Revisiting Subgradient Dominance in Robust MDPs: Counterexamples, Hardness, and Sufficient Conditions

Toshinori Kitamura
University of Alberta
toshinor@ualberta.ca

Arnob Ghosh
New Jersey Institute of Technology
arnob.ghosh@njit.edu

Alex Ayoub
University of Alberta
aayoub@ualberta.ca

Thang D. Chu
University of Alberta
thang@ualberta.ca

Csaba Szepesvári
University of Alberta
szepesva@ualberta.ca

Abstract

Projected subgradient descent (PSD) has gained popularity for solving robust Markov decision processes (RMDPs) because it applies to a broader class of uncertainty sets than traditional dynamic programming. Existing work claims that RMDPs with a general compact uncertainty set satisfy the *subgradient dominance property*, under which exact PSD finds an ε-optimal policy in a polynomial number of updates (e.g., Wang et al. [40]). We show that these claims are incorrect. Even when the uncertainty set has cardinality two, the RMDP objective **need not** be subgradient-dominant and can admit suboptimal strict local minima. Moreover, we prove that finding an ε-optimal policy can be NP-hard even in settings where subgradients are efficiently computable: (i) finite transition uncertainty sets and (ii) sa-rectangular finite transition uncertainty sets with finite non-rectangular cost uncertainty sets. Finally, we identify two conditions under which RMDPs do satisfy subgradient dominance: when, for each policy, either the worst-case transition kernel is unique or a robust action-value function exists.

1 Introduction

Dealing with environmental model uncertainty is crucial for practical decision-making problems [39]. Robust Markov decision processes (RMDPs) provide a framework for designing policies that are robust to such uncertainty, where transition kernels and cost functions are chosen adversarially from an uncertainty set. Recently, projected subgradient descent (PSD) has emerged as a popular approach for solving RMDPs [40, 42, 26, 25]. PSD methods are particularly appealing for their broad applicability: they accommodate general uncertainty sets for which subgradients can be efficiently computed (e.g., finite sets), whereas traditional dynamic programming (DP) methods typically require stronger structural assumptions on the uncertainty set [22, 43].

Existing RMDP literature guarantees the performance of PSD by establishing the *subgradient dominance property*, under which every stationary point is globally optimal (e.g., Wang et al. [40], Kitamura et al. [25]). These results appear mathematically sound, as they sidestep classical NP-hardness results for RMDPs, such as those of Wiesemann et al. [43], by assuming access to exact subgradients. Unfortunately, we show that this line of analysis is flawed and formally refute the claimed PSD guarantees.[1] In particular, we prove that **the RMDP objective is not subgradient-dominant in general**. Even a very simple RMDP with a finite uncertainty set of cardinality two admits a suboptimal strict local minimum, and PSD can get stuck in its neighborhood (Example 3.4).

[1] An erratum in Kitamura et al. [25] acknowledged errors in the proofs of both Kitamura et al. [25] and Wang et al. [40], but did not provide a counterexample to subgradient dominance in general RMDPs.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Setting	Subgrad. dom.?	Example
Finite $\mathcal{P}$ × Singleton $\mathcal{C}$	No, NP-hard (Proposition 3.6)	Example 3.4
sa-rectangular finite $\mathcal{P}$ × Finite non-rectangular $\mathcal{C}$	No†, NP-hard (Proposition 3.7)	Robust constrained MDPs with rectangular $\mathcal{P}$ [25, 29, 9]
Worst-case transition kernel is unique for each policy (Assumption 4.1)	Yes (Theorem 4.3)	MDPs [1], Regularized RMDPs (Section 4.1), Cost-robust MDPs [21], Convex MDPs [46]
Robust Q function exists for each policy (Assumption 4.2)	Yes (Theorem 4.3)	MDPs [1], Regularized RMDPs [45, 48], sa- and r-rectangular RMDPs [22, 17]
s-rectangular RMDPs	Open problem	

Table 1: Summary of the subgradient dominance property across different RMDP settings. Here, $\mathcal{C}$ and $\mathcal{P}$ denote the cost and transition uncertainty sets, respectively. Under Assumption 3.2, PSD is guaranteed to find ε-optimal policies in settings marked "Yes", while those marked "No" are NP-hard even to approximate (Propositions 3.6 and 3.7). †Unless P = NP, subgradient dominance cannot hold with a polynomially bounded constant (Section 3.2). Highlighted settings were previously believed to be subgradient-dominant [40, 27, 41, 25].

Moreover, we establish a stronger negative result: finding an ε-optimal policy can be NP-hard for (i) finite transition uncertainty sets (Proposition 3.6) and (ii) sa-rectangular finite transition uncertainty sets with finite non-rectangular cost uncertainty sets (Proposition 3.7). Unlike prior NP-hardness results on objective evaluation [43] or exact optimization [33, 4], our result shows that approximate policy optimization can remain intractable even when subgradients are efficiently computable, unless P = NP.

Despite these negative results, PSD can succeed under additional structure. We identify two sufficient conditions under which PSD finds ε-optimal policies (Theorem 4.3 and Corollary 4.4): when, for each policy, either the worst-case transition kernel is unique or a robust action-value function exists. These conditions cover several previously studied settings, including regularized RMDPs [45, 48, 11], sa- and r-rectangular uncertainty sets [22, 32, 17], cost-robust MDPs [8, 16], and convex MDPs [47, 46]. A summary of our results is provided in Table 1. Overall, our findings clarify existing misunderstandings about PSD in RMDPs and provide precise conditions under which its guarantees can be restored.

The remainder of the paper is organized as follows. Section 2 introduces the RMDP formulation and the PSD algorithm. Section 3 presents our negative results for finite uncertainty sets. Section 4 establishes sufficient conditions for the success of PSD. Finally, Section 5 situates our contributions within the related literature and discusses the limitations of our work.

2 Preliminaries

Basic notation. The probability simplex over a finite set $\mathcal{S}$ is denoted by $\Delta(\mathcal{S})$. For an integer n, let $[n] \coloneqq \{1, \dots, n\}$. We define $\mathbf{0} \coloneqq (0, \dots, 0)^\top$ and $\mathbf{1} \coloneqq (1, \dots, 1)^\top$. For a set $\mathcal{X} \subset \mathbb{R}^d$, we define $\operatorname{dist}(\mathbf{0}; \mathcal{X}) \coloneqq \inf_{x \in \mathcal{X}} \|x\|_2$. For $\mathcal{X} \subset \mathbb{R}^d$, $\operatorname{conv} \mathcal{X}$ denotes its convex hull. The indicator function $\delta_{\mathcal{X}}$ of a set $\mathcal{X}$ takes value 0 at $x \in \mathcal{X}$ and $+\infty$ otherwise. All scalar operations and inequalities applied to vectors or functions are understood elementwise.

For a function $f : \mathbb{R}^d \to \mathbb{R}$, we let $\partial f(x)$ denote the set of Fréchet subgradients of f at $x \in \mathbb{R}^d$ [36, Definition 8.3]. If $\partial f(x)$ is a singleton, we denote its element by $\nabla f(x)$ and refer to it as the gradient of f at x. A point x is said to be *stationary* for f if $\mathbf{0} \in \partial f(x)$.

2.1 Robust MDPs

An infinite-horizon tabular RMDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, \mu, \gamma, \mathcal{U})$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces, respectively, $\mu \in \Delta(\mathcal{S})$ is the initial state distribution, and $\gamma \in [0, 1)$ is the discount factor. The set $\mathcal{U}$ is a compact uncertainty set of cost functions and transition kernels. We refer to a pair $(c, P) \in \mathcal{U}$ as a *model*. For each $(c, P) \in \mathcal{U}$, the cost function $c : \mathcal{S} \times \mathcal{A} \to [0, 1]$ specifies the cost of taking action a in state s, and the transition kernel P satisfies $P(\cdot \mid s, a) \in \Delta(\mathcal{S})$.

When the uncertainty set decomposes as $\mathcal{U} = \mathcal{C} \times \mathcal{P}$, we use $\mathcal{C}$ and $\mathcal{P}$ to denote the cost and transition sets, respectively. If both $\mathcal{C}$ and $\mathcal{P}$ are singletons, the RMDP reduces to a standard MDP. If $\mathcal{P}$ is

a singleton, the problem is known as a *cost-robust MDP* [21, 16], which encompasses a variety of decision-making problems (e.g., convex MDPs in Section 4.2).

A (stationary Markov) policy π is a probability kernel such that $\pi(\cdot \mid s) \in \Delta(\mathcal{A})$ specifies the action distribution at state $s \in \mathcal{S}$. The set of all policies is denoted by $\Pi \subset \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$, which corresponds to the *direct parameterization* policy class [1].

Given a transition kernel P, the occupancy measure $d_P^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ represents the γ-discounted visitation frequency of each state–action pair under policy π:

$$d_P^\pi(s,a) = (1-\gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_P^\pi(s_h = s, a_h = a) \ ,$$

where the probability is taken over trajectories generated by $s_0 \sim \mu$, $a_h \sim \pi(\cdot \mid s_h)$, and $s_{h+1} \sim P(\cdot \mid s_h, a_h)$. The induced state occupancy measure is defined as $\overline{d}_P^\pi(s) \coloneqq \sum_{a \in \mathcal{A}} d_P^\pi(s,a)$.

The value function $V_{c,P}^\pi : \mathcal{S} \to \mathbb{R}$ (and the action-value function $Q_{c,P}^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$) represents the expected total cost under policy π starting from a state s (and from a state–action pair (s,a), respectively). They are the unique solution to the following *Bellman equations* [34]:

$$V_{c,P}^\pi(s) = c^\pi(s) + \gamma \sum_{s' \in \mathcal{S}} P^\pi(s' \mid s) V_{c,P}^\pi(s') \quad \forall s \in \mathcal{S}$$
$$\text{and} \quad Q_{c,P}^\pi(s,a) = c(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s,a) V_{c,P}^\pi(s') \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A} \ ,$$

where we use shorthands $c^\pi(s) \coloneqq \sum_a \pi(a \mid s) c(s,a)$ and $P^\pi(s' \mid s) \coloneqq \sum_a P(s' \mid s,a) \pi(a \mid s)$. We define the advantage function as $A_{c,P}^\pi(s,a) \coloneqq V_{c,P}^\pi(s) - Q_{c,P}^\pi(s,a)$.

The robust total cost of a policy π is defined as

$$J_{\mathcal{U}}(\pi) = \max_{(c,P) \in \mathcal{U}} \mathbb{E}_P^\pi \left[\sum_{h=0}^{\infty} \gamma^h c(s_h, a_h) \right] = \max_{(c,P) \in \mathcal{U}} \sum_{s \in \mathcal{S}} \mu(s) V_{c,P}^\pi(s) \ . \tag{1}$$

If $\mathcal{U} = \mathcal{C} \times \mathcal{P}$ and either $\mathcal{C}$ or $\mathcal{P}$ is a singleton, we write $\mathcal{U} = \mathcal{C}$ or $\mathcal{U} = \mathcal{P}$ for simplicity. We denote $J_{c,P} \coloneqq J_{\{(c,P)\}}$ as the total cost under a fixed model and refer to $J_{\mathcal{U}}$ as the robust total cost.

The goal of an RMDP is to identify an optimal policy $\pi^\star$ that minimizes the robust total cost:

$$\text{(Robust MDP)} \qquad \pi^\star \in \arg\min_{\pi \in \Pi} J_{\mathcal{U}}(\pi) \ .$$

We call a policy π ε-*optimal* if it satisfies $J_{\mathcal{U}}(\pi) - J_{\mathcal{U}}(\pi^\star) \le \varepsilon$.

2.2 Rectangularity for Transition Uncertainty Sets

Without any assumptions on the transition set $\mathcal{P}$, computing the robust total cost $J_{\mathcal{P}}(\pi)$ is NP-hard, even to approximate [43]. A widely adopted regularity condition is sa-rectangularity [22, 32]:

$$sa\text{-rectangularity} \quad \mathcal{P} = \times_{s,a} \mathcal{P}_{s,a} \quad \text{where} \quad \mathcal{P}_{s,a} \subseteq \Delta(\mathcal{S}) \ .$$

Here, $\times_{s,a}$ denotes the Cartesian product over all state–action pairs. Intuitively, sa-rectangularity allows the adversary to choose worst-case transitions independently for each state–action pair.

Because this assumption can yield overly conservative policies, several works have proposed weaker notions of rectangularity, most notably s-*rectangularity* [43] and r-*rectangularity* [17]:

$$s\text{-rectangular} \quad \mathcal{P} = \times_s \mathcal{P}_s \quad \text{where} \quad \mathcal{P}_s \subseteq \Delta(\mathcal{S})^{|\mathcal{A}|} \ ,$$
$$r\text{-rectangular} \quad \mathcal{P} = \left\{ P \;\middle|\; P(s' \mid s,a) = \sum_{i=1}^{r} \phi_i(s,a) w_i(s') \right\} \ ,$$
$$\text{where } (w_1, \dots, w_r) \in \mathcal{W}_1 \times \cdots \times \mathcal{W}_r \ .$$

Here, $\phi(s,a) \in \Delta([r])$ and $\mathcal{W}_i \subseteq \Delta(\mathcal{S})$ for each $i \in [r]$. Unlike sa-rectangularity, s-rectangular uncertainty sets allow the adversary to select worst-case transitions independently across states, but

not across actions. The r-rectangular structure resembles low-rank MDP models [23], in which the transition kernel admits a linear decomposition via a feature map ϕ of dimension r. Although both notions relax sa-rectangularity, they are not nested: neither s-rectangularity nor r-rectangularity contains the other [17, Proposition 2.3].

These forms of rectangularity enable efficient computation of the robust total cost $J_{\mathcal{P}}(\pi)$ using dynamic programming (DP) methods. Moreover, an ε-optimal policy can be computed in time polynomial in $|\mathcal{S}|$, $|\mathcal{A}|$, $(1-\gamma)^{-1}$, $\log \varepsilon^{-1}$, and the description length of $\mathcal{P}$ [43, 17].

2.3 Projected Subgradient Descent for Robust MDPs

Beyond DP methods for rectangular uncertainty sets, an alternative line of work has explored *projected subgradient descent* (PSD) for solving RMDPs [40, 25, 27, 41]. Unlike DP-based approaches, PSD does not rely on strong structural assumptions on the uncertainty sets $\mathcal{C}$ or $\mathcal{P}$. Instead, it requires access only to a worst-case policy subgradient. This flexibility makes PSD applicable to classes of RMDPs that are not amenable to DP methods. For instance, when the transition set is finite (e.g., $\mathcal{P} = \{P_1, P_2\}$), PSD is applicable whereas DP methods generally are not.

