

Initial Distribution Sensitivity of Constrained Markov Decision Processes

Alperen Tercan and Necmiye Ozay

Abstract—Constrained Markov Decision Processes (CMDPs) are notably more complex to solve than standard MDPs due to the absence of universally optimal policies across all initial state distributions. This necessitates re-solving the CMDP whenever the initial distribution changes. In this work, we analyze how the optimal value of CMDPs varies with different initial distributions, deriving bounds on these variations using duality analysis of CMDPs and perturbation analysis in linear programming. Moreover, we show how such bounds can be used to analyze the regret of a given policy due to unknown variations of the initial distribution.

I. INTRODUCTION

Constrained Markov Decision Processes (CMDPs) are a fundamental framework for modeling sequential decision-making problems where an agent seeks to maximize cumulative rewards while satisfying certain constraints. These processes extend the classical Markov Decision Processes (MDPs) [11] by incorporating constraints on expected costs or resources, making them highly relevant in fields such as operations research, robotics, telecommunications, and finance [1]. In many real-world applications, agents must not only optimize performance but also adhere to limitations like energy consumption, risk levels, or time budgets [8].

A significant challenge in CMDPs is the lack of uniformly optimal policies across all possible initial state distributions. Unlike standard MDPs, where a stationary policy is optimal regardless of the starting state due to the Markov property and the structure of the optimization problem [5], the presence of constraints means that a policy optimal for one initial distribution may not be optimal for another. This issue becomes particularly problematic in environments where the initial state distribution is uncertain or varies over time, leading to potential suboptimal performance if the policy is not adapted accordingly [7].

Traditional approaches often involve solving the CMDP for a specific nominal initial distribution, yielding a policy tailored to that scenario [1]. However, when the actual initial distribution deviates from the nominal one, the performance of this policy can degrade, and re-solving the CMDP at deployment time for the new distribution can be computationally intensive [9]. This is especially impractical in dynamic environments where quick decision-making is crucial. While the continuity and stability of CMDPs under these deviations have been studied [3], [2], there is currently

a lack of numerical methods for quantifying or analyzing the resulting performance impact.

One could instead consider searching for robust policies for uncertain CMDPs, where the uncertainty can be in the initial condition or transition probabilities. However, as opposed to MDPs, robust CMDPs are non-trivial to solve, leading to bilinear optimization problems [13] or heuristic algorithms with local convergence guarantees [15], [12], [14]. We note that when uncertainty is limited to the initial distribution, an optimal policy robust for all possible initial distributions can be found within the class of non-Markovian policies—specifically, by conditioning the policy on the initial state. However, this requires solving a separate problem for each initial state, which becomes intractable in large state spaces.

In this paper, we address this problem by presenting three different methods to derive upper bounds on the expected value function under any arbitrary initial distribution, using the solution obtained from a nominal distribution. Our approach can enable practitioners to estimate the potential performance loss without the need to recompute the optimal policy for each new initial distribution. These upper bounds act as a diagnostic tool to determine whether the existing policy remains sufficiently effective or if re-optimization is warranted under the new conditions. We further show how these bounds can be used to compute sets of initial conditions for which the nominal policy provides a robust solution.

II. PRELIMINARIES

In this section, we introduce the necessary background for our analysis.

A. Constrained Markov Decision Process (CMDP)

A CMDP is defined by the tuple $\mathcal{M}_c = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \mathbf{c}, \tau, \gamma, \beta)$. Here, $\mathcal{S}$ denotes a finite set of states, and $\mathcal{A}$ represents a finite set of actions. The transition probability function $\mathcal{P} : \mathcal{S} \times \mathcal{A} \rightarrow \Delta_{\mathcal{S}}$ ¹ gives the probability $\mathcal{P}(s'|s, a)$ of reaching state s' from state s after action a . The reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ defines the immediate reward $r(s, a)$ received upon taking action a in state s . The vector-valued constraint function $\mathbf{c} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^K$ defines utilities associated with state-action pairs for each of constraints $i = 1, \dots, K$. Similarly, $\tau \in \mathbb{R}^K$ is the constraint thresholds. The discount factor $\gamma \in [0, 1)$ adjusts the weight of future rewards, while $\beta \in \Delta_{\mathcal{S}}$ is the initial distribution over states, with $\beta(s)$ as the probability of starting in state s .

¹ $\Delta_{\mathcal{S}}$ is the probability simplex over $\mathcal{S}$.

This work was supported in part by ONR CLEVR-AI MURI (#N00014-21-1-2431). The authors are with the Electrical Engineering and Computer Science Department at the University of Michigan, Ann Arbor. Emails: {tercan, necmiye}@umich.edu

For a given policy π and an initial distribution β , the expected discounted cumulative sum for any function $f : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is defined as:

$$V_f^\pi(\beta) = \mathbb{E}_{\substack{A_t \sim \pi \\ S_0 \sim \beta}} \left[\sum_{t=0}^{\infty} \gamma^t f(S_t, A_t) \right].$$

This function is linear in β , since it can be written as $\beta^\top V_f^\pi$ where i^{th} entry of $V_f^\pi \in \mathbb{R}^{|\mathcal{S}|}$ defined as:

$$(V_f^\pi)_i = \mathbb{E}_{A_t \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t f(S_t, A_t) \mid S_0 = s_i \right]. \quad (1)$$

The goal in a CMDP is to find a policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ that maximizes the expected discounted cumulative reward while satisfying constraints on the cumulative utility. Formally, the optimization problem is:

$$\begin{aligned} \max_{\pi} \quad & V_r^\pi(\beta) \\ \text{subject to} \quad & V_{c_i}^\pi(\beta) \geq \tau_i, \quad i \in [K], \end{aligned} \quad (2)$$

where $[K]$ denotes the set $\{1, \dots, K\}$. The optimal value of this problem, as a function of the initial distribution β , and constraint threshold vector τ , is denoted by $V^*(\beta, \tau)$. When τ is clear from the context, we simply write $V^*(\beta)$.

