Cancellation-Free Regret Bounds for Lagrangian Approaches in Constrained Markov Decision Processes

Anonymous Author(s)
Affiliation
Address
email

Abstract

Constrained Markov Decision Processes (CMDPs) are one of the common ways to model safe reinforcement learning problems, where the safety objectives are modeled by constraint functions. Lagrangian-based dual or primal-dual algorithms provide efficient methods for learning in CMDPs. For these algorithms, the currently known regret bounds in the finite-horizon setting allow for a *cancellation of errors*; that is, one can compensate for a constraint violation in one episode with a strict constraint satisfaction in another episode. However, in practical applications, we do not consider such a behavior safe.

In this paper, we overcome this weakness by proposing a novel model-based dual algorithm OptAug-CMDP for tabular finite-horizon CMDPs. Our algorithm is motivated by the augmented Lagrangian method and can be performed efficiently. We show that during K episodes of exploring the CMDP, our algorithm obtains a regret of $\tilde{O}(\sqrt{K})$ for both the objective and the constraint violation. Unlike existing Lagrangian approaches, our algorithm achieves this regret without the need for the cancellation of errors.

1 Introduction

In classical reinforcement learning (RL, Sutton and Barto, 2018), the goal is to learn an optimal policy when interacting with an unknown Markov decision process (MDP, Bellman, 1957). In MDPs, an agent aims to minimize the cumulative cost it incurs during an episode of the MDP. However, in practical scenarios, the learned policy often must also adhere to certain safety constraints. For example, when navigating a car on a race track, one would want to avoid crossing the boundaries of the track too often. Such safety requirements are commonly modeled via constrained Markov decision processes (CMDPs, Altman, 1999). We consider the problem of learning an optimal feasible policy in a CMDP. That is, the goal of the agent is to minimize the cost while satisfying the constraints[1]. Since the CMDP is unknown, we formalize these desiderata by considering the regret with respect to an optimal feasible solution for the cost and for the constraint violation, respectively.

Importantly, we do not consider it sufficient to provide an agent whose cumulative cost suboptimality and cumulative constraint violation are sublinear. This is because an agent can have a negative constraint violation (by being very safe but incurring higher cost than an optimal safe policy), or a positive constraint violation (by being unsafe but incurring lower cost than an optimal safe policy). Thus, terms from these two cases can cancel each other out, which we refer to as the so-called *cancellation of errors* (Efroni et al., 2020). An agent for which these cumulative terms are sublinear may violate the safety constraints heavily during learning by oscillating around an optimal safe policy. While such a method converges on average to an optimal safe policy[2], it neither allows for directly

[1] i.e., being feasible for the CMDP, which we also refer to as being *safe*

[2] Here, we refer to the value functions with respect to the underlying CMDP.

Submitted to 16th European Workshop on Reinforcement Learning (EWRL 2023). Do not distribute.

extracting an optimal feasible policy nor does it guarantee safety during learning. We consider a stronger notion of regret, which overcomes this issue by considering the sum of the *positive parts* of the error terms instead. As pointed out by Efroni et al. (2020), it is of major theoretical interest whether Lagrangian approaches can achieve sublinear bounds for this notion of regret.

The approaches to learning CMDPs are split into linear programming (LP) and Lagrangian approaches[3] (Altman, 1999; Efroni et al., 2020). While LP-based algorithms generally allow for sublinear regret bounds without the need for cancellations (Efroni et al., 2020), they can be expensive when dealing with large state-action spaces. In contrast, in Lagrangian methods, we can solve the optimization problem arising in each episode using dynamic programming (DP), offering a computational benefit over solving LPs. However, the currently known bounds for Lagrangian approaches only concern a weaker form of regret that allows for the aforementioned cancellation of errors. As Efroni et al. (2020) pointed out, this is due to the underlying optimization methods rather than a weakness of the analysis. The main goal of this paper is to provide a Lagrangian-based algorithm that guarantees sublinear regret without the cancellation of errors. To achieve this, the key problem we solve is stopping the agent from oscillating around an optimal safe policy. Our contributions can be summarized as follows:

- We propose a novel model-based dual algorithm, OptAug-CMDP, for learning an optimal feasible policy in an unknown CMDP (Section 3). The algorithm is split into a model pre-training phase and an optimistic exploration phase motivated by the augmented Lagrangian method.
- We show that a sub-problem required for OptAug-CMDP can be reformulated as a convex optimization problem. We provide an efficient algorithm to solve it (Section 3), despite the non-linearity introduced by considering the augmented Lagrangian.
- We prove that with high probability, during K episodes OptAug-CMDP achieves regrets for the cost and the constraint violations of $\tilde{O}(\sqrt{K})$ when only highlighting the dependency on K. Notably, we achieve this bound for the stronger notion of regret, which does not allow for the cancellation of errors. This partly settles the open problem posed by Efroni et al. (2020).

1.1 Related Work

The most relevant foundation for our work is the work by Efroni et al. (2020), which reviews model-based algorithms for CMDPs and establishes regret bounds for them. The authors analyze the LP-based algorithms OptCMDP and OptCMDP-bonus that achieve sublinear regret without cancellations. However, the Lagrangian-based algorithms they analyze, OptDual-CMDP and OptPrimalDual-CMDP, only achieve sublinear regret with cancellations. This is because the oscillatory behavior of dual and primal-dual descent methods prevents the individual iterates from being approximately feasible. Therefore, the authors pose the open question of whether one can devise Lagrangian-based algorithms that do not suffer from this issue.

The vast majority of relevant work providing guarantees for Lagrangian approaches to CMDPs is concerned with model-free primal-dual algorithms (Ding et al., 2020; Bai et al., 2022; Ding et al., 2022a,b) or model-based dual algorithms (Liu et al., 2021b; Efroni et al., 2020). However, in both cases, the existing literature does not address the issue of the cancellation of errors when exploring the CMDP and thus does not provide a method for *safely* finding an optimal feasible policy.[4] While there is work on analyzing different forms of regularization to the Lagrangian-based algorithms, their guarantees either require the cancellation of errors (Liu et al., 2021a) or assume access to exact policy gradients (Ying et al., 2022; Li et al., 2021b). Moskovitz et al. (2023) propose a first approach to address the cancellation of errors, by replacing gradients with their optimistic gradient counterparts in well-known Lagrangian-based RL algorithms. While they show the empirical success of their methods, their theoretical analysis is restricted to the case of having access to an exact policy gradient oracle. Stooke et al. (2020) address the underlying problem of oscillations of Lagrangian methods for CMDPs via PID control in the context of deep RL, providing experimental successes but no guarantees. Thus, to the best of our knowledge, none of the existing works addresses the open question of Efroni et al. (2020) in the setup of an unknown CMDP.

While there are mentions of using the augmented Lagrangian method for CMDPs (Li et al., 2021a; Lu, 2022; Krishnamurthy, 2003; Krishnamurthy and Abad, 2011, 2012), all such works are concerned with research questions rather different from ours. The only one similar to ours is that of Li et al.

[3] i.e., dual and primal-dual algorithms

[4] That is, there are no guarantees on the constraint violations of the individual iterates for these methods.

(2021a). The authors propose a surrogate reward inspired by the augmented Lagrangian to promote safety during learning. However, their method significantly differs from ours as it is concerned with *instantaneous* constraints in an infinite-horizon CMDP. Moreover, their analysis only shows that an optimal policy for their surrogate MDP is optimal for the original CMDP (under certain assumptions).

2 Background and Problem Formulation

Notation: For any $n \in \mathbb{N}$, we use the short-hand notation $[n]$ to refer to the set of integers $\{1, \dots, n\}$. For any finite set X, we denote by $\Delta(X)$ the probability simplex over X, i.e., $\Delta(X) = \{v \in [0,1]^X | \sum_{x \in X} v(x) = 1\}$. For $a \in \mathbb{R}$, we set $[a]_+ := \max\{0, a\}$ to be the positive part of a. For a vector $b \in \mathbb{R}^n$, we write $[b]_+$ for the vector whose entries are the positive parts of the corresponding entries of b. Similarly, for two vectors $a, b \in \mathbb{R}^n$, we write $a \leq b$ as a short-hand for $a_i \leq b_i$, for all $i \in [n]$. Throughout the paper, we denote the Euclidean norm by $\| \cdot \|$.

We define a finite-horizon CMDP as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, p, c, (d_i)_{i \in [I]}, (\alpha_i)_{i \in [I]}, s_1)$ with the following components. $\mathcal{S}$ and $\mathcal{A}$ are the state and action space, respectively, and $H > 0$ denotes the horizon. Every episode consists of H steps and starts from the initial state $s_1 \in \mathcal{S}$. At every step $h \in [H]$, $p_h(s'|s,a)$ denotes the probability of transitioning to state s' if the current state and action are s and a. Moreover, $c_h \colon \mathcal{S} \times \mathcal{A} \to [0,1]$ denotes the objective cost function at step h. For $i \in [I]$, $d_{i,h} \colon \mathcal{S} \times \mathcal{A} \to [0,1]$ refers to the cost function of the i-th constraint at step h, and $\alpha_i \in [0, H]$ denotes the threshold for the i-th constraint. We assume that the state and action space are both finite, with cardinalities S and A, respectively. Furthermore, we assume that the agent does not know the transition probabilities, objective costs, or constraint costs beforehand. Whenever the agent takes an action a in state s at time h, it observes costs sampled from random variables $C_h(s,a) \in [0,1]$ and $(D_{i,h}(s,a))_{i \in [I]} \in [0,1]^I$ such that $\mathbb{E}[C_h(s,a)] = c_h(s,a)$ and $\mathbb{E}[D_{i,h}(s,a)] = d_{i,h}(s,a)$, for all $i \in [I]$. The agent interacts with the CMDP by playing a policy $\pi = (\pi_h)_{h \in [H]}$, meaning that if in state s at step $h \in [H]$, the agent samples its next action from $\pi_h(\cdot|s) \in \Delta(\mathcal{A})$. For an arbitrary cost function $l = (l_h)_{h \in [H]}$ and transition probabilities $p' = (p'_h)_{h \in [H]}$, the expected cumulative cost under policy π is measured by the value function defined as follows:

$$V^\pi(l, p') := \mathbb{E}\left[\sum_{h=1}^{H} l_h(s_h, a_h) \mid s_1, \pi, p'\right],$$

where (s_h, a_h) denotes the state-action pair at step h under transitions p' and policy π. We fix an optimal solution of the CMDP, given by a policy π^*, defined as follows:

$$\pi^* \in \arg\min_{\pi \in \Pi} \quad V^\pi(c, p) \quad \text{s.t.} \quad V^\pi(d_i, p) \leq \alpha_i \quad (\forall i \in [I]). \tag{1}$$

For brevity, we write $V^\pi((l_i)_{i \in [I]}, p') := (V^\pi(l_1, p'), \dots, V^\pi(l_I, p'))^T \in \mathbb{R}^I$ in the presence of I different cost functions $l_i = (l_{i,h})_{h \in [H]}$ ($i \in [I]$), and $\alpha := (\alpha_1, \dots, \alpha_I)^T \in \mathbb{R}^I$. Furthermore, we denote by $\Pi := \{\pi = (\pi_h)_{h \in [H]} | \pi_h : \mathcal{S} \to \Delta(\mathcal{A})\}$ the entire policy space.

Strong duality and dual methods: Altman (1999); Paternain et al. (2019) proved that CMDPs possess the *strong duality* property; i.e., given a feasible CMDP $\mathcal{M}$, the following relation holds:

$$V^{\pi^*}(c, p) = \underbrace{\min_{\pi \in \Pi} \max_{\lambda \in \mathbb{R}^I_+} \mathcal{L}(\pi, \lambda)}_{\text{Primal problem}} = \underbrace{\max_{\lambda \in \mathbb{R}^I_+} \min_{\pi \in \Pi} \mathcal{L}(\pi, \lambda)}_{\text{Dual problem}}, \tag{2}$$

where $\mathcal{L}(\pi, \lambda) := V^\pi(c, p) + \lambda^T \left(V^\pi((d_i)_{i \in [I]}, p) - \alpha\right)$ denotes the Lagrangian. The strong duality property gives theoretical justification to dual methods (Altman, 1999; Efroni et al., 2020; Paternain et al., 2019). These methods are popular, as the dual problem can be solved via a sequence of (extended[5]) *unconstrained* MDPs, each of which can be solved efficiently via DP (as opposed to using LPs for solving a sequence of CMDPs, for which the Bellman optimality principle does not hold).

In this work, we solve the min-max problem in Eq. (2) using the augmented Lagrangian method. This is beneficial since the analysis of the augmented Lagrangian method allows for convergence

[5] If the CMDP is unknown, backward induction involves an extra optimization step over the possible transitions (Jin et al., 2019).

guarantees concerning the last iterate, and not just the averaged iterates. Since the occurring subproblems are not MDPs anymore, we justify in Section 3 how they can still be solved efficiently by leveraging a Frank-Wolfe scheme and DP. We are now ready to state the main problem formulation of our work.

Problem formulation: In our setting, the agent interacts with the unknown CMDP over a fixed number of $K > 0$ episodes. In every episode $k \in [K]$, the agent plays a policy $\pi_k \in \Pi$ and its goal is to (simultaneously) minimize its regrets, defined as follows:

$$\mathcal{R}(K; c) := \sum_{k \in [K]} [V^{\pi_k}(c, p) - V^{\pi^*}(c, p)]_+, \qquad \text{(Objective strong regret)}$$

$$\mathcal{R}(K; d) := \max_{i \in [I]} \sum_{k \in [K]} [V^{\pi_k}(d_i, p) - \alpha_i]_+ . \qquad \text{(Constraint strong regret)}$$

For simplicity, we will write *regret* when referring to the strong regret throughout the paper. As we pointed out, existing works on Lagrangian-based algorithms (Liu et al., 2021b; Efroni et al., 2020; Bai et al., 2022; Ding et al., 2022a,b) only prove sublinear guarantees on a *weaker* notion of regret, defined as follows:

$$\mathcal{R}_{\pm}(K; c) := \sum_{k \in [K]} (V^{\pi_k}(c, p) - V^{\pi^*}(c, p)), \qquad \text{(Objective weak regret)}$$

$$\mathcal{R}_{\pm}(K; d) := \max_{i \in [I]} \sum_{k \in [K]} (V^{\pi_k}(d_i, p) - \alpha_i) . \qquad \text{(Constraint weak regret)}$$

The weak regrets allow for the aforementioned cancellation of errors; i.e., even if they are sublinear in K, the agent can continue to compensate for a constraint violation in one episode with a strict constraint satisfaction in another episode. On the other hand, a sublinear bound on the stronger notion of regret guarantees that the agent achieves a low constraint violation in most of the episodes.[6] While this is crucial for practical applications, it is strictly more challenging to provide a bound for the strong regrets instead of the weaker notion.

3 Algorithm and Main Result

In this section, we introduce our algorithm OptAug-CMDP (see Algorithm 1) and state its regret guarantees in Theorem 1. In OptAug-CMDP, the agent interacts with the unknown CMDP over a fixed number of $K > 0$ episodes. To encourage exploration of the CMDP, the agent follows the well-known *optimism in the face of uncertainty* principle (Auer et al., 2008) and builds an optimistic estimate of the CMDP, in every episode $k \in [K]$. That is, in every episode $k \in [K]$, the agent builds optimistic estimates $\tilde{c}_k$ for the objective cost c, optimistic estimates $\tilde{d}_{i,k}$ for the constraint costs d_i, and a set of plausible transition probabilities B_k^p, which we specify in the following paragraph.

Optimistic estimates: Let $n_h^{k-1}(s, a) := \sum_{l=1}^{k-1} \mathbb{1}_{\{s_h^l = s,\ a_h^l = a\}}$ count the number of times that the state-action pair (s, a) has been visited at step h before episode k. Here, $(s_h^l,\ a_h^l)$ denotes the state-action pair that was visited at step h in episode l. First, we compute the empirical averages of the cost and transition probabilities as follows:

$$\bar{c}_h^{k-1}(s, a) := \frac{\sum_{l=1}^{k-1} C_h^l(s, a) \mathbb{1}_{\{s_h^l = s,\ a_h^l = a\}}}{n_h^{k-1}(s, a) \vee 1},$$

$$\bar{d}_{i,h}^{k-1}(s, a) := \frac{\sum_{l=1}^{k-1} D_{i,h}^l(s, a) \mathbb{1}_{\{s_h^l = s,\ a_h^l = a\}}}{n_h^{k-1}(s, a) \vee 1} \quad (\forall i \in [I]),$$

$$\bar{p}_h^{k-1}(s' | s, a) := \frac{\sum_{l=1}^{k-1}, \mathbb{1}_{\{s_h^l = s,\ a_h^l = a,\ s_{h+1}^l = s'\}}}{n_h^{k-1}(s, a) \vee 1},$$

[6] Indeed, fix $\epsilon > 0$ and suppose $\mathcal{R}(K; d) \leq \tilde{O}(\sqrt{K})$. Then there exist at most $\tilde{O}(\sqrt{K}/\epsilon)$ episodes with a constraint violation of at least ϵ. In other words, only a small fraction $\tilde{O}(1/(\epsilon\sqrt{K}))$ of the iterates is not ϵ-safe. In comparison, this is by no means guaranteed by a sublinear bound on $\mathcal{R}_{\pm}(K; d)$.

where $a \vee b := \max\{a, b\}$. With this, we define the optimistic costs and the set of plausible transition probabilities as

$$\begin{aligned}
\tilde{c}_{k,h}(s,a) &:= \bar{c}_h^{k-1}(s,a) - \beta_{k,h}^c(s,a), \\
\tilde{d}_{i,k,h}(s,a) &:= \bar{d}_{i,h}^{k-1}(s,a) - \beta_{i,k,h}^d(s,a) \quad (\forall i \in [I]), \\
B_{k,h}^p(s,a) &:= \{\tilde{p}_h(\cdot|s,a) \in \Delta(\mathcal{S}) \mid \forall s' \in \mathcal{S} \colon |\tilde{p}_h(s'|s,a) - \bar{p}_h^{k-1}(s'|s,a)| \leq \beta_{k,h}^p(s,a,s')\}, \\
B_k^p &:= \{\tilde{p} \mid \forall s,a,h \colon \tilde{p}_h(\cdot|s,a) \in B_{k,h}^p(s,a)\}.
\end{aligned} \tag{3}$$

Here, $\beta_{k,h}^c(s,a) = \beta_{i,k,h}^d(s,a) > 0$ denote the exploration bonus for the costs and $\beta_{k,h}^p(s,a,s') > 0$ denotes the confidence threshold for the transitions. For any $\delta \in (0,1)$, we specify the correct values for those quantities in Appendix E.1, in order to obtain our regret guarantees with probability at least $1-\delta$. In the next paragraph, we describe how the agent computes its policy in episode k.

Policy update: Given the optimistic CMDP at episode k, we derive the next policy π_k using a scheme motivated by the augmented Lagrangian method (cf. Eqs. (4) and (5)). We explain how we can perform the optimization step in Eq. (4) efficiently, up to a specified accuracy ϵ_k, at the end of this section. For now, we treat this part of the algorithm as a black-box subroutine.

Optimistic exploration alone with the augmented Lagrangian, however, is not sufficient to obtain sublinear regret guarantees for our algorithm. For technical reasons, our analysis also requires the optimistic CMDPs with costs $\tilde{c}_k$, $(\tilde{d}_{i,k})_{i\in[I]}$ and transitions $\tilde{p}_k \in B_k^p$ (cf. Eq. (4)) to be *strictly* feasible, in every episode $k \in [K]$. Our analysis in Section 4.2 explains the need for this technical assumption. To address this issue, we propose a pre-training phase *before* the optimistic exploration phase, which we describe in the following paragraph.

Pre-training phase: In this phase, the agent repeatedly executes a fixed policy $\bar{\pi}$ for $K' \leq K$ episodes. The policy $\bar{\pi}$ must be *strictly* feasible with respect to the true CMDP, which we formally state in the following assumption.

