependent upper bound for suboptimality and constraint violation. The upper bound is composed of the optimization error and the intrinsic uncertainty. Moreover, we show that the bound of suboptimality is minimax optimal for linear kernel CMOMDPs up to the optimization error and multiplicative factors of constants related to the CMOMDP. When the trajectories of optimal policy are sufficiently covered, the intrinsic uncertainty is proved to be $\widetilde{\mathcal{O}}(1/\sqrt{N})$ where N is the number of trajectories in the dataset.

Acknowledgments and Disclosure of Funding

We thank all anonymous reviewers for their useful comments. Zhaoran Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, and Two Sigma for their supports. Zhuoran Yang acknowledges Simons Institute (Theory of Reinforcement Learning).

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *NIPS*, volume 11, pages 2312–2320.

Agarwal, R., Schuurmans, D., and Norouzi, M. (2020). An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pages 104–114. PMLR.

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

Antos, A., Szepesvári, C., and Munos, R. (2008). Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129.

Barrett, L. and Narayanan, S. (2008). Learning all optimal policies with multiple criteria. In *Proceedings of the 25th international conference on Machine learning*, pages 41–47.

Bäuerle, N. and Rieder, U. (2010). Markov decision processes. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 112(4):217–243.

Bhatnagar, S. and Lakshmanan, K. (2012). An online actor–critic algorithm with function approximation for constrained markov decision processes. *Journal of Optimization Theory and Applications*, 153(3):688–708.

Brantley, K., Dudik, M., Lykouris, T., Miryoosefi, S., Simchowitz, M., Slivkins, A., and Sun, W. (2020). Constrained episodic reinforcement learning in concave-convex and knapsack settings. *arXiv preprint arXiv:2006.05051*.

Buckman, J., Gelada, C., and Bellemare, M. G. (2020). The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. (2020). Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR.

Castelletti, A., Pianosi, F., and Restelli, M. (2011). Multi-objective fitted q-iteration: Pareto frontier approximation in one single run. In *2011 International Conference on Networking, Sensing and Control*, pages 260–265. IEEE.

Castelletti, A., Pianosi, F., and Restelli, M. (2012). Tree-based fitted q-iteration for multi-objective markov decision problems. In *The 2012 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

Chakraborty, B. and Murphy, S. A. (2014). Dynamic treatment regimes. *Annual review of statistics and its application*, 1:447–464.

Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR.

Cheung, W. C. (2019). Regret minimization for reinforcement learning with vectorial feedback and complex objectives. *Advances in Neural Information Processing Systems*, 32:726–736.

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.

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

Efroni, Y., Mannor, S., and Pirotta, M. (2020). Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*.

Farahmand, A.-m., Ghavamzadeh, M., Szepesvári, C., and Mannor, S. (2016). Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874.

Farahmand, A. M., Munos, R., and Szepesvári, C. (2010). Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*.

Fu, Z., Yang, Z., and Wang, Z. (2020). Single-timescale actor-critic provably finds globally optimal policy. *arXiv preprint arXiv:2008.00483*.

Fujimoto, S., Meger, D., and Precup, D. (2019). Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062. PMLR.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and Celi, L. A. (2019). Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18.

Jaques, N., Ghandeharioun, A., Shen, J. H., Ferguson, C., Lapedriza, A., Jones, N., Gu, S., and Picard, R. (2019). Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020a). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR.

Jin, Y., Yang, Z., and Wang, Z. (2020b). Is pessimism provably efficient for offline rl? *arXiv preprint arXiv:2012.15085*.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. (2020). Morel: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*.

Kumar, A., Fu, J., Tucker, G., and Levine, S. (2019). Stabilizing off-policy q-learning via bootstrapping error reduction. *arXiv preprint arXiv:1906.00949*.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020). Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*.

Lange, S., Gabel, T., and Riedmiller, M. (2012). Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer.

Laroche, R., Trichelair, P., and Des Combes, R. T. (2019). Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pages 3652–3661. PMLR.

Le, H., Voloshin, C., and Yue, Y. (2019). Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR.

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. (2020). Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*.

Miryoosefi, S. and Jin, C. (2021). A simple reward-free approach to constrained reinforcement learning. *arXiv preprint arXiv:2107.05216*.

Nair, A., Dalal, M., Gupta, A., and Levine, S. (2020). Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*.

Natarajan, S. and Tadepalli, P. (2005). Dynamic preferences in multi-criteria reinforcement learning. In *Proceedings of the 22nd International Conference on Machine learning*, pages 601–608.

Paternain, S., Calvo-Fullana, M., Chamon, L. F., and Ribeiro, A. (2019). Safe policies for reinforcement learning via primal-dual methods. *arXiv preprint arXiv:1911.09101*.

Qiu, S., Wei, X., Yang, Z., Ye, J., and Wang, Z. (2020). Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. *arXiv e-prints*, pages arXiv–2003.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*.

Rockafellar, R. T. (1970). *Convex analysis*, volume 36. Princeton university press.

Scherrer, B., Ghavamzadeh, M., Gabillon, V., Lesner, B., and Geist, M. (2015). Approximate modified policy iteration and its application to the game of tetris. *J. Mach. Learn. Res.*, 16:1629–1676.

Shalev-Shwartz, S., Shammah, S., and Shashua, A. (2016). Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*.

Siegel, N. Y., Springenberg, J. T., Berkenkamp, F., Abdolmaleki, A., Neunert, M., Lampe, T., Hafner, R., Heess, N., and Riedmiller, M. (2020). Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *arXiv preprint arXiv:2002.08396*.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al. (2020). Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454.

Tessler, C., Mankowitz, D. J., and Mannor, S. (2018). Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*.

Tosatto, S., Pirotta, M., d'Eramo, C., and Restelli, M. (2017). Boosted fitted q-iteration. In *International Conference on Machine Learning*, pages 3434–3443. PMLR.

Van Moffaert, K., Drugan, M. M., and Nowé, A. (2013). Scalarized multi-objective reinforcement learning: Novel design techniques. In *2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 191–199. IEEE.

Wang, R., Foster, D. P., and Kakade, S. M. (2020a). What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*.

Wang, W. and Sebag, M. (2013). Hypervolume indicator and dominance reward based multi-objective monte-carlo tree search. *Machine learning*, 92(2-3):403–429.

Wang, Z., Novikov, A., Żołna, K., Springenberg, J. T., Reed, S., Shahriari, B., Siegel, N., Merel, J., Gulcehre, C., Heess, N., et al. (2020b). Critic regularized regression. *arXiv preprint arXiv:2006.15134*.

Wu, Y., Tucker, G., and Nachum, O. (2019). Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*.

Yang, L. and Wang, M. (2019). Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR.

Yin, M., Bai, Y., and Wang, Y.-X. (2021). Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1567–1575. PMLR.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., and Ma, T. (2020). Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*.

Yu, T., Tian, Y., Zhang, J., and Sra, S. (2021). Provably efficient algorithms for multi-objective competitive rl. *arXiv preprint arXiv:2102.03192*.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936.

A Generalization of the definition

Adapted from Bäuerle and Rieder (2010), a more general definition of an episodic CMOMDP with horizon $H \in \mathbb{N}$ is given by a sequence of sets of data $\{(\mathcal{S}, \mathcal{A}, \mathcal{X}_h, \mathrm{Pr}_h, \boldsymbol{c}_h)\}_{h=1}^{H}$, where

- $\mathcal{S}$ is the state space equipped with a σ-algebra $\mathcal{F}_{\mathcal{S}}$.
- $\mathcal{A}$ is the action space equipped with a σ-algebra $\mathcal{F}_{\mathcal{A}}$.
- $\mathcal{X}_h$ is a measurable subset of $\mathcal{X} = \mathcal{S} \times \mathcal{A}$, which denotes the set of accessible state-action pairs at time h.
- Pr_h is a stochastic transition kernel at time h. For any fixed $(s, a) \in \mathcal{X}_h$, the mapping $S \mapsto \mathrm{Pr}_h(S \,|\, s, a)$ is a probability measure on $\mathcal{S}$. Moreover, the mapping $(s, a) \mapsto \mathrm{Pr}_h(S \,|\, s, a)$ is measurable with respect to (s, a) for all $S \in \mathcal{F}_{\mathcal{S}}$. Intuitively, $\mathrm{Pr}_h(S \,|\, s, a)$ gives the probability that the next state is in S if the current state is s and action a is taken at time h.
- $\boldsymbol{c}_h : \mathcal{X}_h \to \mathbb{R}^D$ is a measurable function, which defines the expected cost at time h if the current state is s and action a is taken.

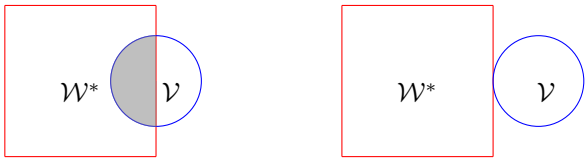
For any $h \in [H]$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$, we define $\mathcal{P}_h(\cdot \,|\, s, a)$ as the density of $\mathrm{Pr}_h(\cdot \,|\, s, a)$ relative to a certain measure μ by the Radon–Nikodym derivative, i.e., $\mathcal{P}_h = \mathrm{d}\,\mathrm{Pr}_h / \mathrm{d}\mu$. We usually let μ be the Lebesgue measure when the state is continuous and be the counting measure when $\mathcal{S}$ is discrete. For example, for measurable real-valued function f, when $\mathcal{S}$ is discrete, $\mathcal{P}_h(\cdot \,|\, s, a)$ is the probability mass function and $\int_{\mathcal{S}} f \,\mathrm{d}\mu$ coincides with $\sum_{s \in \mathcal{S}} f(s)$. When $\mathcal{S}$ is continuous, $\mathcal{P}_h(\cdot \,|\, s, a)$ is the probability density function and $\int_{\mathcal{S}} f \,\mathrm{d}\mu$ denotes the Lebesgue integration. Our proposed method is able to handle both of the above situations. For notational simplicity, we omit the dependence on μ through the paper.

B Explanations and justifications of Assumption 1

In this section, we explain and justify Assumption 1 imposed on the target set. This assumption is a stronger version of the one used in Yu et al. (2021) where $\mathcal{W}^*$ is not assumed to be a lower set.

B.1 Nonsingular intersections

We first explain the meaning of nonsingular intersection, namely the existence of $\gamma_{\max}$. The intuition is that the target set $\mathcal{W}^*$ and the achievable values $\mathcal{V}$ do not share the same support hyperplane.



(a) Nonsingular intersection.

(b) Singular intersection.

Figure 1: The illustration of two types of intersections: the nonsingular intersection and the singular intersection. The target set $\mathcal{W}^*$ is represented by a red square and the achievable values $\mathcal{V}$ is represented by a blue circle. The intersection is in grey.

We illustrate it via a two-dimensional example in Figure 1. The target set $\mathcal{W}^*$ is represented by a red square and the achievable values $\mathcal{V}$ is represented by a blue circle. The intersection is in grey. In (a), the intersection is nonsingular, which guarantees the existence of $\gamma_{\max} < \pi$. However, in (b), $\mathcal{V}$ intersects with $\mathcal{W}$ singularly, leading to the inexistence of an upper bound $\gamma_{\max}$. We notice that Figure 1 also intuitively shows why the Slater's condition implies our assumption, since interior points guarantee the nonsingular intersection.

As we see in related proofs, for any $\boldsymbol{W} \in \mathcal{V}$, nonsingular intersection enables us to reduce the calculation of $\mathrm{dist}(\boldsymbol{W}, \mathcal{W}^*)$ to $\mathrm{dist}(\boldsymbol{W}, \mathcal{W})$ multiplied by the sine of the included angle. We note

that when this geometric requirement is not satisfied (i.e., when the intersection is singular), this reduction may not exist. The reason is that when $\mathcal{V}$ is not of full dimension, $\boldsymbol{W}$ can be arbitrarily close to $\mathcal{W}^*$ while remains far away from $\mathcal{W}$, making the reduction impossible (Yu et al., 2021).

B.2 Being a lower set

Now we discuss the assumption of being a lower set. It is much milder than at first glance and can be easily relaxed. For target set $\mathcal{W}^*$ which is not a lower set, we can transform it into a lower set with at most $D \cdot 2^D$ dimensions. We achieve this by cutting the boundary of $\mathcal{W}^*$ into at most 2^D pieces. We provide a two-dimensional example as shown in Figure 2. The target set is defined as

$$\mathcal{W}^* = \left\{(c^1, c^2) \,|\, c^1 + c^2 \geq H/2, c^1 + c^2 \leq 3H/2, c^1 - c^2 \geq -H/2, c^1 - c^2 \leq H/2\right\}.$$

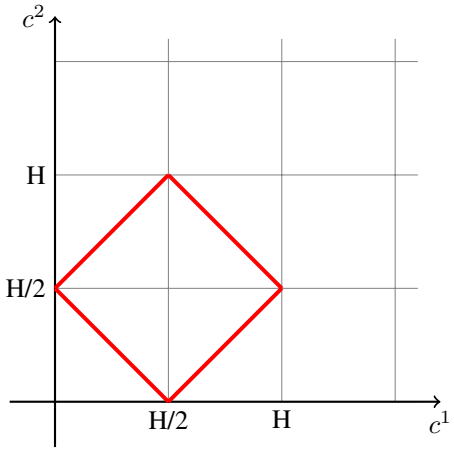


Figure 2: An example of target set $\mathcal{W}^* = \{(c^1, c^2) \,|\, c^1 + c^2 \geq H/2, c^1 + c^2 \leq 3H/2, c^1 - c^2 \geq -H/2, c^1 - c^2 \leq H/2\}$. Although it does not satisfy Assumption 1 of being a lower set, we can cut its boundaries into four pieces, each of which corresponds to a partial constraint. Since each partial constraint can be expressed in a way that is a lower set, the target set $\mathcal{W}^*$ can be reformulated as the combination of those partial constraints to satisfy the lower set requirement at the expense of higher dimension.