For a given policy π , it is useful to have a measure of how much worse it is compared to the optimal policy. *Regret* has been used as the difference between the value of a policy and the optimal policy. In a CMDP setting, we define a generalized notion of regret that also allows quantifying regret for infeasible policies.

Definition 1: A policy π for the CMDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \mathbf{c}, \tau, \gamma, \beta)$ is said to incur (δ, ϵ) -regret if it satisfies the following conditions:

$$\begin{aligned} \tau_i - V_{c_i}^\pi(\beta) &\leq \delta_i \quad \forall i \in [K], & (\delta\text{-Violation}) \\ V^*(\beta, \tau - \delta) - V_r^\pi(\beta) &\leq \epsilon. & (\epsilon\text{-Regret}) \end{aligned}$$

We define the minimal regret of a policy π as the pair (δ^*, ϵ^*) satisfying $(\delta^*, \epsilon^*) \leq (\delta, \epsilon)$ for every (δ, ϵ) such that π incurs (δ, ϵ) -regret. Here, $(\delta, \epsilon) \leq (\delta', \epsilon')$ means that $\delta_i \leq \delta'_i$ for all i and $\epsilon \leq \epsilon'$. Although this partial order is not total, a minimum always exists: each component of δ can be minimized independently under $(\delta\text{-Violation})$, and the minimal ϵ satisfying $(\epsilon\text{-Regret})$ is non-decreasing in each component of δ . This follows from the fact that $V^*(\beta, \tau - \delta)$ depends on δ in a non-decreasing manner—relaxing any constraint does not reduce the optimal value. Moreover, (δ^*, ϵ^*) is the unique solution to the conditions in Definition 1 when the inequalities are replaced with equalities.

This definition allows π to violate the constraints up to δ , but effectively “penalizes” such violations by using $V^*(\beta, \tau - \delta)$ instead of $V^*(\beta, \tau)$ in Equation $(\epsilon\text{-Regret})$ as higher δ values increase the overall regret. In this paper, we focus on the set of initial distributions that are feasible under π ; hence, we consider only $(0, \epsilon)$ -regret policies for ease of exposition. However, we will remark on the extensions for the non-zero δ case whenever appropriate.

B. Linear Program (LP) formulation of CMDPs

In addition to the policy-space formulation in Equation (2), a CMDP can be expressed as an LP in terms of occupation-measures $\rho \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ as follows:

$$\begin{aligned} \max_{\rho \geq 0} \quad & r^\top \rho \\ \text{s.t.} \quad & C\rho \geq \tau, \\ & \Psi\rho = \beta; \end{aligned} \quad (\text{P})$$

where:

$$\begin{aligned} C &= [c^{a_1} \mid c^{a_2} \mid \dots \mid c^{a_{|\mathcal{A}|}}], \\ \Psi &= [I - \gamma T^{a_1} \mid I - \gamma T^{a_2} \mid \dots \mid I - \gamma T^{a_{|\mathcal{A}|}}], \end{aligned}$$

with $(c^{a_k})_{i,j} = c_i(s_j, a_k)$ and $T_{ij}^{a_k} = \mathcal{P}(s_i | s_j, a_k)$.

Once an optimal occupation measure ρ^* for Problem (P) is computed, a policy π inducing ρ^* can be obtained by setting $\pi(a|s) = \rho^*(s, a) / \sum_{a' \in \mathcal{A}} \rho^*(s, a')$. For states with $\sum_{a' \in \mathcal{A}} \rho^*(s, a') = 0$, π can be chosen arbitrarily as such states would be unreachable.

We can also write the dual LP corresponding to Problem (P) as:

$$\begin{aligned} \min_{W, \lambda \geq 0} \quad & \beta^\top W - \lambda^\top \tau \\ \text{s.t.} \quad & \Psi^\top W \geq r + \lambda^\top C; \end{aligned} \quad (\text{D})$$

where $W \in \mathbb{R}^{|\mathcal{S}|}$ and $\lambda \in \mathbb{R}^K$. We denote an optimal solution corresponding to the problem instance with β as $W^*(\beta)$ and $\lambda^*(\beta)$.

III. OBTAINING BOUNDS FOR $V^*(\beta_1)$

In this section, we describe three approaches for obtaining bounds on $V^*(\beta_1)$ without explicitly solving an optimization problem for the initial distribution β_1 .

A. LP Dual Feasibility Bound

Our first bound uses the fact that the solution of Problem (D) at a given nominal initial distribution can be used to bound the optimal performance at different distributions.