At iteration $t \in \mathbb{N}$, PSD updates π_t to a new policy π_{t+1} by:

$$\pi_{t+1} \coloneqq \mathrm{Proj}_{\Pi}(\pi_t - \eta g_t) \quad \text{where } \; g_t \in \partial J_{\mathcal{U}}(\pi_t) \quad \forall t \in [T] \; , \tag{2}$$

where $\eta > 0$ is the learning rate and Proj_{Π} denotes the Euclidean projection onto Π, which can be implemented efficiently [13]. The subgradient g_t can be efficiently computed when a worst-case model in Eq. (1) is available:

Lemma 2.1. *Let $\mathcal{U}^\pi \coloneqq \arg\max_{(c,P)\in\mathcal{U}} J_{c,P}(\pi)$ denote the set of worst-case models under policy π. For any $\pi \in \Pi$, it holds that*

$$\partial J_{\mathcal{U}}(\pi) = \mathrm{conv}\{\nabla J_{c,P}(\pi) \mid (c,P) \in \mathcal{U}^\pi\} \tag{3}$$

$$\textit{where} \quad (\nabla J_{c,P}(\pi))(s,a) = \frac{1}{1-\gamma} \overline{d}_P^\pi(s) Q_{c,P}^\pi(s,a) \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A} \; . \tag{4}$$

Eq. (3) follows from a version of Danskin's theorem (Lemma A.4 in Appendix A) and Eq. (4) is known as the *policy gradient theorem* [44].

The convergence behavior of the PSD update in Eq. (2) can be analyzed using the *Moreau envelope* of *weakly convex* functions.

Definition 2.2 (Moreau envelope)**.** For a function $f : \mathbb{R}^d \to \mathbb{R}$ and a parameter $\nu > 0$, the Moreau envelope is defined by $\mathrm{M}_\nu \circ f : \mathbb{R}^d \to \mathbb{R}$ such that

$$\mathrm{M}_\nu \circ f(x) = \min_{y\in\mathbb{R}^d} \left\{ f(y) + \frac{1}{2\nu}\|x-y\|_2^2 \right\} .$$

The minimal point is called the proximal point: $\mathrm{Prox}_{\nu f}(x) = \arg\min_{y\in\mathbb{R}^d} \left\{ f(y) + \frac{1}{2\nu}\|x-y\|_2^2 \right\}$.

Definition 2.3 (Weak convexity [3, Definition 2.1])**.** For $\omega \geq 0$, a function $f : \mathbb{R}^d \to \mathbb{R}$ is called ω-weakly convex if $f(\cdot) + \frac{\omega}{2}\|\cdot\|_2^2$ is convex.

Weak convexity is a mild regularity condition satisfied by many functions in optimization. In particular, the pointwise maximum of smooth functions is weakly convex [3, Proposition 2.4], and hence the robust total cost $J_{\mathcal{U}}$ is weakly convex. The Moreau envelope provides a smooth approximation of a weakly convex function. Specifically, if f is proper and ω-weakly convex, then for any $\nu \in (0, 1/\omega)$, the Moreau envelope $\mathrm{M}_\nu \circ f$ is differentiable, with gradient given by [12, Lemma 4.3]:

$$\nabla \mathrm{M}_\nu \circ f(x) = \frac{1}{\nu}(x - \mathrm{Prox}_{\nu f}(x)) \quad \forall x \in \mathbb{R}^d \; . \tag{5}$$

Combining Eq. (5) with the definition of the Moreau envelope yields

$$\mathrm{dist}(\mathbf{0}; \partial f(\mathrm{Prox}_{\nu f}(x))) \leq \|\nabla \mathrm{M}_\nu \circ f(x)\|_2 = \frac{1}{\nu}\|x - \mathrm{Prox}_{\nu f}(x)\|_2 \; . \tag{6}$$

Thus, a small Moreau-envelope gradient implies that x is close to a point $\mathrm{Prox}_{\nu f}(x)$ that is nearly stationary for f. For further discussion of weak convexity and the Moreau envelope, see [10, 35, 36].

Applying these tools to RMDPs yields convergence guarantees for PSD. Define the extended objective $\bar{J}_{\mathcal{U}} : \pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \mapsto J_{\mathcal{U}}(\pi) + \delta_{\Pi}(\pi)$ which extends the domain of the robust total cost to $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. Let $\rho \coloneqq \frac{2|\mathcal{A}|}{(1-\gamma)^3}$. Since each total cost $J_{c,P}$ is $\frac{2\gamma|\mathcal{A}|}{(1-\gamma)^3}$-smooth (Lemma A.2) and $\gamma \leq 1$, the extended objective $\bar{J}_{\mathcal{U}}$ is ρ-weakly convex. Standard analyses of PSD for weakly convex functions imply that the iterates produced by Eq. (2) converge to near-stationary points of the Moreau envelope (see Appendix C).

Lemma 2.4. *When $\eta = 1/\sqrt{T}$, the PSD iterates in Eq.* (2) *satisfy*

$$\min_{t\in[T]} \left\| \nabla(\mathrm{M}_{\frac{1}{2\rho}} \circ \bar{J}_{\mathcal{U}})(\pi_t) \right\|_2 \leq T^{-1/4} \sqrt{\frac{4}{1-\gamma} + \frac{4|\mathcal{A}|^2}{(1-\gamma)^7}} \,.$$

Together with Eq. (6), Lemma 2.4 implies that there exists an iterate π_t that lies close to a proximal point which is nearly stationary for $\bar{J}_{\mathcal{U}}$.

3 Failure of Projected Subgradient Descent in Robust MDPs

Because the total cost function is nonconvex in the policy parameterization [1], the stationarity guarantee in Lemma 2.4 alone does not imply near-optimality. To bridge this gap, prior works have attempted to establish the *subgradient dominance* property for the robust total cost $J_{\mathcal{U}}$ [40, 25]:[2]

Definition 3.1 (Subgradient dominance)**.** The robust total cost $J_{\mathcal{U}}$ is said to be *subgradient-dominant* with constant $D > 0$ if, for every $\pi \in \Pi$,

$$J_{\mathcal{U}}(\pi) - J_{\mathcal{U}}(\pi^\star) \leq DG(\pi) \;\text{ where }\; G(\pi) \coloneqq \min_{g\in\partial J_{\mathcal{U}}(\pi)} \max_{\pi'\in\Pi} \sum_{s,a} (\pi(a|s) - \pi'(a|s)) g(s,a) \,. \quad (7)$$

In the standard MDP setting where $\mathcal{U}$ is a singleton, this condition reduces to the classical *gradient dominance* property, which is known to hold.

Assumption 3.2. The initial distribution μ has full support, i.e., $\mu(s) > 0$ for all $s \in \mathcal{S}$.

Lemma 3.3 (gradient dominance [1])**.** *When $\mathcal{U}$ is a singleton and Assumption 3.2 holds*[3]*, the total cost function $J_{\mathcal{U}}$ satisfies Eq.* (7) *with constant $D = ((1-\gamma)\min_s \mu(s))^{-1}$.*

Intuitively, the quantity $G(\pi)$, which is the smallest Frank–Wolfe gap among the subgradients at π, measures the first-order stationarity of π with respect to the robust objective. If $J_{\mathcal{U}}$ is subgradient-dominant, the Lipschitz continuity of $J_{\mathcal{U}}$ relates $G(\pi)$ to the gradient norm of the Moreau envelope and yields the following inequality (see Appendix C for details):

$$J_{\mathcal{U}}(\pi) - \min_{\pi'\in\Pi} J_{\mathcal{U}}(\pi') \leq D' \left\| \nabla(\mathrm{M}_{\frac{1}{2\rho}} \circ \bar{J}_{\mathcal{U}})(\pi) \right\|_2 \;\text{ where }\; D' \coloneqq 2D\sqrt{|\mathcal{S}|} + \frac{1-\gamma}{4\sqrt{|\mathcal{A}|}} \,. \quad (8)$$

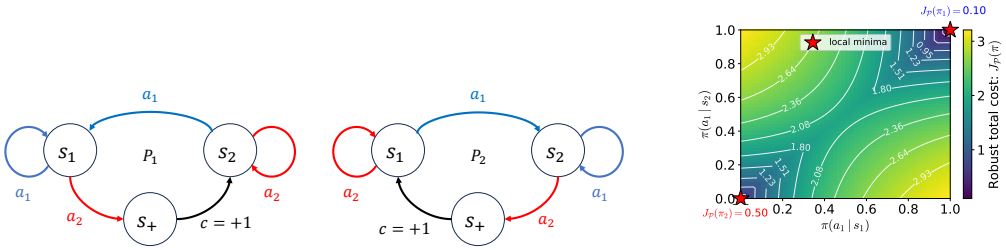
Consequently, under subgradient dominance, combining Lemma 2.4 with Eq. (8) guarantees that PSD finds an ε-optimal policy within $O(\varepsilon^{-4})$ policy updates.

3.1 A Counterexample Where PSD Fails

Consider RMDPs with uncertainty only in the transition kernel, i.e., $\mathcal{U} = \mathcal{P}$. Several prior works claim that, under the full-support assumption on the initial distribution (Assumption 3.2), the robust total cost $J_{\mathcal{P}}$ is subgradient-dominant for any compact uncertainty set $\mathcal{P}$ (e.g., Theorem 3.2 in Wang et al. [40] and Theorem 4 in Kitamura et al. [25]). We show that these claims are incorrect. **In fact, the robust total cost need not be subgradient-dominant even for a simple transition set.** Specifically, when $|\mathcal{P}| = 2$, the function $J_{\mathcal{P}}$ can already admit a suboptimal strict local minimum. The technical errors in the prior analyses are explained in Appendix B.

[2] Wang et al. [40] adopt the formulation in Eq. (8), while Kitamura et al. [25] use the stronger form in Eq. (7). Both notions are sufficient to guarantee that PSD efficiently finds an ε-optimal policy.

[3] Such a coverage condition is necessary to ensure the global convergence of policy gradient methods [31].



(a) A deterministic RMDP with $\mathcal{P} = \{P_1, P_2\}$. Transitioning from state s_+ incurs a cost of $+1$.

(b) Robust total cost landscape with $\gamma = 0.9$.

Figure 1: An RMDP where PSD fails (**left**) and its robust total cost landscape (**right**). The policy π_2, which always selects action a_2, is a strictly suboptimal local minimum. Details are provided in Example 3.4.

Example 3.4 (An RMDP where PSD fails)**.** Consider a deterministic RMDP with three states (s_1, s_2, s_+), two actions (a_1, a_2), and two transition kernels $\mathcal{P} = \{P_1, P_2\}$, as illustrated in Figure 1 (**left**). We set the discount factor to $\gamma = 0.9$ and the initial distribution to $\mu(s_1) = \mu(s_2) = 0.45$, which satisfies Assumption 3.2. Let $\tilde{\pi}_1$ and $\tilde{\pi}_2$ denote the policies that always choose a_1 and a_2, respectively. The policy $\tilde{\pi}_2$ visits the costly state s_+ and therefore incurs a larger robust total cost than $\tilde{\pi}_1$. For this instance, the following proposition holds. The proof is given in Appendix D.

Proposition 3.5. *The following two claims hold in the example instance:*

- *$\tilde{\pi}_2$ is a suboptimal strict local minimum of $J_{\mathcal{P}}$ satisfying $0.505 = J_{\mathcal{P}}(\tilde{\pi}_2) > J_{\mathcal{P}}(\tilde{\pi}_1) = 0.1$.*
- *For all sufficiently small $\eta > 0$, when initialized at $\pi_1 = \tilde{\pi}_2$, the iterates of the PSD update $\pi_{t+1} = \mathrm{Proj}_{\Pi}(\pi_t - \eta g_t)$ with any $g_t \in \partial J_{\mathcal{P}}(\pi_t)$ satisfy $\|\pi_t - \tilde{\pi}_2\|_\infty \leq 3\eta$ for all $t \geq 1$.*

This result formally shows that $J_{\mathcal{P}}$ is not subgradient-dominant: since $\tilde{\pi}_2$ is a local minimum, the first-order optimality condition gives $G(\tilde{\pi}_2) = 0$, whereas $J_{\mathcal{P}}(\tilde{\pi}_2) - J_{\mathcal{P}}(\pi^\star) \geq 0.405$, violating Eq. (7) for every $D > 0$. Moreover, PSD initialized at $\tilde{\pi}_2$ remains in an $O(\eta)$-neighborhood of this suboptimal policy. These facts are also illustrated empirically in Figure 1 (**right**), which shows the landscape of the robust total cost function.

3.2 NP-hardness of ε-Optimal Policy Identification

Since PSD fails to find an ε-optimal policy under general uncertainty sets, a natural question is whether any algorithm can efficiently identify an ε-optimal policy in such settings. Unfortunately, the answer is negative. Even when the transition set is finite, identifying an ε-optimal policy is NP-hard.

Proposition 3.6 (Hardness for finite $\mathcal{P}$ and singleton $\mathcal{C}$)**.** *For any fixed $\gamma \in (0, 1)$ and any $\varepsilon < 0.5\gamma^3(1-\gamma)^{-1}$, finding an ε-optimal policy in RMDPs with a finite transition set and a singleton cost set is NP-hard.*

The proof is based on a reduction from the 3-SAT problem (see Appendix E). Unlike prior NP-hardness results on objective evaluation [43] or exact optimization [33, 4], this result shows that approximate policy optimization can remain NP-hard even when policies can be evaluated efficiently. Indeed, in this setting, since the uncertainty set is finite, both the robust total cost and its subgradients can be approximated to δ accuracy in time polynomial in $|\mathcal{S}|$, $|\mathcal{A}|$, $(1-\gamma)^{-1}$, $\log \delta^{-1}$, and $|\mathcal{U}|$ [26]. A more detailed comparison with prior NP-hardness results is provided in Section 5.

As a byproduct, we obtain an additional hardness result for RMDPs with a rectangular transition set and a finite cost set. This setting is especially relevant for *robust constrained MDPs* (RCMDPs), in which a policy must satisfy multiple robust constraints [25, 37]:

$$\text{(RCMDP)} \quad \min_{\pi \in \Pi} J_{c_0, \mathcal{P}}(\pi) \quad \text{such that} \quad \max_{n \in [N]} J_{c_n, \mathcal{P}}(\pi) \leq b \,.$$

Here, $c_0, \ldots, c_N$ are the objective and constraint cost functions, and $b > 0$ is a constraint threshold. When $\mathcal{P}$ is rectangular, the constraint term $\max_{n \in [N]} J_{c_n, \mathcal{P}}(\pi)$ corresponds to an RMDP with a rectangular transition set and a non-rectangular (finite) cost set $\mathcal{C} = \{c_1, \ldots, c_N\}$.