We cut its boundaries into four pieces: the upper left one, the upper right one, the bottom left one, and the bottom right one. For the upper right boundary, it corresponds to the partial constraint:

$$\left\{(c^1, c^2) \,|\, c^1 + c^2 \leq 3H/2\right\}.$$

It is a little tricky to represent the other three. We take the upper left boundary, $c^2 - c^1 \leq H/2$, for an example. To that end, we notice that it is equivalent to $c^2 + (H - c^1) \leq 3H/2$. Hence, we set $c^3 = H - c^1$ and $c^4 = c^2$. Then we can represent this boundary by $c^3 + c^4 \leq 3H/2$, which satisfies Assumption 1 of being a lower set with respect to c^3 and c^4. Analogously, we introduce c_5 to c_8 for the other two boundaries. Finally, we obtain a new target set with 8 objectives, $c_1, \ldots, c_8$, that meet Assumption 1. Hence, we successfully relax the constraint of being a lower set.

B.3 Summaries

In a word, our assumption is mild, and thus target sets that satisfy Assumption 1 are very common in applications. For example, we can define $\mathcal{W}^*$ as $\mathcal{W}^* = \{\boldsymbol{W} \in [0, H]^D : \boldsymbol{W} \leq \boldsymbol{b}\}$ for $\boldsymbol{b} \in \mathbb{R}^D$, which is equivalent to the formation of constraints in CMDPs.

C Reducing CMOMDPs to CMDPs

C.1 Proof of Proposition 1

Proof of Proposition 1. We first show that the CMDP is a special case of the CMOMDP. Then, we point out the correspondence between the assumptions imposed on CMDPs and on CMOMDPs. To be specific, we show that Slater's condition imposed on CMDPs is equivalent to the existence of $\gamma_{\max}$ in Assumption 1. Hence, the CMOMDP is a reasonable generalization of CMDPs.

Reducing to CMDPs. An episodic CMDP is usually given by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, c, \boldsymbol{u})$, where $\mathcal{S}$ is a state space, $\mathcal{A}$ is an action space, H is the horizon, $\mathcal{P} = \{\mathcal{P}_h\}_{h=1}^H$ is a collection of transition kernels $\mathcal{P}_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, $c = \{c_h\}_{h=1}^H$ is a collection of cost functions $c_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$, and $\boldsymbol{u} = \{\boldsymbol{u}_h\}_{h=1}^H$ is a collection of utility functions $\boldsymbol{u}_h : \mathcal{S} \times \mathcal{A} \to [0, 1]^D$.

The CMDP aims to solve the following problem:

$$\min_{\pi \in \Delta(\mathcal{A} \mid \mathcal{S}, H)} V_1^{c,\pi}(\underline{s}) \quad \text{s.t.} \quad \boldsymbol{V}_1^{\boldsymbol{u},\pi}(\underline{s}) \leq \boldsymbol{b}, \tag{11}$$

where $V^{c,\pi}$ and $\boldsymbol{V}^{\boldsymbol{u},\pi}$ are the state-value functions for cost c and utility $\boldsymbol{u}$, respectively. Here we assume $\boldsymbol{b} \in [0, H]^D$ to avoid triviality.

Now we construct a CMOMDP $\mathcal{M}'$ that is equivalent to $\mathcal{M}$. To that end, we set $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, \boldsymbol{c}')$ where $\mathcal{S}, \mathcal{A}, H$, and $\mathcal{P}$ are same as $\mathcal{M}$. We define $\boldsymbol{c}' = \{\boldsymbol{c}'_h\}_{h=1}^H$ as the concatenation of c and $\boldsymbol{u}$, i.e., $\boldsymbol{c}'_h = (c_h, \boldsymbol{u}_h^\top)^\top \in [0, 1]^{D+1}$ for $h \in [H]$.

The preference function $g : \mathbb{R}^{D+1} \to \mathbb{R}$ is defined as the first coordinate map, i.e., we define $g(x_1, \ldots, x_{D+1}) = x_1$.

We set the target set $\mathcal{W}^*$ as $\mathcal{W}^* = \{\boldsymbol{W} \in [0, H]^{D+1} : \boldsymbol{W}_{2:(D+1)} \leq \boldsymbol{b}\}$ where $\boldsymbol{W}_{2:(D+1)}$ denotes the D-dimensional vector obtained by removing the first coordinate of $\boldsymbol{W}$.

The CMOMDP $\mathcal{M}'$ seeks to solve the following optimization problem

$$\min_{\pi \in \Delta(\mathcal{A} \mid \mathcal{S}, H)} g\big(V_1^{\boldsymbol{c}',\pi}(\underline{s})\big) \quad \text{s.t.} \quad \boldsymbol{V}_1^{\boldsymbol{c}',\pi}(\underline{s}) \in \mathcal{W}^*. \tag{12}$$

A careful analysis will reveal that, by the construction of $\mathcal{M}'$, (12) is equivalent to (11). Hence, CMDP is a special case of CMOMDP.

Correspondence of assumptions. Under the above definition, it is clear that g is 1-Lipschitz and convex and that $g(\boldsymbol{x}) \geq g(\boldsymbol{x}')$ as long as $\boldsymbol{x} \geq \boldsymbol{x}'$ for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{D+1}$. In addition, $\mathcal{W}^*$ is a lower set, and it is close and convex.

It remains to show that the existence of $\gamma_{\max}$ in Assumption 1 is satisfied. However, this is equivalent to Slater's condition on the CMDPs. Recall that Slater's condition supposes the existence of an interior point, which is equivalent to a nonsingular intersection between the target set and the set of achievable values. Hence, it is exactly the geometric analog of Slater's condition. □

C.2 PEDI on CMDPs

Now we analyze PEDI on CMDPs. Recall that the core of PEDI is solving (7). Based on the analysis in Appendix C.1, as we have specified g as the first coordinate map, we can derive its convex conjugate,

$$g^*(\boldsymbol{x}^*) = \sup_{\boldsymbol{x}} \left(\boldsymbol{x}^\top \boldsymbol{x}^* - g(\boldsymbol{x})\right) = \sup_{\boldsymbol{x}} \left(\boldsymbol{x}^\top \boldsymbol{x}^* - x_1\right) = \begin{cases} 0, & \boldsymbol{x}^* = (1, 0, \ldots, 0)^\top \\ +\infty, & \text{otherwise} \end{cases}. \tag{13}$$

Therefore, solving (7) is equivalent to solving

$$\mathsf{p}^* = \min_{\boldsymbol{V} \in \mathcal{V}} \max_{\boldsymbol{\alpha} \in \mathcal{B}^D} V^1 - V^{1,*} + \nu \boldsymbol{\alpha}^\top \boldsymbol{V} - \nu \max_{\boldsymbol{x} \in \mathcal{W}^*} \boldsymbol{\alpha}^\top \boldsymbol{x}, \tag{14}$$

since the objective attains the maximum with respect to $\boldsymbol{\beta}$ when $\boldsymbol{\beta} = (1, 0, \ldots, 0)^\top$. Here, by V^1 and $V^{1,*}$ we mean the first coordinates of $\boldsymbol{V}$ and $\boldsymbol{V}^*$, respectively. Now we consider $\max_{\boldsymbol{x} \in \mathcal{W}^*} \boldsymbol{\alpha}^\top \boldsymbol{x}$. Recall that we construct $\mathcal{W}^*$ by

$$\mathcal{W}^* = \{\boldsymbol{W} \in [0, H]^{D+1} : \boldsymbol{W} \leq \boldsymbol{b}'\}, \tag{15}$$

where we suppose $\boldsymbol{b}' = (H, \boldsymbol{b}^\top)^\top$ for simplicity. We observe this constraint is element-wise, i.e., different coordinates of $\boldsymbol{W}$ are independent. Hence, we have

$$\max_{\boldsymbol{x}\in\mathcal{W}^*} \boldsymbol{\alpha}^\top \boldsymbol{x} = \max_{\boldsymbol{x}\in\mathcal{W}^*} \sum_{i=1}^{D+1} \alpha_i x_i = \max_{\substack{0\le x_i\le b_i,\\ i\in[D+1]}} \sum_{i=1}^{D+1} \alpha_i x_i = \sum_{i=1}^{D+1} \max_{\substack{0\le x_i\le b_i,\\ i\in[D+1]}} \alpha_i x_i = \boldsymbol{\alpha}_{i,+}^\top \boldsymbol{b}', \tag{16}$$

where $\boldsymbol{\alpha}_{i,+}^\top$ denotes $(\boldsymbol{\alpha}_i^\top)_+$. Then, we have

$$\mathsf{p}^* = \min_{\boldsymbol{V}\in\mathcal{V}} \max_{\boldsymbol{\alpha}\in\mathcal{B}^D} V^1 - V^{1,*} + \nu\big(\boldsymbol{\alpha}^\top \boldsymbol{V} - \boldsymbol{\alpha}_+^\top \boldsymbol{b}'\big). \tag{17}$$

A good observation is that $\boldsymbol{\alpha} \ge 0$ always holds (otherwise, we can increase the negative component of $\boldsymbol{\alpha}$ to make the objective larger). Therefore, the optimization problem is actually

$$\begin{aligned} \mathsf{p}^* &= \min_{\boldsymbol{V}\in\mathcal{V}} \max_{\boldsymbol{\alpha}\in\mathcal{B}_+^D} V^1 - V^{1,*} + \nu\boldsymbol{\alpha}(\boldsymbol{V} - \boldsymbol{b}') \\ &= \min_{\boldsymbol{V}\in\mathcal{V}} \max_{\boldsymbol{\alpha}\in\mathcal{B}_+^D} V_1^{c,\pi}(\underline{s}) - V_1^{c,*}(\underline{s}) + \nu\boldsymbol{\alpha}\big(\boldsymbol{V}_1^{\boldsymbol{u},\pi}(\underline{s}) - \boldsymbol{b}\big), \end{aligned} \tag{18}$$

where the second equality is by definition and the fact that $\boldsymbol{b}'_1 = H \ge V^1$. We notice that $\nu\boldsymbol{\alpha}$ can be considered the Lagrangian multiplier. Hence, we conclude that the objective reduces to the Lagrangian formulation of the CMDP optimization problem, and our algorithm reduces to the dual gradient method.

D Equivalence of problems

In this section, we show that (1) and (4) share the same solution.

For convenience, we denote by p_1^* the original problem (1), and by p_2^* the new problem (4). We first substitute their variables π with the state-value function.

For p_1^*, we have

$$\begin{aligned} \mathsf{p}_1^* &= \min_{\pi} g\big(\boldsymbol{V}_1^\pi(\underline{s})\big) \quad \text{s.t.} \quad \boldsymbol{V}_1^\pi(\underline{s}) \in \mathcal{W}^* \\ &= \min_{\boldsymbol{V}\in\mathcal{V}} g\big(\boldsymbol{V}\big) \quad \text{s.t.} \quad \boldsymbol{V} \in \mathcal{W}^* \end{aligned} \tag{19}$$

while for p_2^*, we have

$$\begin{aligned} \mathsf{p}_2^* &= \min_{\pi} \big(\operatorname{SubOpt}(\pi) + \nu \operatorname{Violation}(\pi)\big) \\ &= \min_{\pi} \Big(g\big(\boldsymbol{V}_1^\pi(\underline{s})\big) - g\big(\boldsymbol{V}_1^*(\underline{s})\big) + \nu \operatorname{dist}\big(\boldsymbol{V}_1^\pi(\underline{s}), \mathcal{W}^*\big)\Big) \\ &= \min_{\boldsymbol{V}\in\mathcal{V}} \big(g(\boldsymbol{V}) - g\big(\boldsymbol{V}_1^*(\underline{s})\big) + \nu \operatorname{dist}(\boldsymbol{V}, \mathcal{W}^*)\big) \\ &= \min_{\boldsymbol{V}\in\mathcal{V}} g(\boldsymbol{V}) + \nu \operatorname{dist}(\boldsymbol{V}, \mathcal{W}^*) \end{aligned} \tag{20}$$

We suppose that $\boldsymbol{V}^\dagger$ is the solution of p_2^*. To show that $\boldsymbol{V}^\dagger$ is also the solution of p_1^*, it suffices to verify $\operatorname{dist}(\boldsymbol{V}^\dagger, \mathcal{W}^*) = 0$.