Assumption 1 (Strictly feasible policy)**.** *We have access to a policy $\bar{\pi}$ such that $V^{\bar{\pi}}(d_i, p) < \alpha_i$ for all $i \in [I]$. Furthermore, we assume that the slack γ, defined below, is known*[7]*:*

$$\gamma := \min_{i\in[I]} (\alpha_i - V^{\bar{\pi}}(d_i, p)) \in (0, H].$$

Note that this is stronger than only assuming the *existence* of a strictly feasible policy. However, it is realistic to make this assumption in many practical setups (Liu et al., 2021b; Bura et al., 2022). For example, in the case of a race car that should not exceed the boundary of a track, it would be sufficient to have access to the policy of a car that strictly stays within the boundaries but may be arbitrarily slow. To address the technical issue mentioned earlier, we need to set K' such that the following condition holds for some $\nu \in (0,1)$, with high probability:

$$V^{\bar{\pi}}(\tilde{d}_{i,k}, \tilde{p}_k) \leq \alpha_i - \nu\gamma \quad (\forall i \in [I]\ \forall k \in \{K', \dots, K\}),$$

where $\tilde{p}_k$ is defined by the update in Eq. (4).[8] In particular, the fixed policy $\bar{\pi}$ is strictly feasible with respect to the optimistic CMDP at every episode $k \geq K'$. Indeed, if the agent plays $\bar{\pi}$ for the first K' episodes of the algorithm with a large enough constant K', then for all future episodes, the constraint value function of $\bar{\pi}$ under the estimated model is close to the constraint value function of $\bar{\pi}$ under the true model. Thus, we can ensure the above condition. Leveraging an adaption of an on-policy error bound (see Appendix C), we prove that it is sufficient to set K' as follows:

Lemma 1. *Suppose that Assumption 1 holds, i.e., the agent has access to a strictly feasible $\bar{\pi}$ and its slack $\gamma > 0$. Fix any $\nu \in (0,1)$, and suppose the agent executes $\bar{\pi}$ for $K' = \tilde{O}\left(\max\left\{\frac{(H\sqrt{\mathcal{N}}+S)SAH}{(1-\nu)\gamma}, \frac{\mathcal{N}SAH^3}{(1-\nu)^2\gamma^2}\right\}\right)$ episodes, where $\mathcal{N} := \max_{s,a,h} |\{s' \mid p_h(s'|s,a) > 0\}|$ denotes the maximum number of transitions. Then, if the agent updates the optimistic CMDP based on the observations from those episodes (cf. Eq.* (3)*), with probability at least $1-\delta$ the following condition is satisfied for every $k \in \{K', \dots, K\}$:*

$$V^{\bar{\pi}}(\tilde{d}_{i,k}, \tilde{p}_k) \leq \alpha_i - \nu\gamma \quad (\forall i \in [I]).$$

[7] In other words, there is a Slater point for the constraint set of the true CMDP. Note that knowing a lower bound instead of the exact slack γ is sufficient as well.

[8] For $k = K'$, we just take any $\tilde{p}_{K'} \in B_{K'}^p$

The resulting OPTAUG-CMDP algorithm is described in Algorithm 1.

Algorithm 1 OPTAUG-CMDP

Require: K (total number of episodes), $K' \leq K$ (number of pre-training episodes), $(\eta_k)_{k \geq K'+1}$ (step sizes), $(\epsilon_k)_{k \geq K'+1}$ (accuracies), $\bar{\pi}$ (strictly feasible policy), α (constraint thresholds), $\lambda_{K'+1} := 0 \in \mathbb{R}^I$

// Phase 1: Pre-training the model
for $k = 1, \dots, K'$ **do**
  Play policy $\pi_k = \bar{\pi}$, update estimates of the costs $\tilde{c}_{k+1}$, $(\tilde{d}_{i,k+1})_{i \in [I]}$ and transitions B^p_{k+1} (Eq. (3)).

// Phase 2: Optimistic exploration with pre-trained model
for $k = K'+1, \dots, K$ **do**
  Update policy (by finding $\pi_k, \tilde{p}_k$ such that the objective is ϵ_k-close to the minimum):

$$\pi_k, \tilde{p}_k := \arg \min_{\substack{\pi \in \Pi \\ p' \in B^p_k}} \left(V^\pi(\tilde{c}_k, p') + \frac{1}{2\eta_k} \| [\lambda_k + \eta_k (V^\pi((\tilde{d}_{i,k})_{i \in [I]}, p') - \alpha)]_+ \|^2 \right) \tag{4}$$

  Update dual variables:

$$\lambda_{k+1} := [\lambda_k + \eta_k (V^{\pi_k}((\tilde{d}_{i,k})_{i \in [I]}, \tilde{p}_k) - \alpha)]_+ \tag{5}$$

  Play π_k, update estimates of the costs $\tilde{c}_{k+1}$, $(\tilde{d}_{i,k+1})_{i \in [I]}$ and transitions B^p_{k+1} (Eq. (3)).

We are now ready to state the regret guarantees for OPTAUG-CMDP.

Theorem 1. *Suppose that Assumption 1 holds, let $\delta \in (0,1)$ and $\nu > 0$. Then there exist $K' = \tilde{O}\left(\max\left\{\frac{(H\sqrt{\mathcal{N}}+S)SAH}{(1-\nu)\gamma}, \frac{\mathcal{N}SAH^3}{(1-\nu)^2\gamma^2}\right\}\right)$ and η_k, ϵ_k such that with probability at least $1-\delta$,* OPTAUG-CMDP *achieves a total regret of*

$$\mathcal{R}(K; c) = \tilde{O}\left(\sqrt{\mathcal{N}SAH^3K} + (H\sqrt{\mathcal{N}} + S)SAH + K'H\right),$$
$$\mathcal{R}(K; d) = \tilde{O}\left(\sqrt{\mathcal{N}SAH^3K} + (H\sqrt{\mathcal{N}} + S)SAH\right).$$

We remark that to achieve this bound, using step sizes $\eta_{K'+k} = \Theta(k^{2.5})$ and accuracies $\epsilon_{K'+k} = \Theta(1/\eta_{K'+k})$ (when only highlighting the dependency on k) is sufficient, as we discuss in Appendix D.3.

Comparison with OPTDUAL-CMDP: Crucially, our bound holds for the stronger notion of regret, whereas the one for the related OPTDUAL-CMDP algorithm (see Appendix A.2) by Efroni et al. (2020) only concerns the weak regret, which allows for the cancellation of errors. Apart from this, the bound we obtain is similar in spirit. However, our regret bound does not depend on the number I of constraints, up to polylogarithmic factors. Moreover, note that we get slightly different (but not worse) constants and the constant extra term $K'H$ that is due to the model pre-training phase. In addition, it is important to note that we can choose η_k, ϵ_k in terms of γ (Assumption 1) such that in the leading term of the regret bound, there is no dependency on $\min_{i \in [I]}(\alpha_i - V^{\bar{\pi}}(d_i, p))$, as opposed to OPTDUAL-CMDP.

Solving the inner problem: We now elaborate on the subroutine for solving the optimization problem in Eq. (4) that defines the policy update in Algorithm 1. Importantly, we can reformulate Eq. (4) as a constrained optimization problem that is convex in the state-action-state occupancy measure (see Appendix B.1). However, the resulting problem is neither an LP nor an extended MDP (due to the nonlinear objective), which prevents solving it with a single call of a DP or LP solver. Albeit related, it is not a standard convex RL problem either (due to the additional optimization over the constraint set B^p_k). Moreover, computing projections onto the high-dimensional domain is prohibitive in practice, making it impossible to simply run projected gradient descent.

The projection-free method we propose in Appendix B.2 overcomes this difficulty by combining a Frank-Wolfe scheme with DP in a sequence of (extended) MDPs. In every iteration of this inner method, we consider the linear minimization step needed for a Frank-Wolfe iteration. When switching back to an optimization over Π and B_k^p, we can then perform this minimization step by solving an extended but unconstrained MDP via DP.[9] The smoothness properties of the objective of Eq. (4) then determine the iteration complexity of the Frank-Wolfe scheme. Formally, we have the following.

Proposition 1. *In episode k, fix any accuracy of $\epsilon_k > 0$. There exists an algorithm for solving Eq.* (4) *such that the objective at its output $(\pi_k, \tilde{p}_k)$ is ϵ_k-close to the optimum of Eq.* (4)*, by solving $O\left(\frac{\eta_k I S^2 A H}{\epsilon_k}\right)$ (extended) MDPs via DP.*

4 Sketch of the Regret Analysis

In this section, we outline the key steps in our proof of Theorem 1 and defer the detailed proofs to Appendix E. We will condition our regret analysis on a *success event* G, which we formally define in Appendix E.1. G ensures that (a) the optimistic cost estimates in Eq. (3) are in fact optimistic and (b) the true transitions are contained in the set of plausible models from Eq. (3), i.e.:

$$\tilde{c}_k \leq c, \quad \tilde{d}_{i,k} \leq d_i \ (\forall i \in [I]), \quad p \in B_k^p,$$

for every episode $k \in [K]$. In the following lemma, we prove that G occurs with high probability.

Lemma 2. *Fix $\delta \in (0,1)$ and define the optimistic model in Eq.* (3) *accordingly. Then the success event G occurs with probability at least $1-\delta$, i.e., $P[G] \geq 1-\delta$.*

We proceed with the regret analysis and first split the regrets between the two phases of the algorithm:

$$\mathcal{R}(K;c) = \underbrace{\sum_{k=1}^{K'} [V^{\pi_k}(c,p) - V^{\pi^*}(c,p)]_+}_{\text{Pre-Training}} + \underbrace{\sum_{k=K'+1}^{K} [V^{\pi_k}(c,p) - V^{\pi^*}(c,p)]_+}_{\text{Optimistic Exploration}},$$

$$\mathcal{R}(K;d) \leq \underbrace{\max_{i\in[I]} \sum_{k=1}^{K'} [V^{\pi_k}(d_i,p) - \alpha_i]_+}_{\text{Pre-Training}} + \underbrace{\max_{i\in[I]} \sum_{k=K'+1}^{K} [V^{\pi_k}(d_i,p) - \alpha_i]_+}_{\text{Optimistic Exploration}}.$$

Then, applying Lemma 1, we can trivially bound the objective regret during the pre-training phase by $K'H$. Since $\bar{\pi}$ is strictly feasible, there is no constraint regret during pre-training. We now focus on the regrets incurred in the optimistic exploration phase. For this, we further decompose the regrets as follows (see Appendix E.1):

$$\mathcal{R}(K;c) \leq K'H + \underbrace{\sum_{k=K'+1}^{K} [V^{\pi_k}(c,p) - V^{\pi_k}(\tilde{c}_k,\tilde{p}_k)]_+}_{\text{Estimation Error}} + \underbrace{\sum_{k=K'+1}^{K} [V^{\pi_k}(\tilde{c}_k,\tilde{p}_k) - V^{\pi^*}(c,p)]_+}_{\text{Optimization Error}},$$

$$\mathcal{R}(K;d) \leq \underbrace{\max_{i\in[I]} \sum_{k=K'+1}^{K} [V^{\pi_k}(d_i,p) - V^{\pi_k}(\tilde{d}_{i,k},\tilde{p}_k)]_+}_{\text{Estimation Error}} + \underbrace{\max_{i\in[I]} \sum_{k=K'+1}^{K} [V^{\pi_k}(\tilde{d}_{i,k},\tilde{p}_k) - \alpha_i]_+}_{\text{Optimization Error}}.$$

We have thus decomposed the regrets into

(A) *estimation errors* that are due to the estimated model, and

(B) *optimization errors* that we can analyze via the underlying optimization method.

Conditioning on the success event G, we will obtain bounds sublinear in K for both parts of the decomposition. Note that we cannot adapt the analysis by Efroni et al. (2020) to achieve this goal, since it only allows for bounds on the averages of the *signed* optimization errors. We proceed with bounding the estimation errors in the next section.

[9] Formally, this is because a version of the Bellman optimality principle applies after dualizing the constraints, even if we need to optimize over the confidence intervals for the transitions during backward induction.

4.1 Estimation Errors (Optimistic Exploration)

Leveraging on-policy error bounds for optimistic exploration in MDPs (Appendix C), we establish the desired bound on the estimation errors.

Lemma 3 (Estimation errors)**.** *Let $(\pi_k)_{k=K'+1}^{K}$ be the sequence of policies obtained by* OPTAUG-CMDP. *Then, conditioned on G, we can bound the estimation errors as follows:*

$$\sum_{k=K'+1}^{K} [V^{\pi_k}(c,p) - V^{\pi_k}(\tilde{c}_k, \tilde{p}_k)]_+ \leq \tilde{O}\left(\sqrt{\mathcal{N}SAH^3K} + (H\sqrt{\mathcal{N}} + S)SAH\right),$$

$$\max_{i\in[I]} \sum_{k=K'+1}^{K} [V^{\pi_k}(d_i,p) - V^{\pi_k}(\tilde{d}_{i,k}, \tilde{p}_k)]_+ \leq \tilde{O}\left(\sqrt{\mathcal{N}SAH^3K} + (H\sqrt{\mathcal{N}} + S)SAH\right).$$

We refer to Appendix E.2 for the proof. Lemma 3 proves that the estimation errors for both the objective and constraints can indeed be bounded by a term that is sublinear in K. In the next section, we provide a bound for the optimization errors.

4.2 Optimization Errors (Optimistic Exploration)

Recall that by Proposition 1, our solver for the inner problem (Eq. (4)) has the following guarantee, for every episode $k \geq K' + 1$:

$$\mathcal{L}_k(\pi_k, \tilde{p}_k) \leq \min_{\substack{\pi\in\Pi \\ p'\in B_k^p}} \mathcal{L}_k(\pi, p') + \epsilon_k,$$

where $\mathcal{L}_k(\pi, p') := V^{\pi}(\tilde{c}_k, p') + \frac{1}{2\eta_k}\|[\lambda_k + \eta_k(V^{\pi}((\tilde{d}_{i,k})_{i\in[I]}, p') - \alpha)]_+\|^2$ denotes the objective at episode k. If the true CMDP is known, i.e., no exploration is required, then Xu (2021) proves a sublinear regret bound for the optimization error if the step sizes η_k and accuracies ϵ_k are chosen suitably.[10] They obtain this result by bounding the dual variables λ_k across the iterations k. In our setting, however, since the objective and constraint set of the optimization problem (Eq. (4)) *change* in every episode, we require a novel type of analysis.

As a first step, we show that we can bound the optimization errors in episode k by expressions that depend on the dual variables λ_k and λ_{k+1}.

Lemma 4. *Conditioned on G, for each $k \in \{K'+1, \ldots, K\}$, in* OPTAUG-CMDP *we have*

$$V^{\pi_k}(\tilde{c}_k, \tilde{p}_k) - V^{\pi^*}(c,p) \leq \epsilon_k + \frac{\|\lambda_k\|^2 - \|\lambda_{k+1}\|^2}{2\eta_k},$$

$$V^{\pi_k}(\tilde{d}_{i,k}, \tilde{p}_k) - \alpha_i \leq \frac{\lambda_{k+1}(i) - \lambda_k(i)}{\eta_k} \quad (\forall i \in [I]).$$

To further bound the norm of the dual iterates, for each episode $k \geq K'$ we consider the k-th optimistic CMDP, which we define as follows:

$$\min_{\pi\in\Pi} \quad V^{\pi}(\tilde{c}_k, \tilde{p}_k) \quad \text{s.t.} \quad V^{\pi}(\tilde{d}_{i,k}, \tilde{p}_k) \leq \alpha_i \quad (\forall i \in [I]). \tag{6}$$

Note that Eq. (6) indeed is a CMDP. By Lemma 1, $\bar{\pi}$ is strictly feasible for Eq. (6) for all $k \geq K'$, with a slack of $\geq \nu\gamma$ uniformly bounded away from zero. By strong duality (Paternain et al., 2019), there exist primal-dual pairs (π_k^*, λ_k^*) satisfying

$$V^{\pi_k^*}(\tilde{c}_k, \tilde{p}_k) = \min_{\pi\in\Pi}\left(V^{\pi}(\tilde{c}_k, \tilde{p}_k) + (\lambda_k^*)^T(V^{\pi}((\tilde{d}_{i,k})_{i\in[I]}, \tilde{p}_k) - \alpha)\right).$$

We formalize this with Lemma 18 in Appendix E.3.2, by using the fact that we can formulate Eq. (6) as a convex optimization problem using the LP formulation of CMDPs (Appendices A.3 and A.4). With this, we can establish the following bound on the dual iterates.

[10] That is, such that $\sum_{k=K'+1}^{K} 1/\eta_k = o(K)$ and $\sum_{k=K'+1}^{K} \epsilon_k = o(K)$, see Xu (2021, Remark 7).

Lemma 5. *Let $k \in \{K'+1, \dots, K\}$ and suppose Eq.* (6) *is strictly feasible for every $k' \in \{K', \dots, K\}$. Let $(\pi^*_{k'}, \lambda^*_{k'})$ be pairs of primal-optimal and dual-optimal solutions for Eq.* (6)*. Then the iterates of* OPTAUG-CMDP *satisfy*

$$\|\lambda_{k+1}\| \leq 2 \sum_{t=K'}^{k} \|\lambda^*_t\| + \sum_{t=K'+1}^{k} \sqrt{2\eta_t \epsilon_t}.$$

Having achieved a bound on the dual iterates λ_{k+1} in terms of the dual maximizers $\lambda^*_{k'}$ ($k' \in \{K', \dots, k\}$), we can now aim to provide bounds for the latter. Indeed, we can leverage results from constrained convex optimization (Appendix A.4) to arrive at the following lemma.

Lemma 6. *Suppose Assumption 1 holds. Let $\nu \in (0,1)$ and choose K' as in Lemma 1. Let $k \in \{K', \dots, K\}$, and let (π^*_k, λ^*_k) be a pair of primal-optimal and dual-optimal solutions for Eq.* (6)*. Then, conditioned on G, we have*

$$\|\lambda^*_k\| \leq \|\lambda^*_k\|_1 \leq \frac{H}{\nu\gamma}.$$

Plugging Lemma 5 into the bounds from Lemma 4 and replacing the norms of the λ^*_k using the bound from Lemma 6, we obtain sublinear optimization errors when choosing η_k, ϵ_k correctly:

Lemma 7 (Optimization errors)**.** *Suppose Assumption 1 holds. Let $\nu \in (0,1)$ and choose K' as in Lemma 1. Suppose that the event G occurs. When using step sizes $\eta_{K'+k} = \Theta(k^{2.5})$ and $\epsilon_{K'+k} = \Theta(1/\eta_{K'+k})$, we have*

$$\sum_{k=K'+1}^{K} [V^{\pi_k}(\tilde{c}_k, \tilde{p}_k) - V^{\pi^*}(c,p)]_+ \leq \sum_{k=K'+1}^{K} \left(\frac{(O(\sigma k) + \sum_{t=K'+1}^{k} \sqrt{2\eta_t \epsilon_t})^2}{2\eta_k} + \epsilon_k \right) \leq O(\sqrt{K}),$$

$$\max_{i \in [I]} \sum_{k=K'+1}^{K} [V^{\pi_k}(\tilde{d}_{i,k}, \tilde{p}_k) - \alpha_i]_+ \leq \sum_{k=K'+1}^{K} \frac{O(\sigma k) + \sum_{t=K'+1}^{k} \sqrt{2\eta_t \epsilon_k}}{\eta_k} \leq O(\sqrt{K}),$$

where $\sigma = \frac{H}{\nu\gamma}$ and in fact $O(\sigma k)$ can be replaced by $(2 + 2(k - K'))\sigma$.

Remark: We need to choose η_k large enough (increasing) and ϵ_k small enough (decreasing) to ensure a sublinear error bound. At the same time, we do not want to choose η_k larger than necessary or ϵ_k smaller than necessary for computational reasons (see Proposition 1). We refer to Appendix D.3 for a discussion. In the case of an exact subroutine, we can plug in $\epsilon_k = 0$ to achieve an analogous result.

According to our regret decomposition, by adding up the errors of the pre-training phase, the estimation errors in the second phase, and the optimization errors in the second phase, we can indeed deduce our main result (Theorem 1). We showed how to bound the estimation errors using the optimism paradigm (Lemma 3). For the analysis of the optimization errors, we had to generalize the convergence analysis of the inexact augmented Lagrangian method (Lemma 7). We showed that if we have access to a safe baseline policy, the pre-training phase guarantees all assumptions required for this, and adds a constant term to the regret (Lemma 1).