Recent work by Kitamura et al. [25] proposes a PSD-based algorithm for RMDPs with a general transition set and a finite cost set, suggesting an efficient approach for solving RCMDPs. However, the following Proposition 3.7 shows that finding an ε-optimal policy in an RCMDP is NP-hard, even when the transition set is sa-rectangular and finite.

Proposition 3.7 (Hardness for sa-rectangular $\mathcal{P}$ and finite non-rectangular $\mathcal{C}$)**.** *For any fixed $\gamma \in (0,1)$ and any $\varepsilon < \frac{\gamma^2}{2(2-\gamma)}$, finding an ε-optimal policy in RMDPs with an sa-rectangular finite transition set and a finite non-rectangular cost set is NP-hard. Moreover, determining the feasibility of RCMDPs with sa-rectangular $\mathcal{P}$ is NP-hard.*

The proof constructs an RMDP with an sa-rectangular transition set and a finite cost set that simulates the RMDP instance in Proposition 3.6. Full details are provided in Appendix E.

In the settings of Propositions 3.6 and 3.7, the uncertainty set $\mathcal{U}$ is finite. Thus, for rational γ and model parameters, the robust total cost and its subgradients can be computed exactly in polynomial time. Unless P = NP, PSD therefore cannot find an ε-optimal policy within polynomially many updates, contrary to the claims of Wang et al. [40] and Kitamura et al. [25].[4] In particular, by Lemma 2.4 and Eq. (8), $J_{\mathcal{U}}$ cannot be subgradient-dominant with a polynomially bounded D.

4 Sufficient Conditions When Projected Subgradient Descent Succeeds

Although PSD can fail for general RMDPs, it succeeds under additional conditions. In this section, we identify two sufficient conditions under which the subgradient dominance property (7) holds.

Assumption 4.1. Let $\mathcal{P}^\pi := \{P \mid (c,P) \in \mathcal{U}^\pi\}$ be the set of worst-case transition kernels under policy π. The set $\mathcal{P}^\pi$ is a singleton for any $\pi \in \Pi$.

Assumption 4.2. For any $\pi \in \Pi$, the set $\mathcal{Q}^\pi := \{Q^\pi_{c,P} \mid (c,P) \in \mathcal{U}^\pi\}$ of worst-case action-value functions contains a *robust action-value function* $Q^\pi_{\mathcal{U}}$, that is, $Q \le Q^\pi_{\mathcal{U}}$ for all $Q \in \mathcal{Q}^\pi$.

Our assumptions are motivated by the failure mode of PSD in Example 3.4. Intuitively, suboptimal local minima can arise when different worst-case models prescribe conflicting local directions toward the global optimum. Our assumptions rule out such conflict. Assumption 4.2 (e.g., a singleton $\mathcal{Q}^\pi$) ensures that all worst-case models agree on a local improvement direction, since any action that improves on π under $Q^\pi_{\mathcal{U}}$ also improves on π under every worst-case model. Assumption 4.1 requires agreement only at the transition level, since conflict in the cost is known not to create this pathology [5, 15]. The following theorem shows that either condition is sufficient for subgradient dominance.

Theorem 4.3 (Sufficient conditions for subgradient dominance)**.** *Under the full-support initial distribution (Assumption 3.2), the robust total cost $J_{\mathcal{U}}$ is subgradient-dominant with constant $D = ((1-\gamma)\min_s \mu(s))^{-1}$ if either Assumption 4.1 or Assumption 4.2 holds.*

Proof. Let $g^\star \in \partial J_{\mathcal{U}}(\pi)$ attain the minimum in $G(\pi)$. By Lemma 2.1 and Carathéodory's theorem, there exist active models $(c_1,P_1),\dots,(c_k,P_k) \in \mathcal{U}^\pi$ with finite $k \in \mathbb{N}$, and $\alpha^\star \in \Delta([k])$ such that $g^\star = \sum_{i=1}^k \alpha_i^\star \nabla J_{c_i,P_i}(\pi)$. Thus, restricting the subgradient minimization in $G(\pi)$ to the convex hull of these gradients $\nabla J_{c_i,P_i}(\pi)$ leaves $G(\pi)$ unchanged. For these active models, it holds that

$$J_{\mathcal{U}}(\pi) - J_{\mathcal{U}}(\pi^\star) \overset{(a)}{\le} \min_{i\in[k]} J_{c_i,P_i}(\pi) - J_{c_i,P_i}(\pi^\star) \overset{(b)}{=} \min_{i\in[k]} \frac{1}{1-\gamma} \sum_s \overline{d}^{\pi^\star}_{P_i}(s) \sum_a \pi^\star(a|s) A^\pi_{c_i,P_i}(s,a)$$

$$\overset{(c)}{=} \min_{\alpha\in\Delta([k])} \frac{1}{1-\gamma} \sum_{i=1}^k \alpha_i \sum_s \overline{d}^{\pi^\star}_{P_i}(s) \sum_a \pi^\star(a|s) A^\pi_{c_i,P_i}(s,a) =: ① \,, \tag{9}$$

where (a) uses $J_{\mathcal{U}}(\pi^\star) \ge J_{c,P}(\pi^\star)$ for any (c,P), (b) uses the performance difference lemma for MDPs (Lemma A.1), and (c) holds since the minimum is attained at an extreme point of the simplex.

[4] The constructions use $\mu(s_{\text{ini}}) = 1$, which violates Assumption 3.2. However, since $J_{\mathcal{U}}$ is $(1-\gamma)^{-1}$-Lipschitz in μ, mixing μ with the uniform distribution with a small weight depending only on γ preserves the hardness for a smaller ε.

Under Assumption 4.1. Let $P \in \mathcal{P}^\pi$ be the unique worst-case transition kernel under π. The right-hand side of Eq. (9) is bounded as

$$\begin{aligned}
\textcircled{1} &\le \frac{1}{1-\gamma} \min_{\alpha \in \Delta([k])} \sum_s \overline{d}_P^{\pi^\star}(s) \underbrace{\max_a \sum_{i=1}^k \alpha_i A^\pi_{c_i,P}(s,a)}_{\ge 0} \\
&\overset{(a)}{\le} \frac{1}{1-\gamma} \min_{\alpha \in \Delta([k])} \sum_s \frac{1}{(1-\gamma)\mu(s)} \overline{d}_P^{\pi}(s) \max_a \sum_{i=1}^k \alpha_i A^\pi_{c_i,P}(s,a) \\
&\overset{(b)}{\le} D \min_{\alpha \in \Delta([k])} \max_{\pi' \in \Pi} \sum_{s,a} (\pi(a|s) - \pi'(a|s)) \frac{1}{1-\gamma} \overline{d}_P^{\pi}(s) \sum_{i=1}^k \alpha_i Q^\pi_{c_i,P}(s,a) \overset{(c)}{=} DG(\pi) \ ,
\end{aligned}$$

where (a) uses $\overline{d}_P^{\pi^\star}(s)/\overline{d}_P^{\pi}(s) \le 1/((1-\gamma)\mu(s))$ and (b) uses $A^\pi_{c,P}(s,a) = \sum_{a'} \pi(a'|s) Q^\pi_{c,P}(s,a') - Q^\pi_{c,P}(s,a)$ and introduces $D := ((1-\gamma)\min_s \mu(s))^{-1}$. The equality (c) follows from the choice of active models above. This proves Theorem 4.3 under Assumption 4.1.

Under Assumption 4.2. Choose (c_1, P_1) above to attain $Q^\pi_{\mathcal{U}}$ (with $\alpha^\star_1 = 0$ if necessary), and let $A^\pi_{\mathcal{U}}$ be its advantage function. Since $Q^\pi_{c_i,P_i} \le Q^\pi_{\mathcal{U}}$ and $J_{c_i,P_i}(\pi) = J_{c_1,P_1}(\pi)$, Assumption 3.2 gives $V^\pi_{c_i,P_i} = V^\pi_{c_1,P_1}$, so $A^\pi_{c_i,P_i} \ge A^\pi_{\mathcal{U}}$ and

$$\begin{aligned}
\textcircled{1} &\overset{(a)}{\le} \frac{1}{1-\gamma} \sum_s \overline{d}_{P_1}^{\pi^\star}(s) \underbrace{\max_a A^\pi_{\mathcal{U}}(s,a)}_{\ge 0} \\
&\overset{(b)}{\le} \frac{1}{1-\gamma} \min_\alpha \sum_s \sum_{i=1}^k \alpha_i \frac{1}{(1-\gamma)\mu(s)} \overline{d}_{P_i}^{\pi}(s) \max_a A^\pi_{\mathcal{U}}(s,a) \\
&\overset{(c)}{\le} D \min_\alpha \max_{\pi' \in \Pi} \sum_{s,a} (\pi(a|s) - \pi'(a|s)) \sum_{i=1}^k \alpha_i \frac{1}{1-\gamma} \overline{d}_{P_i}^{\pi}(s) Q^\pi_{c_i,P_i}(s,a) \overset{(d)}{=} DG(\pi) \ ,
\end{aligned}$$

where (a) takes α to be the first unit vector in ①, (b) uses $\overline{d}_{P_1}^{\pi^\star}(s) \le 1 \le \sum_i \alpha_i \overline{d}_{P_i}^{\pi}(s)/((1-\gamma)\mu(s))$, (c) uses $A^\pi_{\mathcal{U}} \le A^\pi_{c_i,P_i}$ as in the previous case, and (d) follows from the choice of active models. □

Combining Theorem 4.3 with the stationarity guarantee in Lemma 2.4 and the Moreau-envelope bound Eq. (8), we obtain the following convergence guarantee.

Corollary 4.4 (PSD convergence in RMDPs)**.** *Suppose that Assumption 3.2 holds and either Assumption 4.1 or Assumption 4.2 holds. When $\eta = 1/\sqrt{T}$, the PSD iterates in Eq.* (2) *satisfy*

$$\min_{t \in [T]} J_{\mathcal{U}}(\pi_t) - J_{\mathcal{U}}(\pi^\star) \le C T^{-1/4} \ , \quad \textit{where } \ C := D' \sqrt{\frac{4}{1-\gamma} + \frac{4|\mathcal{A}|^2}{(1-\gamma)^7}} \ .$$

Here, D' is defined in Eq. (8). Using the above result, an ε-optimal policy is obtained after $\mathcal{O}(|\mathcal{A}|^4 |\mathcal{S}|^2 \varepsilon^{-4} (\min_s \mu(s))^{-4} (1-\gamma)^{-18})$ iterations. This recovers the $\mathcal{O}(\varepsilon^{-4})$ guarantee claimed by prior work [40, 25], although it is slower than the $\mathcal{O}(\varepsilon^{-1})$ rate for standard MDPs [44]. Whether this rate is tight remains an open question.

In the following subsections, we discuss concrete settings in which each assumption holds. Clearly, the standard MDP setting with a singleton uncertainty set $\mathcal{U} = \{(c, P)\}$ satisfies both assumptions. Nonetheless, the two assumptions are independent and apply to different classes of RMDPs. A formal proof is provided in Appendix F.

Proposition 4.5. *Neither of Assumptions 4.1 and 4.2 implies the other.*

4.1 Regularized RMDPs (Assumptions 4.1 and 4.2)

Both Assumptions 4.1 and 4.2 are trivially satisfied when the worst-case model is unique for every policy π, i.e., when $|\mathcal{U}^\pi| = 1$. A common approach to enforcing $|\mathcal{U}^\pi| = 1$ is to introduce

a concave regularizer that renders the adversary's maximization objective strongly concave, thereby guaranteeing a unique worst-case model. For example, consider the following variant of KL-regularized RMDPs [e.g., 45]:

$$\min_{\pi\in\Pi}\max_{P\in\mathcal{P}}\mathbb{E}_P^\pi\left[\sum_{h=0}^{\infty}\gamma^h\left(c(s_h,a_h)-\tau\sum_{a\in\mathcal{A}}D_{\mathrm{KL}}(P(\cdot\mid s_h,a)\|P_0(\cdot\mid s_h,a))\right)\right],$$

where P_0 is a nominal transition kernel and $\tau>0$ is a regularization parameter. When $\mathcal{P}$ is sa-rectangular with $\mathcal{P}_{s,a}=\Delta(\mathcal{S})$, the worst-case kernel is unique and satisfies

$$\bar{P}(\cdot|s,a)=\operatorname*{arg\,max}_{p\in\Delta(\mathcal{S})}\left\{\pi(a|s)\gamma p^\top V_{\mathcal{U}}^\pi-\tau D_{\mathrm{KL}}(p\|P_0(\cdot|s,a))\right\}\propto P_0(\cdot|s,a)\exp(\pi(a|s)\gamma V_{\mathcal{U}}^\pi/\tau)\,,$$

where $V_{\mathcal{U}}^\pi$ denotes the value function of the worst-case model.

While most existing results focus on rectangular uncertainty sets [45, 48, 11], the assumption $|\mathcal{U}^\pi|=1$ itself does not require rectangularity. As an example, consider the following ℓ_2 regularization on P:

$$\max_{P\in\mathcal{P}}J_P(\pi)-\frac{\tau}{2}\|P-P_0\|_F^2\,,$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Assuming that $\mathcal{P}$ is convex, the total cost $J_P(\pi)$ is weakly concave in P due to its smoothness [3], so the regularized objective is strongly concave for sufficiently large τ. Then, the worst-case model is unique for every policy even when $\mathcal{P}$ is non-rectangular, and both Assumptions 4.1 and 4.2 hold. Characterizing broader classes of non-rectangular uncertainty sets that satisfy $|\mathcal{U}^\pi|=1$ is an interesting direction for future work.

4.2 Cost-robust MDPs and Convex MDPs (Assumption 4.1)

Assumption 4.1 is also satisfied in cost-robust MDPs [21, 16], where uncertainty lies solely in the cost function, that is, when $\mathcal{U}=\mathcal{C}$ for a compact cost set $\mathcal{C}$.

Cost-robust MDPs are closely related to *convex MDPs* [47, 46], which study optimization problems that are convex in the occupancy measure under a fixed P:

$$\min_{\pi\in\Pi}f(d_P^\pi)\quad\text{where } f \text{ is convex and finite on an open set containing } \mathcal{D}\coloneqq\{d_P^\pi\mid\pi\in\Pi\}\,.$$

Convex MDPs generalize diverse decision-making problems, including skill learning [14], inverse reinforcement learning [6], imitation learning [20], pure exploration [19], and constrained MDPs [2]. We refer readers to Zahavy et al. [46] for a comprehensive overview.