Suppose $\operatorname{dist}(\boldsymbol{V}^\dagger, \mathcal{W}^*) \ne 0$, i.e., $\boldsymbol{V}^\dagger \notin \mathcal{W}^*$. We consider its projection on $\mathcal{W}^*$, $\boldsymbol{V}^\S = \prod_{\mathcal{W}^*} \boldsymbol{V}^\dagger$. Since g is 1-Lipschitz, when $\nu > 1$, it holds that

$$g(\boldsymbol{V}^\S) - g(\boldsymbol{V}^\dagger) < \nu \operatorname{dist}(\boldsymbol{V}^\dagger, \boldsymbol{V}^\S). \tag{21}$$

Then, by (21), we have

$$\begin{aligned} &g(\boldsymbol{V}^\S) + \nu \operatorname{dist}(\boldsymbol{V}^\S, \mathcal{W}^*) \\ <&g(\boldsymbol{V}^\S) - g(\boldsymbol{V}^\S) + g(\boldsymbol{V}^\dagger) + \nu\big(\operatorname{dist}(\boldsymbol{V}^\S, \mathcal{W}^*) + \operatorname{dist}(\boldsymbol{V}^\dagger, \boldsymbol{V}^\S)\big) \\ =&g(\boldsymbol{V}^\dagger) + \nu \operatorname{dist}(\boldsymbol{V}^\dagger, \mathcal{W}^*), \end{aligned}$$

where the last equality is for $\operatorname{dist}(\boldsymbol{V}^\S, \mathcal{W}^*) = 0$. It contradicts our assumption that $\boldsymbol{V}^\dagger$ is the solution of p_2^*. Therefore, we conclude that the solution $\boldsymbol{V}^\dagger$ of p_2^* should lie in $\mathcal{W}^*$, which completes the proof of the equivalence between (1) and (4).

E An instantiation of PEDI for linear kernel CMOMDPs

This section serves as complements for Section 3.3.

We now propose a method to estimate the empirical transition kernel $\widehat{\mathcal{P}}$ and cost function $\widehat{c}$ and thereby construct a ξ-uncertainty quantifier to specify Algorithm 2 for linear kernel CMOMDPs. We construct the empirical transition kernel by ridge regression on the offline dataset $\mathcal{D}$ as follows,

$$\widehat{\mathcal{P}}_h(s' \mid s, a) = \psi(s, a, s')^\top \widehat{\theta}_h$$
$$\text{where} \quad \widehat{\theta}_h = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^{d_1}} \sum_{\tau=1}^{N} \int_{\mathcal{S}} \left(\delta_{s_{h+1}^\tau}(s') - \psi(s_h^\tau, a_h^\tau, s')^\top \theta\right)^2 \mathrm{d}s' + \lambda \|\theta\|_2^2.$$

Here $\delta_{s_h^\tau}(s)$ is the Dirac function centered at s_h^τ for continuous space and indicator function for finite space, and $\lambda > 0$ is the regularization parameter. Note that we can obtain the following closed form of $\widehat{\theta}_h$,

$$\widehat{\theta}_h = \Lambda_h^{-1} \sum_{\tau=1}^{N} \psi(s_h^\tau, a_h^\tau, s_{h+1}^\tau) \quad \text{where} \quad \Lambda_h = \sum_{\tau=1}^{N} \int_{\mathcal{S}} \psi(s_h^\tau, a_h^\tau, s') \psi(s_h^\tau, a_h^\tau, s')^\top \mathrm{d}s' + \lambda I. \tag{22}$$

We construct $\widehat{c}$ in an analogous way by ridge regression,

$$\widehat{r}_h^i(s, a) = \varphi(s, a)^\top \widehat{\theta}_h^{c^i} \quad \text{where} \quad \widehat{\theta}_h^{c^i} = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^{d_2}} \sum_{\tau=1}^{N} \left(c_h^{i,\tau} - \varphi(s_h^\tau, a_h^\tau)^\top \theta\right)^2 + \lambda \|\theta\|_2^2.$$

and $\widehat{\theta}_h^{c_i}$ has the closed form

$$\widehat{\theta}_h^{c^i} = \Lambda_{\varphi,h}^{-1} \sum_{\tau=1}^{N} c_h^{i,\tau} \cdot \varphi(s_h^\tau, a_h^\tau) \quad \text{where} \quad \Lambda_{\varphi,h} = \sum_{\tau=1}^{N} \varphi(s_h^\tau, a_h^\tau) \varphi(s_h^\tau, a_h^\tau)^\top + \lambda I. \tag{23}$$

Moreover, we construct the ξ-uncertainty quantifier for $h \in [H]$ below

$$\Gamma_h^{\mathcal{P}}(s, a, s') = \min\left\{\kappa \cdot \left\|\psi(s, a, s')\right\|_{\Lambda_h^{-1}}, 1\right\}, \quad \Gamma_h^{c^i}(s, a) = \min\left\{\kappa \cdot \|\varphi(s, a)\|_{\Lambda_{\varphi,h}^{-1}}, 1\right\}, \tag{24}$$

where $\kappa > 0$ is a scaling parameter to be specified later. By plugging (24) into the pessimistic planning (Algorithm 1), we finish the establishment of PEDI on linear kernel CMOMDPs.

F Pessimism is all you need

This section studies the effectiveness of the pessimistic approach for offline CMOMDPs. To that end, we first introduce the model evaluation error and then develop the decomposition lemma, which decomposes the discrepancy between the state-value functions of the learned policy and the optimal policy into three parts: the spurious correlation, the intrinsic uncertainty, and the optimization error. Then, we show that our proposed method successfully eliminates the spurious correlation, which is the most difficult one to control.

We consider a meta-algorithm that constructs a sequence of policies $\{\pi_k\}_{k=1}^K$ which ideally converges to the optimal policy. At the k-th iteration, the algorithm constructs the estimations $\{\boldsymbol{V}_h^k\}_{h=1}^H$ and $\{\boldsymbol{Q}_h^k\}_{h=1}^H$ such that $\boldsymbol{V}_h^k(s) = \mathbb{D}_{\pi^k}[\boldsymbol{Q}_h^k](s)$. We define the model evaluation error below, which characterizes the error of estimating the Bellman equations.

Definition 3 (Model evaluation error)**.** *The model evaluation error for the k-th iteration is defined as*

$$\boldsymbol{\iota}_h^k(s, a) = \boldsymbol{Q}_h^k(s, a) - \mathbb{B}_h \boldsymbol{V}_{h+1}^k(s, a).$$

We denote its i-th scalar component by $\iota_h^{i,k}(s, a)$, i.e., $\iota_h^{i,k}(s, a) = Q_h^{i,k}(s, a) - \mathbb{B}_h V_{h+1}^{i,k}(s, a)$.

Utilizing the model evaluation error, the discrepancy between any state-value function and the optimal one admits a decomposition as shown in the following lemma.

Lemma 2 (Decomposition of state-value function)**.** *Let $\{\boldsymbol{V}_h^k\}_{h=1}^H$ and $\{\boldsymbol{Q}_h^k\}_{h=1}^H$ be any state-value function and action-value function such that $\boldsymbol{V}_h^k(s) = \mathbb{D}_{\pi^k}[\boldsymbol{Q}_h^k](s)$ for any $s \in \mathcal{S}$ and any $h \in [H]$. Then, it holds that*

$$\begin{aligned}\boldsymbol{V}_1^k(\underline{s}) - \boldsymbol{V}_1^*(\underline{s}) = &\sum_{h=1}^H \mathbb{E}_{\pi^*}[\boldsymbol{\iota}_h^k(s_h, a_h) \mid s_1 = \underline{s}] \\ &+ \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\left\langle \boldsymbol{Q}_h^k(s_h, \cdot), (\pi_h^k - \pi_h^*)(\cdot \mid s_h)\right\rangle_{\mathcal{A}} \,\middle|\, s_1 = \underline{s}\right]\end{aligned}$$

and that

$$\begin{aligned}\boldsymbol{V}_1^{\pi_k}(\underline{s}) - \boldsymbol{V}_1^*(\underline{s}) = &\underbrace{-\sum_{h=1}^H \mathbb{E}_{\pi^k}\left[\boldsymbol{\iota}_h^k(s_h, a_h) \,\middle|\, s_1 = \underline{s}\right]}_{\text{(i) Spurious Correlation}} + \underbrace{\sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\boldsymbol{\iota}_h^k(s_h, a_h) \,\middle|\, s_1 = \underline{s}\right]}_{\text{(ii) Intrinsic Uncertainty}} \\ &+ \underbrace{\sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\left\langle \boldsymbol{Q}_h^k(s_h, \cdot), (\pi_h^k - \pi_h^*)(\cdot \mid s_h)\right\rangle_{\mathcal{A}} \,\middle|\, s_1 = \underline{s}\right]}_{\text{(iii) Optimization Error}},\end{aligned}$$

where $\mathbb{E}_{\pi^k}$ and $\mathbb{E}_{\pi^*}$ are taken with respect to the trajectories induced by π^k and π^* in the underlying CMOMDP, respectively.

Proof of Lemma 2. See Appendix H.6.1 for a detailed proof. □

Lemma 2 is the vectorized analogue of Lemma 3.1 in Jin et al. (2020b). It suggests that we can decompose the discrepancy between the state-value function of learned policy and the optimal one into (i) spurious correlation, (ii) intrinsic uncertainty, and (iii) optimization error. Among them, (i) is the most difficult to control since it depends on both π^k and $\boldsymbol{\iota}_h^k$ that spuriously correlated with each other. As the learner has no control over the data collecting process, this spurious correlation could be large even in a multi-armed bandit setting (Jin et al., 2020b). Term (ii) is easier to control since π^* is intrinsic to the underlying CMOMDP and therefore not spuriously correlated with $\boldsymbol{\iota}_h^k$.

As proved in Section 4.1, our proposed algorithm successfully eliminates term (i) through pessimism. In Section 4.2, we have shown that (ii) is impossible to eliminate as it arises from the information-theoretic lower bound of linear kernel CMOMDPs.

G Reducing linear kernel CMOMDPs to tabular CMOMDPs

All we need is to represent the transition kernel $\mathcal{P}$ and cost function $\boldsymbol{c}$ of tabular CMOMDP in the form of linear kernel CMOMDP.

We set $d_1 = |\mathcal{S}||\mathcal{A}||\mathcal{S}|$ and $d_2 = |\mathcal{S}||\mathcal{A}|$ and set $\psi(s, a, s') = \boldsymbol{e}_{(s,a,s')}, (\theta_h)_{(s,a,s')} = \mathcal{P}(s' \mid s, a), \varphi(s, a) = \boldsymbol{e}_{(s,a)}$, and $(\theta_h^{c_i})_{(s,a)} = c_h^i(s, a)$.

Here we denote by $\boldsymbol{e}$ the canonical basis. It can be verified that the definition of linear kernel MDP (Definition 2) is satisfied with $R = 1$.

H Proofs for Section 4

For notational simplicity, we sometime use the shorthand $\boldsymbol{\iota}_h^k$, $\boldsymbol{Q}_h$, $\boldsymbol{V}_h$, π_h to denote $\boldsymbol{\iota}_h^k(s_h, a_h)$, $\boldsymbol{Q}_h(s_h, a_h)$, $\boldsymbol{V}_h(s_h)$, and $\pi_h(a_h \mid s_h)$ when there is no risk of confusion.

H.1 Proof of Lemma 1

Proof of Lemma 1. The proof is by backward induction. Suppose the inequality holds for Q-values in the $(h+1)$-th step. Then, it holds that for any $s \in \mathcal{S}, i \in [D]$,

$$V_{h+1}^{i,k}(s) = \mathbb{D}_{\pi_h^k}\left[Q_{h+1}^{i,k}\right](s) \geq \mathbb{D}_{\pi_h^k}\left[Q_{h+1}^{i,\pi^k}\right](s) = V_{h+1}^{i,\pi^k}.$$

For any $(s,a) \in \mathcal{S} \times \mathcal{A}, i \in [D]$, when $Q_h^{i,k}(s,a) = H-h+1$, it holds that $Q_h^{i,k}(s,a) \geq Q_h^{i,\pi^k}(s,a)$. Otherwise, by the definition of ξ-uncertainty quantifier defined in Definition 1, it holds that

$$\begin{aligned}
Q_h^{i,k}(s,a) \geq & \widehat{c}_h^i(s,a) + \widehat{\mathbb{P}}_h^k\left[V_{h+1}^{i,k}\right](s,a) + \Gamma_h + \Gamma_h^{\mathbf{c}^i} \\
\geq & c_h^i(s,a) + \mathbb{P}_h^k\left[V_{h+1}^{i,k}\right](s,a) \\
\geq & c_h^i(s,a) + \mathbb{P}_h^k\left[V_{h+1}^{i,\pi^k}\right](s,a) \\
= & Q_h^{i,\pi^k}(s,a)
\end{aligned}$$

under event $\mathcal{E}$. Thus, by induction, it holds for all h that $Q_h^{i,k} \geq Q_h^{i,\pi^k}$.

