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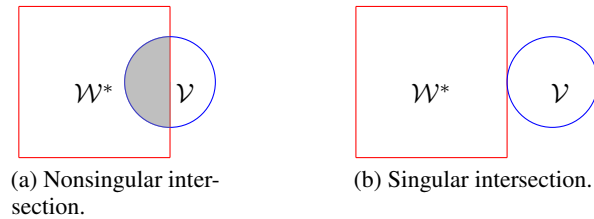
Figure 1: The illustration of two types of intersections: the nonsingular intersection and the singular intersection. The target set $\mathcal{W}^*$ is represented by a red square and the achievable values $\mathcal{V}$ is represented by a blue circle. The intersection is in grey.

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

14

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

$$\mathcal{W}^* = \left\{ (c^1, c^2) \,|\, c^1 + c^2 \geq H/2, c^1 + c^2 \leq 3H/2, c^1 - c^2 \geq -H/2, c^1 - c^2 \leq H/2 \right\}.$$
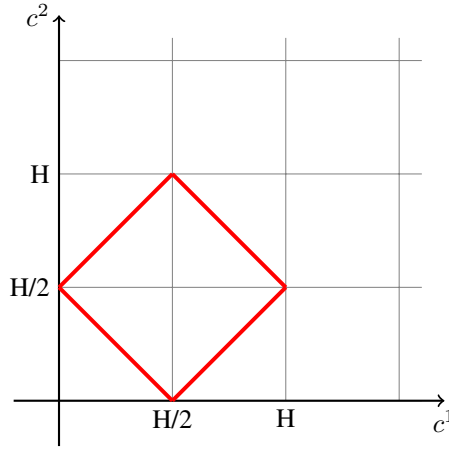
Figure 2: An example of target set $\mathcal{W}^* = \{(c^1, c^2) \,|\, c^1 + c^2 \geq H/2, c^1 + c^2 \leq 3H/2, c^1 - c^2 \geq -H/2, c^1 - c^2 \leq H/2\}$. Although it does not satisfy Assumption 1 of being a lower set, we can cut its boundaries into four pieces, each of which corresponds to a partial 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:

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

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.

15

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. $\qquad\square$

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,

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

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

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

16

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

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

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

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

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

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

where the second equality is by definition and the fact that $b_1' = H \ge V^1$. 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

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

while for p_2^*, we have

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

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 $\mathrm{dist}(\boldsymbol{V}^\dagger, \mathcal{W}^*) = 0$.

Suppose $\mathrm{dist}(\boldsymbol{V}^\dagger, \mathcal{W}^*) \ne 0$, 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

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

Then, by (21), we have

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

where the last equality is for $\mathrm{dist}(\boldsymbol{V}^\S, \mathcal{W}^*) = 0$. 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).

17

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,

$$\widehat{\mathcal{P}}_h(s' \mid s, a) = \psi(s, a, s')^\top \widehat{\theta}_h$$

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

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 $\widehat{\theta}_h$,

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

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

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

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

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

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

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

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.

18

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*

$$\boldsymbol{V}_1^k(\underline{s}) - \boldsymbol{V}_1^*(\underline{s}) = \sum_{h=1}^H \mathbb{E}_{\pi^*}[\boldsymbol{\iota}_h^k(s_h, a_h) \,|\, s_1 = \underline{s}]$$

$$+ \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\left\langle \boldsymbol{Q}_h^k(s_h, \cdot), (\pi_h^k - \pi_h^*)(\cdot \,|\, s_h)\right\rangle_{\mathcal{A}} \,\Big|\, s_1 = \underline{s}\right]$$

and that

$$\boldsymbol{V}_1^{\pi_k}(\underline{s}) - \boldsymbol{V}_1^*(\underline{s}) = \underbrace{-\sum_{h=1}^H \mathbb{E}_{\pi^k}\left[\boldsymbol{\iota}_h^k(s_h, a_h) \,\big|\, s_1 = \underline{s}\right]}_{\text{(i) Spurious Correlation}} + \underbrace{\sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\boldsymbol{\iota}_h^k(s_h, a_h) \,\big|\, s_1 = \underline{s}\right]}_{\text{(ii) Intrinsic Uncertainty}}$$

$$+ \underbrace{\sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\left\langle \boldsymbol{Q}_h^k(s_h, \cdot), (\pi_h^k - \pi_h^*)(\cdot \,|\, s_h)\right\rangle_{\mathcal{A}} \,\Big|\, s_1 = \underline{s}\right]}_{\text{(iii) Optimization Error}},$$

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. $\qquad\square$

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 $\psi(s, a, s') = \boldsymbol{e}_{(s,a,s')}, (\theta_h)_{(s,a,s')} = \mathcal{P}(s' \,|\, s, a), \varphi(s, a) = \boldsymbol{e}_{(s,a)}$, 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 $\pi_h(a_h \,|\, s_h)$ 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 $s \in \mathcal{S}, i \in [D]$,

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

19

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

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

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

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

Thus, we complete the proof of Lemma 1.

