A Unified MDP Framework for Solving Robust, Convex, Multi-Discount Constraints, and Beyond

Toshinori Kitamura[1], **Arnob Ghosh**[5], **Tadashi Kozuno**[2,4],
Wataru Kumagai[2], **Kenta Hoshino**[3], **Yohei Hosoe**[3],
Kazumi Kasaura[2], **Paavo Parmas**[1], **Yutaka Matsuo**[1]
toshinori-k@weblab.t.u-tokyo.ac.jp

[1] **The University of Tokyo**
[2] **OMRON SINIC X Corporation**
[3] **Kyoto University**
[4] **Osaka University**
[5] **New Jersey Institute of Technology**

Abstract

We propose the Composite Robust Markov Decision Process (CompRMDP), a simple framework that unifies a wide range of decision-making problems, including the robust MDP, convex MDP, multi-discount constrained MDP (MD-CMDP), and their combinations. While the CompRMDP objective is non-convex, we prove that, under a mild coverage assumption, such as a full support in the initial distribution, a simple subgradient descent method finds its ε-optimal policy in $\widetilde{\mathcal{O}}(\varepsilon^{-4})$ updates. Furthermore, we introduce a simple technique for ensuring the coverage assumption by perturbing the initial state distribution, while preserving the near-optimality of the resulting policy. This single algorithm solves all the captured settings, including MD-CMDP, which is a long-standing open problem since Feinberg (2000).

1 Introduction

For decades, a vast body of research has studied the Markov Decision Process (MDP) framework (Puterman, 1994), which models a sequential decision-making environment. The typical goal is to find the decision-making policy $\boldsymbol{\pi}$ that minimizes the expected total cost:

$$\text{(MDP)} \quad \min_{\boldsymbol{\pi}} J_{\gamma,\boldsymbol{c},P}(\boldsymbol{\pi}) \coloneqq \mathbb{E}\left[\boldsymbol{c}(s_0,a_0) + \gamma \boldsymbol{c}(s_1,a_1) + \gamma^2 \boldsymbol{c}(s_2,a_2) + \cdots \,\middle|\, P, \boldsymbol{\pi}\right], \tag{1}$$

where $\boldsymbol{c}$ is the cost function, $\gamma \in [0,1)$ the discount factor, P the state transition kernel. Detailed notations are provided in Section 2. An MDP can be represented by the parameters of $(\gamma, \boldsymbol{c}, P)$.

The vanilla MDP in (1), however, is often too simplistic to model real-world problems. To address this limitation, numerous MDP extensions have been developed. Robust MDPs (RMDPs) optimize for the worst-case environmental model within an uncertainty set (Wiesemann et al., 2013), which is crucial when system parameters are not precisely known (Taguchi et al., 1986). Constrained MDPs (CMDPs) introduce additional total cost constraints (Altman, 1999), which is essential for safety-critical tasks such as autonomous driving (Gu et al., 2023) and industrial control (Zhan et al., 2022). ConVex MDPs (CV-MDPs) have emerged to model non-linear objective functions over occupancy measures [1] (Zhang et al., 2020), enabling the formulation of important decision-making problems such as imitation learning (Ho & Ermon, 2016) and pure exploration (Hazan et al., 2019).

[1] The occupancy measure of a policy is its state-action visitation frequencies in the environment. See Section 2.

While each extension admits efficient algorithms, several important MDP classes remain unresolved. As a concrete example, consider the following composite optimization problem.

Example 1 (Composite MDP problem)**.** Suppose you are designing a controller for a daily housekeeping robot. The controller must balance two objectives: quickly completing tasks like dishwashing ($\gamma \ll 1$ and $\boldsymbol{c} < \mathbf{0}$)[2], and continuously patrolling the house to detect abnormalities throughout the day ($\gamma \approx 1$). At the same time, you want its controller to remain close to a baseline implementation while achieving meaningful improvements, a common objective in practice (Rajeswaran et al., 2018). Moreover, the controller must be robust to user variability. These requirements give rise to the following combination of robust, imitation, and Multi-Discount (MD-)CMDP problem:

$$\min_{\boldsymbol{\pi}} \underbrace{\max_{P_0\in\mathcal{P}} J_{\gamma_0,\boldsymbol{c}_0,P_0}(\boldsymbol{\pi})}_{\text{long-term patrol } (\gamma_0 \approx 1)} \quad \text{such that} \quad \underbrace{\max_{P_n\in\mathcal{P}} J_{\gamma_n,\boldsymbol{c}_n,P_n}(\boldsymbol{\pi}) \le 0 \quad \forall n\in\{1,\dots,N\}}_{\text{short-term house chores } (\gamma_n \ll 1)}$$

$$\text{and} \quad \underbrace{\max_{P\in\mathcal{P}} \mathrm{KL}(\boldsymbol{d}^{\boldsymbol{\pi}}_{\gamma,P} \| \boldsymbol{d}^{\boldsymbol{\pi}^{\text{base}}}_{\gamma,P})}_{\text{imitating baseline implementation}} \le \text{threshold}\,,$$

where $\mathrm{KL}(\cdot\|\cdot)$ denotes Kullback-Leibler (KL) divergence, $\boldsymbol{\pi}^{\text{base}}$ is the baseline policy, $\boldsymbol{d}^{\boldsymbol{\pi}}_{\gamma,P}$ is the occupancy measure in (γ, P), $\mathcal{P}$ is the uncertainty set of transition kernels, $\boldsymbol{c}_n$ and γ_n are the cost function and discount factor of the n-th constraint, respectively.

Example 1 illustrates a highly non-trivial MDP problem that remains unsolved in the existing literature. In particular, the MDP-CMDP problem, i.e., CMDP with multiple discount factors, is known to be **NP-hard in general** (Feinberg, 2000), and even the condition under which they become tractable remains unknown. Similarly, the integration of convex, robust, and constrained MDPs, remains an open challenge. While Kitamura et al. (2025) study robust and constrained MDPs, they do not address CV-MDPs. Conversely, Chen et al. (2025) propose an algorithm for robust and convex MDPs but fail to address CMDPs.[3] This motivates the key question of our work: *When and how can we solve composite MDP problems like Example 1?*

1.1 Contributions

We show that most MDP problems can be represented by the following simple extension of the RMDP, which we call the **Composite Robust MDP (CompRMDP)** problem:

$$\textbf{(CompRMDP)} \quad \min_{\boldsymbol{\pi}} F(\boldsymbol{\pi}) \;\text{ where }\; F(\boldsymbol{\pi}) \coloneqq \max_{M\in\mathcal{M}} J_M(\boldsymbol{\pi}) - \psi(M)\;. \tag{2}$$

Here, M is an MDP, $\mathcal{M}$ is a general uncertainty set of MDPs, and $\psi : \mathcal{M} \to \mathbb{R}$ is a *bounded* policy-independent function. Our main contributions are twofold: (i) Section 3 presents a simple algorithm to find a near-optimal policy for CompRMDP, and (ii) Section 4 shows that CompRMDP can represent many key MDP classes, including RMDP, CV-MDP, MD-CMDP, and their combinations.

Tractability of CompRMDP (Section 3). While $F(\boldsymbol{\pi})$ is neither convex nor concave (Agarwal et al., 2021), we prove that an ε-optimal policy can still be found using subgradient descent, thanks to the *subgradient dominance property*. This property guarantees that any first-order stationary point is globally optimal, thereby making CompRMDP tractable via the subgradient descent method.

The MDP objective $J_{\gamma,\boldsymbol{c},P}$ is known to enjoy the aforementioned dominant property Agarwal et al. (2021), and the result has been extended to RMDPs with varying $\boldsymbol{c}$ and P (Wang et al., 2023; Kitamura et al., 2025). We further generalize these results by proving a fundamental theorem: **the dominance property is preserved under pointwise maximization (Theorem 1)**. Combining this with the first-order convergence analysis (Davis & Drusvyatskiy, 2019), we show that under a mild

[2] If $\gamma \approx 1$ or $\boldsymbol{c} \geq 0$ is used for dishwashing, the robot may indefinitely postpone completing the task.

[3] Chen et al. (2025) assume that the objective of a CV-MDP is differentiable with a bounded gradient (see their Assumption 2), which excludes CMDPs where the objective value becomes infinite when taking an infeasible policy.

Table 1: Key instances represented by CompRMDP. Their combinations are also captured by CompRMDP. See Section 4.3. Notably, this is the first result that generalizes and solves MD-CMDP.

	Objective function	Uncertainty set $\mathcal{M}$	Function $\psi : \mathcal{M} \to \mathbb{R}$
MDP (Puterman, 1994)	$J_M(\boldsymbol{\pi})$	$\{M\}$	$\psi(M) = 0$
RMDP [1] (Wiesemann et al., 2013)	$\max_{\boldsymbol{c}, P \in \mathcal{C} \times \mathcal{P}} J_{M_{\boldsymbol{c}, P}}(\boldsymbol{\pi})$	$\mathcal{C} \times \mathcal{P}$	$\psi(M) = 0$
CV-MDP [2] (Zhang et al., 2020)	$F_{\mathrm{cv}}(\boldsymbol{d}_M^{\boldsymbol{\pi}})$	$\mathcal{C} := \mathrm{conv}\{\partial F_{\mathrm{cv}}\}$	$\psi(M_{\boldsymbol{c}}) = F_{\mathrm{cv}}^*(\boldsymbol{c})$
MD-CMDP [3] (Feinberg, 2000)	$J_{M_{\gamma_0, \boldsymbol{c}_0}}(\boldsymbol{\pi})$ s.t. $J_{M_{\gamma_n, \boldsymbol{c}_n}}(\boldsymbol{\pi}) \leq 0 \ \forall n$	$\{(\boldsymbol{c}_0 - j\mathbf{1}, \gamma_0), \ldots, (\boldsymbol{c}_N, \gamma_N)\}$	$\psi(M) = 0$

[1] $\mathcal{C}$ and $\mathcal{P}$ denote the sets of cost functions and transition kernels, respectively.
[2] F_{cv} is a convex function over occupancy measures F_{cv}^* is its conjugate.
[3] Technically, the **feasibility problem** of MD-CMDP is represented by CompRMDP. MD-CMDP can be solved by invoking the feasibility checking for a logarithmic number of times (see Section 4.2).

coverage assumption, such as full support in the initial distribution, subgradient descent identifies an ε-optimal policy after $\mathcal{O}(\varepsilon^{-4})$ updates (Theorem 2).