Clearly, convex MDPs generalize cost-robust MDPs via $f(d)=\max_{c\in\mathcal{C}}\sum_{s,a}d(s,a)c(s,a)$. Conversely, cost-robust MDPs include convex MDPs. Let f^* denote the convex conjugate of f [36, Chapter 11]. Then, for every $\pi\in\Pi$,

$$f(d_P^\pi)\overset{(a)}{=}\max_{g\in\partial f(\mathcal{D})}\sum_{s,a}g(s,a)d_P^\pi(s,a)-f^*(g)\overset{(b)}{=}(1-\gamma)J_{\mathcal{C}}(\pi)\,,$$

where (a) uses the Fenchel–Young inequality, which is tight for $g\in\partial f(d_P^\pi)$, and (b) uses $\sum_{s,a}d_P^\pi(s,a)=1$ with the bounded cost set $\mathcal{C}\coloneqq\{g-f^*(g)\mid g\in\partial f(\mathcal{D})\}$. This equivalence shows that convex MDPs satisfy Assumption 4.1 and thus enjoy the subgradient dominance property.

Unless P = NP, adding transition uncertainty breaks this property even for sa-rectangular $\mathcal{P}$, since $f(d)=\max_m\sum_{s,a}d(s,a)c_m(s,a)$ recovers the NP-hard instances in Proposition 3.7.

4.3 sa- and r-Rectangular RMDPs (Assumption 4.2)

Assumption 4.2 is known to hold for sa-rectangular RMDPs [22, 32]. We show that it extends to the strictly more general class of r-rectangular RMDPs.

Proposition 4.6. *Under Assumption 3.2, Assumption 4.2 holds for r-rectangular $\mathcal{P}$.*

Proof. For simplicity, we consider the case where the cost set $\mathcal{C}$ is a singleton, so that $\mathcal{U}=\mathcal{P}$. The extension with r-rectangular cost sets is straightforward. For any $P\in\mathcal{P}^\pi$, the Bellman equation yields

$$Q_P^\pi(s,a)=c(s,a)+\gamma\sum_{s'}P(s'\mid s,a)V_P^\pi(s')=c(s,a)+\gamma\sum_{i=1}^{r}\phi_i(s,a)\underbrace{\sum_{s'}w_i(s')V_P^\pi(s')}_{=:\beta_i}\,. \quad (10)$$

Let $V^\pi_{\mathcal{U}}(s) := \max_{P\in\mathcal{P}} V^\pi_P(s)$, which is attained at all s simultaneously by the kernel $\bar{P}$ with factors $\bar{w}_i \in \arg\max_{w\in\mathcal{W}_i} \sum_{s'} w(s')V^\pi_{\mathcal{U}}(s')$ [17]. For $P \in \mathcal{P}^\pi$, Assumption 3.2 and $\sum_s \mu(s)V^\pi_P(s) = \sum_s \mu(s)V^\pi_{\mathcal{U}}(s)$ give $V^\pi_P = V^\pi_{\mathcal{U}}$, and hence $\beta_i \le \sum_{s'} \bar{w}_i(s')V^\pi_{\mathcal{U}}(s')$. Since $\phi_i(s,a) \ge 0$, Eq. (10) gives $Q^\pi_P \le Q^\pi_{\bar{P}}$, that is, $Q^\pi_{\bar{P}}$ is the robust action-value function. □

5 Related Work and Open Questions

This paper corrects a prevailing misconception that PSD is guaranteed to find ε-optimal policies in general RMDPs. We construct an explicit counterexample in which PSD remains trapped near a suboptimal policy (Section 3) and identify two sufficient conditions—Assumptions 4.1 and 4.2—under which RMDPs satisfy the subgradient dominance property (Section 4). This section situates our contributions within the existing literature on RMDPs and discusses the limitations of our results.

Hardness of approximate optimization. While the hardness of solving RMDPs has long been recognized, it has remained unclear whether finding an ε-optimal policy is hard even when the robust total cost and its subgradients are efficiently computable. This ambiguity may have contributed to incorrect convergence claims for PSD under general uncertainty sets [40, 25].

The seminal work of Wiesemann et al. [43] shows that evaluating the objective of general RMDPs is strongly NP-hard, but does not provide the hardness of policy optimization. Bagnell et al. [4] establish NP-hardness of policy optimization only for deterministic optimal policies. Mannor et al. [30] claim hardness of approximate optimization under general uncertainty sets. However, the proof is not publicly available, and it is unclear whether the hardness is due to evaluation or optimization. More recently, Ou and Bi [33] prove NP-hardness of finding an optimal policy for finite uncertainty sets, but do not address approximate optimization.

In contrast, our results (Propositions 3.6 and 3.7) establish NP-hardness of ε-optimal policy identification in two settings that admit efficient computation of both the robust total cost and its subgradients: (i) RMDPs with a finite transition set, and (ii) RMDPs with a finite sa-rectangular transition set and a finite non-rectangular cost set. Additionally, our Example 3.4 demonstrates a concrete failure case of PSD in the finite uncertainty set setting, illustrating our hardness results.

First-order methods for RMDPs. Several works have established performance guarantees for first-order methods in specific classes of RMDPs. Li et al. [28] prove global convergence of state-wise mirror descent for sa-rectangular uncertainty sets, while Wang and Zou [42] analyze PSD for R-contamination uncertainty sets, which are also sa-rectangular. More recently, convex MDPs have been shown to satisfy subgradient dominance due to an underlying hidden convexity structure [5, 15].

All of these settings are captured by the sufficient conditions identified in Section 4. In addition, we show that r-rectangular and regularized RMDPs satisfy our conditions, which were not previously known to admit subgradient dominance.

Open questions. Our sufficient conditions do not cover s-rectangular RMDPs, despite the existence of efficient DP-based algorithms for this setting [43, 18]. This gap makes the s-rectangular case particularly intriguing: while DP can efficiently solve such RMDPs, whether PSD converges to globally optimal policies remains open.

We conjecture that s-rectangular RMDPs satisfy the subgradient dominance property and that dominance can hold under conditions weaker than Assumptions 4.1 and 4.2. The following proposition shows that these assumptions are not necessary.

Proposition 5.1. *There exist s-rectangular RMDPs that satisfy the subgradient dominance property but satisfy neither Assumption 4.1 nor Assumption 4.2.*

The proof is provided in Appendix F. Identifying weaker conditions that encompass s-rectangular RMDPs while unifying the settings in Section 4 is an interesting direction for future work.

Acknowledgments and Disclosure of Funding

AG acknowledges NSF EPCL Awards #2534262 and #2621763. CS gratefully acknowledges funding from the Canada CIFAR AI Chairs Program, Amii, and NSERC.

References

[1] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the Theory of Policy Gradient Methods: Optimality, Approximation, and Distribution Shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

[2] Eitan Altman. *Constrained Markov Decision Processes*, volume 7. CRC Press, 1999.

[3] Felipe Atenas, Claudia Sagastizábal, Paulo JS Silva, and Mikhail Solodov. A Unified Analysis of Descent Sequences in Weakly Convex Optimization, Including Convergence Rates for Bundle Methods. *SIAM Journal on Optimization*, 33(1):89–115, 2023.

[4] J Andrew Bagnell, Andrew Y Ng, and Jeff G Schneider. Solving Uncertain Markov Decision Processes. Technical report, Carnegie Mellon University, 2001.

[5] Anas Barakat, Souradip Chakraborty, Peihong Yu, Pratap Tokekar, and Amrit Singh Bedi. On the Global Optimality of Policy Gradient Methods in General Utility Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2025.

[6] Stav Belogolovsky, Philip Korsunsky, Shie Mannor, Chen Tessler, and Tom Zahavy. Inverse Reinforcement Learning in Contextual MDPs. *Machine Learning*, 110(9):2295–2334, 2021.

[7] Dimitri Bertsekas. *Convex Optimization Theory*, volume 1. Athena Scientific, 2009.

[8] Rob Brekelmans, Tim Genewein, Jordi Grau-Moya, Grégoire Delétang, Markus Kunesch, Shane Legg, and Pedro Ortega. Your Policy Regularizer Is Secretly an Adversary. *Transactions on Machine Learning Research*, 2022.

[9] Ziyi Chen, Yan Wen, Zhengmian Hu, and Heng Huang. Robust Reinforcement Learning with General Utility. In *Advances in Neural Information Processing Systems*, 2025.

[10] Damek Davis and Dmitriy Drusvyatskiy. Stochastic Model-Based Minimization of Weakly Convex Functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[11] Esther Derman, Matthieu Geist, and Shie Mannor. Twice Regularized MDPs and the Equivalence between Robustness and Regularization. In *Advances in Neural Information Processing Systems*, 2021.

[12] Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of Minimizing Compositions of Convex Functions and Smooth Maps. *Mathematical Programming*, 178:503–558, 2019.

[13] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient Projections onto the ℓ_1-Ball for Learning in High Dimensions. In *International Conference on Machine Learning*, 2008.

[14] Benjamin Eysenbach, Julian Ibarz, Abhishek Gupta, and Sergey Levine. Diversity Is All You Need: Learning Skills without a Reward Function. In *International Conference on Learning Representations*, 2019.

[15] Ilyas Fatkhullin, Niao He, and Yifan Hu. Stochastic Optimization under Hidden Convexity. *SIAM Journal on Optimization*, 35(4):2544–2571, 2025.

[16] Uri Gadot, Esther Derman, Navdeep Kumar, Maxence Mohamed Elfatihi, Kfir Levy, and Shie Mannor. Solving Non-Rectangular Reward-Robust MDPs via Frequency Regularization. In *AAAI Conference on Artificial Intelligence*, 2024.

[17] Vineet Goyal and Julien Grand-Clement. Robust Markov Decision Processes: Beyond Rectangularity. *Mathematics of Operations Research*, 48(1):203–226, 2023.

[18] Julien Grand-Clément and Christian Kroer. Scalable First-Order Methods for Robust MDPs. In *AAAI Conference on Artificial Intelligence*, 2021.

[19] Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably Efficient Maximum Entropy Exploration. In *International Conference on Machine Learning*, 2019.

[20] Jonathan Ho and Stefano Ermon. Generative Adversarial Imitation Learning. In *Advances in Neural Information Processing Systems*, 2016.

[21] Hisham Husain, Kamil Ciosek, and Ryota Tomioka. Regularized Policies Are Reward Robust. In *International Conference on Artificial Intelligence and Statistics*, 2021.

[22] Garud N Iyengar. Robust Dynamic Programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

[23] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably Efficient Reinforcement Learning with Linear Function Approximation. In *Conference on Learning Theory*, 2020.

[24] Sham Kakade and John Langford. Approximately Optimal Approximate Reinforcement Learning. In *International Conference on Machine Learning*, 2002.

[25] Toshinori Kitamura, Tadashi Kozuno, Wataru Kumagai, Kenta Hoshino, Yohei Hosoe, Kazumi Kasaura, Masashi Hamaya, Paavo Parmas, and Yutaka Matsuo. Near-Optimal Policy Identification in Robust Constrained Markov Decision Processes via Epigraph Form. In *International Conference on Learning Representations*, 2025.

[26] Navdeep Kumar, Esther Derman, Matthieu Geist, Kfir Y Levy, and Shie Mannor. Policy Gradient for Rectangular Robust Markov Decision Processes. In *Advances in Neural Information Processing Systems*, 2024.

[27] Mengmeng Li, Daniel Kuhn, and Tobias Sutter. Policy Gradient Algorithms for Robust MDPs with Non-Rectangular Uncertainty Sets. *arXiv preprint arXiv:2305.19004*, 2023.

[28] Yan Li, Guanghui Lan, and Tuo Zhao. First-Order Policy Optimization for Robust Markov Decision Process. *arXiv preprint arXiv:2209.10579*, 2022.

[29] Shaocong Ma, Ziyi Chen, Yi Zhou, and Heng Huang. Rectified Robust Policy Optimization for Model-Uncertain Constrained Reinforcement Learning without Strong Duality. *Transactions on Machine Learning Research*, 2025.

[30] Shie Mannor, Ofir Mebel, and Huan Xu. Lightning Does Not Strike Twice: Robust MDPs with Coupled Uncertainty. In *International Conference on Machine Learning*, 2012.

[31] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the Global Convergence Rates of Softmax Policy Gradient Methods. In *International Conference on Machine Learning*, 2020.

[32] Arnab Nilim and Laurent El Ghaoui. Robust Control of Markov Decision Processes with Uncertain Transition Matrices. *Operations Research*, 53(5):780–798, 2005.

[33] Wenfan Ou and Sheng Bi. Sequential Decision-Making under Uncertainty: A Robust MDPs Review. *Annals of Operations Research*, 353(3):1239–1285, 2025.

[34] Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

[35] Marien Renaud, Arthur Leclaire, and Nicolas Papadakis. On the Moreau Envelope Properties of Weakly Convex Functions. *arXiv preprint arXiv:2509.13960*, 2025.

[36] R Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

[37] Reazul Hasan Russel, Mouhacine Benosman, and Jeroen Van Baar. Robust Constrained-MDPs: Soft-Constrained Robust Policy Optimization under Model Uncertainty. *arXiv preprint arXiv:2010.04870*, 2020.

[38] Maurice Sion. On General Minimax Theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

[39] Genichi Taguchi. *Introduction to Quality Engineering: Designing Quality into Products and Processes*. Asian Productivity Organization, 1986.

[40] Qiuhao Wang, Chin Pang Ho, and Marek Petrik. Policy Gradient in Robust MDPs with Global Convergence Guarantee. In *International Conference on Machine Learning*, 2023.

[41] Qiuhao Wang, Shaohang Xu, Chin Pang Ho, and Marek Petrik. Policy Gradient for Robust Markov Decision Processes. *arXiv preprint arXiv:2410.22114*, 2024.

[42] Yue Wang and Shaofeng Zou. Policy Gradient Method for Robust Reinforcement Learning. In *International Conference on Machine Learning*, 2022.

[43] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov Decision Processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[44] Lin Xiao. On the Convergence Rates of Policy Gradient Methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022.

[45] Wenhao Yang, Han Wang, Tadashi Kozuno, Scott M Jordan, and Zhihua Zhang. Robust Markov Decision Processes without Model Estimation. *arXiv preprint arXiv:2302.01248*, 2023.

[46] Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward Is Enough for Convex MDPs. In *Advances in Neural Information Processing Systems*, 2021.

[47] Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational Policy Gradient Method for Reinforcement Learning with General Utilities. In *Advances in Neural Information Processing Systems*, 2020.

[48] Runyu Zhang, Yang Hu, and Na Li. Penalized Robust MDPs and Risk-Sensitive MDPs: Equivalence, Policy Gradient, and Sample Complexity. In *International Conference on Learning Representations*, 2024.