5 Conclusion

In this work, we showed how to overcome the problem of the *cancellation of errors*, i.e., the oscillation of standard Lagrangian-based algorithms for CMDPs around an optimal safe policy. We did so by leveraging the augmented Lagrangian method to design an algorithm OPTAUG-CMDP. Unlike the related OPTDUAL-CMDP algorithm of Efroni et al. (2020), this requires a subroutine that solves a non-linear optimization problem in each episode. We devised an efficient algorithm for this, avoiding projections or LP. We then provided a regret analysis that, unlike previous works, does not require the cancellation of errors to arrive at sublinear regret guarantees. This means that in contrast to existing Lagrangian-based algorithms, our algorithm is provably safe *while exploring* the unknown CMDP.

This first partial answer to the open problem posed by Efroni et al. (2020) leads to several further questions: Can we obtain tighter bounds for the inner sub-routine and the regret, as our problem has a richer structure than the general convex optimization setup? While OPTAUG-CMDP enjoys stronger regret guarantees, the proposed inner subroutine has a higher computational cost than the one in OPTDUAL-CMDP, which may be possible to improve. Moreover, it remains open whether one can remove the requirement of access to a strictly feasible policy. Finally, we aim to extend our approach to the more practical function approximation setup.

References

Jacob D Abernethy and Jun-Kun Wang. On frank-wolfe and equilibrium computation. Advances in Neural Information Processing Systems, 30, 2017.

Eitan Altman. Constrained Markov decision processes, volume 7. CRC press, 1999.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. Advances in neural information processing systems, 21, 2008.

Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 3682–3689, 2022.

Amir Beck. First-order methods in optimization. SIAM, 2017.

Richard Bellman. A markovian decision process. Journal of Mathematics and Mechanics, 6(5):679–684, 1957. ISSN 00959057, 19435274. URL http://www.jstor.org/stable/24900506.

Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. Convex analysis and optimization, volume 1. Athena Scientific, 2003.

Dimitri P Bertsekas. Constrained optimization and Lagrange multiplier methods. Academic press, 2014.

Vivek S. Borkar. A convex analytic approach to markov decision processes. Probability Theory and Related Fields, 1988.

Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. Dope: Doubly optimistic and pessimistic exploration for safe reinforcement learning. Advances in Neural Information Processing Systems, 35:1047–1059, 2022.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. Advances in Neural Information Processing Systems, 30, 2017.

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. Advances in Neural Information Processing Systems, 33:8378–8390, 2020.

Dongsheng Ding, Kaiqing Zhang, Tamer Başar, and Mihailo R Jovanović. Convergence and optimality of policy gradient primal-dual method for constrained markov decision processes. In 2022 American Control Conference (ACC), pages 2851–2856. IEEE, 2022a.

Dongsheng Ding, Kaiqing Zhang, Jiali Duan, Tamer Başar, and Mihailo R Jovanović. Convergence and sample complexity of natural policy gradient primal-dual methods for constrained mdps. arXiv preprint arXiv:2206.02346, 2022b.

Yonathan Efroni, Nadav Merlis, Mohammad Ghavamzadeh, and Shie Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. Advances in Neural Information Processing Systems, 32, 2019.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. arXiv preprint arXiv:2003.02189, 2020.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. Naval research logistics quarterly, 3(1-2):95–110, 1956.

Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Olivier Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: the mean-field game viewpoint. arXiv preprint arXiv:2106.03787, 2021.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In International Conference on Machine Learning, pages 2681–2691. PMLR, 2019.

Magnus R Hestenes. Multiplier and gradient methods. Journal of optimization theory and applications, 4(5):303–320, 1969.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In International conference on machine learning, pages 427–435. PMLR, 2013.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial mdps with bandit feedback and unknown transition. arXiv preprint arXiv:1912.01192, 2019.

Vikram Krishnamurthy. Self Learning Control of Constrained Markov Decision Processes: a Gradient Approach. Groupe d'études et de recherche en analyse des décisions, 2003.

Vikram Krishnamurthy and Felisa J Vázquez Abad. Gradient based policy optimization of constrained markov decision processes. In Stochastic Processes, Finance and Control: A Festschrift in Honor of Robert J Elliott, pages 503–547. World Scientific, 2012.

Vikram Krishnamurthy and Felisa Vazquez Abad. Real-time reinforcement learning of constrained markov decision processes with weak derivatives. arXiv preprint arXiv:1110.4946, 2011.

Jingqi Li, David Fridovich-Keil, Somayeh Sojoudi, and Claire J Tomlin. Augmented lagrangian method for instantaneously constrained reinforcement learning problems. In 2021 60th IEEE Conference on Decision and Control (CDC), pages 2982–2989. IEEE, 2021a.

Tianjiao Li, Ziwei Guan, Shaofeng Zou, Tengyu Xu, Yingbin Liang, and Guanghui Lan. Faster algorithm and sharper analysis for constrained markov decision process. arXiv preprint arXiv:2110.10351, 2021b.

Tao Liu, Ruida Zhou, Dileep Kalathil, P. R. Kumar, and Chao Tian. Policy optimization for constrained mdps with provable fast global convergence, 2021a. URL `https://arxiv.org/abs/2111.00552`.

Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained mdps. Advances in Neural Information Processing Systems, 34:17183–17193, 2021b.

Songtao Lu. A single-loop gradient descent and perturbed ascent algorithm for nonconvex functional constrained optimization. In International Conference on Machine Learning, pages 14315–14357. PMLR, 2022.

Ted Moskovitz, Brendan O'Donoghue, Vivek Veeriah, Sebastian Flennerhag, Satinder Singh, and Tom Zahavy. Reload: Reinforcement learning with optimistic ascent-descent for last-iterate convergence in constrained mdps. arXiv preprint arXiv:2302.01275, 2023.

Mojmir Mutny, Tadeusz Janik, and Andreas Krause. Active exploration via experiment design in markov chains. In International Conference on Artificial Intelligence and Statistics, pages 7349–7374. PMLR, 2023.

Santiago Paternain, Luiz FO Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. arXiv preprint arXiv:1910.13393, 2019.

M. J. D. Powell. A method for nonlinear constraints in minimization problems. 1969.

R Tyrrell Rockafellar. Augmented lagrangians and applications of the proximal point algorithm in convex programming. Mathematics of operations research, 1(2):97–116, 1976.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. In International Conference on Machine Learning, pages 5478–5486. PMLR, 2019.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. In International Conference on Machine Learning, pages 9133–9143. PMLR, 2020.

Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. MIT press, 2018.

Jean Tarbouriech and Alessandro Lazaric. Active exploration in markov decision processes. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 974–982. PMLR, 2019.

Jean Tarbouriech, Shubhanshu Shekhar, Matteo Pirotta, Mohammad Ghavamzadeh, and Alessandro Lazaric. Active model estimation in markov decision processes. In Conference on Uncertainty in Artificial Intelligence, pages 1019–1028. PMLR, 2020.

Yangyang Xu. Iteration complexity of inexact augmented lagrangian methods for constrained convex programming. Mathematical Programming, 185(1):199–244, 2021.

Shen Yan and Niao He. Bregman augmented lagrangian method and its acceleration. Journal of Environmental Sciences (China) English Ed, 2020.

Donghao Ying, Yuhao Ding, and Javad Lavaei. A dual approach to constrained markov decision processes with entropy regularization. In International Conference on Artificial Intelligence and Statistics, pages 1887–1909. PMLR, 2022.

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps. Advances in Neural Information Processing Systems, 34:25746–25759, 2021.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In International Conference on Machine Learning, pages 7304–7312. PMLR, 2019.

Appendix

Table of Contents

A Background **13**
A.1 Review of the Augmented Lagrangian Method . . . 13
A.2 Review of OptDual-CMDP . . . 14
A.3 CMDPs and Occupancy Measures . . . 15
A.4 Convex Optimization Preliminaries . . . 16

B Solving the Inner Optimization Problem **17**
B.1 Derivation of the Policy Update via Frank-Wolfe . . . 17
B.2 Pseudocode for the Inner Problem . . . 19

C On-Policy Error Bounds **20**

D Omitted Proofs for Section 3 **24**
D.1 Pre-Training Phase . . . 24
D.2 Iteration Complexity of the Inner Method . . . 25
D.3 Main Result . . . 26

E Omitted Proofs for Section 4 **28**
E.1 Preliminaries . . . 28
E.2 Omitted Proofs for Section 4.1 . . . 30
E.3 Omitted Proofs for Section 4.2 . . . 31

A Background

A.1 Review of the Augmented Lagrangian Method

In this section, we review the fundamentals of the augmented Lagrangian method (Rockafellar, 1976; Bertsekas, 2014), which was first introduced by Hestenes (1969); Powell (1969). Our review is partly inspired by the review by Yan and He (2020).

Consider a closed and convex set $\mathcal{X} \subset \mathbb{R}^n$ and a closed convex function $f\colon \mathcal{X} \to \mathbb{R}$. Furthermore, let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. With this, consider the constrained problem

$$\begin{aligned} \min_{x\in\mathcal{X}} \quad & f(x) \\ \text{s.t.} \quad & Ax \leq b. \end{aligned} \tag{7}$$

After initializing $x_0 \in \mathcal{X}$ and $\lambda_0 \in \mathbb{R}^m_{\geq 0}$, the augmented Lagrangian method performs the following updates in step $k \geq 0$

$$x_{k+1} \in \arg\min_{x\in X} \left(f(x) + \frac{1}{2\eta_k} \|[\lambda_k + \eta_k(Ax - b)]_+\|^2 \right), \tag{8}$$

$$\lambda_{k+1} = [\lambda_k + \eta_k(Ax_{k+1} - b)]_+. \tag{9}$$

It can easily be verified that the augmented Lagrangian method is the proximal point method applied to the Lagrangian dual. Thus, with the Lagrangian $\mathcal{L}(x, \lambda) := f(x) + \lambda^T(Ax - b)$, we have

$$(x_{k+1}, \lambda_{k+1}) \in \arg\max_{\lambda\geq 0} \min_{x\in\mathcal{X}} \left(\mathcal{L}(x, \lambda) - \frac{1}{2\eta_k} \|\lambda - \lambda_k\|^2 \right).$$

Xu (2021) shows a non-asymptotic convergence result for the augmented Lagrangian method, including the case of an inexact subroutine for solving Eq. (8). Notably, the convergence result is concerning the last iterate of the method. That is, if f^* is the optimal value of Eq. (7), then $f(x_k) \to f^*$ as $k \to \infty$, and the convergence rate is determined by the step sizes η_k (and, in the case of an inexact subroutine, by the accuracy parameter ϵ_k).

Alternatively, we could apply the dual projected gradient method to Eq. (7). The updates would then read

$$\tilde{x}_{k+1} \in \arg\min_{x \in X} \left(f(x) + \tilde{\lambda}_k^T (Ax - b) \right), \tag{10}$$

$$\tilde{\lambda}_{k+1} = [\tilde{\lambda}_k + \tilde{\eta}_k (A\tilde{x}_{k+1} - b)]_+. \tag{11}$$

While Eq. (11) coincides with Eq. (9), the update of the primal variable differs from the one in the augmented Lagrangian method. We can view the dual projected gradient method as iterative play between a primal player x_k and a dual player λ_k in a min-max setup, where the objective is the Lagrangian $\mathcal{L}(x, \lambda)$. The primal player here plays best response, while the dual player plays online projected gradient ascent. While this is similar in spirit to applying the proximal point method to the Lagrangian dual, the known non-asymptotic convergence guarantees for this method are ergodic, i.e., they only concern convergence of the averaged iterates. Indeed, simple simulations show that this is not a weakness in the analysis, but that indeed the primal and dual iterates oscillate around an optimal solution pair. This is illustrated in (Beck, 2017, Chapter 8). The iterates of the dual projected gradient method may alternate between satisfying $f(x_k) > f^*$ and $Ax_k < b$ for a couple of iterations and then $f(x_k) < f^*$ and $Ax_k > b$ for a couple of iterations. While the average objective value converges to f^* and the average constraint violation to 0, this is not the case for the individual iterates. Note that these oscillations are not due to estimating the problem, but are present even if the optimization problem is fixed, as in the setup above. Apart from the augmented Lagrangian method, other methods such as extra gradient or optimistic gradient descent-ascent offer solutions to this issue.

A.2 Review of OptDual-CMDP

In this section, we review the related OptDual-CMDP algorithm of Efroni et al. (2020). This model-based dual algorithm is based on the dual projected gradient method (see Appendix A.1) rather than on the augmented Lagrangian method (but builds the model with the same notion of optimism).

Algorithm 2 OptDual-CMDP

Set $\eta_k := \sqrt{\frac{\rho^2}{H^2 IK}}$ and $\lambda_1 := 0 \in \mathbb{R}^I$
for $k = 1, \ldots, K$ **do**
Update policy:

$$\pi_k, \tilde{p}_k := \arg\min_{\substack{\pi \in \Pi \\ p' \in B_k^p}} \left(V^\pi(\tilde{c}_k, p') + \lambda_k^T (V^\pi((\tilde{d}_{i,k})_{i \in [I]}, p') - \alpha) \right) \tag{12}$$

Update dual variables:

$$\lambda_{k+1} := [\lambda_k + \eta_k (V^{\pi_k}((\tilde{d}_{i,k})_{i \in [I]}, \tilde{p}_k) - \alpha)]_+ \tag{13}$$

Play π_k, update estimates of the costs $\tilde{c}_{k+1}$, $(\tilde{d}_{i,k+1})_{i \in [I]}$ and transitions B_{k+1}^p (Eq. (3)).

Dual approaches like the algorithm above turn the CMDP into a series of linearly regularized, (extended) *unconstrained* MDPs (here with objective $V^\pi(\tilde{c}_k, p') + \eta_k \lambda_k^T (V^\pi((\tilde{d}_{i,k})_{i \in [I]}, p') - \alpha)$), that can be solved efficiently with DP. With the augmented Lagrangian approach, the inner problem (Eq. (4)) has a more complicated structure. In Appendix B, we show that the inner problem can be reformulated as a convex optimization problem and provide an efficient method based on Frank-Wolfe and DP to solve it.

For OptDual-CMDP, we have the following guarantee (Efroni et al., 2020, Theorem 5).

Theorem 2. *Suppose there exists a strictly feasible policy π such that for all $i \in [I]$ we have $V^\pi(d_i, p) < \alpha_i$. Set*

$$\rho := \frac{V^\pi(c, p) - V^{\pi^*}(c, p)}{\min_{i \in [I]} (\alpha_i - V^\pi(d_i, p))}.$$

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, OptDual-CMDP *achieves the following regret bounds:*

$$\mathcal{R}_{\pm}(K;c) = \tilde{O}\left(\sqrt{S\mathcal{N}H^4K} + \rho\sqrt{H^2IK} + (\sqrt{\mathcal{N}}+H)H^2SA\right),$$

$$\mathcal{R}_{\pm}(K;d) = \tilde{O}\left((1+\frac{1}{\rho})(\sqrt{IS\mathcal{N}H^4K} + (H\sqrt{\mathcal{N}}+S)\sqrt{I}H^2SA)\right).$$

Notably, this bound only covers the *weak* regret. This is because the dual projected gradient method does not allow for a non-ergodic convergence analysis, and their iterates will generally oscillate around an optimal feasible solution. It is worth mentioning that unlike LP-based approaches (OptCMDP and OptCMDP-bonus of Efroni et al. (2020)), the related primal-dual method OptPrimalDual-CMDP of Efroni et al. (2020) suffers from the same problem as OptDual-CMDP. Moreover, we remark that this is not just a hypothetical issue, but that indeed Lagrangian-based algorithms suffer from the mentioned oscillations in practical applications (Stooke et al., 2020; Moskovitz et al., 2023).

A.3 CMDPs and Occupancy Measures

We summarize the relevant quantities of the CMDP $\mathcal{M}$ as follows.

Discrete state space	$\mathcal{S}$, with cardinality S
Discrete action space	$\mathcal{A}$, with cardinality A
# of constraints	I
Initial state	s_1, same for each episode
Time horizon	H, same for each episode
Transition probability	$P[s_{h+1}=s' \mid s_h=s, a_h=a] = p_h(s'\|s,a)$
Max. # transitions	$\mathcal{N} := \max_{s,a,h} \lvert\{s' \mid p_h(s'\|s,a)>0\}\rvert$
Objective cost	Random variable $C_h(s,a) \in [0,1]$, with $\mathbb{E}[C_h(s,a)] = c_h(s,a)$
Constraint cost ($i \in [I]$)	Random variable $D_{i,h}(s,a)$ with $\mathbb{E}[D_{i,h}(s,a)] = d_{i,h}(s,a)$
Constraint encoding	$(d_i, \alpha_i)_{i\in[I]}$, with $d_i = (d_{i,h})_{h\in[H]}$ and $\alpha_i \in [0,H]$
Policy	$\pi = (\pi_1,\dots,\pi_H) \in \Pi$ with $\pi_h \colon \mathcal{S} \to \Delta(\mathcal{A})$ (non-stationary)
Value function	$V^\pi(s;c,p) = \mathbb{E}[\sum_{h=1}^H c_h(s_h,a_h) \mid s_1,p,\pi]$
Constraint value function ($i \in [I]$)	$V^\pi(s;d_i,p) = \mathbb{E}[\sum_{h=1}^H d_{i,h}(s_h,a_h) \mid s_1,p,\pi]$

Table 1: Summary of CMDP notation

To view the CMDP as a convex optimization problem, we will express it via the common notion of occupancy measures (Borkar, 1988).

Definition 1. *The state-action occupancy measure q^π of a policy π for a CMDP $\mathcal{M}$ is defined as*

$$q_h^\pi(s,a;p) := \mathbb{E}\left[\mathbb{1}_{\{s_h=s,a_h=a\}} \mid s_1,p,\pi\right] = P[s_h=s,a_h=a \mid s_1,p,\pi],$$

for $s\in\mathcal{S}$, $a\in\mathcal{A}$, $h\in[H]$. We denote the stacked vector of these values as $q^\pi(p)\in\mathbb{R}^{SAH}$, with the element at index (s,a,h) being $q_h^\pi(s,a;p)$.

For transition probabilities p', we can now define

$$Q(p') := \left\{q^\pi(p') \in \mathbb{R}^{SAH} \mid \pi\in\Pi\right\}$$

as the *state-action occupancy measure polytope*. Note that $Q(p')$ is indeed a polytope (Altman, 1999; Efroni et al., 2020). Moreover, for any p' we have a surjective map $\pi \mapsto q^\pi(p')$ between Π and $Q(p')$, for which we can explicitly compute an element in the pre-image of $q\in Q(p')$ via $\pi_h(s|a) = q_h(s,a)/(\sum_{a'} q_h(s,a'))$.

We can stack the expected costs $c_h(s,a)$ and constraint costs $d_{ih}(s,a)$ in the same way as $q_h(s,a)$ to obtain vectors $c\in\mathbb{R}^{SAH}$ and $d_i\in\mathbb{R}^{SAH}$. Note that we then have $V^\pi(c,p) = \sum_{h,s,a} q_h^\pi(s,a;p)c_h(s,a) = c^T q^\pi(p)$ by linearity of expectation. Similarly, for all $i\in[I]$, we have $V^\pi(d_i,p) = d_i^T q^\pi(p)$. Moreover, if we stack $D=(d_i)_{i\in[I]}\in\mathbb{R}^{I\times SAH}$ and $\alpha=(\alpha_i)_{i\in[I]}\in\mathbb{R}^I$ as

$$D := \begin{pmatrix} d_1^T \\ \vdots \\ d_I^T \end{pmatrix}, \qquad \alpha := \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_I \end{pmatrix},$$

we obtain $V^\pi(D,p) = Dq^\pi(p) \in [0,H]^I$ for the vector of the constraint value functions. We can thus write

$$\pi^* \in \arg\min_{\pi\in\Pi} \quad V^\pi(c,p) \quad \text{s.t.} \quad V^\pi(d_i,p) \le \alpha_i \quad (\forall i \in [I])$$

equivalently as

$$q^{\pi^*} \in \arg\min_{q^\pi \in Q(p)} \quad c^T q^\pi(p) \quad \text{s.t.} \quad Dq^\pi(p) \le \alpha,$$

which is an LP. In particular, by compactness of the state-action occupancy polytope and continuity of the objective, there exists an optimal solution π^* if we assume feasibility.