Moreover, we introduce a simple technique for ensuring the coverage assumption by perturbing the initial state distribution, while preserving the near-optimality of the resulting policy (Section 3.2). This result does not contradict the NP-hardness of MD-CMDP, as the hardness arises when seeking an exactly optimal policy, whereas our algorithm achieves an ε-optimal policy.

Generality of CompRMDP (Section 4). In addition to the standard MDP and RMDP, CompRMDP generalizes several important MDP classes. First, CompRMDP subsumes CV-MDP via the *biconjugate* representation. Essentially, the CV-MDP problem is formulated as $\min_{\boldsymbol{\pi}} F_{\mathrm{cv}}(\boldsymbol{d}_{\gamma,P}^{\boldsymbol{\pi}})$, where F_{cv} is a convex function over occupancy measures. CompRMDP recovers this formulation by setting ψ to be the convex conjugate of F_{cv} and defining $\mathcal{M}$ as the convex hull of the subdifferentials ∂F_{cv} (Proposition 1). This allows CompRMDP to capture a broad range of non-linear objectives over occupancy measures, including imitation learning and pure exploration.

Moreover, CompRMDP can represent MD-CMDP through the equivalent *epigraph form* of constrained problems (Boyd & Vandenberghe, 2004). This reformulation reduces MD-CMDP to a simple line search problem with CompRMDP as a subroutine (Proposition 2). Thus, by applying *bisection search* to the line search part, MD-CMDP can be solved by invoking CompRMDP a logarithmic number of times. Consequently, an ε-optimal policy of MD-CMDP can be identified by $\widetilde{\mathcal{O}}(\varepsilon^{-4})$ computations of subgradients (Corollary 2). Notably, **this is the first result solving MD-CMDP, which has been remained unresolved since Feinberg (2000).**

Finally, Section 4.3 shows that CompRMDP can also capture the composition of robust, convex, and MD-CMDPs. Using the concrete example of Example 1, we illustrate how CompRMDP encompasses these problems. Table 1 summarizes the range of MDP classes represented by CompRMDP.

2 Preliminaries

Mathematical notations. The probability simplex over a finite set $\mathcal{S}$ is denoted by $\mathscr{P}(\mathcal{S})$. For integers $a \leq b$, let $[\![a, b]\!] := \{a, \ldots, b\}$, and $[\![a, b]\!] := \emptyset$ if $a > b$. For $\boldsymbol{x} \in \mathbb{R}^N$, its n-th element is $\boldsymbol{x}(n)$. For $\boldsymbol{x} \in \mathbb{R}^{MN}$, we denote $\boldsymbol{x}(m, n)$ as its $(m-1)N + n$-th element for $m \in [\![1, M]\!]$ and $n \in [\![1, N]\!]$. We define $\mathbf{0} := (0, \ldots, 0)^\top$ and $\mathbf{1} := (1, \ldots, 1)^\top$. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$, we denote $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_i \boldsymbol{x}(i)\boldsymbol{y}(i)$. For $\mathcal{X} \subset \mathbb{R}^m$, $\mathrm{conv}\, \mathcal{X}$ denotes the convex hull of $\mathcal{X}$. For a function $f : \mathbb{R}^m \to \mathbb{R}$, $\mathrm{dom}\, f := \{\boldsymbol{x} \in \mathbb{R}^m \mid f(\boldsymbol{x}) < \infty\}$. $\mathbb{1}[\mathrm{P}]$ denotes the indicator function of a predicate P, which takes the value of 1 if P is true and 0 otherwise. For a proper function $f : \mathbb{R}^m \to \mathbb{R}$, $\partial f(\boldsymbol{x})$ denotes the Fréchet subdifferential of f at $\boldsymbol{x}$ (see Rockafellar & Wets, 2009, Definition 8.3). If $\partial f(\boldsymbol{x})$ is a singleton, its element is denoted as $\nabla f(\boldsymbol{x})$ and called the gradient of f at $\boldsymbol{x}$.

Markov Decision Process. An infinite-horizon *tabular* MDP is defined by a tuple $M := (\mathcal{S}, \mathcal{A}, \boldsymbol{\mu}, \gamma, \boldsymbol{c}, P)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces, $\boldsymbol{\mu} \in \mathscr{P}(\mathcal{S})$ is the initial distribution, and $\gamma \in [0, 1)$ is the discount factor. $\boldsymbol{c} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ denotes the cost function and $\boldsymbol{c}(s, a)$ is the cost of an action a at a state s. $P(\cdot \mid s, a) \in \mathscr{P}(\mathcal{S})$ is the transition kernel given (s, a).

A (Markovian stationary) policy $\boldsymbol{\pi}$ is a probability kernel such that $\boldsymbol{\pi}(\cdot \mid s) \in \mathscr{P}(\mathcal{A})$ denotes the action distribution at state $s \in \mathcal{S}$. We often treat a policy as a vector, $\boldsymbol{\pi} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. The set of all the policies is denoted as $\Pi \subset \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, which corresponds to the *direct parameterization* policy class (Agarwal et al., 2021). Given an MDP M with $(P, \boldsymbol{\mu}, \gamma)$, the occupancy measure $\boldsymbol{d}_M^{\boldsymbol{\pi}} \in \mathscr{P}(\mathcal{S} \times \mathcal{A})$ represents the expected γ-discounted number of visits to state-action (s, a) under $\boldsymbol{\pi}$, P, and $\boldsymbol{\mu}$: $\boldsymbol{d}_M^{\boldsymbol{\pi}}(s, a) = (1-\gamma)\mathbb{E}\big[\sum_{h=0}^{\infty} \gamma^h \mathbb{1}\{s_h = s, a_h = a\} \bigm| s_0 \sim \boldsymbol{\mu}, \boldsymbol{\pi}, P\big]$, where the expectation is over trajectories with $a_h \sim \boldsymbol{\pi}(\cdot \mid s_h)$ and $s_{h+1} \sim P(\cdot \mid s_h, a_h)$. We define $\mathcal{D}_M := \big\{\boldsymbol{d}_M^{\boldsymbol{\pi}} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \bigm| \boldsymbol{\pi} \in \Pi\big\}$ as the set of all the occupancy measures under M.

$J_M(\boldsymbol{\pi})$ denotes the total cost function expected of $\boldsymbol{\pi}$ under the initial distribution $\boldsymbol{\mu}$, defined as: $J_M : \boldsymbol{\pi} \in \Pi \mapsto \mathbb{E}\big[\sum_{h=0}^{\infty} \gamma^h \boldsymbol{c}(s_h, a_h) \bigm| s_0 \sim \boldsymbol{\mu}, \boldsymbol{\pi}, P\big] = \frac{1}{1-\gamma}\langle \boldsymbol{d}_M^{\boldsymbol{\pi}}, \boldsymbol{c}\rangle$. The goal of an MDP is to identify an optimal policy $\boldsymbol{\pi}^\star$ minimizing the expected total cost:

$$\text{(MDP)} \quad \boldsymbol{\pi}^\star \in \operatorname*{arg\,min}_{\boldsymbol{\pi} \in \Pi} J_M(\boldsymbol{\pi}) \ . \tag{3}$$

Robust Markov Decision Process. Let $\mathcal{M}$ be a finite or infinite compact set of MDPs, sharing a common $\mathcal{S}$ and $\mathcal{A}$ but may differ in other MDP components, i.e., $(\gamma, \boldsymbol{\mu}, \boldsymbol{c}, P)$. $\mathcal{M}$ can be seen as the uncertainty set of the environmental model, for example, fluctuations in component parameters when modeling a real-world dynamical system. To ensure performance for any MDP $M \in \mathcal{M}$, the RMDP problem seeks to minimize the total cost in the worst-case MDP in $\mathcal{M}$:

$$\text{(RMDP)} \quad \min_{\boldsymbol{\pi} \in \Pi} \max_{M \in \mathcal{M}} J_M(\boldsymbol{\pi}) \ . \tag{4}$$

In this paper, we let $\mathcal{C}, \mathcal{P}, \Gamma$ denote the sets of cost functions, transition kernels, and discount factors, respectively. When $\mathcal{M}$ varies only in a subset of MDP components, we represent $\mathcal{M}$ using the corresponding sets. For example, if only the cost and transition vary with fixed γ, we write $\mathcal{M} = \mathcal{C} \times \mathcal{P}$.

3 Unifying Framework: Composite Robust Markov Decision Process

Beyond the standard MDP (3), many MDP classes have emerged to model various decision-making problems, such as RMDP, CV-MDP, and CMDP. This section introduces a simple yet expressive extension of RMDPs which unifies many of these important classes under a single framework.

Specifically, we propose the following **Composite Robust MDP (CompRMDP)** problem, where the worst-case objective is composed of two parts: the expected total cost J_M, and a *penalty function* $\psi : \mathcal{M} \to \mathbb{R}$ that modulates the influence of each MDP in $\mathcal{M}$:

$$\textbf{(CompRMDP)} \quad F^\star := \min_{\boldsymbol{\pi} \in \Pi} F(\boldsymbol{\pi}) \ \text{ where } \ F(\boldsymbol{\pi}) := \max_{M \in \mathcal{M}} J_M(\boldsymbol{\pi}) - \psi(M) \ , \tag{5}$$

where $F^\star$ denotes the optimal value. We assume ψ is **bounded**, which is crucial for the tractability of CompRMDP. As we will see in Section 3.1, this boundedness enables a simple subgradient descent method to find a near-optimal policy of (5), provided that we can evaluate the subgradient of F.

Note that without any assumptions about the MDP set $\mathcal{M}$, evaluating the inner maximization ($\max_{M \in \mathcal{M}} J_M(\boldsymbol{\pi})$) becomes NP-hard due to the hardness result of RMDPs (Wiesemann et al., 2013), making CompRMDP intractable too. A common tractability condition is (s, a)-rectangularity of the uncertainty set $\mathcal{P}$ (Iyengar, 2005; Nilim & El Ghaoui, 2005), defined as $\mathcal{P} = \times_{s,a} \mathcal{P}_{s,a}$, where $\mathcal{P}_{s,a} \subseteq \mathscr{P}(\mathcal{S})$ and $\times_{s,a}$ denotes a Cartesian product over $\mathcal{S} \times \mathcal{A}$. However, enforcing such structure can limit generality of CompRMDP; for instance, rectangularity excludes finite MDP sets like $\mathcal{M} = \{M_1, \ldots, M_m\}$, which are important for CompRMDP to encompass CMDP (see Section 4.2).