Theorem 1: Let β_0 be the nominal initial distribution the CMDP is solved for; with an optimal dual solution $\lambda^*(\beta_0)$ and $W^*(\beta_0)$. For any initial distribution β_1 , the value $V^*(\beta_1)$ satisfies:

$$V^*(\beta_1) \leq \beta_1^\top W^*(\beta_0) - \tau^\top \lambda^*(\beta_0). \quad (3)$$

Proof: As the pair $\lambda^*(\beta_0)$ and $W^*(\beta_0)$ is an optimal solution of Problem (D) for β_0 , it is still feasible when the Problem (D) is solved for β_1 . This is because β only appears in the cost function of Problem (D) but does not impact its feasible set. Hence, $\lambda^*(\beta_0)$ and $W^*(\beta_0)$ are feasible for any initial distribution β_1 . As Problem (D) is a minimization problem, the value of its objective function at any feasible solution would yield an upper bound on its optimal value, $V^*(\beta_1)$. Then, by plugging in $\lambda^*(\beta_0)$ and $W^*(\beta_0)$ into the objective function of Problem (D), we obtain the desired result. ■

Remark 1: To analyze (δ, ϵ) -regret policies, it is necessary to bound $V^*(\beta_1, \tau - \delta)$. The bound derived in this section extends naturally to bound $V^*(\beta_1, \tau - \delta)$. Let $\lambda^*(\beta_0)$ and $W^*(\beta_0)$ be the optimal solution of dual problem of $\mathcal{M}_c = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \mathbf{c}, \tau, \gamma, \beta_0)$. Then, for any initial distribution β_1 and violation δ , the value $V^*(\beta_1, \tau - \delta)$ satisfies:

$$V^*(\beta_1, \tau - \delta) \leq \beta_1^\top W^*(\beta_0) - (\tau - \delta)^\top \lambda^*(\beta_0). \quad (4)$$

B. Bound from Perturbation of LP

An alternative bound can be obtained by treating different initial distributions as perturbations from the nominal initial distribution. Since this perturbation corresponds to a perturbation of the right-hand side vector of equality constraint in Problem (P), we can obtain a bound based on perturbation analysis of LPs. In particular, we adapt the method in [10], summarized with the following theorem.

Theorem 2: (from [10]) Consider Problem (P). Let ρ^* and $[W^*; \lambda^*] \in \mathbb{R}^{|\mathcal{S}|+K}$ to be the optimal solutions of Problem (P) and (D), respectively. Then, the change in optimal solutions, $\Delta\rho^* = \rho_{\text{new}}^* - \rho^*$ and $[\Delta W^*; \Delta\lambda^*] = [W_{\text{new}}^* - W^*; \lambda_{\text{new}}^* - \lambda^*]$, under the new initial distribution $\Delta\beta + \beta$ satisfies the following for any $\Delta\beta$:

$$\sqrt{\|\Delta\rho^*\|^2 + \|[\Delta W^*; \Delta\lambda^*]\|^2} \leq \|R^{-1}\| \|\Delta\beta\|, \quad (5)$$

where $\|\cdot\|$ is the 2-norm for vectors and the spectral norm for matrices. R is defined as:

$$R = \begin{bmatrix} R_1 & 0 & 0 \\ 0 & R_1^\top & 0 \\ 0 & R_2^\top & I \end{bmatrix},$$

where R_1, R_2 constitute a partition of columns of the matrix $\begin{bmatrix} C & -I \\ \Psi & 0 \end{bmatrix}$ such that i^{th} column is in R_1 if $i \leq |\mathcal{S}||\mathcal{A}|$ and $\rho_i^* > 0$, or $i > |\mathcal{S}||\mathcal{A}|$ and $c_j^\top \rho > \tau_j$ for $j = i - |\mathcal{S}||\mathcal{A}|$.

Using Theorem 2, we can prove the following value bound:

Theorem 3: Let β_0 and $V^*(\beta_0)$ be a nominal initial distribution and its optimal value. Then, any initial distribution β_1 satisfies:

$$V^*(\beta_1) \leq V^*(\beta_0) + \|\beta_0 - \beta_1\| \min(\|[\beta_1; -\tau]\| \|R^{-1}\| + \|W_0^*\|, \|R^{-1}\| \|r\|). \quad (6)$$

Proof: We start by noting that Equation (5) gives us the upper bound on the change in the solution but not value. To obtain the bound on value change, we start by setting $\beta_1 = \beta_0 + \Delta\beta$ and $\rho_1^* = \rho_0^* + \Delta\rho$. Then, we have

$$\begin{aligned} |V^*(\beta_1) - V^*(\beta_0)| &= |r^\top \rho_1^* - r^\top \rho_0^*| \\ &\stackrel{(a)}{\leq} \|r\| \|\Delta\rho\| \\ &\stackrel{(b)}{\leq} \|r\| \|R^{-1}\| \|\Delta\beta\| \\ &\leq \|r\| \|R^{-1}\| \|\beta_1 - \beta_0\|, \end{aligned}$$

where inequalities (a) and (b) follow from Cauchy-Schwarz and Equation (5), respectively. Writing the value of the LP

using its dual gives us another bound:

$$\begin{aligned} |V^*(\beta_1) - V^*(\beta_0)| &= |(\beta_1^\top W_1^* - \tau^\top \lambda_1^*) - (\beta_0^\top W_0^* - \tau^\top \lambda_0^*)| \\ &= |\beta_1^\top (W_0^* + \Delta W^*) - \beta_0^\top W_0^* - \tau^\top \Delta\lambda^*| \\ &= |(\beta_1^\top \Delta W^* - \tau^\top \Delta\lambda^*) + \Delta\beta^\top W_0^*| \\ &\stackrel{(a)}{\leq} \|[\beta_1; -\tau]\| \|[\Delta W^*; \Delta\lambda^*]\| + \|\Delta\beta\| \|W_0^*\| \\ &\stackrel{(b)}{\leq} \|[\beta_1; -\tau]\| \|R^{-1}\| \|\Delta\beta\| + \|\Delta\beta\| \|W_0^*\| \\ &= (\|[\beta_1; -\tau]\| \|R^{-1}\| + \|W_0^*\|) \|\beta_1 - \beta_0\|. \end{aligned}$$