$\qquad\square$

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 $\{(\Gamma_h^{\mathcal{P}}, \Gamma_h^{\boldsymbol{c}})\}_{h=1}^H$ in Algorithm 2 is a ξ-uncertainty quantifier. Then under event $\mathcal{E}$, we have*

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

Proof of Lemma 3. See Appendix H.6.3 for a detailed proof. $\qquad\square$

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

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

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

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

We observe these two terms

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

20

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

$$
K\Big[g\big(V_1^{\widehat{\pi}}(\underline{s})\big) - g\big(V_1^*(\underline{s})\big) + \rho\,\mathrm{dist}\big(V_1^{\widehat{\pi}}(\underline{s}), \mathcal{W}^*\big)\Big]
$$
$$
\leq \sum_{k=1}^{K}\big\{(\boldsymbol{\beta}^k)^\top V_1^k(\underline{s}) - g^*(\boldsymbol{\beta}^k)\big\} - Kg\big(V_1^*(\underline{s})\big) + \rho\sum_{k=1}^{K}\big\{(\boldsymbol{\alpha}^k)^\top V_1^{\,k}(\underline{s}) - \max_{\boldsymbol{x}\in\mathcal{W}^*}(\boldsymbol{\alpha}^k)^\top\boldsymbol{x}\big\} \quad (25)
$$
$$
+ C(1+\rho)\sqrt{DH^2K}
$$

where C is a constant. Then, by observing

$$
g^*(\boldsymbol{\beta}^k) = \max_{\boldsymbol{V}}\big\{(\boldsymbol{\beta}^k)^\top\boldsymbol{V} - g(\boldsymbol{V})\big\} \geq (\boldsymbol{\beta}^k)^\top V_1^*(\underline{s}) - g\big(V_1^*(\underline{s})\big),
$$

$$
\max_{\boldsymbol{x}\in\mathcal{W}^*}(\boldsymbol{\alpha}^k)^\top\boldsymbol{x} \geq (\boldsymbol{\alpha}^k)^\top V_1^*(\underline{s}),
$$

we have for (25) that

$$
K\Big[g\big(V_1^{\widehat{\pi}}(\underline{s})\big) - g\big(V_1^*(\underline{s})\big) + \rho\,\mathrm{dist}\big(V_1^{\widehat{\pi}}(\underline{s}), \mathcal{W}^*\big)\Big]
$$
$$
\leq \sum_{k=1}^{K}\Big\{(\boldsymbol{\beta}^k)^\top V_1^k(\underline{s}) - (\boldsymbol{\beta}^k)^\top V_1^*(\underline{s}) + \rho\big((\boldsymbol{\alpha}^k)^\top V_1^k(\underline{s}) - (\boldsymbol{\alpha}^k)^\top V_1^*(\underline{s})\big)\Big\} + C(1+\rho)\sqrt{DH^2K}
$$
$$
= \sum_{k=1}^{K}\Big[(\boldsymbol{\theta}^k)^\top\big(V_1^{\,k}(\underline{s}) - V_1^*(\underline{s})\big)\Big] + C(1+\rho)\sqrt{DH^2K}
$$

$$(26)$$

Applying Lemma 3, we have

$$
K\Big[g\big(V_1^{\widehat{\pi}}(\underline{s})\big) - g\big(V_1^*(\underline{s})\big) + \rho\,\mathrm{dist}\big(V_1^{\widehat{\pi}}(\underline{s}), \mathcal{W}^*\big)\Big]
$$
$$
\leq 2K(1+\rho)\sqrt{D}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\big[\Gamma_h(s_h, a_h) + \|\Gamma_h^c(s_h, a_h)\|_\infty \mid s_1 = \underline{s}\big] + C(1+\rho)\sqrt{DH^2K}
$$
$$
= K(\epsilon_K + \mathrm{IntUncert}_{\mathcal{D}}^{\pi^*}),
$$

where we define

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

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

$$
g\big(V_1^{\widehat{\pi}}(\underline{s})\big) - g\big(V_1^*(\underline{s})\big) \leq \epsilon_K + \mathrm{IntUncert}_{\mathcal{D}}^{\pi^*}.
$$