A Useful Lemmas

Lemma A.1 (The performance difference lemma [24])**.** *For any policies $\pi, \pi' \in \Pi$,*

$$J_{c,P}(\pi) - J_{c,P}(\pi') = \frac{1}{1-\gamma} \sum_{s,a} \overline{d}_P^{\pi'}(s)\pi'(a|s)A_{c,P}^{\pi}(s,a) \ .$$

Lemma A.2 (MDP smoothness [1, Lemma 54])**.** *For any $\pi, \pi' \in \Pi$,*

$$\|\nabla J_{c,P}(\pi) - \nabla J_{c,P}(\pi')\|_2 \leq \frac{2\gamma|\mathcal{A}|}{(1-\gamma)^3}\|\pi - \pi'\|_2 \ .$$

Lemma A.3 (Lipschitzness of RMDP objective)**.** *For any $\pi, \pi' \in \Pi$,*

$$|J_{\mathcal{U}}(\pi) - J_{\mathcal{U}}(\pi')| \leq \frac{\sqrt{|\mathcal{A}|}}{(1-\gamma)^2}\|\pi - \pi'\|_2 \ .$$

Proof. The gradient norm of $J_{c,P}(\pi)$ can be upper bounded as

$$\|\nabla J_{c,P}(\pi)\|_2 = \sqrt{\sum_{s,a}\left(\frac{1}{1-\gamma}\overline{d}_P^{\pi}(s)Q_{c,P}^{\pi}(s,a)\right)^2} \leq \frac{1}{(1-\gamma)^2}\sqrt{\sum_{s,a}\left(\overline{d}_P^{\pi}(s)\right)^2} \leq \frac{\sqrt{|\mathcal{A}|}}{(1-\gamma)^2} \ ,$$

where the last inequality uses $\sum_n p_n^2 \leq \left(\sum_n p_n\right)^2 = 1$ for any probability vector $p \in \Delta([N])$. The claim then follows since bounded gradients imply Lipschitzness. □

Lemma A.4 (A version of Danskin's theorem)**.** *Let $\mathcal{Y} \subset \mathbb{R}^d$ be a compact set. Let $f : \mathbb{R}^n \times \mathcal{Y} \to \mathbb{R}$ be a bivariate jointly continuous function and $f(\cdot, y)$ is L-smooth for all $y \in \mathcal{Y}$. Let $F(x) := \max_{y\in\mathcal{Y}} f(x,y)$. If $\nabla_x f(x,\cdot)$ is continuous on $\mathcal{Y}$ for all x, then the subgradient of F at x is given by*

$$\partial F(x) = \operatorname{conv}\left\{\nabla_x f(x,y) \,\middle|\, y \in \arg\max_{y'\in\mathcal{Y}} f(x,y')\right\} .$$

Proof. The original Danskin's theorem [7, Proposition A.3.2] states that if $f(\cdot, y)$ is convex for all $y \in \mathcal{Y}$, then the statement holds.

For an L-smooth function $f(\cdot, y)$, define $\overline{f}(x, y) \coloneqq f(x, y) + \frac{L}{2}\|x\|_2^2$ and $\overline{F}(x) \coloneqq \max_{y \in \mathcal{Y}} \overline{f}(x, y)$. Since an L-smooth function is L-weakly convex [3, Proposition 2.4], $\overline{f}(\cdot, y)$ is convex. Therefore, the original Danskin's theorem indicates that:

$$\begin{aligned} \partial \overline{F}(x) &= \operatorname{conv}\left\{ \nabla_x f(x, y) + Lx \;\middle|\; y \in \arg\max_{y' \in \mathcal{Y}} \overline{f}(x, y') \right\} \\ \Longrightarrow \partial F(x) &= \operatorname{conv}\left\{ \nabla_x f(x, y) \;\middle|\; y \in \arg\max_{y' \in \mathcal{Y}} f(x, y') \right\} , \end{aligned}$$

where the second line uses $\partial \overline{F}(x) = \partial F(x) + Lx$ [36, Exercise 8.8] and $\arg\max_{y' \in \mathcal{Y}} \overline{f}(x, y') = \arg\max_{y' \in \mathcal{Y}} f(x, y')$. □

Lemma A.5. *Let $f : \mathcal{X} \to \mathbb{R}$ be an ω-weakly convex function where $\mathcal{X}$ is a convex compact set. Define $\bar{f} : x \in \mathbb{R}^d \mapsto f(x) + \delta_{\mathcal{X}}(x)$. Let $x_\nu \coloneqq \operatorname{Prox}_{\nu \bar{f}}(x)$ be the shorthand for the proximal operator. For each $0 < \nu < 1/\omega$, there exists a subgradient $g \in \partial f(x_\nu)$ such that, for any $y \in \mathcal{X}$,*

$$\langle g, x_\nu - y \rangle \leq \left\langle \nabla(\mathrm{M}_\nu \circ \bar{f})(x), x_\nu - y \right\rangle ,$$

where we write $\langle x, y \rangle \coloneqq \sum_{i=1}^d x_i y_i$ for two vectors $x, y \in \mathbb{R}^d$.

Proof. The Moreau envelope $\mathrm{M}_\nu \circ \bar{f}$ satisfies that, for any $x \in \mathbb{R}^d$,

$$\left(\mathrm{M}_\nu \circ \bar{f}\right)(x) = \min_{x' \in \mathbb{R}^d} \left\{ f(x') + \delta_{\mathcal{X}}(x') + \frac{1}{2\nu}\|x - x'\|_2^2 \right\} = \min_{x' \in \mathcal{X}} \left\{ f(x') + \frac{1}{2\nu}\|x - x'\|_2^2 \right\} .$$

It holds that $\nabla\left(\mathrm{M}_\nu \circ \bar{f}\right)(x) = \frac{1}{\nu}(x - x_\nu)$ due to Eq. (5). Note that x_ν is a minimizer of a mapping $x' \mapsto f(x') + \delta_{\mathcal{X}}(x') + \frac{1}{2\nu}\|x - x'\|_2^2$. Therefore,

$$-\frac{1}{\nu}(x_\nu - x) \in \partial(f(y) + \delta_{\mathcal{X}}(y))|_{y = x_\nu} .$$

It is well-known that $\partial \delta_{\mathcal{X}}(x) = N_{\mathcal{X}}(x)$, where $N_{\mathcal{X}}(x)$ is the normal cone of $\mathcal{X}$ at x satisfying

$$N_{\mathcal{X}}(x) \coloneqq \left\{ g \in \mathbb{R}^d \;\middle|\; \langle g, y \rangle \leq \langle g, x \rangle \quad \forall y \in \mathcal{X} \right\} .$$

Therefore, there exists a subgradient $g \in \partial f(x_\nu)$ such that

$$-g - \frac{1}{\nu}(x_\nu - x) \in N_{\mathcal{X}}(x_\nu) .$$

Since any $z \in N_{\mathcal{X}}(x_\nu)$ satisfies $\langle z, y - x_\nu \rangle \leq 0$ for any $y \in \mathcal{X}$, it holds that

$$\langle -g, y - x_\nu \rangle \leq \left\langle \frac{1}{\nu}(x_\nu - x), y - x_\nu \right\rangle , \quad \forall y \in \mathcal{X} .$$

Rearranging the above inequality with Eq. (5) completes the proof. □

B Technical Errors in Prior Work

This section explains the technical errors in the subgradient-dominance proofs of Wang et al. [40] and Kitamura et al. [25].

B.1 The error in Wang et al. [40]

We first note that Wang et al. [40] consider only transition uncertainty. In addition, their Assumption 2.1 requires the cost function to be nonnegative.

A key step in their analysis is proving their Eq. (33), which states that, for some constant $D > 0$,

$$J_{\mathcal{P}}(\pi) - \min_{\pi' \in \Pi} J_{\mathcal{P}}(\pi') \leq D \left\| \nabla(\mathrm{M}_{\frac{1}{2\rho}} \circ \bar{J}_{\mathcal{P}})(\pi) \right\|_2 . \tag{11}$$

This corresponds to our Eq. (8), which ensures that any stationary point at which PSD stops is globally optimal.

Let $\tilde{\pi} = \arg\min_{\pi' \in \Pi} \left\{ J_{\mathcal{P}}(\pi') + \rho \|\pi - \pi'\|_2^2 \right\}$ be the proximal point of π. Wang et al. [40] use their Lemma D.6 to ensure that there exists $\xi \in \partial J_{\mathcal{P}}(\tilde{\pi})$ such that $-\xi \in \mathcal{N}_\Pi(\tilde{\pi}) + 2\rho \|\tilde{\pi} - \pi\| \cdot \mathbb{B}_2$, where $\mathcal{N}_\Pi(\tilde{\pi})$ is the normal cone of Π at $\tilde{\pi}$ and $\mathbb{B}_2$ is the unit ball in an appropriate dimension. Then, using the non-negativity of the cost function, Eq. (30)-(33) in their paper attempt to derive Eq. (11) through the following inequality:

$$0 \leq J_{\mathcal{P}}(\tilde{\pi}) - J_{\mathcal{P}}(\pi^\star) \leq \frac{D}{1-\gamma} \underbrace{\langle \mathbf{1}, -\xi \rangle}_{\leq 0} \leq \frac{D\sqrt{|\mathcal{S}||\mathcal{A}|}}{1-\gamma} \left\| \nabla(\mathrm{M}_{\frac{1}{2\rho}} \circ \bar{J}_{\mathcal{P}})(\pi) \right\|_2 .$$

However, this argument is incorrect. Because the cost is nonnegative, every subgradient of $J_{\mathcal{P}}(\cdot)$ is also non-negative by Eq. (3) and Eq. (4), and hence ξ is a non-negative vector. As a result, the middle term is non-positive, whereas the left-hand side is non-negative. Therefore, the inequality can hold only when $\tilde{\pi}$ is already globally optimal but does not hold in general.

B.2 The error in Kitamura et al. [25]

Kitamura et al. [25] attempt to prove the subgradient dominance property, i.e., Eq. (7). Let $\mathcal{G} \coloneqq \{\nabla J_{c,P}(\pi) \mid (c,P) \in \mathcal{U}^\pi\}$ be the set of gradients of the worst-case models at π, so that $\partial J_{\mathcal{U}}(\pi) = \operatorname{conv} \mathcal{G}$ by Lemma 2.1. In their Eq. (36), using the gradient dominance lemma for MDPs (Lemma 3.3), they first show that

$$J_{\mathcal{U}}(\pi) - \min_{\pi' \in \Pi} J_{\mathcal{U}}(\pi') \leq D \min_{g \in \mathcal{G}} \max_{\pi' \in \Pi} \sum_{s,a} (\pi(a \mid s) - \pi'(a \mid s)) g(s,a) .$$

They then attempt, in their Eq. (37)–(39), to prove that

$$\min_{g \in \mathcal{G}} \max_{\pi' \in \Pi} \sum_{s,a} (\pi(a \mid s) - \pi'(a \mid s)) g(s,a) = \min_{g \in \operatorname{conv} \mathcal{G}} \max_{\pi' \in \Pi} \sum_{s,a} (\pi(a \mid s) - \pi'(a \mid s)) g(s,a) .$$

using Sion's minimax theorem [38] together with the convexity of Π. However, this equality is incorrect.

More generally, their argument requires that, for two compact sets $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$ with $\mathcal{X}$ convex,

$$\min_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}} \langle x, y \rangle = \min_{y \in \operatorname{conv}(\mathcal{Y})} \max_{x \in \mathcal{X}} \langle x, y \rangle$$

holds. However, this is false in general. For example, let $\mathcal{Y} = \{-1, 1\}$ and $\mathcal{X} = [-1, 1]$. Then,

$$1 = \min_{y \in \{-1,1\}} \max_{x \in [-1,1]} xy > \min_{y \in \operatorname{conv}(\{-1,1\})} \max_{x \in [-1,1]} xy = 0 ,$$

so the claimed equality does not hold.

C Convergence Analysis of Projected Subgradient Descent Method

C.1 Proof of Lemma 2.4

We utilize the existing convergence analysis of projected subgradient descent (PSD) for weakly convex functions.

Lemma C.1 ([10, Theorem 3.1])**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be an ω-weakly convex function where $\mathcal{X}$ is a convex compact set. Let $\bar{f} : x \in \mathbb{R}^d \mapsto f(x) + \delta_{\mathcal{X}}(x)$. Assume that there exist constants L and $f_{\max}$ such that $\|g\|_2 \leq L$ and $|f(x)| \leq f_{\max}$ for all $x \in \mathcal{X}$ and $g \in \partial f(x)$. Consider the update:*

$$x_{t+1} = \operatorname{Proj}_{\mathcal{X}}(x_t - \eta g_t) \quad \text{where } g_t \in \partial f(x_t) .$$

Set the step size as $\eta = \frac{1}{\sqrt{T}}$. Then, it holds that

$$\min_{t \in [T]} \left\| \nabla(\mathrm{M}_{\frac{1}{2\omega}} \circ \bar{f})(x_t) \right\|_2^2 \leq \frac{4 f_{\max} + 2\omega L^2}{\sqrt{T}} .$$

By substituting $f = J_{\mathcal{U}}$, $\mathcal{X} = \Pi$, $\omega = \rho = \frac{2|\mathcal{A}|}{(1-\gamma)^3}$, $L = \frac{\sqrt{|\mathcal{A}|}}{(1-\gamma)^2}$, and $f_{\max} = \frac{1}{1-\gamma}$ into Lemma C.1, we obtain Lemma 2.4.

C.2 Derivation of Eq. (8)

Here we provide a derivation of Eq. (8) from the subgradient dominance property Eq. (7). We use a shorthand $\pi_{\frac{1}{2\rho}} \coloneqq \mathrm{Prox}_{\frac{1}{2\rho}\bar{J}_{\mathcal{U}}}(\pi)$. Let $\bar{g} \coloneqq \nabla\Big(\mathrm{M}_{\frac{1}{2\rho}} \circ \bar{J}_{\mathcal{U}}\Big)(\pi)$ and let $g_{\frac{1}{2\rho}} \in \partial J_{\mathcal{U}}(\pi_{\frac{1}{2\rho}})$ be the subgradient guaranteed by Lemma A.5. The right-hand side of Eq. (7) can be upper bounded as

$$\begin{aligned}
J_{\mathcal{U}}(\pi_{\frac{1}{2\rho}}) - \min_{\pi' \in \Pi} J_{\mathcal{U}}(\pi') &\le D \min_{g \in \partial J_{\mathcal{U}}(\pi_{\frac{1}{2\rho}})} \max_{\pi' \in \Pi} \sum_{s,a} \Big(\pi_{\frac{1}{2\rho}}(a \mid s) - \pi'(a \mid s)\Big) g(s,a) \\
&\overset{(a)}{\le} D \max_{\pi' \in \Pi} \sum_{s,a} \Big(\pi_{\frac{1}{2\rho}}(a \mid s) - \pi'(a \mid s)\Big) g_{\frac{1}{2\rho}}(s,a) \\
&\overset{(b)}{\le} D \max_{\pi' \in \Pi} \sum_{s,a} \Big(\pi_{\frac{1}{2\rho}}(a \mid s) - \pi'(a \mid s)\Big) \bar{g}(s,a) \\
&\overset{(c)}{\le} D \max_{\pi' \in \Pi} \Big\|\pi_{\frac{1}{2\rho}} - \pi'\Big\|_2 \|\bar{g}\|_2 \overset{(d)}{\le} 2D\sqrt{|\mathcal{S}|}\|\bar{g}\|_2 \,.
\end{aligned}$$

Here, (a) uses a particular subgradient in the minimum, (b) follows from Lemma A.5, (c) follows from the Cauchy–Schwarz inequality, and (d) uses $\|\pi' - \pi\|_2 \le 2\sqrt{|\mathcal{S}|}$ for any $\pi', \pi \in \Pi$.