Here inequalities (a) and (b) again follow from Cauchy-Schwarz and Equation (5), respectively. To get a tighter upper bound, we compute both of these bounds and take the minimum. This gives us the upper bound in Equation (6). ■

Since the bounds we proved are on $|V^*(\beta_1) - V^*(\beta_0)|$, a lower bound can also be trivially obtained:

$$V^*(\beta_1) \geq V^*(\beta_0) - \|\beta_0 - \beta_1\| \min(\|[\beta_1; -\tau]\| \|R^{-1}\| + \|W_0^*\|, \|R^{-1}\| \|r\|). \quad (7)$$

Remark 2: Theorem 2 can be generalized to obtain following upper bound to $V^*(\beta_1, \tau - \delta)$ for any β_1 initial distribution and δ violation:

$$V^*(\beta_1, \tau - \delta) \leq V^*(\beta_0, \tau) + \phi \min(\|[\beta_1; -\tau]\| \|R^{-1}\| + \|W_0^*\|, \|R^{-1}\| \|r\|), \quad (8)$$

where $\phi = \sqrt{\|\delta\|^2 + \|\Delta\beta\|^2}$ and other quantities are as defined in Theorem 2.

C. Bound from Concavity of LP Values

Finally, we propose an alternative bound based on a well-known concavity result for linear program (LP) values over the set of consistent right-hand side vectors. Specifically, focusing on Problem (P), we define the feasible set of ρ as a function of β :

$$\text{Feas}_\rho(\beta) = \{\rho \geq 0 \mid C\rho \geq \tau, \Psi\rho = \beta\}.$$

We then define the set of consistent β values as

$$\text{Feas}_\beta = \{\beta \mid \exists \rho \text{ s.t. } \text{Feas}_\rho(\beta) \neq \emptyset\}.$$

Equivalently, this set can be expressed as

$$\text{Feas}_\beta = \{\Psi\rho \mid C\rho \geq \tau, \rho \geq 0\},$$

which is convex, since it is the image of a convex set under a linear transformation. For any $\beta \in \text{Feas}_\beta$, define the function

$$F(\beta) = \max_{\rho \in \text{Feas}_\rho(\beta)} r^\top \rho,$$

which represents the optimal value of Problem (P) as a function of β . A direct corollary of Theorem 5.1 in [6] is:

Theorem 4: The function $F(\beta)$ is concave on the domain Feas_β .

Utilizing this result, we obtain an upper bound for the value of the CMDP under an arbitrary initial distribution β_1 .

Theorem 5: Let β_1 be an initial distribution over states, yielding CMDP value $V^*(\beta_1)$. Moreover, let the cardinality

of state space $|\mathcal{S}| = n$ and $\alpha = \max_j \beta_1(j)$ to be largest element of β_1 . Then,

$$V^*(\beta_1) \leq \alpha n V^*(v) - \sum_i (\alpha - \beta_1(i)) V^*(\delta_i), \quad (9)$$

where v is the uniform distribution over state space and $\delta_i(j)$ is non-zero only at $j = i$.

Proof: Using Theorem 4 and definition of concavity, for all λ such that $\lambda_i \geq 0$ for all i and $\sum_i \lambda_i = 1$:

$$V^*\left(\sum_{i=0}^n \lambda_i x_i\right) \geq \sum_{i=0}^n \lambda_i V^*(x_i). \quad (10)$$

Pick $\lambda_0 = \frac{1}{n\alpha}$ and $\lambda_i = \frac{\alpha - \beta_1(i)}{n\alpha}$ for $i > 0$. Then, choose $x_0 = \beta_1$ and $x_i = \delta_i$ for $i > 0$. Observe that:

$$\begin{aligned} \sum_{i=0}^n \lambda_i &= \frac{1}{n\alpha} + \sum_{i=1}^n \frac{\alpha - \beta_1(i)}{n\alpha} \\ &= \frac{1}{n\alpha} + \frac{1}{n\alpha} (n\alpha - \sum_{i=1}^n \beta_1(i)) \\ &= \frac{1}{n\alpha} (1 + n\alpha - 1) = 1. \end{aligned}$$

Moreover, for all i , $\alpha \geq \beta_1(i) \geq 0$, which implies $\lambda_i \geq 0$. Hence, they yield a convex combination of x_i s. Finally, observe that:

$$\begin{aligned} \sum_{i=0}^n \lambda_i x_i &= \frac{1}{n\alpha} \beta_1 + \sum_i \frac{1}{n} \delta_i - \frac{1}{n\alpha} \sum_i \beta_1(i) \delta_i \\ &= v + \frac{1}{n\alpha} \beta_1 - \frac{1}{n\alpha} \beta_1 = v. \end{aligned}$$

Thus, substituting in Equation (11):

$$V^*(v) \geq \frac{1}{n\alpha} V^*(\beta_1) + \sum_{i=0}^n \frac{\alpha - \beta_1(i)}{n\alpha} V^*(\delta_i). \quad (11)$$

Multiplying both sides with $n\alpha$ and reorganizing terms yields Equation (9). ■

Note that, choosing $x_i = \delta_i$, $\lambda_i = \beta_1(i)$ for $n \geq i \geq 1$, and $\lambda_0 = 0$ in Equation (11), we can also obtain the following lower bound on the value under β_1 :

$$V^*(\beta_1) \geq \sum_i \beta_1(i) V^*(\delta_i). \quad (12)$$

D. Comparison of Bounds

In this section, we discuss advantages and drawbacks of each bound, as summarized in Table I.

