Truly No-Regret Learning in Constrained MDPs

Adrian Müller [1] **Pragnya Alatur** [2] **Volkan Cevher** [1] **Giorgia Ramponi** [3] **Niao He** [2]

Abstract

Constrained Markov decision processes (CMDPs) are a common way to model safety constraints in reinforcement learning. State-of-the-art methods for efficiently solving CMDPs are based on primal-dual algorithms. For these algorithms, all currently known regret bounds allow for *error cancellations* — one can compensate for a constraint violation in one round with a strict constraint satisfaction in another. This makes the online learning process unsafe since it only guarantees safety for the final (mixture) policy but not during learning. As Efroni et al. (2020) pointed out, it is an open question whether primal-dual algorithms can provably achieve sublinear regret if we do not allow error cancellations. In this paper, we give the first affirmative answer. We first generalize a result on last-iterate convergence of regularized primal-dual schemes to CMDPs with multiple constraints. Building upon this insight, we propose a model-based primal-dual algorithm to learn in an unknown CMDP. We prove that our algorithm achieves sublinear regret without error cancellations.

1. Introduction

Classical reinforcement learning (RL, Sutton & Barto, 2018) aims to solve sequential decision-making problems under uncertainty. It involves learning a policy while interacting with an unknown Markov decision process (MDP, Bellman, 1957). However, in many real-world situations, RL algorithms need to solve the task while respecting certain safety constraints. For example, in autonomous driving and drone navigation, we must avoid collisions and adhere to traffic rules to ensure safe behavior (Brunke et al., 2022). Such safety requirements are commonly described by constrained Markov decision processes (CMDPs, Altman, 1999). In CMDPs, the goal is to maximize the expected cumulative reward while subject to multiple safety constraints, each modeled by a different expected cumulative reward signal that needs to lie above a respective threshold. We consider the finite-horizon setting, in which an algorithm chooses a policy in each episode, plays it for one episode, and observes the random transitions, rewards, and constraint rewards along its trajectory.

In the literature, there are three standard approaches for finding an optimal policy in a known CMDP: linear programming (LP, Altman, 1999), primal-dual (Paternain et al., 2022), and dual algorithms (Paternain et al., 2019). If the CMDP is unknown, a common approach to handle the uncertainty is the classical paradigm of optimism in the face of uncertainty (Auer et al., 2008). In their influential paper, Efroni et al. (2020) established comprehensive regret guarantees for all three types of optimistic algorithms in the online setup. In practice, especially primal-dual algorithms are preferred due to their high computational efficiency and flexibility for policy parameterization, thereby scaling to high-dimensional problems (Chow et al., 2017; Achiam et al., 2017; Tessler et al., 2018). Thus, it is important to rigorously understand the fundamental properties of this algorithm class. Indeed, there has been a large number of studies on primal-dual (and dual) approaches for CMDPs (Ding et al., 2020; 2022b; Liu et al., 2021b; Ding & Jovanović, 2022; Ghosh et al., 2022; Ding & Lavaei, 2022; Qiu et al., 2020; Liu et al., 2021a; Bai et al., 2022, to list just a few) since the work of Efroni et al. (2020).

However, unlike for LP-based algorithms, the known bounds for primal-dual (and dual) algorithms suffer from the fundamental limitation pointed out by Efroni et al. (2020, Section 2.2): they concern a *weaker*, less safe notion of regret. More precisely, the known guarantees bound the sum of the suboptimalities and the sum of the constraint violations across episodes, where one episode corresponds to one round of learning. However, a policy can have a negative constraint violation (by being very safe but obtaining a lower return than an optimal safe policy) or a positive constraint violation (by being unsafe but obtaining a higher return than an optimal safe policy). Thus, terms from these two cases can cancel each other out when summing the violations across episodes, a phenomenon referred to as *error cancellations* (Efroni et al., 2020). An algorithm with sublinear weak regret may heavily violate safety constraints during learning. For example, if the policies alternate between the two cases

[1]EPFL [2]ETH Zürich [3]University of Zürich. Correspondence to: Adrian Müller <adrian.luis.mueller@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

above in every other episode, the algorithm may even obtain *zero* regret despite being unsafe every second episode. This weak notion of regret falls short of capturing safety in a setup with no simulator and where the algorithm must adhere to constraints during learning. In fact, these cancellations are not a weakness in the analysis but rather due to oscillations of the underlying optimization method, which converges on average but not in the last-iterate (Efroni et al., 2020; Beck, 2017). Indeed, these oscillations are observed in practice (Stooke et al., 2020; Moskovitz et al., 2023).

We thus consider a stronger notion of regret that concerns the sum of the *positive parts* of the error terms instead. This regret does not allow for error cancellations, and we refer to it as *strong regret*. The results of this paper address the research question pointed out by Efroni et al. (2020):

Can we design an efficient primal-dual algorithm that achieves sublinear strong regret in an unknown CMDP?