Additionally, using the Lipschitzness of $J_{\mathcal{U}}(\cdot)$ (Lemma A.3), we have

$$J_{\mathcal{U}}(\pi) - J_{\mathcal{U}}(\pi_{\frac{1}{2\rho}}) \le \frac{\sqrt{|\mathcal{A}|}}{(1-\gamma)^2}\Big\|\pi - \pi_{\frac{1}{2\rho}}\Big\|_2 = \frac{1-\gamma}{4\sqrt{|\mathcal{A}|}}\|\bar{g}\|_2 \,,$$

where the equality follows from Eq. (5). By combining the above two inequalities, we obtain

$$\begin{aligned}
J_{\mathcal{U}}(\pi) - \min_{\pi' \in \Pi} J_{\mathcal{U}}(\pi') &\le J_{\mathcal{U}}(\pi) - J_{\mathcal{U}}(\pi_{\frac{1}{2\rho}}) + J_{\mathcal{U}}(\pi_{\frac{1}{2\rho}}) - \min_{\pi' \in \Pi} J_{\mathcal{U}}(\pi') \\
&\le \left(2D\sqrt{|\mathcal{S}|} + \frac{1-\gamma}{4\sqrt{|\mathcal{A}|}}\right)\|\bar{g}\|_2 \,.
\end{aligned}$$

Setting $D' \coloneqq 2D\sqrt{|\mathcal{S}|} + \frac{1-\gamma}{4\sqrt{|\mathcal{A}|}}$ yields Eq. (8).

D Proof of PSD Failure (Proposition 3.5)

Proposition D.1 (Restatement of Proposition 3.5)**.** *Recall the RMDP instance defined in Example 3.4 with the transition uncertainty set $\mathcal{P} = \{P_1, P_2\}$. The following two claims hold:*

- *The policy $\tilde{\pi}_2$ that always chooses action a_2 is a suboptimal strict local minimum of $J_{\mathcal{P}}(\cdot)$ satisfying $0.505 = J_{\mathcal{P}}(\tilde{\pi}_2) > J_{\mathcal{P}}(\tilde{\pi}_1) = 0.1$.*
- *For all sufficiently small $\eta > 0$, when initialized at $\pi_1 = \tilde{\pi}_2$, the PSD update $\pi_{t+1} = \mathrm{Proj}_{\Pi}(\pi_t - \eta g_t)$ with any $g_t \in \partial J_{\mathcal{P}}(\pi_t)$ remains trapped in a neighborhood of $\tilde{\pi}_2$ such that*

$$\|\pi_t - \tilde{\pi}_2\|_\infty \le 3\eta \qquad \forall t \ge 1 \,.$$

Proof. Throughout the proof, we use shorthands $\xi_1 \coloneqq \nabla J_{P_1}(\tilde{\pi}_2)$ and $\xi_2 \coloneqq \nabla J_{P_2}(\tilde{\pi}_2)$. Using the symmetry that P_2 matches P_1 by swapping the states s_1 and s_2, it is straightforward to verify that

$$\begin{aligned}
&\xi_1(s_1,a_1) = d_1\gamma^2,\ \xi_1(s_1,a_2) = d_1\gamma, \quad \xi_1(s_2,a_1) = d_2\gamma^2,\ \xi_1(s_2,a_2) = 0 \\
&\xi_2(s_1,a_1) = d_2\gamma^2,\ \xi_2(s_1,a_2) = 0, \qquad \xi_2(s_2,a_1) = d_1\gamma^2,\ \xi_2(s_2,a_2) = d_1\gamma,
\end{aligned} \tag{12}$$

where we defined $d_1 \coloneqq \frac{1}{1-\gamma}\overline{d}^{\tilde{\pi}_2}_{P_1}(s_1) = \frac{1}{1-\gamma}\overline{d}^{\tilde{\pi}_2}_{P_2}(s_2) < \frac{1}{1-\gamma}\overline{d}^{\tilde{\pi}_2}_{P_1}(s_2) = \frac{1}{1-\gamma}\overline{d}^{\tilde{\pi}_2}_{P_2}(s_1) =: d_2$.

Strict local minimality. The suboptimality $0.505 = J_{\mathcal{P}}(\tilde{\pi}_2) > J_{\mathcal{P}}(\tilde{\pi}_1) = 0.1$ can be verified by direct calculations. We show the strict local minimality of $\tilde{\pi}_2$.

Let $\pi_{x,y}(a_1 \mid s_1) = x$ and $\pi_{x,y}(a_1 \mid s_2) = y$ be a policy that perturbs $\tilde{\pi}_2$ by x and y at states s_1 and s_2, respectively. The first-order expansion of J_{P_1} and J_{P_2} around $\tilde{\pi}_2$ yields

$$\begin{aligned} J_{P_1}(\pi_{x,y}) &= J_{P_1}(\tilde{\pi}_2) + f_1 x + f_2 y + o(x+y), \\ J_{P_2}(\pi_{x,y}) &= J_{P_2}(\tilde{\pi}_2) + f_2 x + f_1 y + o(x+y), \end{aligned}$$

where

$$\begin{aligned} f_1 &\coloneqq \xi_1(s_1, a_1) - \xi_1(s_1, a_2) = -d_1\gamma(1-\gamma), \\ f_2 &\coloneqq \xi_1(s_2, a_1) - \xi_1(s_2, a_2) = d_2\gamma^2 \,. \end{aligned}$$

Note that $f_1 + f_2 = d_2\gamma^2 - d_1\gamma(1-\gamma) > 0$ because $d_1 < d_2$ and we set $\gamma = 0.9 > 0.5$ in Example 3.4. Since $J_{P_1}(\tilde{\pi}_2) = J_{P_2}(\tilde{\pi}_2) = J_{\mathcal{P}}(\tilde{\pi}_2)$, it holds that

$$\begin{aligned} J_{\mathcal{P}}(\pi_{x,y}) - J_{\mathcal{P}}(\tilde{\pi}_2) &= \max\{f_1 x + f_2 y,\ f_2 x + f_1 y\} + o(x+y) \\ &\geq \frac{(f_1 x + f_2 y + f_2 x + f_1 y)}{2} + o(x+y) = \frac{(f_1+f_2)(x+y)}{2} + o(x+y) \,. \end{aligned}$$

Since $f_2 + f_1 > 0$ and $x, y \geq 0$, we have $J_{\mathcal{P}}(\pi_{x,y}) > J_{\mathcal{P}}(\tilde{\pi}_2)$ for all sufficiently small $(x, y) \neq (0, 0)$. This proves that $\tilde{\pi}_2$ is a strict local minimum of $J_{\mathcal{P}}$.

PSD convergence. We next prove that PSD can get trapped in a neighborhood of $\tilde{\pi}_2$. Note that, for any vector $(a, b) \in \mathbb{R}^2$ and $p \in [0, 1]$, the Euclidean projection onto $\Delta([2])$ satisfies

$$\Pi_{\Delta([2])}\big((p, 1-p) - \eta(a, b)\big) = \Big(\mathrm{clip}_{[0,1]}\Big(p - \frac{\eta}{2}(a-b)\Big),\ 1 - \mathrm{clip}_{[0,1]}\Big(p - \frac{\eta}{2}(a-b)\Big)\Big),$$

where $\mathrm{clip}_{[0,1]}(u) \coloneqq \min\{1, \max\{0, u\}\}$.

Let $\pi_t(a_1 \mid s_1) = x_t$ and $\pi_t(a_1 \mid s_2) = y_t$. Additionally, for $g \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, define $F_1(g) \coloneqq g(s_1, a_1) - g(s_1, a_2)$ and $F_2(g) \coloneqq g(s_2, a_1) - g(s_2, a_2)$. Then, the PSD update can be expressed as

$$x_{t+1} = \mathrm{clip}_{[0,1]}\Big(x_t - \frac{\eta}{2} F_1(g_t)\Big) \quad \text{and} \quad y_{t+1} = \mathrm{clip}_{[0,1]}\Big(y_t - \frac{\eta}{2} F_2(g_t)\Big). \tag{13}$$

Step 1: the initial update. Consider the first PSD update from $\tilde{\pi}_2$ with $g_1 \in \partial J_{\mathcal{P}}(\tilde{\pi}_2)$. We show that the first PSD update lands in an $O(\eta)$-neighborhood of $\tilde{\pi}_2$ that either satisfies $x_2 = 0$ or $y_2 = 0$.

Since $\partial J_{\mathcal{P}}(\tilde{\pi}_2) = \mathrm{conv}\{\xi_1, \xi_2\}$ where ξ_1, ξ_2 are defined in (12), any $g_1 \in \partial J_{\mathcal{P}}(\tilde{\pi}_2)$ can be expressed as $g_1 = \lambda\xi_1 + (1-\lambda)\xi_2$ for some $\lambda \in [0, 1]$. Hence,

$$\begin{aligned} & F_1(g_1) = \lambda F_1(\xi_1) + (1-\lambda) F_1(\xi_2) = \lambda f_1 + (1-\lambda) f_2 \\ & F_2(g_1) = \lambda F_2(\xi_1) + (1-\lambda) F_2(\xi_2) = \lambda f_2 + (1-\lambda) f_1 \\ \Longrightarrow \quad & F_1(g_1) + F_2(g_1) = f_1 + f_2 > 0 \,, \end{aligned}$$

where f_1 and f_2 are defined in the strict local minimality analysis.

Suppose that $F_1(g_1) < 0$ which implies $F_2(g_1) > 0$. Then, the update (13) only changes the policy in s_1 and keeps the policy in s_2 unchanged, i.e., $y_2 = 0$. Since $F_1(g_1) \geq f_1$, the value of x_2 is upper bounded as

$$x_2 = \mathrm{clip}_{[0,1]}\Big(x_1 - \frac{\eta}{2} F_1(g_1)\Big) \leq -\frac{\eta}{2} f_1 \,.$$

Similarly, if $F_1(g_1) > 0$, we have $x_2 = 0$ and $y_2 \leq -\frac{\eta}{2} f_1$. When $F_1(g_1) \geq 0$ and $F_2(g_1) \geq 0$, the policy remains unchanged, i.e., $(x_2, y_2) = (0, 0)$. Since $F_1(g_1) + F_2(g_1) > 0$, it is impossible that $F_1(g_1) < 0$ and $F_2(g_1) < 0$ hold simultaneously. In summary, the next coordinates (x_2, y_2) satisfy

$$(x_2, y_2) \in \{(0,0)\} \cup \Big\{\Big[0, -\frac{\eta}{2} f_1\Big] \times \{0\}\Big\} \cup \Big\{\{0\} \times \Big[0, -\frac{\eta}{2} f_1\Big]\Big\} \,.$$

Step 2: the next update. Consider a time t such that $y_t = 0$ and $x_t > 0$, i.e., the policy at state s_2 always chooses a_2 while the policy at state s_1 has a small positive probability x_t on a_1. From the instance construction, it is clear that the worst-case transition kernel for the policy π_t is P_2 and thus $g_t = \nabla J_{P_2}(\pi_t)$. We show that, when $x_t = O(\eta)$, for a sufficiently small learning rate η, the next PSD update leads to a new policy with coordinates (x_{t+1}, y_{t+1}) satisfying $x_{t+1} = 0$ and $y_{t+1} = O(\eta)$.

Let $\varepsilon := \frac{f_1+f_2}{2}$. Since J_{P_2} is smooth, there exists a constant $r_\varepsilon > 0$ such that, for all $x_t \in (0, r_\varepsilon]$,

$$-F_1(g_t) \le -F_1(\xi_2) + \varepsilon = -f_2 + \varepsilon \quad \text{and} \quad -F_2(g_t) \le -F_2(\xi_2) + \varepsilon = -f_1 + \varepsilon \,. \tag{14}$$

We now set $\eta \le \frac{2r_\varepsilon}{(-f_1+\varepsilon)}$ and assume the value of x_t to satisfy Eq. (15):

$$x_t \le \frac{\eta}{2}(-f_1 + \varepsilon) \le r_\varepsilon \,. \tag{15}$$

Then, the next PSD update leads to a new policy with coordinates (x_{t+1}, y_{t+1}) satisfying

$$\begin{aligned} x_{t+1} &= \mathrm{clip}_{[0,1]}\Big(x_t - \frac{\eta}{2}F_1(g_t)\Big) \le \mathrm{clip}_{[0,1]}\Big(\frac{\eta}{2}(-f_1+\varepsilon) - \frac{\eta}{2}(f_2 - \varepsilon)\Big) = 0 \\ \text{and} \quad y_{t+1} &= \mathrm{clip}_{[0,1]}\Big(0 - \frac{\eta}{2}F_2(g_t)\Big) \le \frac{\eta}{2}(-f_1+\varepsilon), \end{aligned}$$

where the inequalities use Eq. (14).