Computing the duality-based and perturbation bounds relies on solving the CMDP under the nominal initial distribution. Specifically, the duality-based bound depends on the solution to Problem (D), while the perturbation bound requires identifying the zero and strictly positive components of the optimal solution to Problem (P). In contrast, calculating the concavity bound involves solving the CMDP for $|\mathcal{S}| + 1$ distinct initial distributions. This effectively requires determining the optimal policy for every initial state, rendering the computation both intractable and impractical.

Method	Complexity	Empirical Tightness	Linear	Lower Bound
Duality	1 CMDP	✓	Yes	No
Perturbation	1 CMDP	–	No	Yes
Concavity	$(\mathcal{S} + 1)$ CMDPs	×	No	Yes

TABLE I: Comparison of the three bounding methods. The second column indicates how many CMDP instances must be solved to compute each bound. The third column shows each method's overall empirical tightness, i.e., how close the bound is to the actual least upper bound (✓: tightest, ×: least tight, and –: moderately tight). The fourth column specifies if the bound is linear in the initial distribution β . Finally, the last column indicates if the method also provides a lower bound.

While it may be possible to use over-approximations of the optimal values to simplify the process, such an approach often leads to trivial bounds, as the concavity bound is already the loosest among the three.

The obtained bounds are not easy to compare analytically. Hence, in Section V, we compare them numerically by computing the relative gap between the computed upper bound and true value of $V^*(\beta_1)$. We found that the duality-based bounds are the smallest upper bounds and concavity yields the largest bounds. Perturbation bounds are within two order of magnitude from duality-based bounds.

Another important difference between the proposed approaches is whether they provide also a lower bound to $V^*(\beta)$. Both perturbation and concavity-based approaches provide lower bounds as explained in their respective sections. However, duality-based approach does not yield a lower bound. When used in Equation (ϵ -Regret), a lower bound provides an outer-approximation of the true ϵ -regret set, in contrast to inner-approximation obtained by the upper bounds.

Moreover, we compare the geometry of the obtained bounds. Ideally, we would want the bound to be linear in β_1 , reducing Equation (ϵ -Regret) to a halfspace. From the three bounds, only the duality-based bound satisfies this. Perturbation bound involves two-norm of β_1 . Similarly, the concavity bound depends on $\|\beta_1\|_\infty$.

Based on this comparison, we conclude that unless a lower bound is needed, duality-based provides the best trade-off. When a lower bound is needed, perturbation bound provides computational efficiency. The concavity approach is the worst one, both computationally intractable for large state spaces and also cannot provide tight bounds.

Finally, we note an alternative use of these bounds. As discussed in Remark 1 and Remark 2, these bounds can be generalized to provide bounds on $V^*(\beta_1, \tau - \delta)$. If we are interested in sensitivity to constraint threshold instead of initial distribution, choosing β_1 to be the nominal initial distribution gives an upper bound on the optimal value under relaxed or tightened constraints.

IV. APPLICATIONS TO ROBUST REGRET PROBLEM

An important question in the study of the initial distribution sensitivity of CMDPs is determining how robust the

optimal policy under the nominal initial distribution is to changes in the initial distribution. In this section, we show how the bounds obtained in Section III-A can be used to solve some problems related to this question.

Problem 1: Given policy π and regret ϵ , find the set of initial distributions $B \subseteq \Delta_{|S|}$ s.t. π has $(0, \epsilon)$ regret.

Note that for a fixed policy π , $V_{c_i}^\pi(\beta)$ can be expressed as $\beta^\top V_{c_i}^\pi$ where $V_{c_i}^\pi$ is defined as in Equation (1). Hence, each of the feasibility constraints is linear in β and form a half-space in $\mathbb{R}^{|S|}$. However, $V^*(\beta)$ term in Equation (ϵ -Regret) is a hard-to-compute, nonlinear function of β as finding the optimal value requires solving CMDP under initial distribution β ; rendering Problem 1 intractable.

To deal with this issue, we propose computing inner approximations of the true $(0, \epsilon)$ -regret set by replacing $V^*(\beta)$ with an upper bound. If this bound is linear in β , this approximation becomes a halfspace constraint as well.

Proposition 1: The following polytope is an inner-approximation to the set defined in Problem 1:

$$\begin{bmatrix} (V_r^\pi - W^*(\beta_0))^\top \\ (V_{c_1}^\pi)^\top \\ \vdots \\ (V_{c_K}^\pi)^\top \end{bmatrix} \beta \geq \begin{bmatrix} -\tau^\top \lambda^*(\beta_0) - \epsilon \\ \tau_1 \\ \vdots \\ \tau_K \end{bmatrix}, \quad (13)$$

where $W^*(\beta_0)$ and $\lambda^*(\beta_0)$ are the optimal solution of Problem (D) for a nominal distribution β_0 .