Then, by the fact that $g(\boldsymbol{x}) \geq g(\boldsymbol{x}')$ holds as long as $\boldsymbol{x} \geq \boldsymbol{x}'$, we have

$$g\big(\widehat{\boldsymbol{V}}_1(\underline{s})\big) \geq g\big(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s})\big).$$

For the constraint violation, we consider the point $\boldsymbol{W} = \boldsymbol{V}_1^{\pi^k}(\underline{s}) + \prod_{\mathcal{W}^*} \boldsymbol{V}_1^k(\underline{s}) - \boldsymbol{V}_1^k(\underline{s})$. By the fact that $\mathcal{W}^*$ is a lower set in Assumption 1, it holds that $\boldsymbol{W}_+ \in \mathcal{W}^*$ and

$$\operatorname{dist}\left(\boldsymbol{V}_1^{\pi^k}(\underline{s}), \mathcal{W}^*\right) \leq \operatorname{dist}\left(\boldsymbol{V}_1^{\pi^k}(\underline{s}), \boldsymbol{W}\right) \leq \operatorname{dist}\left(\boldsymbol{V}_1^k(\underline{s}), \mathcal{W}^*\right).$$

Thus, we complete the proof of Lemma 1.

□

H.2 Proof of Theorem 1

Proof of Theorem 1. In the following lemma, we show that the difference between any state-value function and the optimal one, as stated in Lemma 2, can be bounded from above by the ξ-uncertainty quantifier with high probability when projected along with $\boldsymbol{\theta}^k$.

Lemma 3 (Upper bound of projected difference of state-value functions)**.** *Suppose $\{(\Gamma_h^{\mathcal{P}}, \Gamma_h^{\mathbf{c}})\}_{h=1}^H$ in Algorithm 2 is a ξ-uncertainty quantifier. Then under event $\mathcal{E}$, we have*

$$(\boldsymbol{\theta}^k)^\top\left(\boldsymbol{V}_1^k(\underline{s}) - \boldsymbol{V}_1^*(\underline{s})\right) \leq 2(1+\rho)\sqrt{D}\sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\Gamma_h(s_h,a_h) + \|\Gamma_h^{\mathbf{c}}(s_h,a_h)\|_\infty \,\middle|\, s_1 = \underline{s}\right]$$

Proof of Lemma 3. See Appendix H.6.3 for a detailed proof. □

In what follows, we suppose the event $\mathcal{E}$ holds, which has probability at least $1-\xi$. Applying Lemma 1, we have

$$\begin{aligned}
& K\left[g\big(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s})\big) - g\big(\boldsymbol{V}_1^*(\underline{s})\big) + \rho \operatorname{dist}\left(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s}), \mathcal{W}^*\right)\right] \\
\leq & K\left[g\big(\widehat{\boldsymbol{V}}_1(\underline{s})\big) - g\big(\boldsymbol{V}_1^*(\underline{s})\big) + \rho \operatorname{dist}\left(\widehat{\boldsymbol{V}}_1(\underline{s}), \mathcal{W}^*\right)\right].
\end{aligned}$$

By the convex conjugate in (6) and the fact that $\widehat{\boldsymbol{V}}_1(\underline{s}) = \sum_{k=1}^K \boldsymbol{V}_1^k(\underline{s})$, we get

$$\begin{aligned}
& K\left[g\big(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s})\big) - g\big(\boldsymbol{V}_1^*(\underline{s})\big) + \rho \operatorname{dist}\left(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s}), \mathcal{W}^*\right)\right] \\
= & \max_{\|\boldsymbol{\beta}\| \leq 1}\left\{\boldsymbol{\beta} \cdot \sum_{k=1}^K \boldsymbol{V}_1^k(\underline{s}) - \sum_{k=1}^K g^*(\boldsymbol{\beta})\right\} - K g\big(\boldsymbol{V}_1^*(\underline{s})\big) \\
& + \rho \max_{\|\boldsymbol{\alpha}\| \leq 1}\left\{\boldsymbol{\alpha} \cdot \sum_{k=1}^K \boldsymbol{V}_1^k(\underline{s}) - \sum_{k=1}^K \max_{\boldsymbol{x} \in \mathcal{W}^*} \boldsymbol{\alpha} \cdot \boldsymbol{x}\right\}
\end{aligned}$$

We observe these two terms

$$\max_{\|\boldsymbol{\beta}\| \leq 1}\left\{\boldsymbol{\beta} \cdot \sum_{k=1}^K \boldsymbol{V}_1^k(\underline{s}) - \sum_{k=1}^K g^*(\boldsymbol{\beta})\right\}, \quad \max_{\|\boldsymbol{\alpha}\| \leq 1}\left\{\boldsymbol{\alpha} \cdot \sum_{k=1}^K \boldsymbol{V}_1^k(\underline{s}) - \sum_{k=1}^K \max_{\boldsymbol{x} \in \mathcal{W}^*} \boldsymbol{\alpha} \cdot \boldsymbol{x}\right\},$$

are the "single best desicion" in hindsight in the projected subgradient method. By setting $\eta^k = 2G^{-1}\sqrt{D/k}$ (or $2G^{-1}\sqrt{D/K}$ if K is predefined), we apply Theorem 5 with $R = 2$ and $G = 2(1+\rho)H\sqrt{D}$ (to verify the conditions, note that g^* is $H\sqrt{D}$-Lipschitz) to get

$$\begin{aligned}
&K\Big[g\big(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s})\big) - g\big(\boldsymbol{V}_1^{*}(\underline{s})\big) + \rho\,\mathrm{dist}\,\big(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s}), \mathcal{W}^*\big)\Big] \\
\leq &\sum_{k=1}^{K}\big\{(\boldsymbol{\beta}^k)^\top \boldsymbol{V}_1{}^k(\underline{s}) - g^*(\boldsymbol{\beta}^k)\big\} - Kg\big(\boldsymbol{V}_1^{*}(\underline{s})\big) + \rho\sum_{k=1}^{K}\big\{(\boldsymbol{\alpha}^k)^\top \boldsymbol{V}_1{}^k(\underline{s}) - \max_{\boldsymbol{x}\in\mathcal{W}^*}(\boldsymbol{\alpha}^k)^\top \boldsymbol{x}\big\} \\
&+ C(1+\rho)\sqrt{DH^2K}
\end{aligned} \tag{25}$$

where C is a constant. Then, by observing

$$g^*(\boldsymbol{\beta}^k) = \max_{\boldsymbol{V}}\big\{(\boldsymbol{\beta}^k)^\top \boldsymbol{V} - g(\boldsymbol{V})\big\} \geq (\boldsymbol{\beta}^k)^\top \boldsymbol{V}_1^{*}(\underline{s}) - g\big(\boldsymbol{V}_1^{*}(\underline{s})\big),$$

$$\max_{\boldsymbol{x}\in\mathcal{W}^*}(\boldsymbol{\alpha}^k)^\top \boldsymbol{x} \geq (\boldsymbol{\alpha}^k)^\top \boldsymbol{V}_1^{*}(\underline{s}),$$

we have for (25) that

$$\begin{aligned}
&K\Big[g\big(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s})\big) - g\big(\boldsymbol{V}_1^{*}(\underline{s})\big) + \rho\,\mathrm{dist}\,\big(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s}), \mathcal{W}^*\big)\Big] \\
\leq &\sum_{k=1}^{K}\Big\{(\boldsymbol{\beta}^k)^\top \boldsymbol{V}_1^{k}(\underline{s}) - (\boldsymbol{\beta}^k)^\top \boldsymbol{V}_1^{*}(\underline{s}) + \rho\big((\boldsymbol{\alpha}^k)^\top \boldsymbol{V}_1^{k}(\underline{s}) - (\boldsymbol{\alpha}^k)^\top \boldsymbol{V}_1^{*}(\underline{s})\big)\Big\} + C(1+\rho)\sqrt{DH^2K} \\
= &\sum_{k=1}^{K}\Big[(\boldsymbol{\theta}^k)^\top\big(\boldsymbol{V}_1{}^k(\underline{s}) - \boldsymbol{V}_1^{*}(\underline{s})\big)\Big] + C(1+\rho)\sqrt{DH^2K}
\end{aligned} \tag{26}$$

Applying Lemma 3, we have

$$\begin{aligned}
&K\Big[g\big(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s})\big) - g\big(\boldsymbol{V}_1^{*}(\underline{s})\big) + \rho\,\mathrm{dist}\,\big(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s}), \mathcal{W}^*\big)\Big] \\
\leq &2K(1+\rho)\sqrt{D}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\big[\Gamma_h(s_h,a_h) + \|\Gamma_h^{\mathbf{c}}(s_h,a_h)\|_\infty \,|\, s_1 = \underline{s}\big] + C(1+\rho)\sqrt{DH^2K} \\
= &K(\epsilon_K + \mathrm{IntUncert}_{\mathcal{D}}^{\pi^*}),
\end{aligned}$$

where we define

$$\epsilon_K = C(1+\rho)\sqrt{\frac{DH^2}{K}}, \quad \mathrm{IntUncert}_{\mathcal{D}}^{\pi^*} = 2(1+\rho)\sqrt{D}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\big[\Gamma_h(s_h,a_h) + \|\Gamma_h^{\mathbf{c}}(s_h,a_h)\|_\infty \,|\, s_1 = \underline{s}\big].$$

Note that $\mathrm{dist}(\boldsymbol{W}^K, \mathcal{W}^*) \geq 0$. Therefore, we can bound the suboptimality from above by

$$g\big(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s})\big) - g\big(\boldsymbol{V}_1^{*}(\underline{s})\big) \leq \epsilon_K + \mathrm{IntUncert}_{\mathcal{D}}^{\pi^*}.$$

To obtain an upper bound of the constraint violation, we employ the following lemma.

Lemma 4. *Let $\boldsymbol{W}^*$ denote a return vector in set $\mathcal{W}$ that achieves the lowest cost, i.e. $\forall \boldsymbol{W} \in \mathcal{W}, g(\boldsymbol{W}) \geq g(\boldsymbol{W}^*)$. Then, under Assumption 1, it holds for any $\boldsymbol{W} \in \mathbb{R}^D$ that*

$$g(\boldsymbol{W}) - g(\boldsymbol{W}^*) \geq -\,\mathrm{dist}(\boldsymbol{W}, \mathcal{W}^*)/\sin(\gamma_{\max}).$$

Proof of Lemma 4. See Lemma 16 in Yu et al. (2021) for a detailed proof. □

We notice that by definition, $\boldsymbol{W}^* = \boldsymbol{V}_1^*(\underline{s})$. Therefore, applying Lemma 4 with $\boldsymbol{W} = \boldsymbol{V}_1^{\widehat{\pi}}(\underline{s})$, we have

$$\begin{aligned}
&\operatorname{dist}\left(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s}), \mathcal{W}^*\right) \\
\leq &\operatorname{dist}\left(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s}), \mathcal{W}^*\right) + \sin(\gamma_{\max})\left[g\big(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s})\big) - g\big(\boldsymbol{V}_1^{*}(\underline{s})\big) + \frac{\operatorname{dist}\left(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s}), \mathcal{W}^*\right)}{\sin(\gamma_{\max})}\right] \\
= &\sin(\gamma_{\max})\Big[g\big(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s})\big) - g\big(\boldsymbol{V}_1^{*}(\underline{s})\big) + \rho \operatorname{dist}\left(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s}), \mathcal{W}^*\right)\Big] \\
\leq &\frac{2}{\rho}(\epsilon_K + \mathrm{IntUncert}_{\mathcal{D}}^{\pi^*}).
\end{aligned}$$

where the last inequality follows from (26). Thus, we complete the proof of Theorem 1. □

H.3 Proof of Theorem 2

Proof of Theorem 2. It suffices to show that $\{(\Gamma_h^{\mathcal{P}}, \Gamma_h^{\mathbf{c}})\}_{h=1}^H$ defined in (24) is a ξ-uncertainty quantifier as stated in the following lemma.

Lemma 5. *Under Assumptions 2, we set*

$$\lambda = 1, \quad \kappa = CR\sqrt{d\log(dN) + \log(DH/\xi)},$$

where $C > 0$ is an absolute constant and $\xi \in (0,1)$ is the confidence parameter. Then, $\{(\Gamma_h^{\mathcal{P}}, \Gamma_h^{\mathbf{c}})\}_{h=1}^H$ in (24) is a ξ-uncertainty quantifier.

Proof of Lemma 5. See Appendix H.6.5 for a detailed proof. □

By Theorem 1, we have that

$$\mathrm{SubOpt}(\widehat{\pi}) \leq \epsilon_K + \mathrm{IntUncert}_{\mathcal{D}}^{\pi^*}, \quad \mathrm{Violation}(\widehat{\pi}) \leq \frac{2}{\rho}(\epsilon_K + \mathrm{IntUncert}_{\mathcal{D}}^{\pi^*})$$

where we define

$$\epsilon_K = C(1+\rho)\sqrt{\frac{DH^2}{K}},\ \mathrm{IntUncert}_{\mathcal{D}}^{\pi^*} = 2(1+\rho)\sqrt{D}\sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\Gamma_h(s,a) + \|\Gamma_h^{\mathbf{c}}(s,a)\|_\infty \,\middle|\, s_1 = \underline{s}\right].$$

By Lemma 5 and the ξ-uncertainty quantifier defined in (24), we finish the proof. □

H.4 Proof of Theorem 3

Proof of Theorem 3. The following lemma is adopted from Jin et al. (2020b), which characterizes the information-theoretic lower bound of offline RL.