Therefore, we study the general setup where we can evaluate the subgradient of the objective F:

Assumption 1. We have an algorithm that computes a subgradient $\boldsymbol{g} \in \partial F(\boldsymbol{\pi})$ for any $\boldsymbol{\pi} \in \Pi$.

The following Assumption 2 and Lemma 1 ensure that the subgradient is well-defined:

Assumption 2. $\mathcal{M}$ is compact, and the total cost $J_{\cdot}(\cdot)$ is jointly continuous on $\mathcal{M} \times \Pi$.

Lemma 1. *Under Assumption 2, the subdifferential of F at $\boldsymbol{\pi} \in \Pi$ is given by*

$$\partial F(\boldsymbol{\pi}) = \operatorname{conv}\left\{ \nabla_{\boldsymbol{\pi}} J_{M'}(\boldsymbol{\pi}) \;\middle|\; M' \in \operatorname*{arg\,max}_{M \in \mathcal{M}} J_M(\boldsymbol{\pi}) - \psi(M) \right\} . \tag{6}$$

The proof is deferred to Lemma 4 in Appendix B. Canonical examples of Assumption 2 include a discrete set $\mathcal{M} = \{M_1, \ldots, M_m\}$ or a closed and bounded subset of $\mathbb{R}^d$. When ψ is a constant, Kumar et al. (2022; 2024); Wang & Zou (2022) have developed efficient subgradient evaluation algorithms under some structural assumptions on $\mathcal{M}$. The penalty function ψ becomes non-constant, for example, in the CV-MDP setting (see, Section 4.1). In such convex cases, ψ is typically defined only on the cost function, which should allow us for efficient subgradient computation (Zahavy et al., 2021). **We leave the general subgradient evaluation method for non-constant ψ to future work**.

3.1 First-Order Algorithm for CompRMDP

When the subgradient is available, the *projected subgradient method* is a generic algorithm for solving $\min_{\boldsymbol{\pi}} F(\boldsymbol{\pi})$. With a learning rate $\eta > 0$, it updates policies as follows:

$$\boldsymbol{\pi}_{t+1} = \operatorname{Proj}_{\Pi}(\boldsymbol{\pi}_t - \eta \nabla J_{M_t}(\boldsymbol{\pi}_t)) \;\text{ where }\; M_t \in \operatorname*{arg\,max}_{M \in \mathcal{M}} J_M(\boldsymbol{\pi}_t) - \psi(M) . \tag{7}$$

By using the standard subgradient method analysis under mild regularity conditions (e.g., Theorem 3.1 of Davis & Drusvyatskiy, 2019), we can show that (7) converges to a stationary point such that $\mathbf{0} \in \partial F(\boldsymbol{\pi})$. Due to the space limitation, we defer the detailed convergence analysis to Appendix B

However, while it converges, due to the non-convexity of $J_M(\cdot)$ in $\boldsymbol{\pi}$ (Agarwal et al., 2021, Lemma 3.1), it does not directly indicate convergence to an optimal policy. For the MDP setting (3), Agarwal et al. (2021) addresses this challenge by showing that $J_M(\cdot)$ satisfies the *gradient dominance* property, which guarantees that any stationary point is indeed optimal of (3).

Definition 1 (Gradient dominance)**.** A function $f : \mathcal{X} \to \mathbb{R}$ is said to be gradient dominant with a constant D if there exists $D > 0$ such that for all $\boldsymbol{x} \in \mathcal{X}$, $f(\boldsymbol{x}) - \min_{\boldsymbol{x}' \in \mathcal{X}} f(\boldsymbol{x}') \leq D \max_{\boldsymbol{x}' \in \mathcal{X}} \langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}' \rangle$.

We extend their result with the following key theorem: **applying a pointwise maximum preserves the dominance property.**

Definition 2 (Subgradient domination)**.** A function $F : \mathcal{X} \to \mathbb{R}$ is said to be subgradient dominant with a constant D if there exists $D > 0$ such that, for any $\boldsymbol{x} \in \mathcal{X}$, $F(\boldsymbol{x}) - \min_{\boldsymbol{x}' \in \mathcal{X}} F(\boldsymbol{x}') \leq D \max_{\boldsymbol{x}' \in \mathcal{X}} \langle \boldsymbol{g},\, \boldsymbol{x} - \boldsymbol{x}' \rangle \quad \forall \boldsymbol{g} \in \partial F(\boldsymbol{x})$.

Theorem 1 (Pointwise maximum preserves dominance)**.** *Let $\mathcal{X} \subset \mathbb{R}^m$ be a compact convex set and $\mathcal{Y} \subset \mathbb{R}^n$ be a compact set. Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a fuunction differentiable in $\boldsymbol{x}$ and $f(\cdot, \boldsymbol{y})$ is gradient dominant with a constant $D > 0$ for each $\boldsymbol{y} \in \mathcal{Y}$. Then, $F : \boldsymbol{x} \in \mathcal{X} \mapsto \max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y})$ is subgradient dominant with D.*

Since $J_M(\cdot)$ is gradient dominant for each $M \in \mathcal{M}$, $F(\cdot)$ inherits the dominance property:

Corollary 1 (Dominance of CompRMDP)**.** *Define $D_{\mathcal{M}} := \max_{M \in \mathcal{M}} \max_{\boldsymbol{\pi} \in \Pi} \|\boldsymbol{d}_M^{\boldsymbol{\pi}^\star} / \boldsymbol{d}_M^{\boldsymbol{\pi}}\|_\infty$. Under Assumption 2, for any $\boldsymbol{\pi} \in \Pi$, $F(\boldsymbol{\pi}) - F^\star \leq D_{\mathcal{M}} \max_{\boldsymbol{\pi}' \in \Pi} \langle \boldsymbol{g}, \boldsymbol{\pi} - \boldsymbol{\pi}' \rangle \;\forall \boldsymbol{g} \in \partial F(\boldsymbol{\pi})$.*

Using Corollary 1 with the subgradient convergence result (Davis & Drusvyatskiy, 2019), we show that the best-iteration of (7) converges to an optimal policy in the rate of $O(T^{-1/4})$ as follows.

Theorem 2 (Policy subgradient convergence)**.** *Let $\gamma_{\max}$ and $c_{\max}$ be the maximum discount factor and absolute value of the cost function in $\mathcal{M}$. Let $F_{\max} \coloneqq \max_{\boldsymbol{\pi}\in\Pi} |F(\boldsymbol{\pi})|$. Suppose $c_{\max}$ and $D_{\mathcal{M}}$ are bounded. When $\eta = 1/\sqrt{T}$, the update by* (7) *satisfies*

$$\min_{t\in[\![1,T]\!]} F(\boldsymbol{\pi}_t) - F^\star \leq C^4 T^{-1/4} \ ,$$

where $C \coloneqq \left(2D_{\mathcal{M}}\sqrt{|\mathcal{S}|} + \frac{\ell}{2L}\right)\sqrt{4F_{\max} + 2L^3}$, $L \coloneqq \frac{2\gamma_{\max}c_{\max}|\mathcal{A}|}{(1-\gamma_{\max})^3}$, and $\ell \coloneqq \frac{c_{\max}\sqrt{A}}{(1-\gamma_{\max})^2}$.

3.2 $\boldsymbol{\mu}$-Perturbation Trick

Note that Theorem 2 requires that $D_{\mathcal{M}}$ is finite, which may not hold in general. This section introduces a simple trick to ensure finite $D_{\mathcal{M}}$ by slightly perturbing the initial state distributions in $\mathcal{M}$. We call this the *$\boldsymbol{\mu}$-perturbation trick.*

Let $\varepsilon_{\boldsymbol{\mu}} \in (0, 1)$ be a small positive value. For each $M \in \mathcal{M}$, we perturb its initial distribution $\boldsymbol{\mu}$ by the following linear mixture with the uniform distribution $\mathbf{1}/|\mathcal{S}|$:

$$\widetilde{\boldsymbol{\mu}} \coloneqq (1 - \varepsilon_{\boldsymbol{\mu}})\boldsymbol{\mu} + \varepsilon_{\boldsymbol{\mu}}\frac{\mathbf{1}}{|\mathcal{S}|} \ . \tag{8}$$

Let $\widetilde{M}$ be the MDP with the perturbed initial state distribution $\widetilde{\boldsymbol{\mu}}$ and the other components the same as M. Let $\widetilde{\mathcal{M}}$ be the set of perturbed MDPs for all $M \in \mathcal{M}$. We then replace the original objective function F of the CompRMDP (5) with the following perturbed objective:

$$\text{(Perturbed CompRMDP)} \quad \widetilde{F}(\boldsymbol{\pi}) \coloneqq \max_{\widetilde{M}\in\widetilde{\mathcal{M}}} J_{\widetilde{M}}(\boldsymbol{\pi}) - \psi(\widetilde{M}) \ , \tag{9}$$

Note that $\min_s \widetilde{\boldsymbol{\mu}}(s) \geq \frac{\varepsilon_{\boldsymbol{\mu}}}{|\mathcal{S}|}$ for any $M \in \mathcal{M}$. Therefore, the value of $D_{\widetilde{\mathcal{M}}}$ becomes finite as follows: $D_{\widetilde{\mathcal{M}}} \leq |\mathcal{S}|(1 - \gamma_{\max})^{-1}\varepsilon_{\boldsymbol{\mu}}^{-1}$.