A.4 Convex Optimization Preliminaries

We state some well-known results from constrained convex optimization that will be useful to bound the dual iterates λ_k appearing in Lemma 4. The results are standard and we refer, for example, to the work by Beck (2017).

Consider the (primal) optimization problem

$$\begin{aligned} f^* := \min \quad & f(x) \\ \text{s.t.} \quad & g(x) \le 0 \\ & x \in X \end{aligned} \tag{14}$$

with the following assumptions.

Assumption 2 (Assumption 8.41, Beck (2017))**.** *In Eq.* (14)*,*

(a) $X \subset \mathbb{R}^n$ *is convex*

(b) $f\colon \mathbb{R}^n \to \mathbb{R}$ *is convex*

(c) $g(\cdot) := (g_1(\cdot), \dots, g_m(\cdot))^T$ *with* $g_i\colon \mathbb{R}^n \to \mathbb{R}$ *convex*

(d) Eq. (14) *has a finite optimal value* f^**, which is attained by exactly the elements of* $X^* \neq \emptyset$

(e) There exists $\bar{x} \in X$ *such that* $g(\bar{x}) < 0$

(f) For all $\lambda \in \mathbb{R}^m_{\ge 0}$*,* $\min_{x\in X}(f(x) + \lambda^T g(x))$ *has an optimal solution*

In this setup, we define the *dual objective* as

$$q(\lambda) := \min_{x\in X} \left(f(x) + \lambda^T g(x)\right),$$

where $\mathcal{L}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $\mathcal{L}(x;\lambda) := f(x) + \lambda^T g(x)$ is the *Lagrangian* of the problem in Eq. (14). The *dual problem* is then defined as

$$\begin{aligned} q^* := \max \quad & q(\lambda) \\ \text{s.t.} \quad & \lambda \ge 0. \end{aligned}$$

In this setup, we have the following results connecting the primal and the dual problem.

Theorem 3 (Theorem A.1, Beck (2017))**.** *Under Assumption 2, strong duality holds in the following sense: We have*

$$f^* = q^*$$

and the optimal solution of the dual problem is attained, with the set of optimal solutions $\Lambda^* \neq \emptyset$*.*

Proof. Proposition 6.4.4 of Bertsekas et al. (2003) gives a proof of the more general Theorem A.1 of Beck (2017). We remark that if we assume affine constraints g and X being a polytope, then we can drop assumption (e) (Beck, 2017, Theorem A.1). □

Theorem 4. *Suppose Assumption 2 holds. Let* $x^* \in X^*$*,* $\lambda^* \in \Lambda^*$ *and* $x \in X$*. Then*

$$f(x) - f(x^*) + (\lambda^*)^T g(x) \ge 0.$$

Proof. We have

$$\begin{aligned} f(x) =& f(x) + (\lambda^*)^T g(x) - (\lambda^*)^T g(x) \\ \geq& q(\lambda^*) - (\lambda^*)^T g(x) && \text{(definition of } q(\cdot)) \\ =& f(x^*) - (\lambda^*)^T g(x) && \text{(since by Theorem 3, } q^* = f^*) \end{aligned}$$

and rearranging this proves the claim. Again, we see that we can drop assumption (e) if we consider affine constraints g and a polytope X. □

Theorem 5. *Under Assumption 2, for all $\lambda^* \in \Lambda^*$ and $\bar{x}$ as in (e), we have*

$$\|\lambda^*\| \leq \|\lambda^*\|_1 \leq \frac{f(\bar{x}) - f^*}{\min_{i\in[m]}(-g_i(\bar{x}))}.$$

Proof. The first relation holds since $\lambda^* \geq 0$. We show the second relation as follows (cf. Beck (2017, Theorem 8.42)). We have

$$\begin{aligned} f(x^*) =& q(\lambda^*) && \text{(Theorem 3)} \\ \leq& f(\bar{x}) + (\lambda^*)^T g(\bar{x}) && \text{(definition of } q(\cdot)) \\ \leq& f(\bar{x}) + \|\lambda^*\|_1 \max_{i\in[m]} g_i(\bar{x}) && \text{(since } \lambda^* \geq 0) \\ =& f(\bar{x}) - \|\lambda^*\|_1 \min_{i\in[m]}(-g_i(\bar{x})) \end{aligned}$$

and rearranging this proves the claim. We remark that for this theorem, we do need assumption (e) even in the affine case. □

B Solving the Inner Optimization Problem

There are numerous works leveraging Frank-Wolfe schemes for RL (including planning) with convex objectives, most commonly in the context of pure/active exploration (Hazan et al., 2019; Tarbouriech and Lazaric, 2019; Tarbouriech et al., 2020; Mutny et al., 2023).

To the best of our knowledge, only Tarbouriech and Lazaric (2019, Appendix A.3) remark that a combination of Frank-Wolfe UCB and planning in extended MDPs would be possible in the presence of a convex objective and plausible transitions. Tarbouriech et al. (2020) follow this approach but solve an extended LP rather than an extended MDP in each Frank-Wolfe iteration. Therefore, we make the former idea explicit and show how we can devise an efficient algorithm for our case.

B.1 Derivation of the Policy Update via Frank-Wolfe

In the following, we provide an efficient algorithm for solving the inner optimization problem in Eq. (4) based on the extended LP formulation of CMDPs, Frank-Wolfe, and DP. Recall the first update in OptAug-CMDP, when rewriting the value functions in terms of occupancy measures (see Appendix A.3):

$$\pi_k, \tilde{p}_k := \arg\min_{\substack{\pi\in\Pi \\ p'\in B_k^p}} \left(\tilde{c}_k^T q^\pi(p') + \frac{1}{2\eta_k} \left\| [\lambda_k + \eta_k(\tilde{D}_k q^\pi(p') - \alpha)]_+ \right\|^2 \right), \tag{15}$$

where $q^\pi(p') \in \mathbb{R}^{SAH}$, $\tilde{D}_k \in \mathbb{R}^{I\times SAH}$ and $\tilde{c}_k \in \mathbb{R}^{SAH}$ are defined as in Appendix A.3.

We first use the extended LP trick (Rosenberg and Mansour, 2019; Efroni et al., 2020) to switch to a convex optimization problem in the state-action-state occupancy measure. That is, in the problem above substitute

$$z_h(s,a,s') := z_h^\pi(s,a,s';p') := p'_h(s'|s,a) q_h^\pi(s,a;p') \tag{16}$$

and note that

$$q_h^\pi(s,a;p') = \sum_{s'} z_h^\pi(s,a,s';p'). \tag{17}$$

Stacked across $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, this simply reads

$$q^\pi(p') = \sum_{s'} z^\pi(s'; p') \in \mathbb{R}^{SAH},$$

for all $s' \in \mathcal{S}$. The objective of Eq. (15) then reads

$$f(z) := \sum_{s'} \tilde{c}_k^T z(s') + \frac{1}{2\eta_k} \|[\lambda_k + \eta_k(\sum_{s'} \tilde{D}_k z(s') - \alpha)]_+\|^2,$$

and we need to minimize it over the set $Z \subset \mathbb{R}^{S^2AH}$ which is given by the constraints

$$\begin{cases} \sum_{a,s'} z_h(s,a,s') = \sum_{s',a'} z_{h-1}(s',a',s) & (\forall h > 1, s) \\ \sum_{a,s'} z_1(s,a,s') = \mu(s) & (\forall s) \\ z_h(s,a,s') \geq 0 & (\forall s, a, s', h) \\ z_h(s,a,s') - (\bar{p}_h^{k-1}(s'|s,a) + \beta_{k,h}^p(s,a,s')) \sum_{s''} z_h(s,a,s'') \leq 0 & (\forall s, a, s', h) \\ -z_h(s,a,s') + (\bar{p}_h^{k-1}(s'|s,a) - \beta_{k,h}^p(s,a,s')) \sum_{s''} z_h(s,a,s'') \leq 0 & (\forall s, a, s', h), \end{cases}$$

where $\mu(s) = 1$ if $s = s_1$ and 0 otherwise. Note that Z is a bounded polytope and thus compact and convex. We can thus equivalently solve

$$\min_{z \in Z} f(z), \tag{18}$$

which is a standard convex optimization problem. Note that even with the convex formulation in Eq. (18), Eq. (4) cannot be written as an LP due to the nonlinear objective. It cannot be rewritten as a standard convex RL problem (Zahavy et al., 2021; Geist et al., 2021) either since the transition probabilities are part of the optimization and we thus had to switch to the extended convex program in the state-action-state occupancy measure z. Note that for efficiency reasons, we aim to avoid LP (which is needed for solving CMDPs due to the lack of the Bellman optimality principle), and projections onto the high-dimensional constraint set Z. Instead, we will make use of DP and gradient-based methods.

If we can approximately solve Eq. (18), we can later retrieve transitions and policy via

$$\tilde{p}_{k,h}(s'|s,a) := \frac{z_h(s,a,s')}{\sum_{s''} z_h(s,a,s'')}, \tag{19}$$

$$\pi_{k,h}(a|s) := \frac{\sum_{s'} z_h(s,a,s')}{\sum_{a',s'} z_h(s,a',s')}. \tag{20}$$

To solve Eq. (18), we make use of Frank-Wolfe. Frank-Wolfe is a well-known iterative method for constrained nonlinear optimization, which allows to minimize a smooth convex function over a convex domain and avoids projections. To apply it to Eq. (18), we need a linear minimization oracle (LMO) that in step t (of episode k) solves

$$\min_{g \in Z} g^T \nabla f(z^t) \tag{21}$$

and then, after finding a minimizer g, updates

$$z^{t+1} := (1 - \gamma_t) z^t + \gamma_t g,$$

where $\gamma_t := 2/(t+2)$. It is well known that Frank-Wolfe converges with a rate of $O(1/T)$ for smooth objectives. This is given here, but the smoothness parameter depends on η_k. We refer to Abernethy and Wang (2017); Jaggi (2013); Frank and Wolfe (1956) for the relevant background.

It remains to show that we can provide an efficient LMO that uses DP instead of LP and avoids projections onto Z. Let $(\nabla f(z))(s,a,s',h) := \frac{\partial f}{\partial z_h(s,a,s')}(z)$ be the gradient of f with respect to z at index (s,a,s',h). Then the s'-th component of the gradient of f with respect to z is

$$(\nabla f(z))(\cdot,\cdot,s',\cdot) = \tilde{c}_k + \tilde{D}_k^T [\lambda_k + \eta_k(\tilde{D}_k \sum_{s''} z(s'') - \alpha)]_+ \in \mathbb{R}^{SAH}, \tag{22}$$

which we can compute explicitly and efficiently. We note that this gradient does not depend on s' and thus the gradient of f with respect to the whole vector $z \in \mathbb{R}^{S^2AH}$ is simply the vector above, repeatedly stacked S times.

We now show how switching back to the optimization over $\Pi \times B_k^p$ allows for an efficient LMO via DP. For $g \in Z$, there are π, p' such that

$$g_h(s,a,s') = g_h^\pi(s,a,s';p') = p_h(s'|a,s)q_h^\pi(s,a;p),$$

via Eqs. (19) and (20). The LMO then needs to minimize

$$\begin{aligned} g^T \nabla f(z^t) &= \sum_{s,a,s',h} g_h(s,a,s')(\nabla f(z^t))(s,a,1,h) \\ &= \sum_{s,a,s',h} p'_h(s'|a,s)q_h^\pi(s,a;p')(\nabla f(z^t))(s,a,1,h) \\ &= \sum_{s,a,h} q_h^\pi(s,a;p')(\nabla f(z^t))(s,a,1,h) \end{aligned} \tag{23}$$

over $\pi \in \Pi$ and $p' \in B_k^p$. This corresponds to solving the extended MDP $\mathcal{M}^+ := \{(M = (\mathcal{S}, \mathcal{A}, r^+, p^+)) \mid \forall s,a,h\colon r_h^+(s,a) := \nabla f(z^t)(s,a,1,h),\ p_h^+(\cdot|s,a) \in B_{k,h}^p(s,a)\}$. We can do so via backward induction (i.e., DP) that optimizes

$$Q_h^k(s,a) := r_h^+(s,a) + \min_{p'(\cdot|s,a) \in B_{k,h}^p(s,a)} \sum_{s'} p'(s'|s,a) \min_{a'} Q_{h+1}^k(s',a') \tag{24}$$

and starts with $Q_{H+1}^k(s,a) = 0$. We can retrieve the transitions and the policy by storing the minimizers in each step of the DP. To compute the solution g of the LMO from π, p', one can then use a simple and efficient DP scheme to compute $q_h^\pi(p')$, which is explained in Appendix B.2. We can use this, in turn, to retrieve g using the substitution in Eq. (16). Then, we perform the second step of Frank-Wolfe to get a convex combination of g and z^t, which concludes the Frank-Wolfe iteration.

We can already see that the computational complexity of the k-th step of OptAug-CMDP is larger than the one of OptDual-CMDP by a factor of $O(1/\epsilon_k)$ (and a dependency on η_k) because we need this many Frank-Wolfe iterations, which is the price we pay for the stronger regret bound. We provide a full analysis of the iteration complexity in Appendix D.2.

B.2 Pseudocode for the Inner Problem

More formally, the algorithm for solving the inner problem (Eq. (4)) reads as follows.

Algorithm 3 InnerOpt-FW

Input: Current estimates $\tilde{c}_k$, $\tilde{D}_k$, B_k^p, and λ_k
Output: Next policy π_k and $\tilde{p}_k$ according to Eq. (15)

Set $T = \frac{2\eta_k I S^2 AH}{\epsilon_k}$, $\gamma_t = 2/(2+t)$
Initialize $z^0 \in Z$ arbitrarily
for $t = 0, \dots, T-1$ **do**
  Compute $\nabla f(z^t)$ according to Eq. (22)
  Minimize Eq. (23) via DP algorithm from Eq. (24) to find π^t, p^t
  Retrieve minimizer g^t from π^t, p^t via DP algorithm from Eq. (26) and Eq. (16)
  Set $z^{t+1} := \gamma_t g^t + (1-\gamma_t) z^t$
Construct π_k, $\tilde{p}_k$ from z^T via Eqs. (19) and (20)
return π_k, $\tilde{p}_k$

As a final step, we now describe how to retrieve $q := q^\pi(p')$ from π and p' in Line 6 (Jin et al., 2019). For $s \in \mathcal{S}$, set $q_h(s) := P[s_h = s|s_1; p', \pi] = \sum_{a' \in \mathcal{A}} q_h(s,a')$. Given a policy π and transition probabilities p', we need to compute

$$q_h(s,a) = q_h(s) \cdot \pi_h(a|s), \tag{25}$$

for every $h \in [H]$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$. This is easily achieved via DP. Indeed, we have

$$q_1(s) = \begin{cases} 1 & (s = s_1) \\ 0 & (\text{else}), \end{cases}$$

and for $h > 1$

$$\begin{aligned} q_h(s) =& P[s_h = s | s_1; p', \pi] \\ =& \sum_{s' \in \mathcal{S}} P[s_h = s | s_{h-1} = s'; s_1, p', \pi] P[s_{h-1} = s' | s_1; p'] \\ =& \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} \pi_{h-1}(a|s') p_{h-1}(s|s', a) P[s_{h-1} = s' | s_1, ; p'] \\ =& \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} \pi_{h-1}(a|s') p_{h-1}(s|s', a) q_{h-1}(s'), \end{aligned} \tag{26}$$

which together with Eq. (25) enables us to retrieve the state-action occupancy measure. Clearly, the above DP scheme can be performed efficiently.

C On-Policy Error Bounds

We consider arbitrary polices $(\pi_k)_{k \in [K]}$. We suppose that in the CMDP $\mathcal{M}$, the agent plays π_k in episode $k \in [K]$ and uses it to update the optimistic model according to Eq. (3), with some fixed $\delta = 3\delta' > 0$.

We first establish Lemmas 8 to 10, which will allow us to bound the estimation errors (Lemma 3). For a definition of the occupancy measure $q^\pi(s, a; p)$, see Appendix A.3. We write $\lesssim$ for an inequality up to polylogarithmic factors.

Lemma 8 (Lemma 36, Efroni et al. (2020))**.** *Suppose for all s, a, h, $k \in [K]$ we have*

$$n_h^{k-1}(s, a) > \frac{1}{2} \sum_{j<k} q_h^{\pi_j}(s, a; p) - H \log\left(\frac{SAH}{\delta'}\right).$$

Then for all $K' \leq K$

$$\sum_{k'=1}^{K'} \sum_{h=1}^{H} \mathbb{E}\left[\frac{1}{\sqrt{n_h^{k'-1}(s_h^{k'}, a_h^{k'})}} \mid \mathcal{F}_{k'-1}\right] \leq \tilde{O}(\sqrt{SAHK'} + SAH),$$

where $\mathcal{F}_{k'-1}$ is the σ-algebra induced by all random variables up to and including episode $k' - 1$.

Proof. We refer to Efroni et al. (2019, Lemma 38) for a proof of the statement. □

Lemma 9 (Lemma 37, Efroni et al. (2020))**.** *Suppose for all s, a, h, $k \in [K]$ we have*

$$n_h^{k-1}(s, a) > \frac{1}{2} \sum_{j<k} q_h^{\pi_j}(s, a; p) - H \log\left(\frac{SAH}{\delta'}\right).$$

Then for all $K' \leq K$

$$\sum_{k'=1}^{K'} \sum_{h=1}^{H} \mathbb{E}\left[\frac{1}{n_h^{k'-1}(s_h^{k'}, a_h^{k'})} \mid \mathcal{F}_{k'-1}\right] \leq \tilde{O}(SA),$$

where $\mathcal{F}_{k'-1}$ is the σ-algebra induced by all random variables up to and including episode $k' - 1$.

Proof. We refer to Zanette and Brunskill (2019, Lemma 13) for a proof of the statement. □

The following lemma provides an on-policy error bound based on the value difference lemma. Together with the preliminaries from Appendix E.1, it allows us to bound the estimation error.