Theorem 4. *For the output $\widehat{\pi}$ of any offline RL algorithm, there exists a tabular MDP $\mathcal{M}$ with initial state $\underline{s} \in \mathcal{S}$ and a dataset $\mathcal{D}$ compliant with $\mathcal{M}$, such that*

$$\mathbb{E}_{\mathcal{D}}\left[\frac{\mathrm{SubOpt}(\widehat{\pi})}{\sum_{h=1}^H \mathbb{E}_{\pi^*}\left[1/\sqrt{1+n_h(s_h,a_h)} \,\middle|\, s_1 = \underline{s}\right]}\right] \geq C$$

where $C > 0$ is an absolute constant. Here $n_h(s_h,a_h) = \sum_{\tau=1}^N \mathbb{1}\{s_h^\tau = s_h, a_h^\tau = a_h\}$ for $(s_h,a_h) \in \mathcal{S}\times\mathcal{A}$.

Proof of Theorem 4. See Theorem 4.6 in Jin et al. (2020b) for a detailed proof. □

Note that we can view the MDP as a special case of CMOMDP with the target set $\mathcal{W}^* = \mathbb{R}^D$, $D = 1$ and g being the identity function. Hence the hard instance in Theorem 4 is also a hard instance of CMOMDP. It remains to reduce the tabular MDP there to a linear kernel MDP defined in Definition 2. To that end, we set $d_1 = |\mathcal{S}||\mathcal{A}||\mathcal{S}|$ and $d_2 = |\mathcal{S}||\mathcal{A}|$ and set $\psi(s,a,s') = \boldsymbol{e}_{(s,a,s')}$, $(\theta_h)_{(s,a,s')} =$

$\mathcal{P}(s' \mid s, a), \varphi(s, a) = \boldsymbol{e}_{(s,a)}$, and $(\theta_h^{c_i})_{(s,a)} = c_h^i(s, a)$. Here $\boldsymbol{e}$ denotes the canonical basis. It can be verified that Definition 2 is satisfied with $R = 1$.

Then, it holds that

$$\Lambda_h = \lambda I + \sum_{\tau=1}^{N} \sum_{s' \in \mathcal{S}} \psi(s_h^\tau, a_h^\tau, s') \psi(s_h^\tau, a_h^\tau, s')^\top = \lambda I + \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \sum_{s' \in \mathcal{S}} n_h(s, a) E_{(s,a,s'),(s,a,s')}$$

where $E_{(s,a,s'),(s,a,s')}$ is the matrix in which entries at $\big((s, a, s'), (s, a, s')\big)$ is 1 and other entries are all 0. We note that $\lambda = 1$ and Λ_h is diagonal and thus we have

$$\big\|\psi(s, a, s')\big\|_{\Lambda_h^{-1}} \leq \frac{1}{\sqrt{1 + n_h(s, a)}} \tag{27}$$

Following the same derivation we get

$$\|\varphi(s, a)\|_{\Lambda_{\varphi,h}^{-1}} \leq \frac{1}{\sqrt{1 + n_h(s, a)}} \tag{28}$$

Then, by Theorem 4, (27), and (28), we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\frac{\mathrm{SubOpt}(\widehat{\pi})}{\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\|\varphi(s_h, a_h)\|_{\Lambda_{\varphi,h}^{-1}} + |\mathcal{S}|^{-1} \int_{\mathcal{S}} \|\psi(s_h, a_h, s')\|_{\Lambda_h^{-1}} \,\mathrm{d}s' \,\middle|\, s_1 = \underline{s}\right]}\right] \\
\geq &\mathbb{E}_{\mathcal{D}}\left[\frac{\mathrm{SubOpt}(\widehat{\pi})}{2\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[1/\sqrt{1 + n_h(s_h, a_h)} \,\middle|\, s_1 = \underline{s}\right]}\right] \\
\geq &c/2.
\end{aligned}$$

which completes the proof of Theorem 3. □

H.5 Proof of Corollary 1

Proof of Corollary 1. By the property of visitation measure, we have that

$$\begin{aligned}
&\mathbb{E}_{\pi^*}\left[\Gamma_h(s_h, a_h) + \Gamma_h^{c_i}(s_h, a_h) \,|\, s_1 = \underline{s}\right] \\
=&\mathbb{E}_{\mu_h^*}\left[\Gamma_h(s, a)\right] + \mathbb{E}_{\mu_h^*}\left[\Gamma_h^{c_i}(s, a)\right] \\
\leq&\left(\sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{\mu_h^*(s, a)}{\mu_h^{\mathrm{b},\tau}(s, a)}\right) \cdot \left(\mathbb{E}_{\mu_h^{\mathrm{b},\tau}}\left[\Gamma_h(s, a)\right] + \mathbb{E}_{\mu_h^{\mathrm{b},\tau}}\left[\Gamma_h^{c_i}(s, a)\right]\right) \\
\leq&\varsigma \cdot \left(\mathbb{E}_{\mu_h^{\mathrm{b},\tau}}\left[\Gamma_h(s, a)\right] + \mathbb{E}_{\mu_h^{\mathrm{b},\tau}}\left[\Gamma_h^{c_i}(s, a)\right]\right),
\end{aligned} \tag{29}$$

where the first inequality follows from Hölder's inequality and the last inequality follows from the condition. We will upper bound the term $\mathbb{E}_{\mu_h^{\mathrm{b},\tau}}\left[\Gamma_h(s, a)\right]$ and $\mathbb{E}_{\mu_h^{\mathrm{b},\tau}}\left[\Gamma_h^{c_i}(s, a)\right]$ respectively.

Let $X_h^{c_i,\tau} = \mathbb{E}_{\mu_h^{\mathrm{b},\tau}}\left[\Gamma_h^{c_i}(s, a)\right] - \Gamma_h^{c_i}(s_h^\tau, a_h^\tau)$, which is a martingale difference process with respect to the filtration $\{\mathcal{F}_h^\tau\}_{\tau=1}^N$. To see this, we have

$$\mathbb{E}[X_h^{c_i,\tau} \,|\, \mathcal{F}_h^{\tau-1}] = \mathbb{E}[X_h^{c_i,\tau}] = 0$$

Note that $|X_h^{c_i,\tau}|$ is bounded by 2, and thus Azuma's inequality implies for all $h \in [H]$ that

$$\sum_{\tau=1}^{N} X_h^{c_i,\tau} \leq C\sqrt{N \log(DH/\xi)} \tag{30}$$

holds with probability at least $1 - \xi/(D + 1)$. Here C denotes an absolute constant.

Moreover, by the Cauchy-Schwarz inequality, it holds that

$$\sum_{\tau=1}^{N} \Gamma_h^{c_i}(s_h^\tau, a_h^\tau) \leq \sum_{\tau=1}^{N} \kappa\sqrt{\varphi(s_h^\tau, a_h^\tau)^\top \Lambda_{\varphi,h}^{-1} \varphi(s_h^\tau, a_h^\tau)} \leq \kappa\sqrt{N \sum_{\tau=1}^{N} \varphi(s_h^\tau, a_h^\tau)^\top \Lambda_{\varphi,h}^{-1} \varphi(s_h^\tau, a_h^\tau)}, \tag{31}$$

and by the property of trace, we have

$$\begin{aligned}
&\sum_{\tau=1}^{N} \varphi(s_h^\tau, a_h^\tau)^\top \Lambda_{\varphi,h}^{-1} \varphi(s_h^\tau, a_h^\tau) \\
=&\operatorname{tr}\big(\sum_{\tau=1}^{N} \varphi(s_h^\tau, a_h^\tau)^\top \Lambda_{\varphi,h}^{-1} \varphi(s_h^\tau, a_h^\tau)\big) \\
=&\operatorname{tr}\big(\sum_{\tau=1}^{N} \varphi(s_h^\tau, a_h^\tau) \varphi(s_h^\tau, a_h^\tau)^\top \Lambda_{\varphi,h}^{-1}\big) \\
=&\operatorname{tr}\big((\Lambda_{\varphi,h} - \lambda I)\Lambda_{\varphi,h}^{-1}\big) \\
=&\operatorname{tr}\big(I - \lambda \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{d_2})^{-1}\big) \\
\leq& d_2,
\end{aligned} \tag{32}$$

where in the last equality we denote by $\lambda_1, \lambda_2, \ldots, \lambda_{d_2}$ the eigenvalues of $\Lambda_{\varphi,h}$. By plugging (32) into (31) we have

$$\sum_{\tau=1}^{N} \Gamma_h^{c_i}(s_h^\tau, a_h^\tau) \leq \kappa\sqrt{dN}. \tag{33}$$

Combining (30) and (33), we have

$$\sum_{\tau=1}^{N} \mathbb{E}_{\mu_h^{\mathrm{b},\tau}}\big[\Gamma_h^{c_i}(s,a)\big] = \sum_{\tau=1}^{N} X_h^{c_i,\tau} + \sum_{\tau=1}^{N} \Gamma_h^{c_i}(s_h^\tau, a_h^\tau) \leq C\kappa\sqrt{dN\log(DH/\xi)} \tag{34}$$

with probability at least $1 - \xi/(D+1)$.

For term $\mathbb{E}_{\mu_h^{\mathrm{b},\tau}}\big[\Gamma_h(s,a)\big]$, we follow a similar derivation. Let $X_h^\tau = \mathbb{E}_{\mu_h^{\mathrm{b},\tau}}\big[\Gamma_h(s,a)\big] - \Gamma_h(s_h^\tau, a_h^\tau)$ be a martingale difference sequence with bound $2H$ for each X_h^τ. Then, by the Azuma's inequality, it holds that

$$\sum_{\tau=1}^{N} X_h^\tau \leq CH\sqrt{N\log(DH/\xi)} \tag{35}$$

with probability at least $1 - \xi/(D+1)$. Similarly, we have

$$\begin{aligned}
\sum_{\tau=1}^{N} \Gamma_h(s_h^\tau, a_h^\tau) &\leq \sum_{\tau=1}^{N} H \int_{\mathcal{S}} \kappa\sqrt{\psi(s_h^\tau, a_h^\tau, s')^\top \Lambda_h^{-1} \psi(s_h^\tau, a_h^\tau, s')}\ \mathrm{d}s' \\
&\leq \kappa H \sqrt{|\mathcal{S}| N \sum_{\tau=1}^{N} \int_{\mathcal{S}} \psi(s_h^\tau, a_h^\tau, s')^\top \Lambda_h^{-1} \psi(s_h^\tau, a_h^\tau, s')\ \mathrm{d}s'}
\end{aligned} \tag{36}$$

and note that

$$\begin{aligned}
&\sum_{\tau=1}^{N} \int_{\mathcal{S}} \psi(s_h^\tau, a_h^\tau, s')^\top \Lambda_h^{-1} \psi(s_h^\tau, a_h^\tau, s')\, \mathrm{d}s' \\
=&\operatorname{tr}\left(\sum_{\tau=1}^{N} \int_{\mathcal{S}} \psi(s_h^\tau, a_h^\tau, s')^\top \Lambda_h^{-1} \psi(s_h^\tau, a_h^\tau, s')\, \mathrm{d}s'\right) \\
=&\operatorname{tr}\left(\sum_{\tau=1}^{N} \int_{\mathcal{S}} \psi(s_h^\tau, a_h^\tau, s') \psi(s_h^\tau, a_h^\tau, s')^\top \Lambda_h^{-1}\ \mathrm{d}s'\right) \\
\leq& d_1.
\end{aligned} \tag{37}$$

Combining (35), (36) and (37), we have

$$\sum_{\tau=1}^{N} \mathbb{E}_{\mu_h^{\mathrm{b},\tau}}\big[\Gamma_h(s,a)\big] = \sum_{\tau=1}^{N} X_h^\tau + \sum_{\tau=1}^{N} \Gamma_h(s_h^\tau, a_h^\tau) \leq C\kappa H\sqrt{dN|\mathcal{S}|\log(DH/\xi)} \tag{38}$$

holds with probability at least $1-\xi/(D+1)$.

Taking the union bound for (34) and (38), by (29), we have

$$\begin{aligned}\text{IntUncert}\,\tfrac{\pi^*}{\mathcal{D}} &= 2(1+\rho)\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\Gamma_h(s_h,a_h)+\|\Gamma_h^{\mathbf{c}}(s_h,a_h)\|_\infty \,\middle|\, s_1=\underline{s}\right]\\ &\le 2\varsigma(1+\rho)\sum_{h=1}^{H}N^{-1}\sum_{\tau=1}^{N}\left(\mathbb{E}_{\mu_h^{\mathrm{b},\tau}}\left[\Gamma_h(s,a)\right]+\mathbb{E}_{\mu_h^{\mathrm{b},\tau}}\left[\|\Gamma_h^{\mathbf{c}}(s,a)\|_\infty\right]\right)\\ &\le C\varsigma\kappa(1+\rho)H^2\sqrt{d|\mathcal{S}|/N\log(DH/\xi)}\end{aligned}$$

with probability at least $1-\xi$. Thus, we complete the proof of Corollary 1. □

H.6 Supporting lemmas and proofs

H.6.1 Proof of Lemma 2

Proof of Lemma 2.