While Problem 1 helps finding the set of initial distributions that satisfy some predetermined robustness conditions, it cannot directly tell how robust a predetermined set of initial distributions is. We formulate this problem as below.

Problem 2: Given policy π and a set $B \subseteq \Delta_{|S|}$ of initial distributions feasible under π , find minimum ϵ such that π has $(0, \epsilon)$ regret for all $\beta \in B$. More formally, we solve

$$\begin{aligned} \min \quad & \epsilon \\ \text{s.t.} \quad & V^*(\beta) - V_r^\pi(\beta) \leq \epsilon \quad \forall \beta \in B. \end{aligned} \quad (14)$$

As formulated in Equation (14), Problem 2 involves an infinite number of constraints and $V^*(\beta)$ is highly nonlinear and makes the problem hard to solve. Using the upper bound in Equation (3), we can approximate the constraints with a linear constraint. Substituting $V^*(\beta)$ with its upper bound in the constraints yields:

$$\beta^\top W^*(\beta_0) - \tau^\top \lambda^*(\beta_0) - V_r^\pi(\beta) \leq \epsilon \quad \forall \beta \in B,$$

where β_0 is the nominal distribution used in the upper bound. Reorganizing the terms, we obtain:

$$\beta^\top (W^*(\beta_0) - V_r^\pi) \leq \tau^\top \lambda^*(\beta_0) + \epsilon \quad \forall \beta \in B. \quad (15)$$

Note that Equation (15) implies the original constraint in Problem 2, representing a stricter condition. Therefore, the problems presented below will yield an upper bound on the true solution.

Next, we propose reformulations that eliminate the infinite constraints under two common uncertainty scenarios.

One such scenario arises when β is known to lie within the convex hull of a finite set of candidate distributions $\{\beta_1, \beta_2, \dots, \beta_L\}$; that is, $B = \text{conv}\{\beta_1, \beta_2, \dots, \beta_L\}$.

Proposition 2: If $B = \text{conv}\{\beta_1, \beta_2, \dots, \beta_L\}$, then the following optimization problem provides an upper bound to true solution of Problem 2:

$$\begin{aligned} \min \quad & \epsilon \\ \text{s.t.} \quad & \beta_i^\top (W^*(\beta_0) - V_r^\pi) \leq \tau^\top \lambda^*(\beta_0) + \epsilon \quad i \in [L]. \end{aligned} \quad (16)$$

Proof: This result follows from the fact that the constraint is linear in β . Therefore, if it holds for all extreme points of the convex hull (i.e., the β_i), it holds for any convex combination of them. ■

One might argue that when L is sufficiently small, Problem 2 could be solved directly by evaluating the constraint at each β_i , without resorting to an upper bound. However, due to Theorem 4, the original constraint is concave in β , and thus verifying it only at the vertices is insufficient to guarantee feasibility over the entire set.

Next, we consider the case where representing B via its vertices is not preferable. For example, Wasserstein-1 balls and box constraints often have too many vertices, making enumeration impractical. In such cases, a reformulation that avoids vertex enumeration is preferred. To achieve this, we leverage the results from [4], which provide equivalent, tractable reformulations for the robust counterparts of many common nonlinear constraints.

Proposition 3: The problem below finds an upper bound to true solution of Problem 2 when $B = \{\beta \mid H\beta \leq h\} \subseteq \Delta_S$:

$$\begin{aligned} \min_{z, \epsilon} \quad & \epsilon \\ \text{s.t.} \quad & h^\top z - \tau^\top \lambda^*(\beta_0) \leq \epsilon, \\ & H^\top z = W^*(\beta_0) - V_r^\pi, \\ & z \geq 0. \end{aligned} \quad (17)$$

Proof: Since Equation (15) is a linear constraint with polytopic uncertainty, the result follows directly from [4]. ■

Note that while we focus on polytopic B sets here, results from [4] allow generalizing Proposition 3 to many common uncertainty sets. Moreover, the analysis above focuses on $(0, \epsilon)$ -regret, but it naturally extends to the more general notion of (δ, ϵ) -regret, where the initial distribution π may violate the constraints by at most δ . In this setting, our bounds continue to hold with appropriate adjustments, as indicated in Remark 1. For example, Proposition 1 generalizes to the following polytope, which provides an inner approximation to the set $B \subseteq \Delta_S$ of initial distributions for which π incurs (δ, ϵ) -regret:

$$\begin{bmatrix} (V_r^\pi - W^*(\beta_0))^\top \\ (V_{c_1}^\pi)^\top \\ \vdots \\ (V_{c_K}^\pi)^\top \end{bmatrix} \beta \geq \begin{bmatrix} -(\tau - \delta)^\top \lambda^*(\beta_0) - \epsilon \\ \tau_1 - \delta \\ \vdots \\ \tau_K - \delta \end{bmatrix}. \quad (18)$$

V. EXPERIMENTS

In this section, we validate our analysis through illustrative examples.

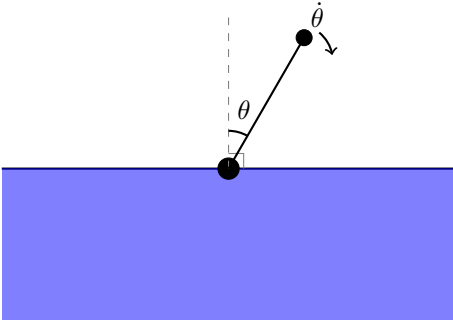


Fig. 1: Water pendulum. The objective is to stabilize the pendulum in the upright position, perpendicular to the surface, with both angular position θ and angular velocity $\dot{\theta}$ equal to zero. The blue region represents water.