Finally, we show that a near-optimal policy for the perturbed objective is also near-optimal for the original CompRMDP problem (5) if ψ satisfies a mild Lipschitz continuity condition:

Assumption 3 (ψ is continuous in $\boldsymbol{\mu}$)**.** For an MDP $M \in \mathcal{M}$, let $M_{\boldsymbol{\mu}}$ be the MDP with the initial state distribution $\boldsymbol{\mu}$ and the other components the same as M. We assume that ψ is continuous in $\boldsymbol{\mu}$ such that, for any $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathscr{P}(\mathcal{S})$, there exists a constant $\ell_\psi > 0$ such that

$$|\psi(M_{\boldsymbol{\mu}_1}) - \psi(M_{\boldsymbol{\mu}_2})| \leq \ell_\psi \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_1 \quad \forall M \in \mathcal{M} \ .$$

Theorem 3 ($\boldsymbol{\mu}$-perturbation trick)**.** *Let $\widetilde{\boldsymbol{\pi}}$ be an ε-optimal policy for the perturbed CompRMDP problem such that: $\widetilde{F}(\widetilde{\boldsymbol{\pi}}) - \min_{\boldsymbol{\pi}\in\Pi}\widetilde{F}(\boldsymbol{\pi}) \leq \varepsilon$. If Assumption 3 holds and we set $\varepsilon_{\boldsymbol{\mu}}$ sufficiently small such that $\varepsilon_{\boldsymbol{\mu}} \leq \frac{1}{4}\left(\frac{c_{\max}}{1-\gamma_{\max}} + \ell_\psi\right)^{-1}$, then $\widetilde{\boldsymbol{\pi}}$ is 2ε-optimal for the original problem* (5) *such that:*

$$F(\widetilde{\boldsymbol{\pi}}) - \min_{\boldsymbol{\pi}\in\Pi} F(\boldsymbol{\pi}) \leq 2\varepsilon \ .$$

We defer the proof to Appendix C.

4 Relation to Existing MDP Problems

This section presents problem examples that can be formulated as CompRMDP.

4.1 ConVex MDP (CV-MDP)

CV-MDP is a general framework that allows for non-linear objective functions over occupancy measures. Specifically, for an MDP M, CV-MDP considers the following optimization:

$$\text{(CV-MDP)} \quad \min_{\boldsymbol{\pi} \in \Pi} F_{\mathrm{cv}}(\boldsymbol{d}_M^{\boldsymbol{\pi}}) \ . \tag{10}$$

Here, $F_{\mathrm{cv}} : \mathcal{D}_M \to \mathbb{R}$ is a bounded[4] continuous and convex function. To show that CompRMDP (5) generalizes CV-MDP, we use the *biconjugate* of F_{cv}:

Definition 3 (Rockafellar & Wets, 2009, Chapter 11)**.** For $f : \mathbb{R}^m \to \mathbb{R}$, its conjugate f^* is defined as $f^*(\boldsymbol{y}) = \sup_{\boldsymbol{x} \in \mathrm{dom}\, f} \langle \boldsymbol{y}, \boldsymbol{x} \rangle - f(\boldsymbol{x})$. We call f^{**} the biconjugate of f.

Let $\mathcal{G} := \{\boldsymbol{g} \mid \boldsymbol{g} \in \partial F_{\mathrm{cv}}(\boldsymbol{d}),\ \boldsymbol{d} \in \mathcal{D}_M\}$ be the set of F_{cv}'s subgradients. Using the properties of the conjugate function (Fact 1 in Appendix A), we have

$$\min_{\boldsymbol{\pi} \in \Pi} F_{\mathrm{cv}}(\boldsymbol{d}_M^{\boldsymbol{\pi}}) \overset{(a)}{=} \min_{\boldsymbol{d} \in \mathcal{D}_M} F_{\mathrm{cv}}(\boldsymbol{d}) \overset{(b)}{=} \min_{\boldsymbol{d} \in \mathcal{D}_M} \max_{\boldsymbol{g} \in \mathrm{dom}\, F_{\mathrm{cv}}^*} \langle \boldsymbol{g}, \boldsymbol{d} \rangle - F_{\mathrm{cv}}^*(\boldsymbol{g}) \overset{(c)}{=} \min_{\boldsymbol{\pi} \in \Pi} \max_{\boldsymbol{g} \in \mathcal{G}} J_{M_{\boldsymbol{g}}}(\boldsymbol{\pi}) - F_{\mathrm{cv}}^*(\boldsymbol{g}) \ ,$$

where (a) follows from the one-to-one mapping between $\boldsymbol{d}_M^{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}$ (Puterman, 1994) and (b) uses definition of the convex conjugate. In (c), $M_{\boldsymbol{g}}$ denotes an MDP with the cost $\boldsymbol{c} = \boldsymbol{g}$. Since $F_{\mathrm{cv}}(\boldsymbol{g})$ is a bounded and continuous, F_{cv}^* is also bounded and continuous in $\mathcal{G}$. The equation above shows that CV-MDP is generalized by CompRMDP, and thus can be solved by the subgradient method (7).

Proposition 1 (CV-MDP $\subset$ CompRMDP)**.** *Let $\mathcal{C} = \mathcal{G}$ be the cost set. Define $\mathcal{M} = \mathcal{C}$ and $\psi : M_{\boldsymbol{g}} \in \mathcal{M} \mapsto F_{\mathrm{cv}}^*(\boldsymbol{g})$. Then, by Theorem 2 with these $\mathcal{M}$ and ψ, applying the update (7) for $\mathcal{O}(\varepsilon^{-4})$ iterations yields an ε-optimal policy $\boldsymbol{\pi}_\varepsilon$ satisfying $F_{\mathrm{cv}}(\boldsymbol{d}_M^{\boldsymbol{\pi}_\varepsilon}) \leq \min_{\boldsymbol{\pi} \in \Pi} F_{\mathrm{cv}}(\boldsymbol{d}_M^{\boldsymbol{\pi}}) + \varepsilon$.*

4.2 Multi-Discount Constrained MDP (MD-CMDP) and Feasibility Problem

MD-CMDP is a setting which seeks to minimize the total cost while satisfying N constraints, with each constraint associated with a different cost function and discount factor. Define cost functions $\mathcal{C} := \{\boldsymbol{c}_0 \dots \boldsymbol{c}_N\}$ and discount factors $\Gamma := \{\gamma_0, \dots, \gamma_N\}$. Without loss of generality, we assume that $\boldsymbol{c}_n \in [0,1]^{|\mathcal{S}||\mathcal{A}|}$ for all n. An MD-CMDP considers the following problem:

$$\text{(MD-CMDP)} \quad j^\star := \min_{\boldsymbol{\pi} \in \Pi} J_{M_{\boldsymbol{c}_0, \gamma_0}}(\boldsymbol{\pi}) \ \text{ such that } \ J_{M_{\boldsymbol{c}_n, \gamma_n}}(\boldsymbol{\pi}) \leq 0 \quad \forall n \in [\![1, N]\!] \ . \tag{11}$$

where $M_{\boldsymbol{c},\gamma}$ denotes an MDP parameterized by a cost function $\boldsymbol{c}$ and a discount factor γ. We assume that (11) is feasible and we denote $j^\star$ as the optimal objective value.

Since Equation (11) is a constrained optimization problem, it can be rewritten as the equivalent *epigraph form* (e.g., Stein, 2025):

$$\begin{aligned} \text{(MD-CMDP-Epigraph)} \quad j^\star &= \min_{j \in [0, (1-\gamma)^{-1}]} j \ \text{ such that } \ \min_{\boldsymbol{\pi} \in \Pi} F_j(\boldsymbol{\pi}) \leq 0 \\ \text{where} \quad F_j(\boldsymbol{\pi}) &:= \max_{n \in [\![0, N]\!]} J_{M_{\boldsymbol{c}_n, \gamma_n}}(\boldsymbol{\pi}) - j \mathbb{1}[n = 0] \ . \end{aligned} \tag{12}$$

For convenience, we call the subproblem $\min_{\boldsymbol{\pi} \in \Pi} F_j(\boldsymbol{\pi})$ inside (12) the **feasibility problem**. It is easy to see that a policy $\boldsymbol{\pi}^\star$ is optimal in (11) if and only if $\boldsymbol{\pi}^\star \in \arg\min_{\boldsymbol{\pi} \in \Pi} F_{j^\star}(\boldsymbol{\pi})$, and $\min_{\boldsymbol{\pi} \in \Pi} F_j(\boldsymbol{\pi})$ is a monotonically decreasing in j. Consequently, a simple bisection search on $[0, (1-\gamma)^{-1}]$ with feasibility subroutine will converge to $j^\star$ and an optimal policy $\boldsymbol{\pi}^\star$:

$$\textbf{(Bisection Search)} \quad \text{Increase } j \text{ if } \min_{\boldsymbol{\pi} \in \Pi} F_j(\boldsymbol{\pi}) > 0 \ \text{ and } \ \text{decrease } j \text{ otherwise.} \tag{13}$$

The subproblem $\min_{\boldsymbol{\pi} \in \Pi} F_j(\boldsymbol{\pi})$ is clearly a CompRMDP instance. In other words, MD-CMDP can be solved by iteratively solving CompRMDP problems. We remark that **this is the first result solving MD-CMDP, which has remained unresolved since Feinberg (2000).**

[4] Due to the boundedness of F_{cv}, CMDP is excluded from CV-MDP. We deal with CMDP in Section 4.2.

Proposition 2 (Feasibility problem $\subset$ CompRMDP)**.** *Define $\psi(\cdot) = 0$ and an MDP set as $\mathcal{M} = \{(\boldsymbol{c}_0 - j\mathbf{1}, \gamma_0), (\boldsymbol{c}_1, \gamma_1), \dots, (\boldsymbol{c}_N, \gamma_N)\}$. Then, by Theorem 2 with these $\mathcal{M}$ and ψ, applying the update (7) for $\mathcal{O}(\varepsilon^{-4})$ iterations yields a policy $\boldsymbol{\pi}_\varepsilon$ satisfying $F_j(\boldsymbol{\pi}_\varepsilon) \le \min_{\boldsymbol{\pi}\in\Pi} F_j(\boldsymbol{\pi}) + \varepsilon$.*

Corollary 2 (Tractability of MD-CMDP)**.** *Suppose $D_{\mathcal{M}}$ is bounded. After $\widetilde{\mathcal{O}}(\varepsilon^{-4})$ total policy updates by (7), the bisection search (13) converges to an ε-optimal policy $\boldsymbol{\pi}_\varepsilon$ for MD-CMDP, such that*

$$J_{M_{\boldsymbol{c}_0,\gamma_0}}(\boldsymbol{\pi}_\varepsilon) \le j^\star + \varepsilon \ , \quad J_{M_{\boldsymbol{c}_n,\gamma_n}}(\boldsymbol{\pi}_\varepsilon) \le \varepsilon \ , \quad \forall n \in [\![1, N]\!] \ . \tag{14}$$