Lemma 6 (Extended value difference (Cai et al., 2020)). *Let $\pi=\{\pi_h\}_{h=1}^H$ and $\pi'=\{\pi'_h\}_{h=1}^H$ be two arbitrary policies and let $\{Q_h\}_{h=1}^H$ be any given Q-functions. For any $h\in[H]$, we define a value function $V_h:\mathcal{S}\to\mathbb{R}$ by letting $V_h(s)=\langle Q_h(s,\cdot),\pi_h(\cdot\,|\,s)\rangle_{\mathcal{A}}$ for all $s\in\mathcal{S}$. Then, we have*

$$\begin{aligned}V_1(\underline{s})-V_1^{\pi'}(\underline{s}) &= \sum_{h=1}^{H}\mathbb{E}_{\pi'}\left[\left\langle Q_h(s_h,\cdot),\pi_h(\cdot\,|\,s_h)-\pi'_h(\cdot\,|\,s_h)\right\rangle_{\mathcal{A}}\,\middle|\, s_1=\underline{s}\right]\\ &\quad+\sum_{h=1}^{H}\mathbb{E}_{\pi'}\left[Q_h(s_h,a_h)-(\mathbb{B}_hV_{h+1})(s_h,a_h)\,\middle|\, s_1=\underline{s}\right].\end{aligned}$$

where $\underline{s}$ is an initial state.

Proof of Lemma 6. See Section B.1 in Cai et al. (2020) for a detailed proof. □

Applying Lemma 6 with $\pi=\pi^k$, $\pi'=\pi^*$, and $Q_h=Q_h^{i,k}$, we get

$$V_1^{i,k}(\underline{s})-V_1^{i,*}(\underline{s})=\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle Q_h^{i,k},\pi_h^k-\pi_h^*\right\rangle\,\middle|\, s_1=\underline{s}\right]+\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\iota_h^{i,k}\,\middle|\, s_1=\underline{s}\right]. \tag{39}$$

Moreover, applying Lemma 6 with $\pi=\pi'=\pi^k$, we get

$$V_1^{i,\pi^k}(\underline{s})-V_1^{i,k}(\underline{s})=-\sum_{h=1}^{H}\mathbb{E}_{\pi^k}\left[\iota_h^{i,k}\,\middle|\, s_1=\underline{s}\right]. \tag{40}$$

Since for any $k\in[K]$,

$$\boldsymbol{V}_1^{\pi^k}(\underline{s})-\boldsymbol{V}_1^*(\underline{s})=\boldsymbol{V}_1^{\pi^k}(\underline{s})-\boldsymbol{V}_1^k(\underline{s})+\boldsymbol{V}_1^k(\underline{s})-\boldsymbol{V}_1^*(\underline{s}), \tag{41}$$

we are done by plugging (39) and (40) into (41).

□

H.6.2 Projected subgradient method

Our algorithm benefits from the online projected subgradient method for the update of dual variables. We formally state it below for compactness.

Online learning. Online learning involves two players: the adversary and the player. The online learning protocol is shown in Algorithm 3.

Algorithm 3 Protocol of Online Learning

1: **for** $t = 1, \ldots, T$ **do**
2: The player chooses an action x_t.
3: The adversary picks a function f_t.
4: The player obtains reward $f_t(x_t)$.
5: The player learns via f_t.
6: **end for**

Note that there is no assumption on how the adversary will pick the function f_t, and it may be adversarially chosen. The player aims to minimize the regret:

$$\text{Regret} = \max_x \sum_{t=1}^{T} f_t(x) - \sum_{t=1}^{T} f_t(x_t), \tag{42}$$

which measures the quality of the player's strategy $x_1, \ldots, x_T$ compared with the single best desicion in hindsight.

Projected subgradient method. The projected subgradient method is a particular case of mirror descent/ascent with Euclidean distance. Applying this method to online learning produces a regret bound of the order $O(\sqrt{T})$.

We suppose that the actions x_t are required to be contained in some convex set $\mathcal{X}$, i.e., $x_t \in \mathcal{X}$. Let $g_t \in \partial f_t(x_t)$ denote a subgradient of f_t at x_t and G and R denote two constant bounds such that $\max_{x,y\in\mathcal{X}} \|x - y\|_2 \leq R$ and $\max_{t\in[T]} \|\partial f_t(x_t)\|_2 \leq G$. We set the step length η_t at the t-th iteration to $\frac{R}{G\sqrt{t}}$ if we do not know the number of iterations T in advance and to $\frac{R}{G\sqrt{T}}$ if we have the knowledge of T. The latter case will leads to an upper bound with a smaller constant multiplicative factor. With these notations, the update rule of projected subgradient method can be expressed as

$$x_{t+1} \leftarrow \arg\max_{x\in\mathcal{X}} \left\{ f_t(x_t) + \langle \eta_t g_t, x - x_t \rangle - \frac{1}{2}\|x - x_t\|_2^2 \right\}.$$

We describe the complete method in Algorithm 4.

Algorithm 4 projected subgradient method

1: Arbitrarily initialize $x_1 \in \mathcal{X}$.
2: **for** $t = 1, \ldots, T-1$ **do**
3: $x_{t+1} \leftarrow \arg\max_{x\in\mathcal{X}} \left\{ f_t(x_t) + \langle \eta_t g_t, x - x_t \rangle - \frac{1}{2}\|x - x_t\|_2^2 \right\}$
4: **end for**

By this method, the regret is guaranteed to increase sublinearly as stated in Theorem 5.

Theorem 5. *Using projected subgradient method mentioned in Algorithm 4, it holds that for the regret* (42) *that*

$$\text{Regret} \leq CRG\sqrt{T},$$

where C is an absolute constant.

Proof of Theorem 5. See Zinkevich (2003) for a detailed proof. □

H.6.3 Proof of Lemma 3

Proof of Lemma 3. The model evaluation error can be upper bounded by the ξ-uncertainty quantifier. We formally state it below.

Lemma 7. *Under event $\mathcal{E}$, it holds that for any $k \in [K]$ and $h \in [H]$,*

$$0 \leq \boldsymbol{\iota}_h^k(s,a) \leq 2\big(\Gamma_h(s,a) \cdot \mathbf{1} + \Gamma_h^{\mathbf{c}}(s,a)\big).$$

Proof of Lemma 7. See Appendix H.6.4 for a detailed proof. □

By Lemma 2, it holds that

$$(\boldsymbol{\theta}^k)^\top\big(\boldsymbol{V}_1^k(\underline{s})-\boldsymbol{V}_1^*(\underline{s})\big)=\underbrace{\sum_{h=1}^H\mathbb{E}_{\pi^*}\Big[(\boldsymbol{\theta}^k)^\top\boldsymbol{\iota}_h^k\,\Big|\,s_1=\underline{s}\Big]}_{\text{(i)}}+\underbrace{\sum_{h=1}^H\mathbb{E}_{\pi^*}\Big[\big\langle(\boldsymbol{\theta}^k)^\top\boldsymbol{Q}_h^k,\pi_h^k-\pi_h^*\big\rangle_{\mathcal{A}}\,\Big|\,s_1=\underline{s}\Big]}_{\text{(ii)}}.$$

We bound these two terms above separately. For (ii), since π^k in Algorithm 2 is greedy (see Line 5), we have (ii)≤ 0. For (i), by Lemma 7, we get

$$\text{(i)}\leq 2|\boldsymbol{\theta}^k|^\top\sum_{h=1}^H\mathbb{E}_{\pi^*}\big[\Gamma_h(s_h,a_h)\cdot\mathbf{1}+\Gamma_h^{\mathbf{c}}(s_h,a_h)\,\big|\,s_1=\underline{s}\big].$$

By plugging them back and applying Hölder's inequality, we have

$$(\boldsymbol{\theta}^k)^\top\big(\boldsymbol{V}_1^k(\underline{s})-\boldsymbol{V}_1^*(\underline{s})\big)\leq 2(1+\rho)\sqrt{D}\sum_{h=1}^H\mathbb{E}_{\pi^*}\big[\Gamma_h(s_h,a_h)+\|\Gamma_h^{\mathbf{c}}(s_h,a_h)\|_\infty\,\big|\,s_1=\underline{s}\big],$$

where we notice that $\|\boldsymbol{\theta}^k\|_1\leq(1+\rho)\sqrt{D}$. Thus, we finish the proof of Lemma 3. □

H.6.4 Proof of Lemma 7

Proof of Lemma 7. For any $i\in[D]$, recall that we have

$$Q_h^{i,k}(s,a)=\min\big\{\overline{Q}_h^{i,k}(s,a),H-h+1\big\}_+.$$

Under event $\mathcal{E}$, we have

$$\begin{aligned}\overline{Q}_h^{i,k}(s,a)=&\widehat{c}_h^i(s,a)+\widehat{\mathbb{P}}_h[V_{h+1}^{i,k}](s,a)+\Gamma_h+\Gamma_h^{c^i}\\ \geq&c_h^i(s,a)+\mathbb{P}_h[V_{h+1}^{i,k}]\\ \geq&0,\end{aligned}$$

where the last inequality follows from $V_{h+1}^{i,k}\in[0,H-h]$. Therefore, it holds that $Q_h^{i,k}(s,a)\leq\overline{Q}_h^{i,k}(s,a)$ and

$$\begin{aligned}Q_h^{i,k}(s,a)=&\min\big\{\overline{Q}_h^{i,k}(s,a),H-h+1\big\}_+\\ \geq&\min\big\{c_h^i(s,a)+\mathbb{P}_h[V_{h+1}^{i,k}],H-h+1\big\}_+\\ =&c_h^i(s,a)+\mathbb{P}_h[V_{h+1}^{i,k}],\end{aligned}$$

which implies

$$\iota_h^{i,k}(s,a)=Q_h^{i,k}(s,a)-\big[c_h^i(s,a)+\mathbb{P}_hV_{h+1}^{i,k}(s,a)\big]\geq 0.$$

It remains to establish an upper bound for $\iota_h^{i,k}(s,a)$. To that end, we have

$$\begin{aligned}\iota_h^{i,k}(s,a)=&Q_h^{i,k}(s,a)-\big[c_h^i(s,a)+\mathbb{P}_hV_{h+1}^{i,k}(s,a)\big]\\ \leq&\overline{Q}_h^{i,k}(s,a)-\big[c_h^i(s,a)+\mathbb{P}_hV_{h+1}^{i,k}(s,a)\big]\\ =&\big[\widehat{c}_h^i(s,a)+\widehat{\mathbb{P}}_h[V_{h+1}^{i,k}](s,a)+\Gamma_h(s,a)+\Gamma_h^{c^i}(s,a)\big]-\big[c_h^i(s,a)+\mathbb{P}_h[V_{h+1}^{i,k}](s,a)\big]\\ =&\big[\Gamma_h^{c^i}(s,a)-c_h^i(s,a)+\widehat{c}_h^i(s,a)\big]+\big[\Gamma_h(s,a)-\mathbb{P}_hV_{h+1}^{i,k}(s,a)+\widehat{\mathbb{P}}_h[V_{h+1}^{i,k}](s,a)\big]\\ \leq&2\big(\Gamma_h(s,a)+\Gamma_h^{c^i}(s,a)\big),\end{aligned}$$

where the last inequality follows from the definition of the ξ-uncertainty quantifier. Thus, we finish the proof of Lemma 7. □

H.6.5 Proof of Lemma 5

Proof of Lemma 5. In what follows, we show that $\{(\Gamma_h^{\mathcal{P}}, \Gamma_h^{\mathbf{c}})\}_{h=1}^H$ defined in (24) are ξ-uncertainty quantifier for linear kernel CMOMDP.

Uncertainty quantifier for $\mathcal{P}$. We first show that $\Gamma_h^{\mathcal{P}}$ is the ξ-uncertainty quantifier of $\mathcal{P}$.