A. Numerical Comparison of Bounds on Random CMDPs

Our first set of experiments evaluates the numerical tightness of the proposed bounds. To this end, we randomly generate 100 CMDPs, each with $\mathcal{S} = 100$ states, $\mathcal{A} = 3$ actions, and $K = 2$ constraints. For each CMDP, the transition probabilities, reward functions, and constraint functions are uniformly sampled, with half of the transitions pruned to avoid fully connected CMDPs.

Moreover, we randomly sample 150 initial distributions β_1 divided into 3 bins according to their total variation (TV) distance from the uniform nominal initial distribution β_0 . For each β_1 -CMDP pair, we compute the true value and the upper bounds based on β_0 . The tightness of the bounds is quantified using *relative looseness*, defined as the gap between the true value and the upper bound as percentage of the true value.

To summarize the results, we compute relative looseness percentage across all CMDPs and β_1 distributions for each distance bin. Then, we report the median of each bin for each method. This provides a comprehensive measure of how the bounds perform across diverse scenarios. The results are summarized in Table II. As expected, the larger distance between β_0 and β_1 leads to larger errors. However, notice that these are percentage errors; hence, it can be said that duality-based bound is very reliable even for furthest β_1 points.

Bounds	TV Distance		
	0.25-0.5	0.5-0.75	0.75-1
Duality	0.014	0.048	0.297
Perturbation	1.988	3.692	9.108
Concavity	9.006	18.703	61.241

TABLE II: Median relative looseness of the bounds within TV distance interval for randomly generated CMDPs. For each interval, we sort the relative looseness values of the bounds corresponding to all initial distributions and CMDPs; and report the median where a value of 0 corresponds to the upper bound being equal to the true value.

B. Simple Water Pendulum

While the experiments in the previous section useful for comparing our bounds across diverse set of CMDPs,

the random CMDPs are difficult to conceptualize. For this purpose, we use the constrained version of the inverted pendulum task, illustrated in Figure 1.

The inverted pendulum task involves balancing a pendulum upright by applying a torque input at its pivot point. The state consists of the pendulum angle $\theta \in (-\pi, \pi)$, with upright position assigned $\theta = 0$, and $\dot{\theta}$, which takes positive values clockwise. The reward function is negative quadratic in deviation from upright position and angular velocity for all $a \in \mathcal{A}$:

$$r([\theta, \dot{\theta}], a) = -(\theta^2 + \frac{1}{10}\dot{\theta}^2),$$

which is between 0 and -16.27 .

In our version, shown in Figure 1, we assume the pendulum is placed on the surface of a water body such that it is immersed in water for $\theta \geq \pi/2$ or $\theta \leq -\pi/2$. When in water, a buoyant force equal to half of gravity affects the pendulum. We assume two constraint utilities: (1) a -1 utility is incurred every step where pendulum is submerged, (2) an energy budget constraint negatively quadratic of applied torque, $-\frac{1}{4}|a|^2$ where a is the input to system, i.e applied torque. These constraints make the task more challenging and align it with the CMDP framework. For the given mass m and length l of the pendulum, the dynamics of this system is:

$$\ddot{\theta} = \begin{cases} \frac{3}{2} \left(\frac{g}{l} \sin(\theta) \right) + \frac{3}{ml^2} u, & \text{if } |\theta| \geq \frac{\pi}{2}, \\ 3 \left(\frac{g}{l} \sin(\theta) \right) + \frac{3}{ml^2} u, & \text{otherwise,} \end{cases} \quad (19)$$

where g denotes the acceleration due to gravity.

We discretize the state space by dividing the range of each state dimension into 50 bins. Similarly, the action range is divided into seven. Transition probabilities are obtained by uniformly sampling 100 points from each bin, simulating them forward under the deterministic continuous dynamics, assigning the end points to bins, and defining probabilities as the fraction of points landing in each target bin. This results in a two constraint CMDP with $|\mathcal{S}| = 2500$ and $|\mathcal{A}| = 7$. We run the experiments described in Section V-A. The results are summarized in Table III. We do not use Concavity bound in this setup as large $|\mathcal{S}|$ makes it computationally too expensive.

Bounds	TV Distance		
	0.25-0.5	0.5-0.75	0.75-1
Duality	0.000	0.001	0.004
Perturbation	0.123	0.254	0.342

TABLE III: Median relative looseness of the bounds within each TV distance interval for Water Pendulum. For each interval, we sort the relative looseness values of the bounds corresponding to all initial distributions and report the median.

C. Measuring Policy Robustness by $(0, \epsilon)$ -Regret Sets

Next, we compare the robustness of optimal policies for three initial distributions: β_T , β_B , and β_U . The distribution β_T assigns high probability mass when $\theta = 0$, β_B when $\theta =$

$\pm\pi$, and β_U is the uniform distribution. We propose the size of the $(0, \epsilon)$ -regret sets of the optimal policies corresponding to these distributions as a measure of robustness of such policies to changes in the initial distribution.

Using Proposition 1, we compute, for each initial distribution, an inner approximation of the set of β distributions for which the corresponding optimal policy π incurs at most $(0, 0.01)$ regret. As a proxy for the size of these sets, we sample 100 random distributions from the probability simplex using a Dirichlet distribution and compute the ratio of samples that fall within each polytope. This ratio, or *hit rate*, serves as a surrogate for the size.