4.3 Composition of Robust, Convex, and Multi-Discount CMDPs

CompRMDP can represent the composition of all the MDP classes described above. We demonstrate this with a concrete example, Example 1 from Section 1, which is a composition of MD-CMDP, RMDP, and CV-MDP. Let $\mathcal{P}$ be a set of transition kernels, $\mathcal{C} = \{\boldsymbol{c}_0, \boldsymbol{c}_1, \dots, \boldsymbol{c}_N\}$ be $N+1$ cost functions, and $\Gamma = \{\gamma_0, \gamma_1, \dots, \gamma_N, \gamma\}$ be $N+2$ discount factors. Let $\widetilde{\boldsymbol{\pi}} \in \Pi$ be a base policy to imitate. Define $M_{\gamma,\boldsymbol{c},P}$ as an MDP parameterized by $(\gamma, \boldsymbol{c}, P)$, and $M_{\gamma,P}$ as the corresponding MDP without a cost function. Then, Example 1 considers the following constrained optimization problem:

$$\begin{aligned}\min_{\boldsymbol{\pi}} \underbrace{\max_{P_0\in\mathcal{P}} J_{M_{\gamma_0,\boldsymbol{c}_0,P_0}}(\boldsymbol{\pi})}_{\textcircled{1}} \quad \text{such that} \quad & \underbrace{\max_{P_n\in\mathcal{P}} J_{M_{\gamma_n,\boldsymbol{c}_n,P_n}}(\boldsymbol{\pi})}_{\textcircled{2}} \le 0 \quad \forall n \in \{1, \dots, N\} \\ \text{and} \quad & \underbrace{\max_{P\in\mathcal{P}} \mathrm{KL}(\boldsymbol{d}^{\boldsymbol{\pi}}_{M_{\gamma,P}} \| \boldsymbol{d}^{\widetilde{\boldsymbol{\pi}}}_{M_{\gamma,P}}) - \rho}_{\textcircled{3}} \le 0\end{aligned}$$

where $\rho \ge 0$ is a threshold for the imitation. Each component ①, ②, and ③ can be instantiated as a CompRMDP with the following MDP sets and penalty functions:

① $\mathcal{M}_0 = \{(\gamma_0, \boldsymbol{c}_0, P_0) \mid P_0 \in \mathcal{P}\}$ and $\psi_0(M) = 0$

② $\mathcal{M}_n = \{(\gamma_n, \boldsymbol{c}_n, P_n) \mid P_n \in \mathcal{P}\}$ and $\psi_n(M) = 0$ for all $n \in [\![1, N]\!]$.

③ Define $f_P : \boldsymbol{d} \in \mathcal{D}_{M_{\gamma,P}} \mapsto \mathrm{KL}(\boldsymbol{d} \| \boldsymbol{d}^{\widetilde{\boldsymbol{\pi}}}_{M_{\gamma,P}})$. Let $\mathcal{C}_P := \mathrm{conv}\big\{\nabla f_P(\boldsymbol{d}) \bigm| \boldsymbol{d} \in \mathcal{D}_{M_{\gamma,P}}\big\}$ be the cost set under P. According to Section 4.1, ② can be represented as a CompRMDP with $\mathcal{M}_{\mathrm{KL}} := \{(\gamma, \boldsymbol{c}, P) \mid \boldsymbol{c} \in \mathcal{C}_P, P \in \mathcal{P}\}$ and $\psi_{\mathrm{KL}}(M_{\gamma,\boldsymbol{c},P}) = f^*_P(\boldsymbol{c}) - \rho$ for all $M_{\gamma,\boldsymbol{c},P} \in \mathcal{M}_{\mathrm{KL}}$.

Define $F_n(\boldsymbol{\pi}) := \max_{M\in\mathcal{M}_n} J_M(\boldsymbol{\pi}) - \psi_n(M)$ be the corresponding CompRMDP objective function for each $\mathcal{M}_n$ and ψ_n. We denote the one for ③ as F_{N+1}. Using the same epigraph technique as in Section 4.2, we can solve the composite problem by the following bisection search:

$$\text{Increase } j \text{ if } \min_{\boldsymbol{\pi}\in\Pi} \max_{n\in[\![0,N+1]\!]} F_n(\boldsymbol{\pi}) - j\mathbb{1}[n=0] > 0 \ \text{ and } \ \text{decrease } j \text{ otherwise} \ .$$

Clearly, the subproblem $\min_{\boldsymbol{\pi}\in\Pi} \max_{n\in[\![0,N+1]\!]} F_n(\boldsymbol{\pi}) - j\mathbb{1}[n=0]$ is a CompRMDP instance, which can be solved by the subgradient method (7). Beyond this example, a more general and formal algorithm for composite problems is provided in Appendix D.

Notably, **this is the first work to solve the combination of convex, robust, and (MD-)CMDPs**. Chen et al. (2025) attempted to address this combination from the CV-MDP perspective but could not fully resolve it. Their analysis assumes that the CV-MDP objective is differentiable with a bounded gradient (see their Assumption 2), which excludes CMDPs where the objective value becomes infinite when taking an infeasible policy. We resolve the boundedness challenge by using the epigraph form of (MD-)CMDP.

5 Conclusion

We proposed CompRMDP, a unifying framework that captures many important MDP problems through a simple extension of RMDP. We show that the subgradient descent method converges to an ε-optimal policy under the coverage condition, which can be satisfied our perturbation technique (Section 3.2). CompRMDP generalizes RMDPs, CV-MDPs, MD-CMDPs, and their combinations. This is the first result that solves MD-CMDP, which has been unresolved since Feinberg (2000).

References

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.

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

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.

Dimitri Bertsekas. *Convex optimization theory*, volume 1. Athena Scientific, 2009.

Stephen P Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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

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

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

Eugene A Feinberg. Constrained Discounted Markov Decision Processes and Hamiltonian Cycles. *Mathematics of Operations Research*, 25(1):130–140, 2000.

Ziqing Gu, Lingping Gao, Haitong Ma, Shengbo Eben Li, Sifa Zheng, Wei Jing, and Junbo Chen. Safe-State Enhancement Method for Autonomous Driving via Direct Hierarchical Reinforcement Learning. *IEEE Transactions on Intelligent Transportation Systems*, 24(9):9966–9983, 2023.

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

Jonathan Ho and Stefano Ermon. Generative Adversarial Imitation Learning. In *Advances in neural information processing systems*, 2016.

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

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.

Navdeep Kumar, Kfir Levy, Kaixin Wang, and Shie Mannor. Efficient Policy Iteration for Robust Markov Decision Processes via Regularization. *arXiv preprint arXiv:2205.14327*, 2022.

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.

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

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

Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations. *Robotics: Science and Systems XIV*, 2018.

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

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

Oliver Stein. A Tutorial on Properties of the Epigraph Reformulation. *EURO Journal on Computational Optimization*, pp. 100109, 2025.

G. Taguchi, G. Taguchi, and Asian Productivity Organization. *Introduction to Quality Engineering: Designing Quality Into Products and Processes*. Asian Productivity Organization, 1986.

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

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

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

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

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

Xianyuan Zhan, Haoran Xu, Yue Zhang, Xiangyu Zhu, Honglei Yin, and Yu Zheng. DeepThermal: Combustion Optimization for Thermal Power Generating Units Using Offline Reinforcement Learning. In *AAAI Conference on Artificial Intelligence*, 2022.

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.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Useful Facts and Lemmas

Fact 1. *Suppose that $f : \mathbb{R}^m \to \mathbb{R}$ is a proper convex function. Then,*

- *(Rockafellar & Wets, 2009, Theorem 11.1) $f^{**} = f$,*
- *(Rockafellar & Wets, 2009, Proposition 11.3) $\partial f(\boldsymbol{x}) = \arg\max_{\boldsymbol{y} \in \mathrm{dom}\, f^*} \{\langle \boldsymbol{y}, \boldsymbol{x} \rangle - f^*(\boldsymbol{y})\}$.*

Fact 2 (Policy gradient theorem; Xiao, 2022, Appendix A.1)**.** *For a fixed MDP M, for any $\boldsymbol{\pi} \in \Pi$,*

$$(\nabla J_M(\boldsymbol{\pi}))(s,a) = \frac{1}{1-\gamma} \boldsymbol{d}^{\boldsymbol{\pi}}_M(s) Q^{\boldsymbol{\pi}}_M(s,a) \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}\ . \tag{15}$$

where $Q^{\boldsymbol{\pi}}_M : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the action-value function, which represents the expected total cost of $\boldsymbol{\pi}$ starting from state s and taking action a:

$$Q^{\boldsymbol{\pi}}_M(s,a) = \boldsymbol{c}(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s,a) \sum_{a' \in \mathcal{A}} \boldsymbol{\pi}(a' \mid s') Q^{\boldsymbol{\pi}}_M(s',a') \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}\ .$$

Fact 3 (MDP continuity; Wang et al., 2023, Lemma 3.1)**.** *Consider an MDP M where the absolute value of the cost function is bounded by $c_{\max}$. For any $\boldsymbol{\pi}, \boldsymbol{\pi}' \in \Pi$,*

$$|J_M(\boldsymbol{\pi}) - J_M(\boldsymbol{\pi}')| \le \ell \left\|\boldsymbol{\pi} - \boldsymbol{\pi}'\right\|_2 \ \ \textit{where} \ \ \ell := \frac{c_{\max}\sqrt{A}}{(1-\gamma)^2}\ . \tag{16}$$

Fact 4 (MDP smoothness; Agarwal et al., 2021, Lemma 54)**.** *Consider an MDP M where the absolute value of the cost function is bounded by $c_{\max}$. For any $\boldsymbol{\pi}, \boldsymbol{\pi}' \in \Pi$,*

$$\left\|\nabla J_M(\boldsymbol{\pi}) - \nabla J_M(\boldsymbol{\pi}')\right\|_2 \le L \left\|\boldsymbol{\pi} - \boldsymbol{\pi}'\right\|_2 \ \ \textit{where} \ \ L := \frac{2\gamma c_{\max}|\mathcal{A}|}{(1-\gamma)^3}\ . \tag{17}$$

Fact 5 (Gradient dominance; Agarwal et al., 2021, Lemma 4)**.** *For a fixed MDP M, for any $\boldsymbol{\pi} \in \Pi$,*

$$J_M(\boldsymbol{\pi}) - J_M(\boldsymbol{\pi}^\star) \le D_M \max_{\bar{\boldsymbol{\pi}} \in \Pi} \langle \nabla J_M(\boldsymbol{\pi}),\ \boldsymbol{\pi} - \bar{\boldsymbol{\pi}} \rangle \ \ \textit{where} \ \ D_M := \max_{\boldsymbol{\pi} \in \Pi} \left\| \frac{\boldsymbol{d}^{\boldsymbol{\pi}^\star}_M}{\boldsymbol{d}^{\boldsymbol{\pi}}_M} \right\|_\infty\ . \tag{18}$$

B Proof of Theorem 2

We prove Theorem 2 by considering the following general optimization problem:

$$(\textbf{CompMax}) \quad \min_{\boldsymbol{x} \in \mathcal{X}} F(\boldsymbol{x}) \ \text{ where } \ F : \boldsymbol{x} \in \mathcal{X} \mapsto \max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y})\ . \tag{19}$$

For convenience, we call this problem Composite optimization with Maximum (CompMax). Here, $\mathcal{X} \subset \mathbb{R}^m$, $\mathcal{Y} \subset \mathbb{R}^n$, and $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ satisfy the following properties:

Assumption 4. f is bounded and jointly continuous on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is a compact convex set and $\mathcal{Y}$ is a compact set. There exists $B > 0$ such that $\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|_2 \le B$ for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$. Moreover, for any $\boldsymbol{y} \in \mathcal{Y}$, $f(\cdot, \boldsymbol{y})$ is L-smooth and ℓ-continuous on $\mathcal{X}$.