By definition, we have

$$\begin{aligned}\mathcal{P}_h(s' \,|\, s, a) &= \psi(s, a, s')^\top \theta_h = \psi(s, a, s')^\top \Lambda_h^{-1} \Lambda_h \theta_h \\ &= \psi(s, a, s')^\top \Lambda_h^{-1} \bigg(\sum_{\tau=1}^N \int_{\mathcal{S}} \psi(s_h^\tau, a_h^\tau, s') \mathcal{P}_h(s' \,|\, s_h^\tau, a_h^\tau) \,\mathrm{d}s' + \lambda \theta_h \bigg).\end{aligned}$$

Thus, by the closed form of $\widehat{\theta}_h$ in (22), we have

$$\begin{aligned}&\mathcal{P}_h(s' \,|\, s, a) - \widehat{\mathcal{P}}_h(s' \,|\, s, a) \\ &\quad = \mathcal{P}_h(s' \,|\, s, a) - \psi(s, a, s')^\top \widehat{\theta}_h \\ &\quad = \underbrace{\psi(s, a, s')^\top \Lambda_h^{-1} \bigg(\sum_{\tau=1}^N \Big(\int_{\mathcal{S}} \psi(s_h^\tau, a_h^\tau, s') \mathcal{P}_h(s' \,|\, s_h^\tau, a_h^\tau) \,\mathrm{d}s' - \psi(s_h^\tau, a_h^\tau, s_{h+1}^\tau) \Big) \bigg)}_{\text{(i)}} \\ &\qquad + \underbrace{\lambda \cdot \psi(s, a, s')^\top \Lambda_h^{-1} \theta_h}_{\text{(ii)}}.\end{aligned} \tag{43}$$

For term (i), by Cauchy-Schwartz inequality, we have

$$\begin{aligned}|\text{(i)}| &\le \big\|\psi(s, a, s')\big\|_{\Lambda_h^{-1}} \cdot \bigg\| \sum_{\tau=1}^N \Big(\int_{\mathcal{S}} \psi(s_h^\tau, a_h^\tau, s') \mathcal{P}_h(s' \,|\, s_h^\tau, a_h^\tau) \,\mathrm{d}s' - \psi(s_h^\tau, a_h^\tau, s_{h+1}^\tau) \Big) \bigg\|_{\Lambda_h^{-1}} \\ &\le C_1 R \cdot \sqrt{d \log(dN) + \log(DH/\xi)} \cdot \big\|\psi(s, a, s')\big\|_{\Lambda_h^{-1}}, \quad \forall (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}, h \in [H]\end{aligned}$$

with probability at least $1 - \xi/(D+1)$. Now we prove the last inequality of the above derivation. To that end, we need the following lemma that generalizes the Theorem 1 in Abbasi-Yadkori et al. (2011) to function-valued process.

Lemma 8 (Self-normalized bound for function-valued Process)**.** *Let Ω be a probability space and $\{\eta_t\}_{t=1}^\infty$ be a function-valued stochastic process with a filtration $\{\mathcal{G}_t\}_{t=0}^\infty$, i.e., $\eta_t : \Omega \times \mathcal{S} \to \mathbb{R}$. We assume that $\eta_t \,|\, \mathcal{G}_{t-1}$ is zero-mean and $\sigma-$sub-Gaussian, i.e.,*

$$\mathbb{E}\big[\eta_t(s) \,\big|\, \mathcal{G}_{t-1}\big] = 0, \quad \forall s \in \mathcal{S},$$

$$\log \mathbb{E}\Big[\exp\big(\langle f, \eta_t \rangle \big) \,\Big|\, \mathcal{G}_{t-1} \Big] \le \|f\|_\infty^2 \cdot \sigma^2/2, \quad \forall f : \mathcal{S} \to \mathbb{R}.$$

Let $\{X_t\}_{t=0}^\infty$ be an $\mathbb{R}^d$-function-valued stochastic process, i.e., $X_t : \Omega \times \mathcal{S} \to \mathbb{R}^d$, and suppose X_t is $\mathcal{G}_{t-1}$-measurable. We further assume that

$$\|\lambda^\top X_t\|_\infty \le R \cdot \|\lambda^\top X_t\|_2 \tag{44}$$

almost surely for any $\lambda \in \mathbb{R}^d$. Let $V \in \mathbb{R}^{d \times d}$ be a positive definite matrix. We define

$$\overline{V}_t = V + \sum_{\tau=1}^t \int_{\mathcal{S}} X_\tau(s) X_\tau(s)^\top \,\mathrm{d}s$$

and

$$S_t = \sum_{\tau=1}^t \langle X_\tau, \eta_\tau \rangle_{\mathcal{S}}.$$

Then for any $\delta > 0$, with probability at least $1 - \delta$, it holds that for any $t \ge 0$,

$$\|S_t\|_{\overline{V}_t^{-1}}^2 \le 2\sigma^2 R^2 \log\bigg(\frac{\det(\overline{V}_t)^{1/2}}{\delta \det(V)^{1/2}} \bigg).$$

Proof of Lemma 8. See Appendix H.6.6 for a detailed proof. □

We consider the filtration $\{\mathcal{F}_h^\tau\}_{h,\tau=1}^{H,N}$ defined in Assumption 2. Then it holds that

$$\mathbb{E}\big[\delta_{s_{h+1}^\tau}(s') \,|\, \mathcal{F}_h^\tau\big] = \mathcal{P}_h(s' \,|\, s_h^\tau, a_h^\tau).$$

For any $f : \mathcal{S} \to \mathbb{R}$, by Hölder's inequality, it holds that $\langle f, \mathcal{P}_h(\cdot \,|\, s_h^\tau, a_h^\tau) - \delta_{s_{h+1}^\tau}\rangle_{\mathcal{S}} \le 2\|f\|_\infty$, which implies

$$\log \mathbb{E}\Big[\exp\big(\langle f, \mathcal{P}_h(\cdot \,|\, s_h^\tau, a_h^\tau) - \delta_{s_{h+1}^\tau}\rangle\big) \,\Big|\, \mathcal{F}_{h,\tau}\Big] \le 4\|f\|_\infty^2/2.$$

It corresponds to the conditional 2-sub-Gaussianity. Moreover, noticing that $\psi(s_h^\tau, a_h^\tau, s')$ is $\mathcal{F}_h^\tau$-measurable and both $\mathcal{P}(\cdot \,|\, s_h^\tau, a_h^\tau)$ and $\delta_{s_{h+1}^\tau}$ are $\mathcal{F}_{h+1}^\tau$-measurable, we apply Lemma 8 with $X_\tau = \psi(s_h^\tau, a_h^\tau, \cdot)$, $\eta_h = \mathcal{P}_h(\cdot \,|\, s_h^\tau, a_h^\tau) - \delta_{s_{h+1}^\tau}$ and $V = \lambda I$ to get

$$\begin{aligned}\Big\|\sum_{\tau=1}^N \big(\int_{\mathcal{S}} \psi(s_h^\tau, a_h^\tau, s')\mathcal{P}_h(s' \,|\, s_h^\tau, a_h^\tau)\,\mathrm{d}s' - \psi(s_h^\tau, a_h^\tau, s_{h+1}^\tau)\big)\Big\|_{\Lambda_h^{-1}}^2 \\ \le 8R^2 \cdot \log\big(H/p \cdot \det(\Lambda_h)^{1/2}\det(\lambda I)^{-1/2}\big)\end{aligned} \tag{45}$$

with probability at least $1 - p/H$. It remains to upper bound $\det(\Lambda_h)$.

By Definition 2, we have

$$y^\top \Lambda_h y = \lambda\|y\|_2^2 + \sum_{\tau=1}^N \big\langle y^\top \psi(s_h^\tau, a_h^\tau, \cdot), y^\top \psi(s_h^\tau, a_h^\tau, \cdot)\big\rangle \le \lambda \cdot \|y\|_2^2 + dN \cdot \|y\|_2^2,$$

which implies $\|\Lambda_h\|_2 \le \lambda + dN$, and therefore,

$$\det(\Lambda_h) \le \|\Lambda_h\|_2^d \le (\lambda + dN)^d. \tag{46}$$

Setting $\lambda = 1$ and plugging (46) back into (45), we get

$$\begin{aligned}&\Big\|\sum_{\tau=1}^N \bigg(\int_{\mathcal{S}} \psi(s_h^\tau, a_h^\tau, s')\mathcal{P}_h(s' \,|\, s_h^\tau, a_h^\tau)\,\mathrm{d}s' - \psi(s_h^\tau, a_h^\tau, s_{h+1}^\tau)\bigg)\Big\|_{\Lambda_h^{-1}}^2 \\ &\le 8R^2 \cdot \big(1/2 \cdot d\log(1 + dN) + \log(H/p)\big) \\ &\le CR^2 \cdot \big(d\log(dN) + \log(H/p)\big)\end{aligned} \tag{47}$$

holds with probability at least $1 - p/H$. Here C is an absolute constant. By the union bound, (47) holds for all $h \in [H]$ with probability at least $1 - p$.

For term (ii) in (43), by setting $\lambda = 1$, we have

$$|\text{(ii)}| \le \big\|\psi(s, a, s')\big\|_{\Lambda_h^{-1}} \cdot \|\theta_h\|_{\Lambda_h^{-1}} \le \sqrt{d} \cdot \big\|\psi(s, a, s')\big\|_{\Lambda_h^{-1}} \tag{48}$$

where the last inequality is due to the definition of linear kernel MDP (Definition 2) and $\|\Lambda_h^{-1}\|_2 \le 1$.

By plugging (47) and (48) into (43), we get, for all $h \in [H], (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,

$$\begin{aligned}|\mathcal{P}_h(s' \,|\, s, a) - \widehat{\mathcal{P}}_h(s' \,|\, s, a)| &\le CR\sqrt{d\log(dN) + \log(DH/\xi)} \cdot \big\|\psi(s, a, s')\big\|_{\Lambda_h^{-1}} \\ &\le \kappa \cdot \big\|\psi(s, a, s')\big\|_{\Lambda_h^{-1}}\end{aligned} \tag{49}$$

holds with probability at least $1 - \xi/(D + 1)$.

Uncertainty quantifier for c. Due to the closed form solution of $\widehat{\theta}_h^{c^i}$ in (23), we have

$$\begin{aligned}
&\big|c_h^i(s,a) - \widehat{c}_h^i(s,a)\big| \\
=&|\varphi(s,a)^\top(\theta_h^{c^i} - \widehat{\theta}_h^{c^i})| \\
=&|\varphi(s,a)^\top \Lambda_{\varphi,h}^{-1}\big(\sum_{\tau=1}^N c_h^{i,\tau}\cdot\varphi(s_h^\tau,a_h^\tau) - \Lambda_{\varphi,h}\theta_h^{c^i}\big)| \\
=&|\varphi(s,a)^\top \Lambda_{\varphi,h}^{-1}\Big(\sum_{\tau=1}^N \varphi(s_h^\tau,a_h^\tau)\big(c_h^{i,\tau} - \varphi(s_h^\tau,a_h^\tau)^\top\theta_h^{c^i}\big) - \lambda\theta_h^{c^i}\Big)| \\
\le&\|\varphi(s,a)\|_{\Lambda_{\varphi,h}^{-1}}\cdot\|\sum_{\tau=1}^N \varphi(s_h^\tau,a_h^\tau)\big(c_h^{i,\tau} - \varphi(s_h^\tau,a_h^\tau)^\top\theta_h^{c^i}\big)\|_{\Lambda_{\varphi,h}^{-1}} + \lambda\|\theta_h^{c^i}\|_{\Lambda_{\varphi,h}^{-1}}
\end{aligned}$$

Following a similar argument as we did for $\mathcal{P}$, we obtain a result analogous to (49),

$$\|c_h^i(s,a) - \widehat{c}_h^i(s,a)\| \le \kappa\cdot\|\varphi(s,a)\|_{\Lambda_{\phi,h}^{-1}}.$$

with probability at least $1-p$.

Finally, by setting $p = \xi/(D+1)$ and taking union bound for $\mathcal{P}$ and c^i ($i\in[D]$), we complete the proof of Lemma 5. □

H.6.6 Proof of Lemma 8

Proof of Lemma 8. We generalize the proof of Abbasi-Yadkori et al. (2011) as follows.

Lemma 9. *Let $\lambda\in\mathbb{R}^d$ be an arbitrary vector and*

$$M_t^\lambda = \exp\left[\sum_{\tau=1}^t\left(\frac{\langle\lambda^\top X_\tau,\eta_\tau\rangle}{\sigma^2R^2} - \frac{\|\lambda^\top X_\tau\|_2^2}{2}\right)\right].$$

Let T be a stopping time with respect to $\{\mathcal{G}_t\}_{t=1}^\infty$. Then M_T^λ is almost surely well-defined and $\mathbb{E}[M_T^\lambda]\le 1$.