Lemma 10 (On-policy errors; Lemma 29, Efroni et al. (2020))**.** *Consider an MDP with transition dynamics p and arbitrary estimated transition dynamics $\hat{p}_k$ (for $k \in [K]$, each forming a conditional probability measure). Consider policy iterates $(\pi_k)_{k\in[K]}$ and suppose π_k is played in episode $k \in [K]$ and used to update the counters. Let $l_h(s,a)$, $\tilde{l}_{k,h}(s,a)$ be the cost and corresponding optimistic cost with $l = c$ or $l = d_i$ as discussed in Eq.* (3)*. For a policy π, let $V_h^\pi(s;l,p)$, $V_h^\pi(s;\tilde{l}_k,\hat{p}_k)$ be the values of π according to the true resp. estimated model. Assume that for all s,a,h,k we have*

$$n_h^{k-1}(s,a) > \frac{1}{2}\sum_{j<k} q_h^{\pi_j}(s,a;p) - H\log\left(\frac{SAH}{\delta'}\right), \tag{a}$$

$$|\tilde{l}_{k,h}(s,a) - l_h(s,a)| \le \tilde{O}\left(\frac{1}{\sqrt{n_h^{k-1}(s,a)\vee 1}}\right), \tag{b}$$

$$|\hat{p}_{k,h}(s'|s,a) - p_h(s'|s,a)| \le \tilde{O}\left(\sqrt{\frac{p_h(s'|s,a)L_\delta^p}{n_h^{k-1}(s,a)\vee 1}} + \frac{L_\delta^p}{n_h^{k-1}(s,a)\vee 1}\right). \tag{c}$$

Then we have

$$\sum_{k=1}^K |V^{\pi_k}(l,p) - V^{\pi_k}(\tilde{l}_k,\hat{p}_k)| \le \tilde{O}\left(\sqrt{\mathcal{N}SAH^3K} + (H\sqrt{\mathcal{N}} + S)SAH\right). \tag{27}$$

Proof. From the value difference lemma (Dann et al., 2017, Lemma E.15), we get

$$\begin{aligned}
&\sum_{k=1}^K |V^{\pi_k}(l,p) - V^{\pi_k}(\tilde{l}_k,\hat{p}_k)| \\
\le &\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}[|l_h(s_h,a_h) - \tilde{l}_{k,h}(s_h,a_h)| \mid s_1,p,\pi_k] \\
+ &\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}\Big[\sum_{s'} |p_h(s|s_h,a_h) - \hat{p}_k(s|s_h,a_h)| V_{h+1}^{\pi_k}(s';\tilde{l}_k,\hat{p}_k) \mid s_1,p,\pi_k\Big].
\end{aligned}$$

We can bound the first of the two terms as follows.

$$\begin{aligned}
&\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}[|l_h(s_h,a_h) - \tilde{l}_{k,h}(s_h,a_h)| \mid s_1,p,\pi_k] \\
\overset{(b)}{\lesssim} &\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}\Big[\frac{1}{\sqrt{n_h^{k-1}(s_h,a_h)\vee 1}} \mid s_1,p,\pi_k\Big] \\
= &\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}\Big[\frac{1}{\sqrt{n_h^{k-1}(s_h^k,a_h^k)\vee 1}} \mid \mathcal{F}_{k-1}\Big] \\
\overset{\text{Lemma 8}}{\le} &\tilde{O}\left(\sqrt{SAHK} + SAH\right),
\end{aligned}$$

where the second to last inequality holds since π_k is played in episode k, and Lemma 8 applies due to assertion (a). For the second term, we first note that $|V_{h+1}^{\pi_k}(s';\tilde{l}_k,\hat{p}_k)| \lesssim H$ since $|\tilde{l}_{k,h}(s,a)| \le$

$l_h(s,a) + \frac{1}{\sqrt{n_h^{k-1}(s,a)\vee 1}}$ by (b). Hence

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\Big[\sum_{s'}|p_h(s|s_h,a_h)-\hat{p}_k(s|s_h,a_h)|V_{h+1}^{\pi_k}(s';\tilde{l}_k,\hat{p}_k)\mid s_1,p,\pi_k\Big]\\
&\overset{(c)}{\lesssim} H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\Big[\frac{1}{\sqrt{n_h^{k-1}(s_h,a_h)\vee 1}}\big(\sum_{s'}\sqrt{p_h(s'|s_h,a_h)}\big)+\frac{S}{n_h^{k-1}(s_h,a_h)\vee 1}\mid s_1,p,\pi_k\Big]\\
&\overset{\text{Jensen}}{\leq} H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\Big[\frac{1}{\sqrt{n_h^{k-1}(s_h,a_h)\vee 1}}\sqrt{\mathcal{N}}\sqrt{\sum_{s'}p_h(s'|s_h,a_h)}+\frac{S}{n_h^{k-1}(s_h,a_h)\vee 1}\mid s_1,p,\pi_k\Big]\\
&= H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\Big[\frac{\sqrt{\mathcal{N}}}{\sqrt{n_h^{k-1}(s_h^k,a_h^k)\vee 1}}+\frac{S}{n_h^{k-1}(s_h^k,a_h^k)\vee 1}\mid \mathcal{F}_{k-1}\Big]\\
&\lesssim H\sqrt{\mathcal{N}}\cdot\sqrt{SAHK}+H\sqrt{\mathcal{N}}\cdot SAH+HS\cdot SA\\
&= \sqrt{\mathcal{N}SAH^3K}+(H\sqrt{\mathcal{N}}+S)SAH,
\end{aligned}$$

where the second to last relation holds due to Lemma 8 and Lemma 9 (which in turn apply due to assertion (a)), and the one before holds since π_k is played in episode k. □

Next, we prove Lemma 13, which is a variation of Lemma 10 and will allow us to establish the bound on the pre-training duration in Lemma 1. For this, we first need to establish Lemmas 11 and 12, which are a consequence of Lemmas 8 and 9, respectively.

Lemma 11. *Let $K' \in [K]$. Assume that for all s, a, h, $k \in [K]$ we have*

$$n_h^{k-1}(s,a) > \frac{1}{2}\sum_{j<k} q_h^{\pi_j}(s,a;p) - H\log\left(\frac{SAH}{\delta'}\right).$$

Let π be any fixed policy and suppose that in episode $k \in [K]$, policy π_k is played, with $\pi_{k'} = \pi$ for all $k' \in [K']$. Then for all $k \geq K'$ we have

$$\sum_{h=1}^{H}\mathbb{E}\left[\frac{1}{\sqrt{n_h^{k-1}(s_h,a_h)\vee 1}}\mid s_1,\pi,p\right] \leq \tilde{O}\left(\sqrt{SAH}(K')^{-1/2}+SAH(K')^{-1}\right).$$

Proof. Note that for any realization of all random variables up to and including episode K, we have (for $k \in \{0,\dots,K-1\}$)

$$n_h^{k+1}(s,a) \geq n_h^k(s,a),$$

since the counters can only increase across episodes, and thus for all $k' \leq K' < k$ we have

$$\frac{1}{\sqrt{n_h^{k-1}(s_h,a_h)\vee 1}} \leq \frac{1}{\sqrt{n_h^{k'-1}(s_h,a_h)\vee 1}}.$$

For $k \geq K'$, we thus find

$$\begin{aligned}
\sum_{h=1}^{H}\mathbb{E}\Big[\frac{1}{\sqrt{n_h^{k-1}(s_h^k,a_h^k)}}\mid s_1,\pi,p\Big] &\leq \frac{1}{K'}\sum_{k'=1}^{K'}\sum_{h=1}^{H}\mathbb{E}\Big[\frac{1}{\sqrt{n_h^{k'-1}(s_h,a_h)}}\mid s_1,\pi,p\Big]\\
&= \frac{1}{K'}\sum_{k'=1}^{K'}\sum_{h=1}^{H}\mathbb{E}\Big[\frac{1}{\sqrt{n_h^{k'-1}(s_h^{k'},a_h^{k'})}}\mid \mathcal{F}_{k'-1}\Big]\\
&\leq \frac{1}{K'}\tilde{O}(\sqrt{SAHK'}+SAH),
\end{aligned}$$

where $\mathcal{F}_{k'-1}$ is the σ-algebra induced by all random variables up to and including episode $k'-1$ and where the first relation holds by monotonicity of the counters, the second relation holds since π is played in the k'-th episode, and the final relation holds by Lemma 8. □

Lemma 12. *Let $K' \in [K]$. Assume that for all s, a, h, $k \in [K]$ we have*

$$n_h^{k-1}(s,a) > \frac{1}{2}\sum_{j<k} q_h^{\pi_j}(s,a;p) - H\log\left(\frac{SAH}{\delta'}\right).$$

Let π be any fixed policy and suppose that in episode $k \in [K]$, policy π_k is played, with $\pi_{k'} = \pi$ for all $k' \in [K']$. Then for all $k \geq K'$ we have

$$\sum_{h=1}^{H} \mathbb{E}\left[\frac{1}{n_h^{k-1}(s_h,a_h)\vee 1} \mid s_1,\pi,p\right] \leq \tilde{O}\left(SA(K')^{-1}\right).$$

Proof. Note that for any realization of all random variables up to and including episode K, we have (for $k \in \{0,\dots,K-1\}$)

$$n_h^{k+1}(s,a) \geq n_h^{k}(s,a),$$

since the counters can only increase across episodes, and thus for all $k' \leq K' < k$ we have

$$\frac{1}{n_h^{k-1}(s_h,a_h)\vee 1} \leq \frac{1}{n_h^{k'-1}(s_h,a_h)\vee 1}.$$

For $k \geq K'$, we thus find

$$\begin{aligned}
\sum_{h=1}^{H} \mathbb{E}\left[\frac{1}{n_h^{k-1}(s_h,a_h)} \mid s_1,\pi,p\right] &\leq \frac{1}{K'}\sum_{k'=1}^{K'}\sum_{h=1}^{H} \mathbb{E}\left[\frac{1}{n_h^{k'-1}(s_h,a_h)} \mid s_1,\pi,p\right] \\
&= \frac{1}{K'}\sum_{k'=1}^{K'}\sum_{h=1}^{H} \mathbb{E}\left[\frac{1}{n_h^{k'-1}(s_h^{k'},a_h^{k'})} \mid \mathcal{F}_{k'-1}\right] \\
&\leq \frac{1}{K'}\tilde{O}(SA),
\end{aligned}$$

where $\mathcal{F}_{k'-1}$ is the σ-algebra induced by all random variables up to and including episode $k'-1$ and where the first relation holds by monotonicity of the counters, the second relation holds since π is played in the k'-th episode, and the final relation holds by Lemma 9. □

We are now ready to prove the needed variation of Lemma 10.

Lemma 13. *(Last-iterate fixed policy errors) Consider an MDP with transition dynamics p and arbitrary estimated transition dynamics $\hat{p}_k$ (for $k \in [K]$, each forming a probability measure). Consider policy iterates $(\pi_k)_{k\in[K]}$ and suppose π_k is played in episode $k \in [K]$ and used to update the counters. Let $l_h(s,a)$, $\tilde{l}_{k,h}(s,a)$ be cost and corresponding optimistic cost with $l = c$ or $l = d_i$ as discussed in Eq. (3). Consider $K' \in [K]$ and a policy π. Suppose that for all $k' \in [K']$, $\pi_{k'} = \pi$. Let $V_h^{\pi}(s;l,p)$, $V_h^{\pi}(s;\tilde{l}_k,\hat{p}_k)$ be the values of π according to the true and estimated model, respectively. Assume that for all s,a,h,k we have*

$$n_h^{k-1}(s,a) > \frac{1}{2}\sum_{j<k} q_h^{\pi_j}(s,a;p) - H\log\left(\frac{SAH}{\delta'}\right), \tag{a}$$

$$|\tilde{l}_{k,h}(s,a) - l_h(s,a)| \leq \tilde{O}\left(\frac{1}{\sqrt{n_h^{k-1}(s,a)\vee 1}}\right), \tag{b}$$

$$|\hat{p}_{k,h}(s'|s,a) - p_h(s'|s,a)| \leq \tilde{O}\left(\sqrt{\frac{p_h(s'|s,a)L_\delta^p}{n_h^{k-1}(s,a)\vee 1}} + \frac{L_\delta^p}{n_h^{k-1}(s,a)\vee 1}\right). \tag{c}$$

Then for all $k \geq K'$ we have

$$|V^{\pi}(l,p) - V^{\pi}(\tilde{l}_k,\hat{p}_k)| \leq \tilde{O}\left(\sqrt{\mathcal{N}SAH^3}(K')^{-1/2} + (H\sqrt{\mathcal{N}} + S)SAH(K')^{-1}\right).$$

In words, if the agent plays π for the first K' episodes, then for all future episodes, the value differences for π are upper bounded by the value on the RHS.

Proof. From the value difference lemma (Dann et al., 2017, Lemma E.15) we get

$$\begin{aligned}
&|V^\pi(l,p) - V^\pi(\tilde{l}_k, \hat{p}_k)| \\
\leq &\sum_{h=1}^{H} \mathbb{E}[|l_h(s_h,a_h) - \tilde{l}_h(s_h,a_h)| \mid s_1, p, \pi] \\
+ &\sum_{h=1}^{H} \mathbb{E}\Big[\sum_{s'} |p_h(s|s_h,a_h) - \hat{p}_k(s|s_h,a_h)| V^\pi_{h+1}(s'; \tilde{l}_k, \hat{p}_k) \mid s_1, p, \pi\Big].
\end{aligned}$$

We can bound the first of the two terms as follows.

$$\begin{aligned}
&\sum_{h=1}^{H} \mathbb{E}[|l_h(s_h,a_h) - \tilde{l}_h(s_h,a_h)| \mid s_1, p, \pi] \\
&\overset{(b)}{\lesssim} \sum_{h=1}^{H} \mathbb{E}\Big[\frac{1}{\sqrt{n_h^{k-1}(s_h,a_h) \vee 1}} \mid s_1, p, \pi\Big] \\
&\overset{\text{Lemma 11}}{\leq} \tilde{O}\left(\sqrt{SAH}(K')^{-1/2} + SAH(K')^{-1}\right),
\end{aligned}$$

where Lemma Lemma 11 applies due to assertion (a). For the second term, we first note that $|V^\pi_{h+1}(s'; \tilde{l}_k, \hat{p}_k)| \lesssim H$ since $|\tilde{l}_{k,h}(s,a)| \leq l_h(s,a) + \frac{1}{\sqrt{n_h^{k-1}(s,a) \vee 1}}$ by (b). Hence

$$\begin{aligned}
&\sum_{h=1}^{H} \mathbb{E}\Big[\sum_{s'} |(p_h(s|s_h,a_h) - \hat{p}_k(s|s_h,a_h)) V^\pi_{h+1}(s'; \tilde{l}_k, \hat{p}_k) \mid s_1, p, \pi\Big] \\
&\overset{(c)}{\lesssim} H \sum_{h=1}^{H} \mathbb{E}\Big[\frac{1}{\sqrt{n_h^{k-1}(s_h,a_h) \vee 1}} \big(\sum_{s'} \sqrt{p_h(s'|s_h,a_h)}\big) + \frac{S}{n_h^{k-1}(s_h,a_h) \vee 1} \mid s_1, p, \pi\Big] \\
&\overset{\text{Jensen}}{\leq} H \sum_{h=1}^{H} \mathbb{E}\Big[\frac{1}{\sqrt{n_h^{k-1}(s_h,a_h) \vee 1}} \sqrt{\mathcal{N}} \sqrt{\sum_{s'} p_h(s'|s_h,a_h)} + \frac{S}{n_h^{k-1}(s_h,a_h) \vee 1} \mid s_1, p, \pi\Big] \\
&\lesssim H\sqrt{\mathcal{N}} \cdot \sqrt{SAH}(K')^{-1/2} + H\sqrt{\mathcal{N}} \cdot SAH(K')^{-1} + HS \cdot SA(K')^{-1} \\
&= \sqrt{\mathcal{N} SAH^3}(K')^{-1/2} + (H\sqrt{\mathcal{N}} + S) SAH(K')^{-1},
\end{aligned}$$

where the second to last relation holds due to Lemma 11 and Lemma 12 (which in turn apply due to assertion (a)). □

D Omitted Proofs for Section 3

D.1 Pre-Training Phase

We now establish which constant duration K' of playing $\bar{\pi}$ is sufficient to guarantee the desired strict feasibility of Eq. (6) (with the choice of K' favoring readability over tightness). Note that the conclusion holds conditioned on the success event G.

Lemma 1. *Suppose that Assumption 1 holds, i.e., the agent has access to a strictly feasible $\bar{\pi}$ and its slack $\gamma > 0$. Fix any $\nu \in (0,1)$, and suppose the agent executes $\bar{\pi}$ for $K' = \tilde{O}\left(\max\left\{\frac{(H\sqrt{\mathcal{N}}+S)SAH}{(1-\nu)\gamma}, \frac{\mathcal{N}SAH^3}{(1-\nu)^2\gamma^2}\right\}\right)$ episodes, where $\mathcal{N} := \max_{s,a,h} |\{s' \mid p_h(s'|s,a) > 0\}|$ denotes the maximum number of transitions. Then, if the agent updates the optimistic CMDP based on the observations from those episodes (cf. Eq.* (3)*), with probability at least $1-\delta$ the following condition is satisfied for every $k \in \{K', \dots, K\}$:*

$$V^{\bar{\pi}}(\tilde{d}_{i,k}, \tilde{p}_k) \leq \alpha_i - \nu\gamma \quad (\forall i \in [I]).$$

Proof. Condition on the success event G, which by Lemma 2 happens with probability $\geq 1-\delta$. Then by construction of G, the assumptions (a) and (b) in Lemma 13 are met for $l = d_i$ for all constraints i. By Lemma 14 also the assumption (c) in Lemma 13 is met with $\hat{p}_k := \tilde{p}_k$ from the k-th iteration of OPTAUG-CMDP. Thus, the conclusion from Lemma 13 holds for $\pi = \bar{\pi}$ and $l_h(s,a) = d_{i,h}(s,a)$ (for all $i \in [I]$) and we have for all $i \in [I]$ and $k \geq K'$

$$|V^{\pi}(d_i, p) - V^{\pi}(\tilde{d}_{i,k}, \tilde{p}_k)| \leq b(K')^{-1/2} + a(K')^{-1},$$

where

$$b = \tilde{O}\left(\sqrt{\mathcal{N}SAH^3}\right), \qquad a = \tilde{O}\left((H\sqrt{\mathcal{N}} + S)SAH\right).$$

Thus, if $K' \geq \max\left\{\frac{2(H\sqrt{\mathcal{N}}+S)SAH}{(1-\nu)\gamma}, \frac{4\mathcal{N}SAH^3}{(1-\nu)^2\gamma^2}\right\}$, we have

$$|V^{\bar{\pi}}(d_i, p) - V^{\bar{\pi}}(\tilde{d}_{i,k}, \tilde{p}_k)| \leq \tilde{O}((1-\nu)\gamma).$$

By further correcting for the missing polylogarithmic terms, this shows that there exists $K' = \tilde{O}\left(\max\left\{\frac{(H\sqrt{\mathcal{N}}+S)SAH}{(1-\nu)\gamma}, \frac{\mathcal{N}SAH^3}{(1-\nu)^2\gamma^2}\right\}\right)$ such that for all $i \in [I]$ and $k \geq K'$, we have

$$|V^{\bar{\pi}}(d_i, p) - V^{\bar{\pi}}(\tilde{d}_{i,k}, \tilde{p}_k)| \leq (1-\nu)\gamma.$$

At the same time, Assumption 1 guarantees

$$V^{\bar{\pi}}(d_i, p) \leq \alpha_i - \gamma.$$

Hence

$$\begin{aligned} V^{\bar{\pi}}(\tilde{d}_{i,k}, \tilde{p}_k) =& V^{\bar{\pi}}(\tilde{d}_{i,k}, \tilde{p}_k) - V^{\bar{\pi}}(d_i, p) + V^{\bar{\pi}}(d_i, p) \\ \leq& (1-\nu)\gamma + \alpha_i - \gamma \\ =& \alpha_i - \nu\gamma, \end{aligned}$$

which proves the claim. □

Remark 1. *Notice that we chose K' such that both the terms $b(K')^{-1/2}$ and $a(K')^{-1}$ are less than $1/2$, which is sufficient. For a tighter but less concise bound, we can alternatively solve the quadratic equation*

$$b(K')^{-1/2} + a(K')^{-1} \leq (1-\nu)\gamma \tag{28}$$

for K' to obtain the smallest possible pre-training time, up to polylogarithmic factors. Furthermore, note that since a, b are only specified in $\tilde{O}$-notation, we only have an asymptotic bound up to polylogarithmic factors on how large to choose K'. An exact bound could be established by carrying along all factors dropped in the $\tilde{O}$-notation, but for brevity we only specify how large to choose K' asymptotically. In practice, one can take the bound above and choose K' slightly larger.

D.2 Iteration Complexity of the Inner Method

Recall Proposition 1:

Proposition 1. *In episode k, fix any accuracy of $\epsilon_k > 0$. There exists an algorithm for solving Eq.* (4) *such that the objective at its output $(\pi_k, \tilde{p}_k)$ is ϵ_k-close to the optimum of Eq.* (4)*, by solving $O\left(\frac{\eta_k I S^2 AH}{\epsilon_k}\right)$ (extended) MDPs via DP.*

In Corollary 1, we prove Proposition 1 by showing that the method proposed in Appendix B achieves the desired iteration complexity for solving Eq. (4). As seen in Appendix B.1, every iteration of the Frank-Wolfe scheme INNEROPT-FW (Appendix B.2) above can be performed efficiently via DP, i.e., in (low-degree) polynomial time in the parameters defining the CMDP. To analyze the number of such Frank-Wolfe iterations needed to reach an ϵ_k-close solution, we recall the following result due to Jaggi (2013).

Theorem 6. *The iterates $(z_k)_{k\geq 0}$ of the Frank-Wolfe algorithm with an exact LMO applied to a convex objective $f\colon \mathcal{D} \to \mathbb{R}$ (with closed, convex and bounded domain $\mathcal{D} \subset \mathbb{R}^d$) satisfy*

$$f(z_k) - f(z^*) \leq \frac{2C_f}{k+2}, \tag{29}$$

where z^* is a minimizer of f over $\mathcal{D}$ and C_f is the curvature constant of f. Moreover, if ∇f is L-Lipschitz continuous w.r.t. an arbitrary norm $\|\cdot\|_c$ on $\mathbb{R}^d$, then

$$C_f \leq diam_{\|\cdot\|_c}(\mathcal{D})^2 L. \tag{30}$$

In our case, this yields the following bound on the iteration complexity of the inner optimization procedure.