We find the hit rates to be 0.26, 0.01, and 0.51 for β_T , β_B , and β_U , respectively. For comparison, if we evaluate the hit rates for the true $(0, 0.01)$ -regret sets instead of the inner approximations given by Proposition 1, the values are 0.33, 0.01, and 0.51. This close agreement suggests that our inner approximation is tight. Note that hit rates for the true regret sets are computed by re-solving the CMDP for each sample point and computing true regret, which might in general be prohibitive. For instance, for this example, hit rate computation for the true regret sets takes ~ 16 minutes while our method takes only 0.03 seconds.

Upon inspection, we observe that, in general, the optimal policy for β_T becomes infeasible under most other distributions, as it tends to apply excessive torque and thereby violates the second constraint. This behavior arises because, when the pole starts near the top, only minimal torque is required to maintain balance. Since the constraint is enforced in expectation, the policy optimized for β_T can afford to use large amounts of torque in other states without violating the constraint under β_T itself.

In contrast, the optimal policies for β_B and β_U are more broadly feasible across different distributions, though they tend to exhibit slightly higher regret. This is because they are more conservative in their torque usage—sometimes overly so—in order to robustly satisfy the constraint across a wider range of scenarios.

D. Finding Minimal Regret over a Set of Distributions

Based on our earlier observation that the policy induced by β_T is not $(0, 0.01)$ -regret in a large part of the simplex, we evaluate its transferability to a broader set of initial distributions. Specifically, we consider the convex hull of three vertex distributions sampled from the initial distributions that did not fall into true $(0, 0.01)$ -regret set in the previous section. We then compute the minimal regret of the policy optimal for β_T over this set, as described in Problem 2.

By using Proposition 2, we find 0.29 to be an upper bound to true minimal regret. Next, we evaluate the tightness of this upper bound by comparing it with the solution of Equation (14) by sampling the set B . Note that this approach requires exponential number of samples; hence, we keep the number of vertex distributions low. This approach yields a lower bound to the true regret as we are effectively sampling some constraints from infinite number of constraints. This approach gives 0.24, so the true minimal regret is within the range $[0.24, 0.29]$.

VI. CONCLUSION

In this work, we studied how the performance of a CMDP changes under different initial state distributions and proposed three bounding methods—based on dual feasibility, LP perturbation theory, and concavity—to estimate this variation without solving the CMDP repeatedly. Among these, the duality-based bound was consistently the tightest, while the LP perturbation approach offered both upper and lower bounds with moderate accuracy. The concavity-based method was looser and more computationally demanding. We further demonstrated how these bounds enable practical analyses of policy robustness to changing initial distributions. Specifically, we introduced the notion of (δ, ϵ) -regret to assess robustness, and showed how our bounds efficiently approximate (i) the set of initial distributions where a policy remains near-optimal, and (ii) the policy’s suboptimality over a given set.

Looking ahead, we are interested in extending our analysis to account for uncertainty or changes in the transition dynamics. In this setting, a promising direction is to develop robust and adaptive policies that can accommodate such uncertainties, leveraging the bounds and sensitivity analysis introduced in this work.

Acknowledgments: The authors thank N.S. Aybat (Penn State) and M. Gurbuzbalaban (Rutgers) for providing valuable pointers on LP perturbation analysis.

REFERENCES

- [1] E. Altman. *Constrained Markov Decision Processes*. CRC Press, 1999.
- [2] E. Altman and V. A. Gaitsgory. Stability and singular perturbations in constrained markov decision problems. *IEEE transactions on automatic control*, 38(6):971–975, 2002.
- [3] E. Altman and A. Shwartz. Sensitivity of constrained markov decision processes. *Annals of Operations Research*, 32(1):1–22, 1991.
- [4] A. Ben-Tal, D. Den Hertog, and J.-P. Vial. Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical programming*, 149(1):265–299, 2015.
- [5] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 1995.
- [6] D. Bertsimas and J. N. Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.
- [7] Y. Chow, A. Tamar, S. Mannor, and M. Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18:1–51, 2017.
- [8] E. A. Feinberg and A. Shwartz. Constrained markov decision models with weighted discounted rewards. *Mathematics of Operations Research*, 25(1):130–140, 2000.
- [9] S. Feyzabadi and S. Carpin. Planning using hierarchical constrained markov decision processes. *Autonomous Robots*, 41:1589–1607, 2017.
- [10] M. V. Nayakkankuppam and M. L. Overton. Conditioning of semidefinite programs. *Mathematical programming*, 85(3), 1999.
- [11] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.
- [12] R. H. Russel, M. Benosman, and J. Van Baar. Robust constrained-mdps: Soft-constrained robust policy optimization under model uncertainty. *arXiv preprint arXiv:2010.04870*, 2020.
- [13] V. Varagapriya, V. V. Singh, and A. Lissner. Constrained markov decision processes with uncertain transition probabilities. In *2024 European Control Conference (ECC)*, pages 3003–3008. IEEE, 2024.
- [14] Y. Wang, F. Miao, and S. Zou. Robust constrained reinforcement learning. *arXiv preprint arXiv:2209.06866*, 2022.
- [15] Z. Zhang, K. Panaganti, L. Shi, Y. Sui, A. Wierman, and Y. Yue. Distributionally robust constrained reinforcement learning under strong duality. *arXiv preprint arXiv:2406.15788*, 2024.