Assumption 5 (Pointwise gradient dominance)**.** f satisfies the following *pointwise gradient dominance* property: For any $\boldsymbol{y} \in \mathcal{Y}$, there exists a constant $D > 0$ such that

$$f(\boldsymbol{x}, \boldsymbol{y}) - \min_{\boldsymbol{x}' \in \mathcal{X}} f(\boldsymbol{x}', \boldsymbol{y}) \le D \max_{\boldsymbol{x}' \in \mathcal{X}} \langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}),\ \boldsymbol{x} - \boldsymbol{x} \rangle \ \forall \boldsymbol{x} \in \mathcal{X}\ . \tag{20}$$

The following lemma formally shows that CompRMDP is subsumed by CompMax:

Lemma 2 (CompRMDP $\subset$ CompMax)**.** *When $\mathcal{X} = \Pi$, $\mathcal{Y} = \mathcal{M}$ satisfying Assumption 2, and $f(\boldsymbol{\pi}, M) = J_M(\boldsymbol{\pi}) + \psi(M)$, CompMax (19) with Assumptions 4 and 5 encompasses CompRMDP (5) with Assumption 2.*

Proof. The claim holds by showing that $J_M(\cdot)$ satisfies the requirements of Assumptions 4 and 5. It is easy to verify that $\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2 \leq 2\sqrt{|\mathcal{S}|}$ for any $\boldsymbol{\pi}, \boldsymbol{\pi}' \in \Pi$. Additionally, due to Fact 3 and 4, $J_M(\cdot)$ is L-smooth and ℓ-continuous, where L and ℓ are defined in Fact 3 and 4. Thus, f in the setting of Lemma 2 satisfies Assumption 4. Assumption 5 is clearly satisfied by Fact 5. The convexity and compactness of $\mathcal{X}$ is trivial since Π is a probability simplex. This concludes the proof. □

Consequently, Theorem 2 can be proven by analyzing the following general subgradient method for CompMax (19). At iteration $k \in \mathbb{N}$, the method updates $\boldsymbol{x}_t$ to a new point $\boldsymbol{x}_{t+1}$ as follows:

$$\boldsymbol{x}_{t+1} \coloneqq \operatorname{Proj}_{\mathcal{X}}(\boldsymbol{x}_t - \eta \boldsymbol{g}_t) \text{ where } \boldsymbol{g}_t \in \partial F(\boldsymbol{x}_t)\ . \tag{21}$$

Here, $\eta > 0$ is the step size. To facilitate the subsequent analysis, we first introduce the properties on F and f in (19).

B.1 Properties of the composite problem

While we do not require convexity on f in (19), it satisfies the following *weak convexity*:

Definition 4 (Weak convexity; Atenas et al., 2023, Definition 2.1)**.** $f : \mathbb{R}^m \to \mathbb{R}$ is ω-weakly convex if there exists $\omega \geq 0$ such that $f(\cdot) + \frac{\omega}{2}\|\cdot\|_2^2$ is a convex function.

Lemma 3 ($f(\cdot, \boldsymbol{y})$ and F are weakly convex)**.** *Under Assumption 4, $f(\cdot, \boldsymbol{y})$ is L-weakly convex for any $\boldsymbol{y} \in \mathcal{Y}$, and F is L-weakly convex.*

Proof. For an ℓ-continuous convex function g and an L-smooth function f, the composition $g(f(\cdot))$ is known to be L-weakly convex (Atenas et al., 2023, Proposition 2.4). □

Using Lemma 3, the subdifferential of F is given as follows:

Lemma 4. *Under Assumption 4, the subdifferential of F at $\boldsymbol{x} \in \mathcal{X}$ is given by*

$$\partial F(\boldsymbol{x}) = \operatorname{conv}\left\{ \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}) \;\middle|\; \boldsymbol{y} \in \operatorname*{arg\,max}_{\boldsymbol{y}' \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}') \right\}\ . \tag{22}$$

Proof. Due to Lemma 3, $\bar{f}(\cdot, \boldsymbol{y}) \coloneqq f(\cdot, \boldsymbol{y}) + \frac{L}{2}\|\cdot\|_2^2$ is convex. Let $\bar{F}(\boldsymbol{x}) \coloneqq \max_{\boldsymbol{y} \in \mathcal{Y}} \bar{f}(\boldsymbol{x}, \boldsymbol{y})$. Since $\bar{f}(\cdot, \boldsymbol{y})$ is convex, the Danskin's theorem (Bertsekas, 2009, Proposition A.3.2) implies

$$\partial \bar{F}(\boldsymbol{x}) = \operatorname{conv}\left\{ \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}) + \ell \boldsymbol{x} \;\middle|\; \boldsymbol{y} \in \operatorname*{arg\,max}_{\boldsymbol{y}' \in \mathcal{Y}} \bar{f}(\boldsymbol{x}, \boldsymbol{y}') \right\}\ .$$

Since $\partial \bar{F}(\boldsymbol{x}) = \partial F(\boldsymbol{x}) + L\boldsymbol{x}$ holds (e.g., Rockafellar & Wets, 2009, Exercise 8.8), the claim immediately follows from the above equation. □

The following Theorem 4 is the key to establish the tractability of CompRMDP, which shows that **the dominance property is preserved under pointwise maximization**:

Definition 5 (Subgradient domination)**.** Consider a proper, weakly convex function $F : \mathbb{R}^m \to \mathbb{R}$ and a compact set $\mathcal{X} \subset \mathbb{R}^m$. F is said to be **subgradient dominant** for $\mathcal{X}$ if there exists a constant $D > 0$ such that, for any $\boldsymbol{x} \in \mathcal{X}$,

$$F(\boldsymbol{x}) - \min_{\boldsymbol{x}' \in \mathcal{X}} F(\boldsymbol{x}') \leq D \max_{\boldsymbol{x}' \in \mathcal{X}} \langle \boldsymbol{g},\ \boldsymbol{x} - \boldsymbol{x}' \rangle \quad \forall \boldsymbol{g} \in \partial F(\boldsymbol{x})\ . \tag{23}$$

Theorem 4 (Restatement of Theorem 1)**.** *Let $\mathcal{X} \subset \mathbb{R}^m$ be a compact convex and $\mathcal{Y} \subset \mathbb{R}^n$ be a compact set. Consider a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ differentiable in $\boldsymbol{x}$ that satisfies Assumption 5. Then, $F : \boldsymbol{x} \in \mathcal{X} \mapsto \max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y})$ is subgradient dominant for $\mathcal{X}$.*

Proof. Consider a $\bar{\boldsymbol{x}} \in \mathcal{X}$. Since $f(\cdot, \boldsymbol{y})$ is gradient dominant, for any $\bar{\boldsymbol{y}} \in \arg\max_{\boldsymbol{y} \in \mathcal{Y}} f(\bar{\boldsymbol{x}}, \boldsymbol{y})$, we have

$$F(\bar{\boldsymbol{x}}) - \min_{\boldsymbol{x}' \in \mathcal{X}} F(\boldsymbol{x}') \leq f(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}) - \min_{\boldsymbol{x}' \in \mathcal{X}} f(\boldsymbol{x}', \bar{\boldsymbol{y}}) \leq D \max_{\boldsymbol{x}' \in \mathcal{X}} \langle \nabla_{\boldsymbol{x}} f(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}),\ \bar{\boldsymbol{x}} - \boldsymbol{x}' \rangle \ ,$$

Thus, by letting $\mathcal{G}_{\bar{\boldsymbol{x}}} := \left\{ \nabla_{\boldsymbol{x}} f(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}) \;\middle|\; \bar{\boldsymbol{y}} \in \arg\max_{\boldsymbol{y} \in \mathcal{Y}} f(\bar{\boldsymbol{x}}, \boldsymbol{y}) \right\}$, we have

$$F(\bar{\boldsymbol{x}}) - \min_{\boldsymbol{x}' \in \mathcal{X}} F(\boldsymbol{x}') \leq D \min_{\boldsymbol{g} \in \mathcal{G}_{\bar{\boldsymbol{x}}}} \max_{\boldsymbol{x}' \in \mathcal{X}} \langle \boldsymbol{g},\ \bar{\boldsymbol{x}} - \boldsymbol{x}' \rangle =: D① \ . \tag{24}$$

Due to Lemma 4, we have $\partial F(\bar{\boldsymbol{x}}) = \operatorname{conv} \mathcal{G}_{\bar{\boldsymbol{x}}}$. Therefore,

$$① = \min_{\boldsymbol{g} \in \mathcal{G}_{\bar{\boldsymbol{x}}}} \max_{\boldsymbol{x}' \in \mathcal{X}} \langle \boldsymbol{g},\ \bar{\boldsymbol{x}} - \boldsymbol{x}' \rangle \geq \min_{\boldsymbol{g} \in \operatorname{conv} \mathcal{G}_{\bar{\boldsymbol{x}}}} \max_{\boldsymbol{x}' \in \mathcal{X}} \langle \boldsymbol{g},\ \bar{\boldsymbol{x}} - \boldsymbol{x}' \rangle = \min_{\boldsymbol{g} \in \partial F(\bar{\boldsymbol{x}})} \max_{\boldsymbol{x}' \in \mathcal{X}} \langle \boldsymbol{g},\ \bar{\boldsymbol{x}} - \boldsymbol{x}' \rangle =: ② \ .$$