Corollary 1. *After $T = \frac{2\eta_k I S^2 AH}{\epsilon_k}$ Frank-Wolfe steps,* INNEROPT-FW *returns an ϵ_k-close solution to the inner problem (Eq.* (4)*).*

Proof. Consider the problem of minimizing f over Z, as defined in Appendix B.1. Let $y, z \in Z$. With respect to $\|\cdot\|_c = \|\cdot\|_\infty$ we have, using the expression of the gradients from Eq. (22),

$$\begin{aligned}
&\|\nabla f(y) - \nabla f(z)\|_\infty \\
=&\|(\nabla f(y))(\cdot,\cdot,1,\cdot) - (\nabla f(z))(\cdot,\cdot,1,\cdot)\|_\infty \\
=&\|\tilde{D}_k^T[\lambda_k + \eta_k(\tilde{D}_k \sum_{s''} y(s'') - \alpha)]_+ - \tilde{D}_k^T[\lambda_k + \eta_k(\tilde{D}_k \sum_{s''} z(s'') - \alpha)]_+\|_\infty \\
\leq&\|\tilde{D}_k^T\|_\infty \cdot \|[\lambda_k + \eta_k(\tilde{D}_k \sum_{s''} y(s'') - \alpha)]_+ - [\lambda_k + \eta_k(\tilde{D}_k \sum_{s''} z(s'') - \alpha)]_+\|_\infty \\
&\text{and using} [a]_+ - [b]_+ \leq [a-b]_+ \leq |a-b| \\
\leq&\|\tilde{D}_k^T\|_1 \cdot \|[\lambda_k + \eta_k(\tilde{D}_k \sum_{s''} y(s'') - \alpha)] - [\lambda_k + \eta_k(\tilde{D}_k \sum_{s''} z(s'') - \alpha)]\|_\infty \\
=&\|\tilde{D}_k^T\|_\infty \cdot \eta_k \|\tilde{D}_k \sum_{s''} (y(s'') - z(s''))\|_\infty \\
\leq&\|\tilde{D}_k^T\|_\infty \cdot \eta_k \|\tilde{D}_k\|_\infty \sum_{s''} \|y(s'') - z(s'')\|_\infty \\
\leq&SAH \cdot \eta_k \cdot I \cdot S \|y - z\|_\infty,
\end{aligned}$$

where the final bound holds since the entries of $\tilde{D}_k$ are in $[0,1]$. Thus, $\nabla f(\cdot)$ is $(S^2AH \cdot \eta_k \cdot I)$-Lipschitz continuous and thus f is $(S^2AH \cdot \eta_k \cdot I)$-smooth. Moreover, for all $z \in Z$ we have $0 \leq z_h(s,a,s') \leq 1$ in every component, so for $y, z \in Z$ we have $\|z - y\|_\infty \leq 1$. Thus $\text{diam}_{\|\cdot\|_\infty}(Z)^2 \leq 1$. Plugging both into Theorem 6 yields the result. □

Note that it may be possible to improve this bound by using a norm different from $\|\cdot\|_c = \|\cdot\|_\infty$ for the analysis.

D.3 Main Result

Our main result shows that with probability $1-\delta$, OPTAUG-CMDP achieves a regret that is sublinear in the number of episodes K and polylogarithmic in $1/\delta$.

Theorem 1. *Suppose that Assumption 1 holds, let $\delta \in (0,1)$ and $\nu > 0$. Then there exist $K' = \tilde{O}\left(\max\left\{\frac{(H\sqrt{\mathcal{N}}+S)SAH}{(1-\nu)\gamma}, \frac{\mathcal{N}SAH^3}{(1-\nu)^2\gamma^2}\right\}\right)$ and η_k, ϵ_k such that with probability at least $1-\delta$,* OPTAUG-CMDP *achieves a total regret of*

$$\mathcal{R}(K;c) = \tilde{O}\left(\sqrt{\mathcal{N}SAH^3K} + (H\sqrt{\mathcal{N}} + S)SAH + K'H\right),$$
$$\mathcal{R}(K;d) = \tilde{O}\left(\sqrt{\mathcal{N}SAH^3K} + (H\sqrt{\mathcal{N}} + S)SAH\right).$$

Proof. According to our regret decomposition (Observation 2), we have

$$\mathcal{R}(K;c) \le K'H + \underbrace{\sum_{k=K'+1}^{K} [V^{\pi_k}(c,p) - V^{\pi_k}(\tilde{c}_k,\tilde{p}_k)]_+}_{\text{Estimation Error}} + \underbrace{\sum_{k=K'+1}^{K} [V^{\pi_k}(\tilde{c}_k,\tilde{p}_k) - V^{\pi^*}(c,p)]_+}_{\text{Optimization Error}},$$

$$\mathcal{R}(K;d) \le \underbrace{\max_{i\in[I]} \sum_{k=K'+1}^{K} [V^{\pi_k}(d_i,p) - V^{\pi_k}(\tilde{d}_{i,k},\tilde{p}_k)]_+}_{\text{Estimation Error}} + \underbrace{\max_{i\in[I]} \sum_{k=K'+1}^{K} [V^{\pi_k}(\tilde{d}_{i,k},\tilde{p}_k) - \alpha_i]_+}_{\text{Optimization Error}}.$$

Suppose the success event G occurs, which happens with probability at least $1-\delta$ as shown in Lemma 2. By Lemma 3, the estimation errors satisfy

$$\sum_{k=K'+1}^{K} [V^{\pi_k}(c,p) - V^{\pi_k}(\tilde{c}_k,\tilde{p}_k)]_+ \le \tilde{O}\left(\sqrt{\mathcal{N}SAH^3K} + (H\sqrt{\mathcal{N}}+S)SAH\right),$$

$$\max_{i\in[I]} \sum_{k=K'+1}^{K} \left[V^{\pi_k}(d_i,p) - V^{\pi_k}(\tilde{d}_{i,k},\tilde{p}_k)\right]_+ \le \tilde{O}\left(\sqrt{\mathcal{N}SAH^3K} + (H\sqrt{\mathcal{N}}+S)SAH\right).$$

By Lemma 7, the optimization errors satisfy

$$\sum_{k=K'+1}^{K} [V^{\pi_k}(\tilde{c}_k,\tilde{p}_k) - V^{\pi^*}(c,p)]_+ \le \sum_{k=K'+1}^{K} \left(\frac{(O(\sigma k) + \sum_{t=K'+1}^{k}\sqrt{2\eta_t\epsilon_t})^2}{2\eta_k} + \epsilon_k\right) \le O(\sqrt{K}),$$

$$\max_{i\in[I]} \sum_{k=K'+1}^{K} [V^{\pi_k}(\tilde{d}_{i,k},\tilde{p}_k) - \alpha_i]_+ \le \sum_{k=K'+1}^{K} \frac{O(\sigma k) + \sum_{t=K'+1}^{k}\sqrt{2\eta_t\epsilon_k}}{\eta_k} \le O(\sqrt{K}),$$

where $\sigma = \frac{H}{\nu\gamma}$ and the term $O(\sigma k)$ can be replaced by $(2+2(k-K'))\sigma$. As shown in Lemma 7, we can choose step sizes $\eta_{K'+k} = \Theta(k^{2.5})$ and $\epsilon_{K'+k} = \Theta(1/\eta_{K'+k})$ such that both of the above bounds are $O(\sqrt{K})$ up to constant factors. Summing up concludes the proof. □

Remark 2. *Note that while the bound for the estimation errors does not depend on the choices for η_k and ϵ_k, the one for the optimization errors does. For any such choice, the general regret bound we obtain reads*

$$\begin{aligned}\mathcal{R}(K;c) &\le \tilde{O}\left(\sqrt{\mathcal{N}SAH^3K} + (H\sqrt{\mathcal{N}}+S)SAH + K'H\right)\\ &\quad + \sum_{k=K'+1}^{K}\left(\frac{(O(\sigma k) + \sum_{t=K'+1}^{k}\sqrt{2\eta_t\epsilon_t})^2}{2\eta_k} + \epsilon_k\right), \qquad (31)\\ \mathcal{R}(K;d) &\le \tilde{O}\left(\sqrt{\mathcal{N}SAH^3K} + (H\sqrt{\mathcal{N}}+S)SAH\right) + \sum_{k=K'+1}^{K}\frac{O(\sigma k) + \sum_{t=K'+1}^{k}\sqrt{2\eta_t\epsilon_k}}{\eta_k},\end{aligned}$$

where $\sigma = \frac{H}{\nu\gamma}$ and the term $O(\sigma k)$ can be replaced by $(2+2(k-K'))\sigma$.

We remark that there are multiple ways to choose η_k, ϵ_k in the bound above that yield different regret guarantees and complexities of the inner algorithm InnerOpt-FW. *One possible way is to choose a decreasing approximation error ϵ_k in every episode and η_k large enough so that the parameter-dependent terms above are of order $O(\sqrt{K})$. This yields vanishing average regret for both the objective and the constraints. One such choice is*[11]

$$\epsilon_{K'+k} := \frac{1}{2\eta_{K'+k}} \qquad (k \ge 1),$$
$$\eta_{K'+k} := ((2+3k)\sigma)^{2.5} \qquad (k \ge 1),$$

[11]Note that if γ is not known, having an estimate of $\sigma = \frac{H}{\nu\gamma}$ that is larger than the true value is clearly sufficient.

as we discuss in Lemma 7.

It is worth noting that, when choosing η_k very large and ϵ_k very small, the bound we get on the parameter-dependent terms in Eq. (31) *gets better, and even becomes constant when choosing η_k, ϵ_k exponential in $k - K'$. However, the inner problem (Eq.* (4)*) that we solve in every episode becomes computationally harder. This is due to the changing smoothness of the objective (with respect to the occupancy measure). Indeed, Proposition 1 quantifies the resulting iteration complexity, formalizing this tradeoff between statistical guarantees and computational efficiency. Thus, in practice, it is advised to choose η_k as large (ϵ_k as small) as needed to ensure sublinear regret but as small as possible to avoid a higher computational cost.*

E Omitted Proofs for Section 4

E.1 Preliminaries

E.1.1 Preliminary Confidence Bounds

Fix an arbitrary number of episodes K and $0 < \delta < 1$, corresponding to a confidence of $1 - \delta$. Moreover, let $(\pi_k)_{k=1}^K$ be the policies according to which the optimistic estimates are updated (Section 3).

Define the following confidence sets

$$\begin{aligned}
B^p_{k,h}(s,a) &:= \left\{\tilde{p}_h(\cdot|s,a) \in \Delta(\mathcal{S})\, S \mid \forall s' \in \mathcal{S}\colon |\tilde{p}_h(s'|s,a) - \bar{p}_h^{k-1}(s'|s,a)| \le \beta^p_{k,h}(s,a,s')\right\},\\
B^c_{k,h}(s,a) &:= \left[\bar{c}_h^{k-1}(s,a) - \beta^c_{k,h}(s,a), \bar{c}_h^{k-1}(s,a) + \beta^c_{k,h}(s,a)\right],\\
B^d_{i,k,h}(s,a) &:= \left[\bar{d}_{i,h}^{k-1}(s,a) - \beta^d_{i,k,h}(s,a), \bar{d}_{i,h}^{k-1}(s,a) + \beta^d_{i,k,h}(s,a)\right],
\end{aligned}$$

where (derived using Hoeffding for the costs and empirical Bernstein for the transitions)

$$\begin{aligned}
\beta^p_{k,h}(s,a,s') :=& 2\sqrt{\frac{\bar{p}_h^{k-1}(s'|s,a)(1-\bar{p}_h^{k-1}(s'|s,a))L^p_\delta}{n_h^{k-1}(s,a)\vee 1}} + \frac{\frac{14}{3}L^p_\delta}{n_h^{k-1}(s,a)\vee 1},\\
\beta^c_{k,h}(s,a) := \beta^d_{i,k,h}(s,a) :=& \sqrt{\frac{L_\delta}{n_h^{k-1}(s,a)\vee 1}},
\end{aligned}$$

with

$$L^p_\delta := \log\left(\frac{6SAHK}{\delta}\right) \qquad L_\delta := \log\left(\frac{6SAH(I+1)K}{\delta}\right).$$

We want that, with probability at least $1 - \delta$, the true MDP is contained within these bounds across all $k \in [K]$. This will be needed for bounding the optimization error.

With the same thresholds $\beta^p_{k,h}$, $\beta^c_{k,h}$, $\beta^d_{i,k,h}$ as above we define the following *failure events*

$$\begin{aligned}
F^p_k &:= \left\{\exists s,a,s',h\colon |p_h(s'|s,a) - \bar{p}_h^{k-1}(\cdot|s,a)| > \beta^p_{k,h}(s,a,s')\right\},\\
F^c_k &:= \left\{\exists s,a,h\colon |\bar{c}_h^{k-1}(s,a) - c_h(s,a)| > \beta^c_{k,h}(s,a)\right\},\\
F^d_k &:= \left\{\exists s,a,i,h\colon |\bar{d}_{i,h}^{k-1}(s,a) - d_{i,h}(s,a)| > \beta^d_{k,h}(s,a)\right\},\\
F^N_k &:= \left\{\exists s,a,h\colon n_h^{k-1}(s,a) \le \sum_{j<k} q_h^{\pi_j}(s,a;p) - H\log\left(\frac{SAH}{\delta'}\right)\right\},
\end{aligned}$$

where we set $\delta' := \delta/3$. With this notation, set $F^p := \cup_{k=1}^K F^p_k$, $F^c := \cup_{k=1}^K F^c_k$, $F^d := \cup_{k=1}^K F^d_k$, $F^N := \cup_{k=1}^K F^N_k$ and finally denote the event that none of the failure events ever occurs by

$$G := \overline{\left(F^p \bigcup F^c \bigcup F^d \bigcup F^N\right)},$$

which we will refer to as the *success event*.

By construction of G and the optimistic estimates, we immediately have the following.

Observation 1. *Conditioned on the success event G, for every $k \in [K]$ the CMDP with cost $\tilde{c}_k$ and constraint costs $(\tilde{d}_{i,k})_{i\in[I]}$ is optimistic and all $B_k^p(s,a)$ contain the true respective transition probability distribution, i.e.,*

$$\tilde{c}_k \leq c, \qquad \tilde{d}_{i,k} \leq d_{i,k}, \qquad \text{and} \qquad p \in B_k^p,$$

where $B_k^p := \{\tilde{p} \mid \forall s,a,h\colon \tilde{p}_h(\cdot|s,a) \in B_{k,h}^p(s,a)\}$ is the set of plausible transition probabilities.

Concentration bounds on the individual random variables together with a union bound over all indices now allow us to establish the following result. In Appendix D.3, we show that conditioned on G, the regrets are sublinear in K and polylogarithmic in $1/\delta$.

Lemma 2. *Fix $\delta \in (0,1)$ and define the optimistic model in Eq.* (3) *accordingly. Then the success event G occurs with probability at least $1-\delta$, i.e., $P[G] \geq 1-\delta$.*

Proof. Efroni et al. (2020, Appendix A.1) give a proof of this. Note that the proof does not require any specific properties of the used policy iterates $(\pi_k)_{k\in[K]}$ but only uses that the collected costs and transitions are i.i.d.. □

Moreover, we have the following result, which will allow us to bound the estimation error. The absolute constants could in fact be specified via a short calculation, but we omit this for brevity.

Lemma 14 (Lemma 8, Jin et al. (2019))**.** *In the setup above, conditioned on G, there exist absolute constants $C_1, C_2 > 0$ such that for all indices k, h, s, a, s' we have*

$$|p_h(s'|s,a) - \bar{p}_h^{k-1}(s'|s,a)| \leq C_1 \sqrt{\frac{p_h(s'|s,a) L_\delta^p}{n_h^{k-1}(s,a) \vee 1}} + C_2 \frac{L_\delta^p}{n_h^{k-1}(s,a) \vee 1},$$

with $L_\delta^p = \log\left(\frac{6SAHK}{\delta}\right)$ as before.

Proof. We refer to Jin et al. (2019, Lemma 8) for a proof. □

E.1.2 Regret Decomposition

The simple observation that we can decompose the regrets can be seen as follows.

Observation 2 (Regret decomposition)**.**

$$\mathcal{R}(K;c) \leq K'H + \underbrace{\sum_{k=K'+1}^{K} [V^{\pi_k}(c,p) - V^{\pi_k}(\tilde{c}_k,\tilde{p}_k)]_+}_{\textit{Estimation Error}} + \underbrace{\sum_{k=K'+1}^{K} [V^{\pi_k}(\tilde{c}_k,\tilde{p}_k) - V^{\pi^*}(c,p)]_+}_{\textit{Optimization Error}}$$

$$\mathcal{R}(K;d) \leq \underbrace{\max_{i\in[I]} \sum_{k=K'+1}^{K} [V^{\pi_k}(d_i,p) - V^{\pi_k}(\tilde{d}_{i,k},\tilde{p}_k)]_+}_{\textit{Estimation Error}} + \underbrace{\max_{i\in[I]} \sum_{k=K'+1}^{K} [V^{\pi_k}(\tilde{d}_{i,k},\tilde{p}_k) - \alpha_i]_+}_{\textit{Optimization Error}}$$

Proof. We first split the regrets between the two phases of the algorithm.

$$\mathcal{R}(K;c) = \underbrace{\sum_{k=1}^{K'} [V^{\pi_k}(c,p) - V^{\pi^*}(c,p)]_+}_{\text{Pre-Training, } \leq K'H} + \underbrace{\sum_{k=K'+1}^{K} [V^{\pi_k}(c,p) - V^{\pi^*}(c,p)]_+}_{\text{Optimistic Exploration}}$$

$$\mathcal{R}(K;d) \leq \underbrace{\max_{i\in[I]} \sum_{k=1}^{K'} [V^{\pi_k}(d_i,p) - \alpha_i]_+}_{\text{Pre-Training, } =0} + \underbrace{\max_{i\in[I]} \sum_{k=K'+1}^{K} [V^{\pi_k}(d_i,p) - \alpha_i]_+}_{\text{Optimistic Exploration}}$$

We can trivially bound the objective regret during the pre-training phase by $K'H$ (since the expected costs are in $[0,1]$ and the time horizon is H, so the value functions are in $[0,H]$). Since $\bar{\pi}$ is strictly feasible, there is no constraint regret during pre-training.

We now focus on the regrets incurred in the optimistic exploration phase. We split the sum and use that $[a+b]_+ \leq [a]_+ + [b]_+$. For the objective regret we have

$$\begin{aligned}
\mathcal{R}(K;c) =& K'H + \sum_{k=K'+1}^{K} [V^{\pi_k}(c,p) - V^{\pi^*}(c,p)]_+ \\
=& K'H + \sum_{k=K'+1}^{K} [V^{\pi_k}(c,p) - V^{\pi_k}(\tilde{c}_k,\tilde{p}_k) + V^{\pi_k}(\tilde{c}_k,\tilde{p}_k) - V^{\pi^*}(c,p)]_+ \\
\leq& K'H + \sum_{k=K'+1}^{K} [V^{\pi_k}(c,p) - V^{\pi_k}(\tilde{c}_k,\tilde{p}_k)]_+ + \sum_{k=K'+1}^{K} [V^{\pi_k}(\tilde{c}_k,\tilde{p}_k) - V^{\pi^*}(c,p)]_+
\end{aligned}$$

and similarly, for the constraint regret we get

$$\begin{aligned}
\mathcal{R}(K;d) =& 0 + \max_{i\in[I]} \left(\sum_{k=K'+1}^{K} [V^{\pi_k}(d_i,p) - \alpha_i]_+ \right) \\
=& \max_{i\in[I]} \sum_{k=K'+1}^{K} \left[V^{\pi_k}(d_i,p) - \alpha_i - (V^{\pi_k}(\tilde{d}_{i,k},\tilde{p}_k) - \alpha_i) + (V^{\pi_k}(\tilde{d}_{i,k},\tilde{p}_k) - \alpha_i) \right]_+ \\
\leq& \max_{i\in[I]} \left(\sum_{k=K'+1}^{K} \left[V^{\pi_k}(d_i,p) - V^{\pi_k}(\tilde{d}_{i,k},\tilde{p}_k) \right]_+ + \sum_{k=K'+1}^{K} \left[V^{\pi_k}(\tilde{d}_{i,k},\tilde{p}_k) - \alpha_i \right]_+ \right) \\
\leq& \max_{i\in[I]} \sum_{k=K'+1}^{K} \left[V^{\pi_k}(d_i,p) - V^{\pi_k}(\tilde{d}_{i,k},\tilde{p}_k) \right]_+ + \max_{i\in[I]} \sum_{k=K'+1}^{K} \left[V^{\pi_k}(\tilde{d}_{i,k},\tilde{p}_k) - \alpha_i \right]_+ .
\end{aligned}$$

□

E.2 Omitted Proofs for Section 4.1

We leverage the on-policy error bound from Lemma 10 and the preliminaries from Appendix E.1.1 to establish the desired bound on the estimation errors.