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. □

21

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

$$\text{dist}\left(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s}), \mathcal{W}^*\right)$$

$$\leq \text{dist}\left(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s}), \mathcal{W}^*\right) + \sin(\gamma_{\max})\left[g\left(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s})\right) - g\left(\boldsymbol{V}_1^*(\underline{s})\right) + \frac{\text{dist}\left(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s}), \mathcal{W}^*\right)}{\sin(\gamma_{\max})}\right]$$

$$= \sin(\gamma_{\max})\left[g\left(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s})\right) - g\left(\boldsymbol{V}_1^*(\underline{s})\right) + \rho\,\text{dist}\left(\boldsymbol{V}_1^{\widehat{\pi}}(\underline{s}), \mathcal{W}^*\right)\right]$$

$$\leq \frac{2}{\rho}(\epsilon_K + \text{IntUncert}_{\mathcal{D}}^{\pi^*}).$$

where the last inequality follows from (26). Thus, we complete the proof of Theorem 1. $\qquad\square$

H.3 Proof of Theorem 2

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

Proof of Lemma 5. See Appendix H.6.5 for a detailed proof. $\qquad\square$

By Theorem 1, we have that

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

where we define

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

By Lemma 5 and the ξ-uncertainty quantifier defined in (24), we finish the proof. $\qquad\square$

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*

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

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. $\qquad\square$

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 $\psi(s, a, s') = \boldsymbol{e}_{(s,a,s')}$, $(\theta_h)_{(s,a,s')} =$

22

$\mathcal{P}(s' \mid s, a)$, $\varphi(s, a) = e_{(s,a)}$, and $(\theta_h^{c_i})_{(s,a)} = c_h^i(s, a)$. Here 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

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

Following the same derivation we get

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

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

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

$$\geq \mathbb{E}_{\mathcal{D}} \left[\frac{\mathrm{SubOpt}(\widehat{\pi})}{2 \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[1/\sqrt{1 + n_h(s_h, a_h)} \mid s_1 = \underline{s} \right]} \right]$$

$$\geq c/2.$$

which completes the proof of Theorem 3. $\qquad \square$

H.5 Proof of Corollary 1

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

$$\mathbb{E}_{\pi^*} \left[\Gamma_h(s_h, a_h) + \Gamma_h^{c_i}(s_h, a_h) \mid s_1 = \underline{s} \right]$$

$$= \mathbb{E}_{\mu_h^*} \left[\Gamma_h(s, a) \right] + \mathbb{E}_{\mu_h^*} \left[\Gamma_h^{c_i}(s, a) \right]$$

$$\leq \left(\sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{\mu_h^*(s, a)}{\mu_h^{b,\tau}(s, a)} \right) \cdot \left(\mathbb{E}_{\mu_h^{b,\tau}} \left[\Gamma_h(s, a) \right] + \mathbb{E}_{\mu_h^{b,\tau}} \left[\Gamma_h^{c_i}(s, a) \right] \right) \tag{29}$$

$$\leq \varsigma \cdot \left(\mathbb{E}_{\mu_h^{b,\tau}} \left[\Gamma_h(s, a) \right] + \mathbb{E}_{\mu_h^{b,\tau}} \left[\Gamma_h^{c_i}(s, a) \right] \right),$$

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

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

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

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

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

$$\tag{31}$$

23

and by the property of trace, we have

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

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

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

Combining (30) and (33), we have

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

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

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

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

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

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

and note that

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

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

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

24

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

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

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

$$
\leq 2\varsigma(1+\rho)\sum_{h=1}^{H} N^{-1}\sum_{\tau=1}^{N}\Big(\mathbb{E}_{\mu_h^{\mathrm{b},\tau}}\big[\Gamma_h(s,a)\big] + \mathbb{E}_{\mu_h^{\mathrm{b},\tau}}\big[\|\Gamma_h^c(s,a)\|_\infty\big]\Big)
$$

$$
\leq C\varsigma\kappa(1+\rho)H^2\sqrt{d|\mathcal{S}|/N\log(DH/\xi)}
$$

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

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*

$$
V_1(\underline{s}) - V_1^{\pi'}(\underline{s}) = \sum_{h=1}^{H}\mathbb{E}_{\pi'}\Big[\big\langle Q_h(s_h, \cdot), \pi_h(\cdot\,|\,s_h) - \pi'_h(\cdot\,|\,s_h)\big\rangle_{\mathcal{A}} \,\Big|\, s_1 = \underline{s}\Big]
$$

$$
+ \sum_{h=1}^{H}\mathbb{E}_{\pi'}\big[Q_h(s_h, a_h) - (\mathbb{B}_h V_{h+1})(s_h, a_h) \,\big|\, s_1 = \underline{s}\big].
$$

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

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

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

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

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

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

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).