By symmetry, if $x_t = 0$ and $y_t = \delta_y \le \frac{\eta}{2}(-f_1+\varepsilon)$, we can similarly show that the next PSD update leads to a new policy with coordinates (x_{t+1}, y_{t+1}) satisfying

$$x_{t+1} \le \frac{\eta}{2}(-f_1+\varepsilon) \quad \text{and} \quad y_{t+1} = 0 \,.$$

Step 3: repeating argument. Consider a sufficiently small learning rate $\eta \le \frac{2r_\varepsilon}{(-f_1+\varepsilon)}$ where $\varepsilon = \frac{f_1+f_2}{2}$ and r_ε is defined to satisfy Eq. (14). Then, using the results from **Step 1**, the first PSD update from $\tilde{\pi}_2$ leads to a new policy with coordinates (x_2, y_2) satisfying

$$(x_2, y_2) \in \{(0,0)\} \cup \Big\{\Big[0, -\frac{\eta}{2}f_1\Big] \times \{0\}\Big\} \cup \Big\{\{0\} \times \Big[0, -\frac{\eta}{2}f_1\Big]\Big\} \,.$$

Since this (x_2, y_2) satisfies the conditions for (x_t, y_t) in **Step 2** (see Eq. (15) for x_t) or $(x_t, y_t) = (0, 0)$ as in **Step 1**, the next PSD update leads to coordinates (x_3, y_3) satisfying

$$(x_3, y_3) \in \{(0,0)\} \cup \Big\{\Big[0, \frac{\eta}{2}(-f_1+\varepsilon)\Big] \times \{0\}\Big\} \cup \Big\{\{0\} \times \Big[0, \frac{\eta}{2}(-f_1+\varepsilon)\Big]\Big\} \,.$$

This (x_3, y_3) again satisfies the conditions for (x_t, y_t) in **Step 2** or **Step 1**. The claim follows by iterating the above argument for all $t \ge 2$ and noting that

$$\frac{\eta}{2}(-f_1+\varepsilon) = \frac{\eta}{2}\frac{-f_1+f_2}{2} = \frac{\eta}{4}\big(d_2\gamma^2 + d_1\gamma(1-\gamma)\big) \le \frac{\eta}{4\cdot 0.1}0.9 \le 3\eta \,,$$

where we used the fact that $d_1 < d_2 < \frac{1}{1-\gamma}$ and $\gamma = 0.9$ in Example 3.4. □

E Proofs of NP-hardness

In this section, we provide proofs of Proposition 3.6 and Proposition 3.7.

E.1 NP-hardness of finite $\mathcal{P}$ and singleton $\mathcal{C}$

Proposition E.1 (Restatement of Proposition 3.6)**.** *For any fixed $\gamma \in (0,1)$ and any $\varepsilon < 0.5\gamma^3(1-\gamma)^{-1}$, finding an ε-optimal policy in RMDPs with a finite transition uncertainty set is NP-hard.*

Proof. We prove the claim by a reduction from 3-SAT.

Definition E.2 (3-SAT decision problem)**.** A 3-SAT instance consists of N Boolean variables $(x_1, \ldots, x_N) \in \{\text{False}, \text{True}\}^N$ and M clauses $C_1, \ldots, C_M$, where each clause is a disjunction of three literals. A literal is either a variable x_n or its negation $\neg x_n$. For example, a clause C_m may take the form

$$C_m = (x_i \vee \neg x_j \vee x_k) \, .$$

The 3-SAT decision problem asks whether there exists an assignment $(x_1, \ldots, x_N) \in \{\text{False}, \text{True}\}^N$ such that the entire formula evaluates to True:

$$C_1 \wedge C_2 \wedge \cdots \wedge C_M = \text{True}$$

RMDP construction. Without loss of generality, no clause contains both x_n and $\neg x_n$, since such a clause is always satisfied.

We construct an RMDP instance corresponding to a given 3-SAT instance. We begin by introducing the following states:

- the initial state s_{ini};
- the clause states $s_{C_1}, \ldots, s_{C_M}$;
- the variable states $s_{x_1}, \ldots, s_{x_N}$; and
- the absorbing states s_0 and s_{+1}, where the agent incurs $+1$ cost in s_{+1}.

We set the cost to 0 for all state–action pairs except for the absorbing state s_{+1}, and the initial distribution to $\mu(s_{\text{ini}}) = 1$.

For each clause C_m, we define a deterministic transition kernel P_m. Under P_m, the agent transitions from the initial state to the corresponding clause state s_{C_m}:

$$P_m(s_{C_m} \mid s_{\text{ini}}) = 1$$

Let m_1, m_2, m_3 denote the indices of the three literals appearing in clause C_m. For example, $C_m = (x_{m_1} \vee \neg x_{m_2} \vee x_{m_3})$. At the clause state s_{C_m}, the agent chooses an action from the set $\{a_1, a_2, a_3\}$. Selecting action $a_i \in \{a_1, a_2, a_3\}$ deterministically transitions the agent to the corresponding variable state $s_{x_{m_i}}$.

At a variable state s_{x_n}, the agent has two available actions, a_{F} and a_{T}. Under P_m, if x_n appears in C_m as the positive literal x_n, then action a_{T} leads to the absorbing state s_0, while action a_{F} transitions to the cost state s_{+1}. If x_n appears as the negated literal $\neg x_n$, these transitions are reversed: action a_{F} leads to s_0, while action a_{T} transitions to s_{+1}. Figure 2 illustrates the resulting transition structure under P_m for the clause $C_m = (x_{m_1} \vee \neg x_{m_2} \vee x_{m_3})$.

Policy construction. We prove the claim by showing that the optimal robust total cost of this RMDP is 0 when the 3-SAT instance is satisfiable, and is at least $0.5\gamma^3(1-\gamma)^{-1}$ otherwise.

Since the agent reaches a clause state s_{C_m} only under the transition kernel P_m, greedily selecting an action at s_{C_m} always minimizes the robust cost among all policies at that state. Therefore, without loss of generality, we restrict our attention to the policies that satisfy

$$\pi(a_i \mid s_{C_m}) = 1 \quad \text{where} \quad i \in \arg\min_{j \in \{1,2,3\}} V^{\pi}_{P_m}(s_{x_{m_j}}), \text{ breaking ties arbitrarily.}$$

When 3-SAT is unsatisfiable. Using a policy π, we construct a 3-SAT assignment $\beta_\pi : \{1, \cdots N\} \to \{\text{False}, \text{True}\}$ to the variables $x_1 \cdots x_N$ by

$$\beta_\pi(n) = \begin{cases} \text{True} & \text{if } \pi(a_{\text{T}} | s_{x_n}) > 0.5 \\ \text{False} & \text{otherwise} \end{cases}$$

Under P_m, the value $V^{\pi}_{P_m}(s_{x_{m_i}})$ equals $\gamma(1-\gamma)^{-1}\pi(a_{\text{F}} \mid s_{x_{m_i}})$ if the i-th literal is x_{m_i}, and $\gamma(1-\gamma)^{-1}\pi(a_{\text{T}} \mid s_{x_{m_i}})$ if it is $\neg x_{m_i}$. Hence, for a clause C_m, we have the following two cases:

- If i-th literal is x_{m_i}, β_π satisfies the i-th literal if and only if $V^{\pi}_{P_m}(s_{x_{m_i}}) < 0.5\gamma(1-\gamma)^{-1}$.

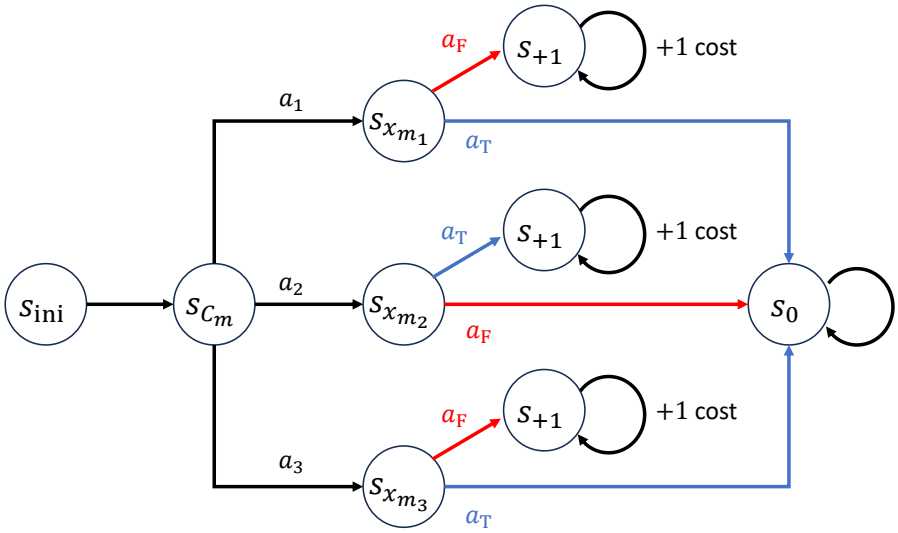


Figure 2: An illustration of the transition kernel P_m for a clause $C_m = (x_{m_1} \vee \neg x_{m_2} \vee x_{m_3})$.

- If i-th literal is $\neg x_{m_i}$, β_π satisfies the i-th literal if and only if $V^\pi_{P_m}(s_{x_{m_i}}) \leq 0.5\gamma(1-\gamma)^{-1}$.

Since the 3-SAT instance is unsatisfiable, for any policy π there exists a clause C_m for which all three literals are unsatisfied by β_π, and hence $V^\pi_{P_m}(s_{x_{m_i}}) \geq 0.5\gamma(1-\gamma)^{-1}$ for every $i \in \{1,2,3\}$. Under P_m, the agent moves from s_{ini} to s_{C_m} and then to one of the three variable states without incurring any cost. Consequently,

$$J_{\mathcal{P}}(\pi) \geq J_{P_m}(\pi) = \gamma^2 \sum_{i=1}^{3} \pi(a_i \mid s_{C_m}) V^\pi_{P_m}(s_{x_{m_i}}) \geq 0.5\gamma^3(1-\gamma)^{-1} \quad \text{for all} \quad \pi \in \Pi\,.$$

When 3-SAT is satisfiable. Let β be an assignment that satisfies the 3-SAT. Define a policy by

$$\pi_\beta(a_T \mid s_{x_n}) = \begin{cases} 1 & \text{if } \beta(n) = \text{True} \\ 0 & \text{otherwise} \end{cases}$$

Since β satisfies the 3-SAT, for any clause C_m, this policy attains $V^{\pi_\beta}_{P_m}(s_{x_{m_i}}) = 0$ for at least one $i \in \{1,2,3\}$. Due to the greedy action selection at the clause state s_{C_m}, the total cost under P_m becomes $J_{P_m}(\pi_\beta) = 0$. Consequently,

$$\min_{\pi \in \Pi} J_{\mathcal{P}}(\pi) = 0\,.$$

Therefore, if one can find an ε optimal policy π_ε with $\varepsilon < 0.5\gamma^3(1-\gamma)^{-1}$, we can solve the 3-SAT problem: It is satisfiable if $J_{\mathcal{P}}(\pi_\varepsilon) < 0.5\gamma^3(1-\gamma)^{-1}$, unsatisfiable otherwise. □

E.2 NP-hardness of sa-rectangular $\mathcal{P}$ and finite non-rectangular $\mathcal{C}$

Proposition E.3 (First claim of Proposition 3.7)**.** *For any fixed $\gamma \in (0,1)$ and any $\varepsilon < \frac{\gamma^2}{2(2-\gamma)}$, finding an ε-optimal policy with an sa-rectangular finite transition set and a finite non-rectangular cost set is NP-hard.*

Proof. We prove the claim by constructing an RMDP instance which is almost equivalent to that in Proposition E.1. Consider the following state space:

- the initial state s_{ini};

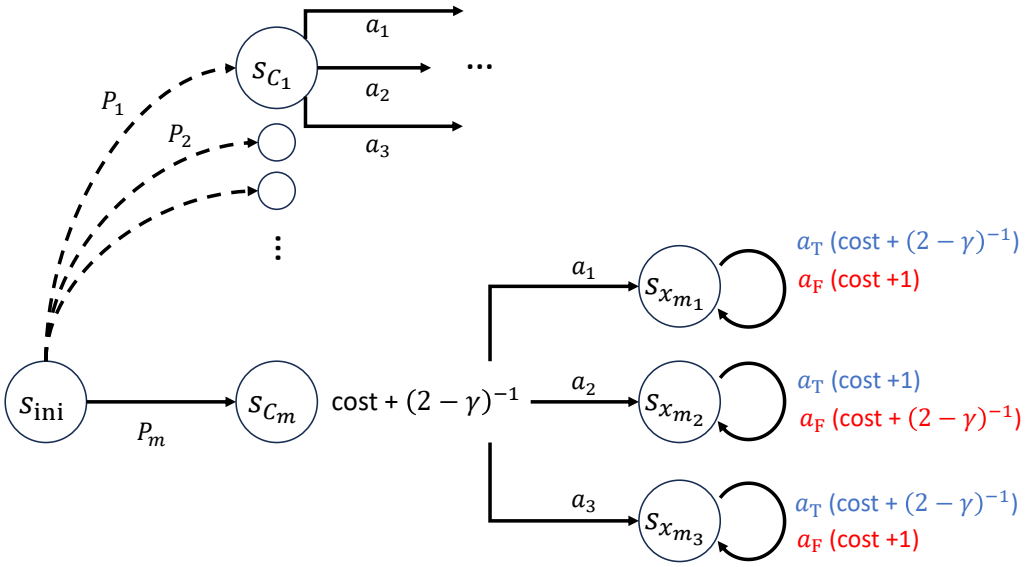


Figure 3: An illustration of the transition and the m-th cost function for $C_m = (x_{m_1} \vee \neg x_{m_2} \vee x_{m_3})$. For any policy, the total cost starting from s_{C_m} is at least $((2-\gamma)(1-\gamma))^{-1}$, whereas that starting from s_{C_i} with $i \neq m$ is at most $\gamma(1-\gamma)^{-1}$. Since $(2-\gamma)^{-1} > \gamma$, the worst-case transition kernel for c_m is P_m.

- the clause states $s_{C_1}, \ldots, s_{C_M}$; and
- the variable states $s_{x_1}, \ldots, s_{x_N}$.

The action space and the initial distribution $\mu(s_{\text{ini}}) = 1$ are identical to those in Proposition E.1.

We define the sa-rectangular transition uncertainty set $\mathcal{P} = \times_{s,a} \mathcal{P}_{s,a}$ as follows:

- Each variable state s_{x_n} is absorbing, that is, any action taken at s_{x_n} transitions to s_{x_n} with probability 1.
- Transitions at the clause states s_{C_m} are the same as in Proposition E.1: choosing an action $a_i \in \{a_1, a_2, a_3\}$ deterministically transitions the agent to the corresponding variable state $s_{x_{m_i}}$.
- At the initial state s_{ini}, the adversary chooses the next clause state independently for each action, that is, $\mathcal{P}_{s_{\text{ini}},a} := \{p_1, \ldots, p_M\}$ for every action a, where $p_m \in \Delta(\mathcal{S})$ puts probability one on s_{C_m}.

The sets $\mathcal{P}_{s,a}$ at all other state–action pairs are singletons, so $\mathcal{P}$ is a finite sa-rectangular set. For each $m \in [M]$, let $P_m \in \mathcal{P}$ denote the kernel that selects p_m for every action at s_{ini}.