Lemma 3 (Estimation errors)**.** *Let $(\pi_k)_{k=K'+1}^{K}$ be the sequence of policies obtained by* OPTAUG-CMDP. *Then, conditioned on G, we can bound the estimation errors as follows:*

$$\begin{aligned}
\sum_{k=K'+1}^{K} [V^{\pi_k}(c,p) - V^{\pi_k}(\tilde{c}_k,\tilde{p}_k)]_+ \leq& \tilde{O}\left(\sqrt{\mathcal{N}SAH^3K} + (H\sqrt{\mathcal{N}} + S)SAH \right), \\
\max_{i\in[I]} \sum_{k=K'+1}^{K} [V^{\pi_k}(d_i,p) - V^{\pi_k}(\tilde{d}_{i,k},\tilde{p}_k)]_+ \leq& \tilde{O}\left(\sqrt{\mathcal{N}SAH^3K} + (H\sqrt{\mathcal{N}} + S)SAH \right).
\end{aligned}$$

Proof. Condition on the success event G. Then by construction of G, the assumptions (a) and (b) in Lemma 10 are met for $l = c$ and $l = d_i$ for all constraints $i \in [I]$. By Lemma 14, also assumption (c) in Lemma 10 is met with $\hat{p}_k := \tilde{p}_k$ from the k-th iteration of OPTAUG-CMDP.

First, consider the terms $[V^{\pi_k}(c,p) - V^{\pi_k}(\tilde{c}_k,\tilde{p}_k)]_+$. For the cost $l = c$, the values of the true and the estimated MDPs are $V^{\pi_k}(c,p)$ and $V^{\pi_k}(\tilde{c}_k,\tilde{p}_k)$, respectively. Thus, invoking Lemma 10 we find

$$\begin{aligned}
\sum_{k=K'+1}^{K} [V^{\pi_k}(c,p) - V^{\pi_k}(\tilde{c}_k,\tilde{p}_k)]_+ \leq& \sum_{k=K'+1}^{K} |V^{\pi_k}(c,p) - V^{\pi_k}(\tilde{c}_k,\tilde{p}_k)| \\
\leq& \sum_{k=1}^{K} |V^{\pi_k}(c,p) - V^{\pi_k}(\tilde{c}_k,\tilde{p}_k)| \\
\leq& \tilde{O}\left(\sqrt{\mathcal{N}SAH^3K} + (H\sqrt{\mathcal{N}} + S)SAH \right).
\end{aligned}$$

Similarly, let $i \in [I]$ and consider the terms $[V^{\pi_k}(d_i, p) - V^{\pi_k}(\tilde{d}_{i,k}, \tilde{p}_k)]_+$. For the cost $l = d_i$, the values of the true and the estimated MDPs are $V^{\pi_k}(d_i, p)$ and $V^{\pi_k}(\tilde{d}_{i,k}, \tilde{p}_k)$, respectively. Thus, invoking Lemma 10 we find

$$\begin{aligned}\sum_{k=K'+1}^{K} [V^{\pi_k}(d_i, p) - V^{\pi_k}(\tilde{d}_{i,k}, \tilde{p}_k)]_+ \leq & \sum_{k=1}^{K} |V^{\pi_k}(d_i, p) - V^{\pi_k}(\tilde{d}_{i,k}, \tilde{p}_k)| \\ \leq & \tilde{O}\left(\sqrt{\mathcal{N}SAH^3K} + (H\sqrt{\mathcal{N}} + S)SAH\right).\end{aligned}$$

□

E.3 Omitted Proofs for Section 4.2

For notational convenience, we write $D = (d_i)_{i\in[I]}$ and $\tilde{D}_k = (\tilde{d}_{i,k})_{i\in[I]}$ throughout this section.

E.3.1 Preliminary Bounds

Recall that in OPTAUG-CMDP, we find an ϵ_k-close solution to the inner problem Eq. (4) in every episode $k \in \{K'+1, \ldots, K\}$.

Lemma 4. *Conditioned on G, for each $k \in \{K'+1, \ldots, K\}$, in* OPTAUG-CMDP *we have*

$$\begin{aligned} V^{\pi_k}(\tilde{c}_k, \tilde{p}_k) - V^{\pi^*}(c, p) &\leq \epsilon_k + \frac{\|\lambda_k\|^2 - \|\lambda_{k+1}\|^2}{2\eta_k}, \\ V^{\pi_k}(\tilde{d}_{i,k}, \tilde{p}_k) - \alpha_i &\leq \frac{\lambda_{k+1}(i) - \lambda_k(i)}{\eta_k} \quad (\forall i \in [I]). \end{aligned}$$

Proof. We first bound the error for the objective cost. Let $k \in \{K'+1, \ldots, K\}$. Conditioned on the success event, we have (componentwise) $c \geq \tilde{c}_k$, $D \geq \tilde{D}_k$ and $p \in B_k^p$ (Observation 1), and thus

$$\begin{aligned}
& V^{\pi^*}(c, p) \\
=& \left(V^{\pi^*}(c, p) + \frac{1}{2\eta_k}\|[\lambda_k + \eta_k(V^{\pi^*}(D, p) - \alpha)]_+\|^2\right) - \frac{1}{2\eta_k}\|[\lambda_k + \eta_k(V^{\pi^*}(D, p) - \alpha)]_+\|^2 \\
\geq& \left(V^{\pi^*}(\tilde{c}_k, p) + \frac{1}{2\eta_k}\|[\lambda_k + \eta_k(V^{\pi^*}(\tilde{D}_k, p) - \alpha)]_+\|^2\right) - \frac{1}{2\eta_k}\|[\lambda_k + \eta_k(V^{\pi^*}(D, p) - \alpha)]_+\|^2 \\
\geq& \min_{\substack{\pi\in\Pi \\ p'\in B_k^p}} \left(V^{\pi}(\tilde{c}_k, p') + \frac{1}{2\eta_k}\|[\lambda_k + \eta_k(V^{\pi}(\tilde{D}_k, p') - \alpha)]_+\|^2\right) \\
& - \frac{1}{2\eta_k}\|[\lambda_k + \eta_k(V^{\pi^*}(D, p) - \alpha)]_+\|^2 \\
\geq& \left(V^{\pi_k}(\tilde{c}_k, \tilde{p}_k) + \frac{1}{2\eta_k}\|[\lambda_k + \eta_k(V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)]_+\|^2 - \epsilon_k\right) \\
& - \frac{1}{2\eta_k}\|[\lambda_k + \eta_k(V^{\pi^*}(D, p) - \alpha)]_+\|^2 \\
=& V^{\pi_k}(\tilde{c}_k, \tilde{p}_k) - \epsilon_k + \frac{1}{2\eta_k}\|[\lambda_k + \eta_k(V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)]_+\|^2 \\
& - \frac{1}{2\eta_k}\|[\lambda_k + \eta_k(V^{\pi^*}(D, p) - \alpha)]_+\|^2,
\end{aligned}$$

where the first inequality is due to optimism and monotonicity of $\|[\cdot]_+\|^2$, the second due to $p \in B_k^p$, and the third due to the ϵ_k-closeness of π_k, $\tilde{p}_k$. Now since π^* is primal-feasible we have $V^{\pi^*}(D, p) - \alpha \leq 0$ and thus $\|[\lambda_k + \eta_k(V^{\pi^*}(D, p) - \alpha)]_+\|^2 \leq \|\lambda_k\|^2$ by monotonicity of $\|[\cdot]_+\|^2$.

Plugging in the update for λ_{k+1} thus shows

$$\begin{aligned}
&\frac{1}{2\eta_k}\|[\lambda_k + \eta_k(V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)]_+\|^2 - \frac{1}{2\eta_k}\|[\lambda_k + \eta_k(V^{\pi^*}(D, p) - \alpha)]_+\|^2 \\
\geq &\frac{1}{2\eta_k}\|[\lambda_k + \eta_k(V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)]_+\|^2 - \frac{1}{2\eta_k}\|\lambda_k\|^2 \\
= &\frac{1}{2\eta_k}\|\lambda_{k+1}\|^2 - \frac{1}{2\eta_k}\|\lambda_k\|^2,
\end{aligned}$$

concluding the proof after plugging this into the previous inequality and rearranging.

We now proceed by proving the bound on the constraint cost. Let $k \in \{K' + 1, \dots, K\}$. Let $i \in [I]$ and recall that $\lambda_{k+1} = [\lambda_k + \eta_k(V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)]_+$. If $\lambda_{k+1}(i) > 0$, then we have $\lambda_{k+1}(i) - \lambda_k(i) = \eta_k(V^{\pi_k}(\tilde{d}_{i,k}, \tilde{p}_k) - \alpha_i)$ with equality. On the other hand, if $\lambda_{k+1}(i) = 0$, then $\lambda_k(i) + \eta_k(V^{\pi_k}(\tilde{d}_{i,k}, \tilde{p}_k) - \alpha_i) \leq 0$ and thus $\lambda_{k+1}(i) - \lambda_k(i) = -\lambda_k(i) \geq (\lambda_k(i) + \eta_k(V^{\pi_k}(\tilde{d}_{i,k}, \tilde{p}_k) - \alpha_i)) - \lambda_k(i) = \eta_k(V^{\pi_k}(\tilde{d}_{i,k}, \tilde{p}_k) - \alpha_i)$. Thus in both cases $\lambda_{k+1}(i) - \lambda_k(i) \geq \eta_k(V^{\pi_k}(\tilde{d}_{i,k}, \tilde{p}_k) - \alpha_i)$, which proves the claim. □

E.3.2 Bounding the Dual Iterates

In order to prove Lemma 19, which implies Lemma 5, we first establish (Lemmas 15 to 17), following the analysis of Xu (2021).

Lemma 15. *Let $k \in \{K' + 1, \dots, K\}$. In the k-th iteration of* OptAug-CMDP *we have, for all $\lambda \in \mathbb{R}^I_{\geq 0}$,*

$$\begin{aligned}
&\frac{1}{2\eta_k}\left(\|\lambda_{k+1} - \lambda\|^2 - \|\lambda_k - \lambda\|^2 + \|\lambda_{k+1} - \lambda_k\|^2\right) \\
&= \sum_{i=1}^{I}(\lambda_{k+1}(i) - \lambda(i)) \max\{-\frac{\lambda_k(i)}{\eta_k}, (V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)_i\}.
\end{aligned}$$

Proof. Using $2u^T v = \|u\|^2 + \|v\|^2 - \|u - v\|^2$ with $u = \lambda_{k+1} - \lambda_k$, $v = \lambda_{k+1} - \lambda$ we find that the LHS in the Lemma reads

$$\frac{1}{2\eta_k}\left(\|\lambda_{k+1} - \lambda\|^2 - \|\lambda_k - \lambda\|^2 + \|\lambda_{k+1} - \lambda_k\|^2\right) = \frac{1}{2\eta_k}2(\lambda_{k+1} - \lambda)^T(\lambda_{k+1} - \lambda_k).$$

The update rule of for λ_{k+1} can equivalently be written as

$$\lambda_{k+1}(i) = \lambda_k(i) + \eta_k \max\{-\frac{\lambda_k(i)}{\eta_k}, (V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)_i\},$$

for all $i \in [I]$, from which we obtain $\lambda_{k+1}(i) - \lambda_k(i) = \eta_k \max\{-\frac{\lambda_k(i)}{\eta_k}, (V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)_i\}$. Plugging this into the equation above proves the Lemma. □

Lemma 16. *Let $k \in \{K' + 1, \dots, K\}$. In the k-th iteration of* OptAug-CMDP *we have for, any $\lambda \in \mathbb{R}^I_{\geq 0}$,*

$$\begin{aligned}
&\sum_{i=1}^{I}(\lambda_{k+1}(i) - \lambda(i)) \max\{-\frac{\lambda_k(i)}{\eta_k}, (V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)_i\} \\
&\leq \sum_{i=1}^{I}(\lambda_{k+1}(i) - \lambda(i))(V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)_i.
\end{aligned}$$

Proof. Fix k and set $I_+ := \{i \in [I] \mid \lambda_k(i) + \eta_k(V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)_i > 0\} = \{i \in [I] \mid -\frac{\lambda_k(i)}{\eta_k} < (V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)_i\}$ and $I_- := [I] \setminus I_+$. Then subtracting the RHS from the LHS in the Lemma we

get

$$
\begin{aligned}
&\sum_{i=1}^{I}(\lambda_{k+1}(i)-\lambda(i))\max\{-\frac{\lambda_k(i)}{\eta_k},(V^{\pi_k}(\tilde{D}_k,\tilde{p}_k)-\alpha)_i\}\\
&-\sum_{i=1}^{I}(\lambda_{k+1}(i)-\lambda(i))(V^{\pi_k}(\tilde{D}_k,\tilde{p}_k)-\alpha)_i\\
=&\sum_{i\in I_+}(\lambda_{k+1}(i)-\lambda(i))((V^{\pi_k}(\tilde{D}_k,\tilde{p}_k)-\alpha)_i-(V^{\pi_k}(\tilde{D}_k,\tilde{p}_k)-\alpha)_i)\\
&+\sum_{i\in I_-}\lambda(i)((V^{\pi_k}(\tilde{D}_k,\tilde{p}_k)-\alpha)_i+\frac{\lambda_k(i)}{\eta_k})\\
=&\sum_{i\in I_-}\lambda(i)((V^{\pi_k}(\tilde{D}_k,\tilde{p}_k)-\alpha)_i+\frac{\lambda_k(i)}{\eta_k}),
\end{aligned}
$$

and using that $\lambda\geq 0$ as well as $(V^{\pi_k}(\tilde{D}_k,\tilde{p}_k)-\alpha)_i+\frac{\lambda_k(i)}{\eta_k}\leq 0$ for $i\in I_-$ shows that this is ≤ 0, which proves the claim. □

Combining the previous two lemmas we find the analogous result of the one-step progress inequality of the standard inexact augmented Lagrangian method analysis:

Lemma 17 (One-step progress of iALM)**.** *Let $k\in\{K'+1,\dots,K\}$. In step k of* OPTAUG-CMDP*, for all $\pi\in\Pi$ and all $\lambda\geq 0$ we have*

$$
\begin{aligned}
V^{\pi_k}(\tilde{c}_k,\tilde{p}_k)+\lambda^T(V^{\pi_k}(\tilde{D}_k,\tilde{p}_k)-\alpha)+\frac{1}{2\eta_k}\|\lambda_{k+1}-\lambda\|^2\\
\leq\tilde{L}_k(\pi,\lambda_k)+\frac{1}{2\eta_k}\|\lambda_k-\lambda\|^2+\epsilon_k,
\end{aligned}
$$

where for notational convenience

$$
\tilde{L}_k(\pi,\lambda):=V^{\pi}(\tilde{c}_k,\tilde{p}_k)+\frac{1}{2\eta_k}\|[\lambda+\eta_k(V^{\pi}(\tilde{D}_k,\tilde{p}_k)-\alpha)]_+\|^2-\frac{1}{2\eta_k}\|\lambda\|^2.
$$

Proof. By ϵ_k-closeness, for all $\pi\in\Pi$ we have

$$
\begin{aligned}
V^{\pi_k}(\tilde{c}_k,\tilde{p}_k)+\frac{1}{2\eta_k}\|\lambda_{k+1}\|\leq&V^{\pi}(\tilde{c}_k,\tilde{p}_k)+\frac{1}{2\eta_k}\|\left[\lambda_k+\eta_k(V^{\pi}(\tilde{D}_k,\tilde{p}_k)-\alpha)\right]_+\|^2+\epsilon_k\\
=&\tilde{L}_k(\pi,\lambda_k)+\frac{1}{2\eta_k}\|\lambda_k\|^2+\epsilon_k.
\end{aligned}
\tag{32}
$$

In addition, again by distinguishing between indices in I_+ and I_- (see proof of Lemma 16), we see

$$
\begin{aligned}
&\frac{1}{2\eta_k}\|[\lambda_k+\eta_k(V^{\pi_k}(\tilde{D}_k,\tilde{p}_k)-\alpha)]_+\|^2-\frac{1}{2\eta_k}\|\lambda_k\|^2\\
&-\sum_{i\in[I]}[\lambda_k(i)+\eta_k(V^{\pi_k}(\tilde{D}_k,\tilde{p}_k)-\alpha)_i]_+(V^{\pi_k}(\tilde{D}_k,\tilde{p}_k)-\alpha)_i\\
=&\frac{1}{2\eta_k}\sum_{i\in I_+}\Bigg((\lambda_k(i)+\eta_k(V^{\pi_k}(\tilde{D}_k,\tilde{p}_k)-\alpha)_i)^2-\lambda_k(i)^2\\
&-2(\lambda_k(i)+\eta_k(V^{\pi_k}(\tilde{D}_k,\tilde{p}_k)-\alpha)_i)\eta_k(V^{\pi_k}(\tilde{D}_k,\tilde{p}_k)-\alpha)_i\Bigg)-\frac{1}{2\eta_k}\sum_{i\in I_-}\lambda_k(i)^2\\
=&\frac{1}{2\eta_k}\sum_{i\in I_+}-(\eta_k(V^{\pi_k}(\tilde{D}_k,\tilde{p}_k)-\alpha)_i)^2-\frac{1}{2\eta_k}\sum_{i\in I_-}\lambda_k(i)^2\\
=&\frac{1}{2\eta_k}\sum_{i\in I_+}-(\lambda_{k+1}(i)-\lambda_k(i))^2-\frac{1}{2\eta_k}\sum_{i\in I_-}(\lambda_{k+1}(i)-\lambda_k(i))^2\\
=&-\frac{1}{2\eta_k}\|\lambda_{k+1}-\lambda_k\|^2,
\end{aligned}
$$

and with Eq. (32) this shows

$$
\begin{aligned}
&V^{\pi_k}(\tilde{c}_k, \tilde{p}_k) + \sum_{i\in[I]} [\lambda_k(i) + \eta_k(V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)_i]_+ (V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)_i \\
&- \frac{1}{2\eta_k}\|\lambda_{k+1} - \lambda_k\|^2 \\
=&V^{\pi_k}(\tilde{c}_k, \tilde{p}_k) + \frac{1}{2\eta_k}\|\lambda_{k+1}\|^2 - \frac{1}{2\eta_k}\|\lambda_k\|^2 \\
\overset{Eq.\ (32)}{\leq}& \tilde{L}_k(\pi, \lambda_k) + \epsilon_k. \qquad (33)
\end{aligned}
$$

Now combining Lemma 15 and Lemma 16, we also have

$$
\begin{aligned}
&\frac{\|\lambda_{k+1} - \lambda\|^2 - \|\lambda_k - \lambda\|^2 + \|\lambda_{k+1} - \lambda_k\|^2}{2\eta_k} \\
\leq& \sum_{i\in[I]} ([\lambda_k(i) + \eta_k(V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)_i]_+ - \lambda(i))(V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)_i.
\end{aligned}
$$

Hence

$$
\begin{aligned}
&V^{\pi_k}(\tilde{c}_k, \tilde{p}_k) + \frac{1}{2\eta_k}\|\lambda_{k+1} - \lambda\|^2 - \frac{1}{2\eta_k}\|\lambda_k - \lambda\|^2 \\
\leq& V^{\pi_k}(\tilde{c}_k, \tilde{p}_k) \\
&+ \sum_{i\in[I]} ([\lambda_k(i) + \eta_k(V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)_i]_+ - \lambda(i))(V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)_i \\
&- \frac{1}{2\eta_k}\|\lambda_{k+1} - \lambda_k\|^2 \\
\overset{Eq.\ (33)}{\leq}& - \sum_{i\in[I]} \lambda(i)(V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)_i + \tilde{L}_k(\pi, \lambda_k) + \epsilon_k,
\end{aligned}
$$

and rearranging this yields the desired inequality. □

Before proving Lemma 5, we finally need to establish Lemma 18. This result is based on standard properties from constrained convex optimization (Appendix A.4) and the LP formulation of CMDPs (Appendix A.3). Recall that we have $V^\pi(\tilde{c}_k, \tilde{p}_k) = \tilde{c}_k^T q^\pi(\tilde{p}_k)$ and $V^\pi(\tilde{D}_k, \tilde{p}_k) = \tilde{D}_k q^\pi(\tilde{p}_k)$, where $q^\pi(\tilde{p}_k) \in \mathbb{R}^{SAH}$, $\tilde{c}_k \in \mathbb{R}^{SAH}$, $\tilde{d}_{i,k} \in \mathbb{R}^{SAH}$, and $\tilde{D}_k \in \mathbb{R}^{I\times SAH}$ are defined as in Appendix A.3. Thus, when switching to occupancy measures, the optimistic problem (Eq. (6)) equivalently reads

$$
\min_{q^\pi \in Q(\tilde{p}_k)} \quad \tilde{c}_k^T q^\pi \quad \text{s.t.} \quad \tilde{D}_k q^\pi - \alpha \leq 0, \qquad (34)
$$

where $Q(\tilde{p}_k) \subset \mathbb{R}^{SAH}$ is defined as in Appendix A.3. This is a convex optimization problem over the set of occupancy measures $Q(\tilde{p}_k)$.