$\qquad\square$

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.

25

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

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

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. $\square$

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 $h \in [H]$,*

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

Proof of Lemma 7. See Appendix H.6.4 for a detailed proof. $\square$

26

By Lemma 2, it holds that

$$(\boldsymbol{\theta}^k)^\top \big(\boldsymbol{V}_1^k(\underline{s}) - \boldsymbol{V}_1^*(\underline{s})\big) = \underbrace{\sum_{h=1}^H \mathbb{E}_{\pi^*}\Big[(\boldsymbol{\theta}^k)^\top \boldsymbol{\iota}_h^k \,\Big|\, s_1 = \underline{s}\Big]}_{(\mathrm{i})}$$

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

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

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

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

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

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

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

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

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

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

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

27

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

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

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

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

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

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

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 $\eta_t \mid \mathcal{G}_{t-1}$ is zero-mean and $\sigma-$sub-Gaussian, i.e.,*

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

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

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 $t \ge 0$,

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

28

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 $\langle f, \mathcal{P}_h(\cdot \,|\, s_h^\tau, a_h^\tau) - \delta_{s_{h+1}^\tau}\rangle_\mathcal{S} \leq 2\|f\|_\infty$, which implies

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

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

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

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

By Definition 2, we have

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

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

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

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

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

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

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

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

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

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

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

29

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

$$
\left| c_h^i(s,a) - \widehat{c}_h^i(s,a) \right|
$$

$$
= |\varphi(s,a)^\top (\theta_h^{c^i} - \widehat{\theta}_h^{c^i})|
$$

$$
= |\varphi(s,a)^\top \Lambda_{\varphi,h}^{-1} \big(\sum_{\tau=1}^N c_h^{i,\tau} \cdot \varphi(s_h^\tau, a_h^\tau) - \Lambda_{\varphi,h} \theta_h^{c^i} \big)|
$$

$$
= |\varphi(s,a)^\top \Lambda_{\varphi,h}^{-1} \Big(\sum_{\tau=1}^N \varphi(s_h^\tau, a_h^\tau) \big(c_h^{i,\tau} - \varphi(s_h^\tau, a_h^\tau)^\top \theta_h^{c^i} \big) - \lambda \theta_h^{c^i} \Big)|
$$

$$
\leq \|\varphi(s,a)\|_{\Lambda_{\varphi,h}^{-1}} \cdot \| \sum_{\tau=1}^N \varphi(s_h^\tau, a_h^\tau) \big(c_h^{i,\tau} - \varphi(s_h^\tau, a_h^\tau)^\top \theta_h^{c^i} \big) \|_{\Lambda_{\varphi,h}^{-1}} + \lambda \|\theta_h^{c^i}\|_{\Lambda_{\varphi,h}^{-1}}
$$

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

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

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. $\qquad\square$

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*

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

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 $\mathbb{E}[M_T^\lambda] \leq 1$.

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

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

$$
\leq M_{t-1}^\lambda \cdot \mathbb{E}\left[\exp(\frac{\|\lambda^\top X_t\|_\infty^2}{2R^2} - \frac{\|\lambda^\top X_t\|_\infty^2}{2R^2}) \,\middle|\, \mathcal{G}_{t-1} \right]
$$

$$
= M_{t-1}^\lambda.
$$

where the inequality is due to the conditional σ-sub-Gaussianity of η_t and the condition in (44). It also implies $\mathbb{E}[M_t^\lambda] \leq 1$ for any $t \geq 0$. 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

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

$\qquad\square$

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 $1 - \delta$*

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

30

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

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

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

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

Hence, we have

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

which completes the proof of Lemma 10. $\qquad\square$

We construct a stopping time as below.

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

Then, we have

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

which completes the proof of Lemma 8. $\qquad\square$

31

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
$|\mathcal{S}|$: cardinality of state space	5
$|\mathcal{A}|$: cardinality of action space	5
$|\mathcal{D}|$: 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.

32

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

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

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.*

33