Define two vectors $\boldsymbol{g}^\star$ and $\boldsymbol{x}^\star$ such that

$$\boldsymbol{g}^\star \in \mathop{\arg\min}_{\boldsymbol{g} \in \operatorname{conv} \mathcal{G}_{\bar{\boldsymbol{x}}}} \max_{\boldsymbol{x}' \in \mathcal{X}} \langle \boldsymbol{g}, \bar{\boldsymbol{x}} - \boldsymbol{x}' \rangle \quad \text{and} \quad \boldsymbol{x}^\star \in \mathop{\arg\max}_{\boldsymbol{x}' \in \mathcal{X}} \min_{\boldsymbol{g} \in \operatorname{conv} \mathcal{G}_{\bar{\boldsymbol{x}}}} \langle \boldsymbol{g}, \bar{\boldsymbol{x}} - \boldsymbol{x}' \rangle \ .$$

The claim holds by showing that there exists $\boldsymbol{g}^\star$ such that $\boldsymbol{g}^\star \in \mathcal{G}_{\bar{\boldsymbol{x}}}$, which ensures ① = ② in the above inequality. For the term ②, we have

$$② = \max_{\boldsymbol{x}' \in \mathcal{X}} \langle \boldsymbol{g}^\star,\ \bar{\boldsymbol{x}} - \boldsymbol{x}' \rangle \overset{(a)}{=} \langle \boldsymbol{g}^\star,\ \bar{\boldsymbol{x}} - \boldsymbol{x}^\star \rangle \overset{(b)}{=} \min_{\boldsymbol{g} \in \operatorname{conv} \mathcal{G}_{\bar{\boldsymbol{x}}}} \langle \boldsymbol{g},\ \bar{\boldsymbol{x}} - \boldsymbol{x}^\star \rangle \overset{(c)}{=} \min_{\boldsymbol{g} \in \mathcal{G}_{\bar{\boldsymbol{x}}}} \langle \boldsymbol{g},\ \bar{\boldsymbol{x}} - \boldsymbol{x}^\star \rangle \ , \tag{25}$$

where (a) and (b) use the Sion's minimax theorem (Sion, 1958, Theorem 3.4), and (c) holds because linear minimization on a convex hull has its minimum at the extreme points (Kitamura et al., 2025, Lemma 15). Since there exists a $\boldsymbol{g}^\star \in \mathcal{G}_{\bar{\boldsymbol{x}}}$ by (25), the claim holds. □

B.2 Convergence to Stationary Points of Moreau Envelope

By leveraging the weak-convexity of F by Lemma 3, we show that the subgradient method (7) converges to a stationary point of F.

Since F in (19) is non-smooth, we establish this property through the *Moreau envelope*, which offers a smooth approximation of a non-smooth function. For $\mathcal{X} \subset \mathbb{R}^m$, a proper function $F : \mathcal{X} \to \overline{\mathbb{R}}$, and $\nu > 0$, the Moreau envelope is defined by $F_\nu : \mathbb{R}^m \to \overline{\mathbb{R}}$ such that

$$F_\nu(\boldsymbol{x}) = \min_{\boldsymbol{y} \in \mathcal{X}} \left\{ F(\boldsymbol{y}) + \frac{1}{2\nu} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \right\} \ .$$

While F is non-smooth, the Moreau envelope with small ν is smooth and differentiable:

Fact 6 (Drusvyatskiy & Paquette, 2019, Lemma 4.3)**.** *Let $F : \mathbb{R}^m \to \overline{\mathbb{R}}$ be a proper ω-weakly convex function. For any $0 < \nu < 1/\omega$, the gradient of the Moreau envelope F_ν is*

$$\nabla F_\nu(\boldsymbol{x}) = \frac{1}{\nu} \left(\boldsymbol{x} - \mathop{\arg\min}_{\boldsymbol{y} \in \mathcal{X}} \left(F(\boldsymbol{y}) + \frac{1}{2\nu} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \right) \right) \quad \forall \boldsymbol{x} \in \mathbb{R}^m \ . \tag{26}$$

Moreover, the stationary points of F_ν are the same as those of F.

This smoothness of the Moreau envelope leads to the following descent property:

Lemma 5 (Descent property of Moreau envelope)**.** *Under Assumption 4, for any constant $\bar{\omega} > L$ and for any $k > 0$, it holds that*

$$F_{1/\bar{\omega}}(\boldsymbol{x}_t) - F_{1/\bar{\omega}}(\boldsymbol{x}_{t+1}) \geq \frac{\eta(\bar{\omega} - L)}{\bar{\omega}} \left\|\nabla F_{1/\bar{\omega}}(\boldsymbol{x}_t)\right\|_2^2 - \frac{\eta^2 \bar{\omega} L^2}{2} \ .$$

Therefore, when $\bar{\omega} = 2L$,

$$F_{1/2L}(\boldsymbol{x}_t) - F_{1/2L}(\boldsymbol{x}_{t+1}) \geq \frac{\eta}{2} \left\|\nabla F_{1/2L}(\boldsymbol{x}_t)\right\|_2^2 - L^3 \eta^2 \ . \tag{27}$$

Proof. As F is L-weakly convex (Lemma 3), the claim directly follows from the standard subgradient method analysis (Theorem 3.1 of Davis & Drusvyatskiy (2019)). □

By taking the telescoping sum of (27) over $t \in [\![1, T]\!]$, the sequence $\{\boldsymbol{x}_t\}$ satisfies that

$$F_{1/2L}(\boldsymbol{x}_1) - F_{1/2L}(\boldsymbol{x}_{t+1}) \geq \frac{\eta T}{2} \min_{t \in [\![1,T]\!]} \left\| \nabla F_{1/2L}(\boldsymbol{x}_t) \right\|_2^2 - T L^3 \eta^2$$

$$\Longrightarrow \sqrt{\frac{4 \max_{\boldsymbol{x} \in \mathcal{X}} |F(\boldsymbol{x})|}{T\eta} + 2L^3\eta} \geq \min_{t \in [\![1,T]\!]} \left\| \nabla F_{1/2L}(\boldsymbol{x}_t) \right\|_2 \ , \tag{28}$$

where the second line uses $F_\rho(\boldsymbol{x}) \leq F(\boldsymbol{x})$ for any $\rho > 0$ (Rockafellar & Wets, 2009, Theorem 1.25). Intuitively, the norm $\left\| \nabla F_{1/2L}(\boldsymbol{x}) \right\|_2$ measures how close $\boldsymbol{x}$ is to a stationary point of F. Indeed, for sufficiently small ν, the stationary points of F_ν coincide with those of F (Drusvyatskiy & Paquette, 2019, Lemma 4.3). Thus, (28) implies that when $\eta = 1/\sqrt{T}$, the projected subgradient method (21) convergences to a stationary point at the rate of $O(T^{-1/4})$.

B.3 Putting Everything Together

We finish the proof of Theorem 2 by combining the subgradient domination (23) to the stationary point convergence (28). To this end, we utilize the following fact:

Fact 7 (Kitamura et al., 2025, Lemma 14)**.** *Consider an ω-weakly convex function $F : \mathcal{X} \to \mathbb{R}$ and a $\boldsymbol{x} \in \mathcal{X}$. For $0 < \nu < 1/\omega$, define*

$$\bar{\boldsymbol{x}}_\nu \in \operatorname*{arg\,min}_{\boldsymbol{x}' \in \mathcal{X}} F\left(\boldsymbol{x}'\right) + \frac{1}{2\nu} \left\| \boldsymbol{x} - \boldsymbol{x}' \right\|_2^2 \ .$$

Then, there exists a subgradient $\boldsymbol{g} \in \partial F\left(\bar{\boldsymbol{x}}_\nu\right)$ such that, for any $\boldsymbol{y} \in \mathcal{X}$,

$$\langle \boldsymbol{g}, \bar{\boldsymbol{x}}_\nu - \boldsymbol{y} \rangle \leq \langle \nabla F_\nu(\boldsymbol{x}), \bar{\boldsymbol{x}}_\nu - \boldsymbol{y} \rangle \ .$$

Let $\bar{\boldsymbol{x}}_t \in \arg\min_{\boldsymbol{x}' \in \mathcal{X}} F\left(\boldsymbol{x}'\right) + L \left\| \boldsymbol{x}_t - \boldsymbol{x}' \right\|_2^2$. Combining Fact 7 and (23), we have

$$F(\bar{\boldsymbol{x}}_t) - F^\star \leq D \min_{\boldsymbol{g} \in \mathcal{G}_{\bar{\boldsymbol{x}}}} \max_{\boldsymbol{x}' \in \mathcal{X}} \langle \boldsymbol{g},\ \bar{\boldsymbol{x}}_t - \boldsymbol{x}' \rangle \leq D \max_{\boldsymbol{x}' \in \mathcal{X}} \langle \nabla F_{1/2L}(\boldsymbol{x}_t), \bar{\boldsymbol{x}}_t - \boldsymbol{x}' \rangle \leq DB \left\| \nabla F_{1/2L}(\boldsymbol{x}_t) \right\|_2 \ , \tag{29}$$

where the last inequality uses the Cauchy-Schwarz inequality with the assumption $\left\| \boldsymbol{x} - \boldsymbol{x}' \right\|_2 \leq B$. On the other hand, Due to the ℓ-continuity of F (Assumption 4),

$$F(\boldsymbol{x}_t) - F(\bar{\boldsymbol{x}}_t) \leq \ell \left\| \boldsymbol{x}_t - \bar{\boldsymbol{x}}_t \right\|_2 = \frac{\ell}{2L} \left\| \nabla F_{1/2L}(\boldsymbol{x}_t) \right\|_2 \ , \tag{30}$$

where the last equality uses the gradient of Moreau envelope (Fact 6).

Finally, by combining (29), (30), and Theorem 4, we have

$$\min_{t \in [\![1,T]\!]} F(\boldsymbol{x}_t) - F^\star \leq \left(DB + \frac{\ell}{2L} \right) \min_{t \in [\![1,T]\!]} \left\| \nabla F_{1/2L}(\boldsymbol{x}_t) \right\|_2 \leq C T^{-1/4} \ , \tag{31}$$

where we defined $C \coloneqq \left(DB + \frac{\ell}{2L} \right) \sqrt{4 \max_{\boldsymbol{x} \in \mathcal{X}} |F(\boldsymbol{x})| + 2L^3}$.