Formally, we make the following assumption. By Lemma 1, the assumption will be guaranteed to hold with (with $\pi^0 = \bar{\pi}$ and $\sigma = \frac{H}{\nu\gamma}$), conditioned on G.

Assumption 3 (Slater points)**.** *There exists $\pi^0 \in \Pi$ such that for all $k \in \{K', \dots, K\}$ and all $i \in [I]$ we have $V^{\pi^0}(\tilde{d}_{i,k}, \tilde{p}_k) < \alpha_i$. In particular, Eq.* (6) *is feasible. Let π_k^* be an optimal solution for Eq.* (6) *and suppose there is a fixed constant $\sigma > 0$ such that, for all $k \in \{K', \dots, K\}$,*

$$
\frac{V^{\pi^0}(\tilde{c}_k, \tilde{p}_k) - V^{\pi_k^*}(\tilde{c}_k, \tilde{p}_k)}{\min_{i\in[I]}(\alpha_i - V^{\pi^0}(\tilde{d}_{i,k}, \tilde{p}_k)} \leq \sigma.
$$

Remark 3. *Note that, under Assumption 3, we can view Eq.* (6) *as the convex optimization problem in Eq.* (34) *over $Q(\tilde{p}_k)$ that satisfies all parts of Assumption 2 from Appendix A.4. Indeed,*

(a) $X := Q(\tilde{p}_k)$ is a polytope and thus convex

(b) the objective $f(\cdot) := \tilde{c}_k^T(\cdot)$ is affine and thus convex

(c) the constraints $g_i(\cdot) := \tilde{d}_{i,k}^T(\cdot) - \alpha_i$ are affine and thus convex

(d) by Assumption 3, Eq. (34) is feasible and thus its minimum is attained (since the domain is compact and the objective continuous)

(e) a Slater point exists by Assumption 3, namely $q^{\pi^0}(\tilde{p}_k)$

(f) all dual problems have an optimal solution since the domain X is compact and the objective $f(\cdot) + \lambda^T g(\cdot)$ is continuous,

where $Q(\tilde{p}_k) \subset \mathbb{R}^{SAH}$, $\tilde{c}_k \in \mathbb{R}^{SAH}$ and $\tilde{d}_{i,k} \in \mathbb{R}^{SAH}$ are defined as in Appendix A.3, and the setup is in line with the general convex optimization setup described in Appendix A.4.

Lemma 18. *If Eq. (34) is stricly feasible, then there exists a point (π_k^*, λ_k^*) with $q^{\pi_k^*} = q^{\pi_k^*}(\tilde{p}_k)$ optimal for Eq. (34) and λ_k^* optimal for its dual problem. For any such pair and any $\pi \in \Pi$ we have*

$$V^{\pi}(\tilde{c}_k, \tilde{p}_k) - V^{\pi_k^*}(\tilde{c}_k, \tilde{p}_k) + (\lambda_k^*)^T (V^{\pi}(\tilde{D}_k, \tilde{p}_k) - \alpha) \geq 0.$$

Proof. By Remark 3, we are in the setup of Eq. (14) under Assumption 2. Thus, by applying Theorem 4, in the notation of Appendix A.3 we immediately get

$$\tilde{c}_k^T q^{\pi}(\tilde{p}_k) - \tilde{c}_k^T q^{\pi_k^*}(\tilde{p}_k) + (\lambda_k^*)^T (\tilde{D}_k q^{\pi}(\tilde{p}_k) - \alpha) \geq 0,$$

which proves the claim by plugging in the value functions. □

The following lemma allows us to deduce Lemma 5.

Lemma 19. *Let $k \in \{K' + 1, \ldots, K\}$ and suppose Eq. (6) is strictly feasible. Let (π_k^*, λ_k^*) be a pair of primal-optimal and dual-optimal solutions for Eq. (6) (see Lemma 18). Then the iterates of* OptAug-CMDP *satisfy*

$$\|\lambda_{k+1} - \lambda_k^*\|^2 \leq \|\lambda_k - \lambda_k^*\|^2 + 2\eta_k \epsilon_k.$$

Proof. Fix an arbitrary point (π_k^*, λ_k^*) as in Lemma 18. Then Lemma 18 with $\pi = \pi_k$ (note that π need not satisfy the constraints) we have

$$0 \leq V^{\pi_k}(\tilde{c}_k, \tilde{p}_k) - V^{\pi_k^*}(\tilde{c}_k, \tilde{p}_k) + (\lambda_k^*)^T (V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha),$$

and by feasibility of π_k^* we have $\tilde{L}_k(\pi_k^*, \lambda_k) \leq V^{\pi_k^*}(\tilde{c}_k, \tilde{p}_k)$, with $\tilde{L}_k(\pi, \lambda)$ as defined in Lemma 17. Thus the above becomes

$$0 \leq V^{\pi_k}(\tilde{c}_k, \tilde{p}_k) - \tilde{L}_k(\pi_k^*, \lambda_k) + (\lambda_k^*)^T (V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha). \tag{35}$$

Moreover, from Lemma 17 we know that

$$\begin{aligned} 2\eta_k (V^{\pi_k}(\tilde{c}_k, \tilde{p}_k) + \lambda^T (V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)) + \|\lambda_{k+1} - \lambda\|^2 \\ \leq 2\eta_k \tilde{L}_k(\pi, \lambda_k) + \|\lambda_k - \lambda\|^2 + 2\eta_k \epsilon_k, \end{aligned}$$

and we can choose $(\pi, \lambda) = (\pi_k^*, \lambda_k^*)$ in this inequality to get

$$\begin{aligned} 2\eta_k (V^{\pi_k}(\tilde{c}_k, \tilde{p}_k) + (\lambda_k^*)^T (V^{\pi_k}(\tilde{D}_k, \tilde{p}_k) - \alpha)) + \|\lambda_{k+1} - \lambda_k^*\|^2 \\ \leq 2\eta_k \tilde{L}_k(\pi_k^*, \lambda_k) + \|\lambda_k - \lambda_k^*\|^2 + 2\eta_k \epsilon_k. \end{aligned}$$

Adding $2\eta_k$ times Eq. (35) to this and cancelling terms yields

$$\|\lambda_{k+1} - \lambda_k^*\|^2 \leq \|\lambda_k - \lambda_k^*\|^2 + 2\eta_k \epsilon_k.$$

□

From Lemma 19 we readily obtain Lemma 5.

Lemma 5. *Let $k \in \{K'+1, \dots, K\}$ and suppose Eq.* (6) *is strictly feasible for every $k' \in \{K', \dots, K\}$. Let $(\pi^*_{k'}, \lambda^*_{k'})$ be pairs of primal-optimal and dual-optimal solutions for Eq.* (6)*. Then the iterates of* OPTAUG-CMDP *satisfy*

$$\|\lambda_{k+1}\| \leq 2 \sum_{t=K'}^{k} \|\lambda^*_t\| + \sum_{t=K'+1}^{k} \sqrt{2\eta_t \epsilon_t}.$$

Proof. Taking square root and using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ in Lemma 19 yields $\|\lambda_{k+1} - \lambda^*_k\| \leq \|\lambda_k - \lambda^*_k\| + \sqrt{2\eta_k \epsilon_k}$. By triangle inequality we get

$$\|\lambda_{k+1} - \lambda^*_k\| \leq \|\lambda_k - \lambda^*_k\| + \sqrt{2\eta_k \epsilon_k} \leq \|\lambda_k - \lambda^*_{k-1}\| + \|\lambda^*_{k-1} - \lambda^*_k\| + \sqrt{2\eta_k \epsilon_k},$$

and thus by induction

$$\begin{aligned}
\|\lambda_{k+1} - \lambda^*_k\| \leq & \|\lambda_{K'+1} - \lambda^*_{K'}\| + \sum_{t=K'+1}^{k} \|\lambda^*_{t-1} - \lambda^*_t\| + \sum_{t=K'+1}^{k} \sqrt{2\eta_t \epsilon_t} \\
\leq & \|\lambda_{K'+1}\| + \|\lambda^*_{K'}\| + \sum_{t=K'+1}^{k} (\|\lambda^*_{t-1}\| + \|\lambda^*_t\|) + \sum_{t=K'+1}^{k} \sqrt{2\eta_t \epsilon_t} \\
= & \|\lambda^*_k\| + 2 \sum_{t=K'}^{k-1} \|\lambda^*_t\| + \sum_{t=K'+1}^{k} \sqrt{2\eta_t \epsilon_t},
\end{aligned}$$

again using the triangle inequality and that $\lambda_{K'+1} = 0$. The claim now follows by invoking the inverse triangle inequality on the LHS and rearranging. □

E.3.3 Bounding the Dual Maximizers

We can now deduce Lemma 6 from the preliminaries shown in Appendix A.4, as follows. See Lemma 18 from the previous section for a formal introduction of the primal-dual pairs (π^*_k, λ^*_k).

Lemma 6. *Suppose Assumption 1 holds. Let $\nu \in (0,1)$ and choose K' as in Lemma 1. Let $k \in \{K', \dots, K\}$, and let (π^*_k, λ^*_k) be a pair of primal-optimal and dual-optimal solutions for Eq.* (6)*. Then, conditioned on G, we have*

$$\|\lambda^*_k\| \leq \|\lambda^*_k\|_1 \leq \frac{H}{\nu\gamma}.$$

Proof. Conditioned on G, the conclusion from Lemma 1 holds. That is, for every $k \in \{K', \dots, K\}$ and every $i \in [I]$ we have

$$V^{\bar{\pi}}(\tilde{d}_{i,k}, \tilde{p}_k) \leq \alpha_i - \nu\gamma,$$

or equivalently, $\min_{i \in [I]}(\alpha_i - V^{\bar{\pi}}(\tilde{d}_{i,k}, \tilde{p}_k)) \geq \nu\gamma$. Moreover, since the expected costs are in $[0,1]$ and the time horizon is H, the value functions are in $[0, H]$, so we have $V^{\bar{\pi}}(\tilde{c}_k, \tilde{p}_k) - V^{\pi^*_k}(\tilde{c}_k, \tilde{p}_k) \leq H$. Hence Assumption 3 holds, with $\pi^0 = \bar{\pi}$ and $\sigma = \frac{H}{\nu\gamma}$.

As discussed in Remark 3, under Assumption 3 we can apply the general bound on dual maximizers from Theorem 5. For this, we write Eq. (6) as a convex optimization problem in the occupancy measure (see Eq. (34)), with $V^{\pi}(\tilde{c}_k, \tilde{p}_k) = \tilde{c}_k^T q^{\pi}(\tilde{p}_k)$ and $V^{\pi}(\tilde{D}_k, \tilde{p}_k) = \tilde{D}_k q^{\pi}(\tilde{p}_k)$. Then, set $X = Q(\tilde{p}_k)$, $\bar{x} = q^{\bar{\pi}}(\tilde{p}_k)$, $f(\cdot) = \tilde{c}_k^T(\cdot)$ and $g_i(\cdot) = \tilde{d}_{i,k}^T(\cdot)$ as in Remark 3. Plugging this into Theorem 5 indeed yields the second of the claimed inequalities since we have shown that Assumption 3 holds with $\pi^0 = \bar{\pi}$, $\sigma = \frac{H}{\nu\gamma}$. The first inequality holds because λ^*_k only has non-negative entries. □

E.3.4 Optimization Error Bound

We are now ready to prove the desired bound on the optimization errors.

Lemma 7 (Optimization errors)**.** *Suppose Assumption 1 holds. Let $\nu \in (0,1)$ and choose K' as in Lemma 1. Suppose that the event G occurs. When using step sizes $\eta_{K'+k} = \Theta(k^{2.5})$ and $\epsilon_{K'+k} = \Theta(1/\eta_{K'+k})$, we have*

$$\sum_{k=K'+1}^{K} [V^{\pi_k}(\tilde{c}_k, \tilde{p}_k) - V^{\pi^*}(c,p)]_+ \leq \sum_{k=K'+1}^{K} \left(\frac{(O(\sigma k) + \sum_{t=K'+1}^{k} \sqrt{2\eta_t \epsilon_t})^2}{2\eta_k} + \epsilon_k \right) \leq O(\sqrt{K}),$$

$$\max_{i\in[I]} \sum_{k=K'+1}^{K} [V^{\pi_k}(\tilde{d}_{i,k}, \tilde{p}_k) - \alpha_i]_+ \leq \sum_{k=K'+1}^{K} \frac{O(\sigma k) + \sum_{t=K'+1}^{k} \sqrt{2\eta_t \epsilon_k}}{\eta_k} \leq O(\sqrt{K}),$$

where $\sigma = \frac{H}{\nu\gamma}$ and in fact $O(\sigma k)$ can be replaced by $(2 + 2(k - K'))\sigma$.

Proof. Recall that conditioned on G, Lemma 4 shows ($k \in \{K'+1, \dots, K\}$)

$$[V^{\pi_k}(\tilde{c}_k, \tilde{p}_k) - V^{\pi^*}(c,p)]_+ \leq \epsilon_k + \frac{\|\lambda_k\|^2}{2\eta_k}, \tag{36}$$

$$[V^{\pi_k}(\tilde{d}_{i,k}, \tilde{p}_k) - \alpha_i]_+ \leq \frac{\|\lambda_{k+1}\|}{\eta_k}, \tag{37}$$

when dropping the negative terms in Lemma 4. Moreover, from Lemma 5 and Lemma 6 we get

$$\begin{aligned}
\|\lambda_{k+1}\| &\leq 2 \sum_{t=K'}^{k} \|\lambda_t^*\| + \sum_{t=K'+1}^{k} \sqrt{2\eta_t \epsilon_t} \\
&\leq 2 \sum_{t=K'}^{k} \sigma + \sum_{t=K'+1}^{k} \sqrt{2\eta_t \epsilon_t} \\
&= (2 + 2(k - K'))\sigma + \sum_{t=K'+1}^{k} \sqrt{2\eta_t \epsilon_t},
\end{aligned} \tag{38}$$

where $\sigma = \frac{H}{\nu\gamma}$. Note that in the bound from Eq. (36), for $k = K'+1$ we have $\|\lambda_{K'+1}\|^2 = 0 \leq ((2 + 2 \cdot 1)\sigma + \sqrt{2\eta_{K'+1}\epsilon_{K'+1}})^2$ since $\lambda_{K'+1} = 0$. And for $k \geq K'+2$, by Eq. (38) we have $\|\lambda_k\|^2 \leq ((2+2((k-1)-K'))\sigma + \sum_{t=K'+1}^{k-1} \sqrt{2\eta_t \epsilon_t})^2 \leq ((2+2(k-K'))\sigma + \sum_{t=K'+1}^{k} \sqrt{2\eta_t \epsilon_t})^2$. Hence, by plugging this into Eq. (36) and summing up, the parameter-dependent bound on the objective errors follows. For the constraint errors, we can directly plug in the bound from Eq. (38) into Eq. (37), thus obtaining the second parameter-dependent bound.

We now show how to obtain bounds of order $O(\sqrt{K})$ from this. Note that we can always choose $\epsilon_k := \frac{1}{2\eta_k}$ so that $\sum_{t=K'+1}^{k} \sqrt{2\eta_t \epsilon_t} = k - K' \leq \sigma(k - K')$, since $\sigma = \frac{H}{\nu\gamma} > 1$. If we do so, then we can loosely bound the term in Eq. (38) by

$$(2 + 2(k - K'))\sigma + \sum_{t=K'+1}^{k} \sqrt{2\eta_t \epsilon_t} \leq (2 + 3(k - K'))\sigma.$$

As seen above, with this, Eq. (36) and Eq. (37) become

$$[V^{\pi_k}(\tilde{c}_k, \tilde{p}_k) - V^{\pi^*}(c,p)]_+ \leq \epsilon_k + \frac{((2 + 3(k - K'))\sigma)^2}{2\eta_k}, \tag{39}$$

$$[V^{\pi_k}(\tilde{d}_{i,k}, \tilde{p}_k) - \alpha_i]_+ \leq \frac{(2 + 3(k - K'))\sigma}{\eta_k}. \tag{40}$$

We can now choose[12]

$$\eta_{K'+k} := ((2 + 3k)\sigma)^{2.5} \qquad (k > 0)$$

[12]Note that if γ is not known, having an estimate of $\sigma = \frac{H}{\nu\gamma}$ that is larger than the true value is clearly sufficient.

and sum each of the two errors above to bound the regret due to optimization. For the objective cost, we get (recall that we had set $\epsilon_k = 1/(2\eta_k)$)

$$\begin{aligned}
&\sum_{k=K'+1}^{K} [V^{\pi_k}(\tilde{c}_k, \tilde{p}_k) - V^{\pi^*}(c,p)]_+ \\
&\overset{Eq.\ (39)}{\leq} \sum_{k=K'+1}^{K} \epsilon_k + +\frac{1}{2} \sum_{k=K'+1}^{K} \frac{((2+3(k-K'))\sigma)^2}{((2+3(k-K'))\sigma)^{2.5}} \\
&\leq \frac{1}{2} \sum_{k=1}^{K} \frac{1}{((2+3k)\sigma)^{2.5}} + \frac{1}{2} \sum_{k=1}^{K} \frac{1}{\sqrt{(2+3k)\sigma}} \\
&= O(K^{1/2}),
\end{aligned}$$

up to absolute constants, where the last step follows by considering the dominating series $\sum_{k\geq 1} 1/k^2 = \pi^2/6$, and using $\sum_{k=1}^{K} 1/\sqrt{k} = \Theta(\sqrt{K})$, as well as $\sigma > 1$. Similarly, for the constraint cost, we get

$$\begin{aligned}
\sum_{k=K'+1}^{K} [V^{\pi_k}(\tilde{d}_{i,k}, \tilde{p}_k) - \alpha_i]_+ \leq & \sum_{k=K'+1}^{K} \frac{((2+3(k-K'))\sigma)}{((2+3k)\sigma)^{2.5}} \\
\leq & \sum_{k=1}^{K} \frac{1}{((2+3k)\sigma)^{1.5}} \\
= & O(K^{1/2}),
\end{aligned}$$

up to absolute constants. □

We remark that the choices for η_k. ϵ_k are not necessarily optimal and there is a trade-off between the regret due to the optimization error and the iteration complexity of the inner loop, as we discuss in Remark 2.