Proof of Lemma 9. We first show that $\{M_t^\lambda\}_{t=0}^\infty$ is a supermartingale. To see this, we have

$$\begin{aligned}
\mathbb{E}\left[M_t^\lambda\,\middle|\,\mathcal{G}_{t-1}\right] &= M_{t-1}^\lambda\cdot\mathbb{E}\left[\exp(\frac{\langle\lambda^\top X_t,\eta_t\rangle}{\sigma^2R^2} - \frac{\|\lambda^\top X_t\|_2^2}{2})\,\middle|\,\mathcal{G}_{t-1}\right] \\
&\le M_{t-1}^\lambda\cdot\mathbb{E}\left[\exp(\frac{\|\lambda^\top X_t\|_\infty^2}{2R^2} - \frac{\|\lambda^\top X_t\|_\infty^2}{2R^2})\,\middle|\,\mathcal{G}_{t-1}\right] \\
&= M_{t-1}^\lambda.
\end{aligned}$$

where the inequality is due to the conditional σ-sub-Gaussianity of η_t and the condition in (44). It also implies $\mathbb{E}[M_t^\lambda]\le 1$ for any $t\ge 0$. By martingale convergence theorem, with $t\to\infty$, M_t^λ almost surely converges to a random variable M_∞^λ with finite expectation, and thus M_T^λ is well-defined almost surely. Applying Fatou's lemma, we have

$$\mathbb{E}\left[M_T^\lambda\right] = \mathbb{E}\left[\liminf_{t\to\infty} M_{T\wedge t}^\lambda\right] \le \liminf_{t\to\infty}\mathbb{E}\left[M_{T\wedge t}^\lambda\right] \le 1.$$

□

Lemma 10. *Let T be a stopping time with respect to $\{\mathcal{G}_t\}_{t=0}^\infty$. Then the following holds with probability at least $1-\delta$*

$$\|S_T\|_{\overline{V}_t^{-1}}^2 \le 2\sigma^2R^2\log\left(\frac{\det(\overline{V}_t)^{1/2}}{\delta\det(V)^{1/2}}\right).$$

Proof of Lemma 10. For notational simplicity, we assume $\sigma \cdot R = 1$. We define

$$V_t = \sum_{\tau=1}^{t} \int_{\mathcal{S}} X_\tau(s) X_\tau(s)^\top \, \mathrm{d}s$$

and therefore $\overline{V}_t = V_t + V$. Then we can write $M_t^\lambda = \exp(\lambda^\top S_t - \|\lambda\|_{V_t}^2/2)$. Let Λ be a $\mathbb{R}^d$-valued Gaussian random variable with covariance V^{-1} and that it is independent of $\{\mathcal{G}_t\}_{t=0}^\infty$. Let $M_t = \mathbb{E}\left[M_t^\Lambda \,\middle|\, \mathcal{G}_\infty\right]$ where $\mathcal{G}_\infty = \sigma(cup_{t=0}^\infty \mathcal{G}_t)$. Let q denote the density of Λ and $v(A) = \int \exp(-x^\top A x/2)\,\mathrm{d}x = \sqrt{(2\pi)^d/\det(A)}$ for any positive definite matrix A.

Then, we have

$$\begin{aligned} M_t &= \int_{\mathbb{R}^d} \exp(\lambda^\top S_t - \|\lambda\|_{V_t}^2/2) q(\lambda)\,\mathrm{d}\lambda \\ &= \int_{\mathbb{R}^d} \exp(-\|\lambda - V_t^{-1} S_t\|_{V_t}^2/2 + \|S_t\|_{V_t^{-1}}^2/2) q(\lambda)\,\mathrm{d}\lambda \\ &= v(V)^{-1} \cdot \exp(\|S_t\|_{V_t^{-1}}^2/2) \cdot \int_{\mathbb{R}^d} \exp\Big(-\big(\|\lambda - V_t^{-1} S_t\|_{V_t}^2 + \|\lambda\|_V^2\big)/2 \Big)\,\mathrm{d}\lambda. \end{aligned} \tag{50}$$

Note that

$$\begin{aligned} \|\lambda - V_t^{-1} S_t\|_{V_t}^2 + \|\lambda\|_V^2 &= \|\lambda - \overline{V}_t^{-1} S_t\|_{\overline{V}_t}^2 + \|V_t^{-1} S_t\|_{V_t}^2 - \|S_t\|_{\overline{V}_t^{-1}}^2 \\ &= \|\lambda - \overline{V}_t^{-1} S_t\|_{\overline{V}_t}^2 + \|S_t\|_{V_t^{-1}}^2 - \|S_t\|_{\overline{V}_t^{-1}}^2. \end{aligned} \tag{51}$$

By plugging (51) into (50), we get

$$\begin{aligned} M_t &= v(V)^{-1} \cdot \exp\big(\|S_t\|_{\overline{V}_t^{-1}}^2/2\big) \cdot \int_{\mathbb{R}^d} \exp\big(-\|\lambda - \overline{V}_t^{-1} S_t\|_{\overline{V}_t}^2/2\big)\,\mathrm{d}\lambda \\ &= \frac{v(\overline{V}_t)}{v(V)} \cdot \exp\big(\|S_t\|_{\overline{V}_t^{-1}}^2/2\big) \\ &= \sqrt{\det(V)/\det(\overline{V}_t)} \cdot \exp\big(\|S_t\|_{\overline{V}_t^{-1}}^2/2\big). \end{aligned}$$

Hence, we have

$$\mathbb{P}\Big(\|S_T\|_{\overline{V}_T^{-1}}^2 > 2\log\big(\frac{\det(\overline{V}_T)^{1/2}}{\delta \det(V)^{1/2}}\big)\Big) = \mathbb{P}(\delta \cdot M_T > 1) \le \mathbb{E}[\delta \cdot M_T] \le \delta,$$

which completes the proof of Lemma 10. □

We construct a stopping time as below.

$$T = \inf\left\{ t \ge 0 : \|S_t\|_{\overline{V}_t^{-1}}^2 > 2\log\Big(\frac{\det(\overline{V}_t)^{1/2}}{\delta \det(V)^{1/2}}\Big) \right\}$$

Then, we have

$$\begin{aligned} &\mathbb{P}\left(\exists t \ge 0, \|S_t\|_{\overline{V}_t^{-1}}^2 > 2\log\left(\tfrac{\det(\overline{V}_t)^{1/2}}{\delta \det(V)^{1/2}}\right)\right) \\ &= \mathbb{P}(T < \infty) \\ &= \mathbb{P}\left(\|S_T\|_{\overline{V}_T^{-1}}^2 > 2\log\left(\tfrac{\det(\overline{V}_T)^{1/2}}{\delta \det(V)^{1/2}}\right), T < \infty\right) \\ &\le \mathbb{P}\left(\|S_T\|_{\overline{V}_T^{-1}}^2 > 2\log\left(\tfrac{\det(\overline{V}_T)^{1/2}}{\delta \det(V)^{1/2}}\right)\right) \\ &\le \delta, \end{aligned}$$

which completes the proof of Lemma 8. □

I Experiments

Experiments are conducted on tabular CMOMDPs as follows. We define the constraint set as $\mathcal{W}^* = \{x \in \mathbb{R}^D : \|x\|_2 \leq 1\}$ for simplicity, and one can verify that it satisfies Assumption 1. The transition kernel $\mathcal{P}$ and cost function $\boldsymbol{c}$ are generated uniformly at random from $[0, 1]$ (and we conduct normalization for $\mathcal{P}$). We make the cost deterministic for simplicity. In addition, we set $\mathcal{P}_h(s_0 \,|\, s_0, a_0) = 1$ and $\boldsymbol{c}_h(s_0, a_0) = 0$ for a certain state action pair $(s_0, a_0) \in \mathcal{S} \times \mathcal{A}$ for all $h \in [H]$, and the initial state is set to s_0. The intuition here is to ensure that the optimal policy, which always takes action a_0, achieves zero total cost and zero constraint violation for simplicity. The dataset is generated by a uniformly random experimenter, i.e., it picks $a \in \mathcal{A}$ uniformly at random at each step. Hyperparameters are listed in Table 1.

Table 1: List of hyperparameters

Hyperparameter	Value
H: horizon	5
D: dimension of cost function	6
$\lvert\mathcal{S}\rvert$: cardinality of state space	5
$\lvert\mathcal{A}\rvert$: cardinality of action space	5
$\lvert\mathcal{D}\rvert$: dataset size	50000
K: number of iteration of PEDI	100
δ: confidence level	0.9
η: step length	0.01
ν: scaling constant	3

In our implementation, PEDI estimates the transition and cost functions by the empirical mean, i.e., $\widehat{\mathcal{P}}_h(s, a) = n_h(s, a, s')/n_h(s, a)$ and $\widehat{c}_h^i(s, a) = f_h^i(s, a)/n_h(s, a)$ for $i \in [D]$ where $n_h(s, a)$ is the number of visits to (s, a) at step h and $f_h^i(s, a)$ is the sum of the i-th cost incurred in the dataset when visiting (s, a) at step h. We construct the Hoeffding-style uncertainty quantifiers, i.e., $\Gamma_h^{\mathcal{P}}(s, a, s') = \sqrt{\log(2H|\mathcal{S}||\mathcal{A}||\mathcal{S}|/\delta)/(2n_h(s, a))}$ and $\Gamma_h^{\boldsymbol{c}} = \sqrt{\log(2DH|\mathcal{S}||\mathcal{A}|/\delta)/(2n_h(s, a))}$. We can verify that they satisfy the definition (Definition 1).

We conduct experiments to see how PEDI converges to the optimal policy with different preference functions: quadratic functions, polynomial functions, and their combinations.

Quadratic Functions. Suppose the interplay of cost functions can be modeled by a positive definite matrix A, a vector b and a constant c, i.e., the preference function is defined as

$$g(x) = \frac{1}{2} x^\top A x + b^\top x + c,$$

where A is positive definite. For simplicity, we assume b is the zero vector and $c = 0$. To guarantee 1-Lipschitzness, it suffices to restrict the spectral radius $\lambda_{\max}$. In particular, we require $\lambda_{\max}(A) \leq 1/(2HD^{1/2})$ since $\|\partial_x g\|_2 = \|2Ax\|_2 \leq 2\lambda_{\max}(A)HD^{1/2}$. For the convex conjugate, we can verify that $g^*(x^*) = \frac{1}{2}(x^* - b)^\top A^{-1}(x^* - b) - c = \frac{1}{2} x^* A^{-1} x^*$. In the numerical experiment, the matrix A is randomly generated with the mentioned spectral radius requirement. The results are given in Table 2.

Table 2: Results of quadratic preference functions

Iteration k	Suboptimality	Constraint Violation
1	0.067	0.880
2	0.505	4.007
3	0.067	0.880
$4, 5, \ldots, 100$	0.000	0.000

As we see, it converges to the optimal policy in mere four iterations and stays optimal permanently.

Polynomial Functions. Suppose the preference function is polynomial, i.e.,

$$g(x) = \sum_{i=1}^{D} c_i |x_i|^{p_i}.$$

For simplicity, we assume $p = p_i = p_j$ and $c = c_i = c_j$ for any $1 \leq i, j \leq D$. To ensure 1-Lipschitzness, it suffices to set $c = 1/(pH^{p-1}D^{1/2})$ for all i which results in $\|\partial_x g\|_2 = cpx^{p-1}D^{1/2} \leq 1$ for $x \geq 0$. Then, we have $g^*(x^*) = \sum_{i=1}^{D} \frac{|x_i^*|^q}{c^{q-1}p^{q-1}q}$ where $\frac{1}{p} + \frac{1}{q} = 1$. In the numerical experiment, we set $p = 2$. The results are shown in Table 3.

Table 3: Results of polynomial preference functions

Iteration k	Suboptimality	Constraint Violation
1	0.165	1.009
2	0.139	0.844
$3, 4, \ldots, 100$	0.000	0.000

As we see, it reaches the optimal solution in only three iterations.

Combination of Quadratic Functions and Polynomial Functions. We consider more complex scenarios where preference functions are combinations of quadratic functions and polynomial functions, i.e., $g(x) = g_1(x_1) + g_2(x_2)$ with

$$g_1(x) = \frac{1}{2}x^\top Ax + b^\top x, \qquad g_2(x) = \sum_{i=1}^{D_2} c_i |x_i|^{p_i}.$$

Here $x = (x_1^\top, x_2^\top)^\top$ with $x_1 \in \mathbb{R}^{D_1}$, $x_2 \in \mathbb{R}^{D_2}$ and $D_1 + D_2 = D$. It is clear that $g^*(x^*) = g_1^*(x_1^*) + g_2^*(x_2^*)$. In experiments we set $D_1 = D_2 = D/2$. Moreover, we impose similar requirements and restrictions as we did previously to ensure 1-Lipschitzness. The numerical results are in Table 4, which show that the suboptimality and constraint violation decrease as the number of iterations increases, and PEDI finds the solution in nine iterations.

Table 4: Results of functions that are a combination of quadratic functions and polynomial preference functions

Iteration k	Suboptimality	Constraint Violation
1	1.517	6.033
2	1.490	6.000
$3, 4, 5$	1.438	5.864
6	1.333	5.609
7	0.811	4.152
8	0.093	0.884
$9, 10, \ldots, 100$	0.000	0.000

The following two remarks discuss (1) possibilities to handle other (even more general) preference functions and (2) some practical variants of PEDI for application, which is left to future work, as this paper is mainly theoretical.

Remark 1 (General Preference Function)**.** *In addition to the above demonstration, PEDI is also easily applicable to preference functions from many function classes such as exponential function, logarithmic function, and entropy function. Even when the exact expression of the preference function g is not good or even unknown, PEDI applies as long as we can approximate g^* by some numerical methods, say, by directly approximating $\sup_x(\langle x^*, x\rangle - g(x))$, which is the definition of the convex conjugate. To obtain the subgradient, we can use certain techniques such as numerical differentiation.*

Remark 2 (General Planning Algorithm)**.** *For a real-world application, the pessimistic planning (*PESSPLANNING*, see Algorithm 1) may seem too heavy. It can be replaced by any algorithms as long as it approximately produces the desired policy π^k and a pessimistic estimation of the value functions V^k at each iteration. For example, we can apply policy iteration algorithms or even any neural network-based approximate algorithms.*