Consequently, Theorem 2 holds by inserting $D = D_{\mathcal{M}}$ from Corollary 1, $B = 2\sqrt{|\mathcal{S}|}$, ℓ and L from Fact 3 and 4, into the definition of C.

C Proof of Theorem 3

For an MDP $M = (P, \boldsymbol{c}, \gamma, \boldsymbol{\mu})$, it is well-known that the total cost $J_M(\boldsymbol{\pi})$ can be represented as (Puterman, 1994)

$$J_M(\boldsymbol{\pi}) = \boldsymbol{\mu}^\top (I - \gamma P^{\boldsymbol{\pi}})^{-1} \boldsymbol{c}^{\boldsymbol{\pi}} \ ,$$

```
Algorithm 1: Bisection Search for CompRCMDP
Input: Iteration length $K$, search space bounds $(\ell_0, u_0)$, MDP set $\{\mathcal{M}_n\}_{n \in [\![0,N]\!]}$, and penalty
       function $\{\psi_n : \mathcal{M} \to \mathbb{R}\}_{n \in [\![0,N]\!]}$
For $j \in \mathbb{R}$, let $F_j(\boldsymbol{\pi}) := \max_{n \in [\![0,N]\!]} \max_{M_n \in \mathcal{M}_n} J_{M_n}(\boldsymbol{\pi}) - \psi_n(M) - j\mathbb{1}[n = 0]$
for $k = 0, \dots, K-1$ do
    Compute a policy $\boldsymbol{\pi}_k$ such that (see Proposition 3)
        $F_{j_k}(\boldsymbol{\pi}_k) \le \min_{\boldsymbol{\pi} \in \Pi} F_{j_k}(\boldsymbol{\pi}) + \varepsilon$  where  $j_k := (\ell_k + u_k)/2$
    Shrink the search space by
        $\ell_{k+1} := \begin{cases} j_k & \text{if } F_{j_k}(\boldsymbol{\pi}_k) > 0 \\ \ell_k & \text{otherwise} \end{cases}$  and  $u_{k+1} := \begin{cases} u_k & \text{if } F_{j_k}(\boldsymbol{\pi}_k) > 0 \\ j_k & \text{otherwise} \end{cases}$
return $\boldsymbol{\pi} \in \arg\min_{\boldsymbol{\pi} \in \Pi} F_{j_K}(\boldsymbol{\pi})$
```

where I is the identity matrix, $P^{\boldsymbol{\pi}} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ and $\boldsymbol{c}^{\boldsymbol{\pi}} \in \mathbb{R}^{|\mathcal{S}|}$ are the probability matrix and vector such that $P^{\boldsymbol{\pi}}(s' \mid s) = \sum_{a \in \mathcal{A}} \boldsymbol{\pi}(a \mid s) P(s' \mid s, a)$ and $\boldsymbol{c}^{\boldsymbol{\pi}}(s) = \sum_{a \in \mathcal{A}} \boldsymbol{\pi}(a \mid s) \boldsymbol{c}(s, a)$. Note that $J_{M_{\boldsymbol{\mu}}}(\boldsymbol{\pi})$ is continuous in $\boldsymbol{\mu}$ since for any $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathscr{P}(\mathcal{S})$,

$$\left| J_{M_{\boldsymbol{\mu}_1}}(\boldsymbol{\pi}) - J_{M_{\boldsymbol{\mu}_2}}(\boldsymbol{\pi}) \right| = \left| (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top (I - \gamma P^{\boldsymbol{\pi}})^{-1} \boldsymbol{c}^{\boldsymbol{\pi}} \right| \le \frac{\|\boldsymbol{c}\|_\infty}{1 - \gamma} \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_1 \,.$$

Combined with Assumption 3, for any $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathscr{P}(\mathcal{S})$,

$$\left| \left(J_{M_{\boldsymbol{\mu}_1}}(\boldsymbol{\pi}) - \psi(M_{\boldsymbol{\mu}_1}) \right) - \left(J_{M_{\boldsymbol{\mu}_2}}(\boldsymbol{\pi}) - \psi(M_{\boldsymbol{\mu}_2}) \right) \right| \le \left(\frac{\|\boldsymbol{c}\|_\infty}{1 - \gamma} + \ell_\psi \right) \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_1 \,. \tag{32}$$

Since $\|\boldsymbol{\mu}_M - \widetilde{\boldsymbol{\mu}}_M\|_1 \le 2\varepsilon_{\boldsymbol{\mu}}$, it holds that for any $\boldsymbol{\pi} \in \Pi$,

$$\left| F(\boldsymbol{\pi}) - \widetilde{F}(\boldsymbol{\pi}) \right| \le 2 \left(\frac{c_{\max}}{1 - \gamma_{\max}} + \ell_\psi \right) \varepsilon_{\boldsymbol{\mu}} \,,$$

where we used (32) with $|\max_i x_i - \max_i y_i| \le \max_i |x_i - y_i|$ for real numbers $\{x_i\}$ and $\{y_i\}$. The claim holds by the following inequality:

$$\begin{aligned} F(\widetilde{\boldsymbol{\pi}}) - \min_{\boldsymbol{\pi} \in \Pi} F(\boldsymbol{\pi}) &= \widetilde{F}(\widetilde{\boldsymbol{\pi}}) + \left(F(\widetilde{\boldsymbol{\pi}}) - \widetilde{F}(\widetilde{\boldsymbol{\pi}}) \right) - \min_{\boldsymbol{\pi} \in \Pi} \left(\widetilde{F}(\boldsymbol{\pi}) + \left(F(\boldsymbol{\pi}) - \widetilde{F}(\boldsymbol{\pi}) \right) \right) \\ &\le \widetilde{F}(\widetilde{\boldsymbol{\pi}}) - \min_{\boldsymbol{\pi} \in \Pi} \widetilde{F}(\boldsymbol{\pi}) + 4 \left(\frac{c_{\max}}{1 - \gamma_{\max}} + \ell_\psi \right) \varepsilon_{\boldsymbol{\mu}} \,. \end{aligned}$$

D General Optimization Form for Composite Problems

Beyond the composite problem example discussed in Section 4.3, this section presents the more general form of the CompRMDP framework that encompasses all composite problems expressible within it. Let N be the number of constraints. Define $N + 1$ general MDP sets $\{\mathcal{M}_0, \dots, \mathcal{M}_N\}$ and $N + 1$ penalty functions $\{\psi_0, \dots, \psi_N\}$. Then, we define the following **Composite Robust Constrained MDP (CompRCMDP)** problem:

$$\begin{aligned} \textbf{(CompRCMDP)} \quad & j^\star = \min_{\boldsymbol{\pi} \in \Pi} \max_{M_0 \in \mathcal{M}_0} J_{M_0}(\boldsymbol{\pi}) - \psi_0(M_0) \\ \text{such that} \quad & \max_{M_n \in \mathcal{M}_n} J_{M_n}(\boldsymbol{\pi}) - \psi_n(M_n) \le 0 \quad \forall n \in [\![1, N]\!] \,. \end{aligned} \tag{33}$$

We assume that (33) is feasible and we denote $j^\star$ as the optimal objective value. It is easy to see that (33) generalizes all the CompRMDP instances in Section 4 and their combinations.

By reformulating (33) into an epigraph form and building on the discussion in Section 4.2, it is easy to see that the following bisection search will find a near-optimal policy for the CompRCMDP:

$$\begin{aligned}&\text{Increase } j \text{ if } \min_{\boldsymbol{\pi}\in\Pi} F_j(\boldsymbol{\pi}) > 0 \ \text{ and } \ \text{decrease } j \text{ otherwise,}\\&\text{where } F_j(\boldsymbol{\pi}) \coloneqq \max_{n\in[\![0,N]\!]} \max_{M_n\in\mathcal{M}_n} J_{M_n}(\boldsymbol{\pi}) - \psi_n(M) - j\mathbb{1}[n=0] \ .\end{aligned} \tag{34}$$

We summarize the bisection search algorithm in Algorithm 1. For convenience, we call the subproblem $\min_{\boldsymbol{\pi}\in\Pi} F_j(\boldsymbol{\pi})$ during the search (34) the **CompRCMDP-feasibility** problem. This subproblem can be represented as a CompRMDP instance.

Proposition 3 (CompRCMDP-feasibility $\subset$ CompRMDP)**.** *Let $\mathcal{M}_0'$ be an MDP set that adds $j\mathbf{1}$ to all the cost functions in $\mathcal{M}_0$. Define $\mathcal{M} = \mathcal{M}_0' \cup \mathcal{M}_1 \cup \cdots \cup \mathcal{M}_N$ and $\psi : M \in \mathcal{M} \mapsto \sum_{n\in[\![0,N]\!]} \psi_n(M)\mathbb{1}[M \in \mathcal{M}_n]$. Then, by Theorem 2 with these $\mathcal{M}$ and ψ, applying the update (7) for $\mathcal{O}(\varepsilon^{-4})$ iterations yields a policy $\boldsymbol{\pi}_\varepsilon$ satisfying*

$$F_j(\boldsymbol{\pi}_\varepsilon) \le \min_{\boldsymbol{\pi}\in\Pi} F_j(\boldsymbol{\pi}) + \varepsilon \ .$$

The following theorem guarantees that by running the CompRMDP algorithm **logarithmic number of times**, we can find a near-optimal policy for the CompRCMDP.

Theorem 5. *Set ℓ_0 and u_0 such that $\ell_0 \le J_M(\boldsymbol{\pi}) - \psi_n(M) \le u_0$ for all $M \in \mathcal{M}_n$, $n \in [\![0,N]\!]$ and $\boldsymbol{\pi} \in \Pi$. Then, Algorithm 1 returns a policy π_K satisfying*

$$\begin{aligned}&\max_{M_0\in\mathcal{M}_0} J_{M_0}(\boldsymbol{\pi}_K) - \psi_0(M_0) \le j^\star + \varepsilon + (u_0-\ell_0)2^{-K}\\ \textit{and}\quad &\max_{M_n\in\mathcal{M}_n} J_{M_n}(\boldsymbol{\pi}_K) - \psi_n(M_n) \le \varepsilon + (u_0-\ell_0)2^{-K} \quad \forall n\in[\![1,N]\!] \ .\end{aligned}$$