We next define the finite cost set $\mathcal{C} = \{c_1, \ldots, c_M\}$ as follows. For each $m \in [M]$,

- At clause state s_{C_i}, we set $c_m(s_{C_i}, \cdot) = 0$ if $i \neq m$, and $c_m(s_{C_m}, \cdot) = (2-\gamma)^{-1}$.
- At variable states $s_{x_{m_i}}$ for $i = 1, 2, 3$, we set $c_m(s_{x_{m_i}}, a_{\text{T}}) = (2-\gamma)^{-1}$ and $c_m(s_{x_{m_i}}, a_{\text{F}}) = 1$ if the i-th literal is x_{m_i}; otherwise, we set $c_m(s_{x_{m_i}}, a_{\text{T}}) = 1$ and $c_m(s_{x_{m_i}}, a_{\text{F}}) = (2-\gamma)^{-1}$.
- All remaining costs are set to zero.

Figure 3 illustrates the resulting transition structure and cost function for the clause $C_m = (x_{m_1} \vee \neg x_{m_2} \vee x_{m_3})$.

Note that, for the cost function c_m, the worst-case transition kernel is P_m. Indeed, for every policy π,

$$V^{\pi}_{c_m,P_m}(s_{C_m}) \geq \frac{1}{(2-\gamma)(1-\gamma)} , \qquad V^{\pi}_{c_m,P_i}(s_{C_i}) \leq \frac{\gamma}{1-\gamma} \quad \forall i \neq m .$$

Since $(2-\gamma)^{-1} > \gamma$ and s_{ini} is never revisited, transitioning to s_{C_m} under every action at s_{ini} is worst-case for c_m. Consequently, for $\mathcal{U} \coloneqq \mathcal{C} \times \mathcal{P}$, we have

$$\max_{m\in[M]} J_{c_m,P_m}(\pi) = J_{\mathcal{U}}(\pi) \quad \forall \pi \in \Pi .$$

Therefore, the constructed RMDP has the same robust objective as the paired finite uncertainty set $\{(c_m, P_m) \mid m \in [M]\}$.

Under the m-th model (c_m, P_m), the total cost of any policy is obtained from that under the corresponding model in Proposition E.1 by adding $\gamma/((2-\gamma)(1-\gamma))$ and scaling its policy-dependent part by $(1-\gamma)/(\gamma(2-\gamma))$. Indeed, the baseline cost $(2-\gamma)^{-1}$ at s_{C_m} and at the absorbing variable state yields the policy-independent shift

$$\frac{\gamma}{2-\gamma} + \frac{\gamma^2}{(2-\gamma)(1-\gamma)} = \frac{\gamma}{(2-\gamma)(1-\gamma)} ,$$

while the policy-dependent cost is reduced to $1-(2-\gamma)^{-1} = (1-\gamma)/(2-\gamma)$ and incurred one step earlier, yielding the stated scaling factor. Hence, by the same argument as in Proposition E.1, finding an ε-optimal policy with $\varepsilon < \gamma^2/(2(2-\gamma))$ is NP-hard: the instance is satisfiable if

$$J_{\mathcal{U}}(\pi_\varepsilon) < \frac{\gamma}{(2-\gamma)(1-\gamma)} + \frac{\gamma^2}{2(2-\gamma)}$$

and is unsatisfiable otherwise. □

Proposition E.4 (Second claim of Proposition 3.7)**.** *Deciding whether an RCMDP with sa-rectangular $\mathcal{P}$ is feasible is NP-hard.*

Proof. Consider the sa-rectangular transition set $\mathcal{P}$ and cost functions $\mathcal{C} = \{c_1, \ldots, c_M\}$ constructed in the proof of Proposition E.3. We consider the following RCMDP problem:

$$\text{(RCMDP)} \quad \min_{\pi\in\Pi} J_{c_0,\mathcal{P}}(\pi) \quad \text{such that} \quad \max_{m\in[M]} J_{c_m,\mathcal{P}}(\pi) \leq b \quad \text{where} \quad b \coloneqq \frac{\gamma}{(2-\gamma)(1-\gamma)} ,$$

where $c_0 \equiv 0$ is the objective cost and $c_1, \ldots, c_M$ are the constraint costs constructed above. The proof of Proposition E.3 shows that $\max_{m\in[M]} J_{c_m,\mathcal{P}}(\pi) = J_{\mathcal{U}}(\pi)$ for every $\pi \in \Pi$, and that $\min_{\pi\in\Pi} J_{\mathcal{U}}(\pi)$ equals b if the 3-SAT instance is satisfiable and is at least $b + \gamma^2/(2(2-\gamma))$ otherwise. Hence, the RCMDP is feasible if and only if the 3-SAT instance is satisfiable. □

F Proof of Propositions 4.5 and 5.1

Proposition F.1 (Restatement of Proposition 4.5)**.** *Neither of the following conditions implies the other:*

- *(Assumption 4.1) Let $\mathcal{P}^\pi \coloneqq \{P \mid (c, P) \in \mathcal{U}^\pi\}$ be the set of worst-case transition kernels under policy π. The set $\mathcal{P}^\pi$ is a singleton for any $\pi \in \Pi$.*
- *(Assumption 4.2) For any $\pi \in \Pi$, the set $\mathcal{Q}^\pi \coloneqq \{Q^\pi_{c,P} \mid (c, P) \in \mathcal{U}^\pi\}$ of worst-case action-value functions contains a robust action-value function $Q^\pi_{\mathcal{U}}$, that is, $Q \leq Q^\pi_{\mathcal{U}}$ for all $Q \in \mathcal{Q}^\pi$.*

Proof. **Example 1: unique worst-case action-value, non-unique worst-case transition.** Consider an RMDP with three states $\{s_1, s_2, s_3\}$ and a single action $\{a_1\}$. The initial state is s_1, and the cost function is identically zero. The uncertainty set consists of two transition kernels:

$$P_1 : P_1(s_2 \mid s_1, a_1) = 1 , \qquad P_2 : P_2(s_3 \mid s_1, a_1) = 1 .$$

Since the cost is zero everywhere, the action-value function is identically zero under both P_1 and P_2 for every policy π, so $\mathcal{Q}^\pi$ is a singleton. However, both P_1 and P_2 are worst-case transitions, so $\mathcal{P}^\pi = \{P_1, P_2\}$. This instance satisfies only the second condition.

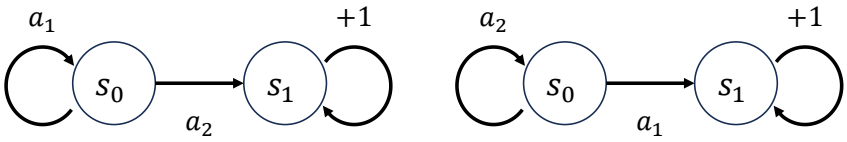


Figure 4: An s-rectangular RMDP example satisfying the subgradient dominance property but satisfying neither Assumption 4.1 nor Assumption 4.2. Left: P_1. Right: P_2.

Example 2: unique worst-case transition, no robust action-value function. Now consider a cost-robust MDP with two states $\{s_1, s_2\}$ and two actions $\{a_1, a_2\}$. The initial state is s_1. After taking any action at s_1, the agent deterministically transitions to s_2 and remains there forever. The transition kernel is therefore unique.

Define two cost functions c_1 and c_2 as

$$c_1 : c_1(s_1, a_1) = 0\ ,\ c_1(s_1, a_2) = 1\ , \qquad c_2 : c_2(s_1, a_1) = 1\ ,\ c_2(s_1, a_2) = 0\ ,$$

and assume zero cost at state s_2 for both cost functions. Clearly, the worst-case transition kernel is unique for every policy, so $|\mathcal{P}^\pi| = 1$.

However, a robust action-value function does not always exist. Consider the policy π satisfying $\pi(a_1 \mid s_1) = \pi(a_2 \mid s_1) = 1/2$. Under this policy, both c_1 and c_2 attain the same worst-case value, and hence both belong to $\mathcal{U}^\pi$. The corresponding action-value functions satisfy

$$Q^\pi_{c_1}(s_1, a_1) = 0\ ,\ Q^\pi_{c_1}(s_1, a_2) = 1\ ,$$
$$Q^\pi_{c_2}(s_1, a_1) = 1\ ,\ Q^\pi_{c_2}(s_1, a_2) = 0\ ,$$

so neither dominates the other and no robust action-value function exists under π. This instance satisfies only the first condition. □

Proposition F.2 (Restatement of Proposition 5.1)**.** *There exist s-rectangular RMDPs that satisfy the subgradient dominance property but satisfy neither Assumption 4.1 nor Assumption 4.2.*

Proof. Consider the RMDP illustrated in Figure 4 with two states $\{s_0, s_1\}$ and two actions $\{a_1, a_2\}$. State s_0 is the initial state and s_1 is absorbing. State s_1 incurs $+1$ cost and s_0 incurs 0 cost.

It has two transition kernels P_1 and P_2 defined as

$$P_1 : \quad P_1(s_0 \mid s_0, a_1) = 1\ , \quad P_1(s_1 \mid s_0, a_2) = 1\ ,$$
$$P_2 : \quad P_2(s_1 \mid s_0, a_1) = 1\ , \quad P_2(s_0 \mid s_0, a_2) = 1\ .$$

The uncertainty set $\mathcal{U} = \{P_1, P_2\}$ is s-rectangular since the uncertainty arises only at state s_0.

Fix any $\gamma \in (0, 1)$, and let π_p denote a policy satisfying $\pi_p(a_1 \mid s_0) = p$ and $\pi_p(a_2 \mid s_0) = 1 - p$ for $p \in [0, 1]$. The action choice at the absorbing state s_1 is immaterial. Since $V^{\pi_p}_{P_i}(s_1) = 1/(1-\gamma)$, the Bellman equations at s_0 give

$$V^{\pi_p}_{P_1}(s_0) = \gamma \left[p V^{\pi_p}_{P_1}(s_0) + \frac{1-p}{1-\gamma} \right],$$
$$V^{\pi_p}_{P_2}(s_0) = \gamma \left[\frac{p}{1-\gamma} + (1-p) V^{\pi_p}_{P_2}(s_0) \right].$$

Consequently,

$$J_{P_1}(\pi_p) = \frac{\gamma(1-p)}{(1-\gamma)(1-\gamma p)}\ , \qquad J_{P_2}(\pi_p) = \frac{\gamma p}{(1-\gamma)(1-\gamma+\gamma p)}\ .$$

Their difference is

$$J_{P_1}(\pi_p) - J_{P_2}(\pi_p) = \frac{\gamma(1-2p)}{(1-\gamma p)(1-\gamma+\gamma p)}\ .$$

The first branch is strictly decreasing and the second is strictly increasing. Thus, writing $\Phi(p) \coloneqq J_{\mathcal{U}}(\pi_p)$, the active model is P_1 for $p < 1/2$ and P_2 for $p > 1/2$. The two branches intersect only at the unique minimizer $p^\star = 1/2$, where

$$\Phi(p^\star) = \frac{\gamma}{(1-\gamma)(2-\gamma)} \ .$$

We next verify the subgradient dominance property directly. For a policy subgradient g, only the difference $\delta \coloneqq g(s_0, a_1) - g(s_0, a_2)$ contributes to the Frank–Wolfe gap in Eq. (7), since the two components at s_1 are equal. The set of possible differences is exactly $\partial\Phi(p)$, and hence

$$G(\pi_p) = \min_{\delta \in \partial\Phi(p)} \max_{q \in [0,1]} (p-q)\delta = \begin{cases} \dfrac{\gamma(1-p)}{(1-\gamma p)^2}, & p < 1/2, \\ 0, & p = 1/2, \\ \dfrac{\gamma p}{(1-\gamma+\gamma p)^2}, & p > 1/2. \end{cases}$$

Here, at $p = 1/2$, the interval

$$\partial\Phi(1/2) = \left[-\frac{4\gamma}{(2-\gamma)^2}, \frac{4\gamma}{(2-\gamma)^2} \right]$$

contains zero. On the other hand, the optimality gap is

$$\Phi(p) - \Phi(1/2) = \begin{cases} \dfrac{\gamma(1-2p)}{(1-\gamma p)(2-\gamma)}, & p < 1/2, \\ 0, & p = 1/2, \\ \dfrac{\gamma(2p-1)}{(1-\gamma+\gamma p)(2-\gamma)}, & p > 1/2. \end{cases}$$

For $p < 1/2$ and $p > 1/2$, respectively, the ratios of this gap to $G(\pi_p)$ satisfy

$$\frac{(1-2p)(1-\gamma p)}{(2-\gamma)(1-p)} \le 1 \ , \qquad \frac{(2p-1)(1-\gamma+\gamma p)}{(2-\gamma)p} \le 1 \ .$$

Therefore, $J_{\mathcal{U}}(\pi_p) - \min_{\pi' \in \Pi} J_{\mathcal{U}}(\pi') \le G(\pi_p)$, so the robust objective is subgradient-dominant with constant $D = 1$.

Finally, at $p = 1/2$, both P_1 and P_2 are worst-case transition kernels, so Assumption 4.1 is not satisfied. Writing $v_0 \coloneqq \gamma/((1-\gamma)(2-\gamma))$ and $v_1 \coloneqq 1/(1-\gamma)$, their action-value functions satisfy

$$Q_{P_1}^{\pi_{1/2}}(s_0, a_1) = \gamma v_0 = \frac{\gamma^2}{(1-\gamma)(2-\gamma)}, \qquad Q_{P_1}^{\pi_{1/2}}(s_0, a_2) = \gamma v_1 = \frac{\gamma}{1-\gamma},$$
$$Q_{P_2}^{\pi_{1/2}}(s_0, a_1) = \gamma v_1 = \frac{\gamma}{1-\gamma}, \qquad Q_{P_2}^{\pi_{1/2}}(s_0, a_2) = \gamma v_0 = \frac{\gamma^2}{(1-\gamma)(2-\gamma)}.$$

Since $v_0 < v_1$ for $\gamma \in (0,1)$, we have $Q_{P_1}^{\pi_{1/2}}(s_0, a_1) < Q_{P_2}^{\pi_{1/2}}(s_0, a_1)$ and $Q_{P_1}^{\pi_{1/2}}(s_0, a_2) > Q_{P_2}^{\pi_{1/2}}(s_0, a_2)$. Hence, neither of the two worst-case action-value functions dominates the other, so no robust action-value function exists under $\pi_{1/2}$, and Assumption 4.2 is not satisfied either. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Example 3.4, Assumptions 4.1 and 4.2, and Theorem 4.3.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The s-rectangularity discussion in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All assumptions (e.g., Assumptions 3.2, 4.1, and 4.2) are stated in the main text, and complete proofs are provided in Section 4 and the appendices.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: No experiments are included.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: No experiments are included.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: No experiments are included.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: No experiments are included.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: No experiments are included.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is a primarily theoretical study on robust Markov decision processes. The social impact of this research is not likely to be significant.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No model or data is released.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: We do not use any existing code assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core of this research is unrelated to LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.