We provide the first affirmative answer for tabular finite-horizon CMDPs. Specifically, we introduce a regularization framework inspired by the recent work of Ding et al. (2023) and derive guarantees in the online setup for a primal-dual algorithm that arises from this formulation.

Contributions Our main contributions are the following:

- We first prove non-asymptotic policy last-iterate convergence (Definition 4.1) of a regularized primal-dual scheme for CMDPs despite the inherent non-concavity, assuming access to a value function oracle (Section 4). Our guarantee generalizes previous results for the strictly easier problem of CMDPs with only a single constraint. This is the first analysis that establishes last-iterate convergence of primal-dual algorithms in *arbitrary* CMDPs.
- Combining this regularized primal-dual scheme with optimistic exploration, we propose an improved model-based primal-dual algorithm (Algorithm 1) for online learning in CMDPs (Section 5). Our algorithm requires no prior knowledge of the CMDP and maintains value-optimism for the *regularized* problem.
- Finally, we establish that our algorithm achieves sublinear *strong* regret when learning an unknown CMDP (Section 5.2). This is the first primal-dual algorithm achieving a sublinear regret guarantee without allowing error cancellations, providing the first answer to the open question posed by Efroni et al. (2020).

The latter is relevant due to the efficiency and practical importance of primal-dual algorithms, which are often preferred over LP-based algorithms in large-scale applications. Additionally, we provide numerical evaluations of our algorithm in simple environments. We illustrate that it exhibits sublinear regret when safety during learning is concerned, while unregularized algorithms do not. We conclude that error cancellations are not merely a hypothetical issue of existing algorithms but bear practical relevance.

1.1. Related Work

Since Efroni et al. (2020) analyzed the vanilla primal-dual (and dual) algorithm, their analysis has been extended in various works, both for the case of an unknown or known CMDP (Ding et al., 2020; 2022b; Liu et al., 2021b; Ding & Jovanović, 2022; Ghosh et al., 2022; Ding & Lavaei, 2022; Qiu et al., 2020; Liu et al., 2021a; Bai et al., 2022). As Calvo-Fullana et al. (2023) pointed out, even the works assuming full knowledge of the CMDP only establish convergence of the averaged iterates. Hence, none of the mentioned works provides a guarantee for the strong regret or is easily amendable to obtain one.

Very recently, Ding et al. (2023) were the first to provide a last-iterate convergence analysis for a primal-dual algorithm in a known discounted infinite-horizon CMDP closely related to our algorithm. However, their analysis is limited to the case of a single constraint, which is non-trivial to generalize to multiple constraints (Section 4). Moreover, the authors left it as an open question whether the algorithm can be generalized to achieve last-iterate convergence in the online setup, when the CMDP is unknown (Section 5). In addition, our analysis holds for an algorithm with closed-form updates (Equations (6) and (7)), while their algorithm involves Bregman projections for technical reasons (Lemma 4.1). Prior, Moskovitz et al. (2023) showed last-iterate convergence of a primal-dual scheme, but their analysis concerns a hypothetical algorithm whose implicit updates do not allow efficient implementation. Li et al. (2021) provided a dual (not primal-dual) algorithm based on regularization like ours but only proved convergence for a history-weighted mixture policy[1] in a known CMDP. Similarly, Ying et al. (2022) derived a dual algorithm with last-iterate convergence but left it open whether an online version is possible. Müller et al. (2023) were the first to prove a sublinear regret guarantee without error cancellations for a dual algorithm in the online setup. However, their algorithm, which is based on the augmented Lagrangian method, lacks the desired computational efficiency. We refer to Appendix B for further comparison with prior results.

2. Problem Formulation

Notation For $n \in \mathbb{N}$, we use $[n]$ to refer to the set of integers $\{1, \dots, n\}$. For a finite set X, we denote the probability simplex over X as $\Delta(X) = \{v \in [0,1]^X \mid \sum_{x \in X} v_x = 1\}$. For $a \in \mathbb{R}$, we set $[a]_+ := \max\{0, a\}$ to be the positive

[1] By mixture policy we refer to a policy randomly drawn from all policy iterates.

part of a. $\|\boldsymbol{b}\|$ denotes the ℓ_2-norm of a vector $\boldsymbol{b} \in \mathbb{R}^n$. $\tilde{O}$-notation refers to asymptotics up to poly-logarithmic factors.

Constrained MDPs A finite-horizon CMDP with state and action spaces $\mathcal{S}$, $\mathcal{A}$ (with finite cardinalities S and A) and horizon $H > 0$ is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, p, r, \boldsymbol{u}, \boldsymbol{c})$. Every episode consists of H steps and starts from an initial state $s_1 \in \mathcal{S}$.[2] At every step 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, $r_h \colon \mathcal{S} \times \mathcal{A} \to [0,1]$, $(s,a) \mapsto r_h(s,a)$ denotes the reward function at step $h \in [H]$. Similarly, $\boldsymbol{u}_h \colon \mathcal{S} \times \mathcal{A} \to [0,1]^I$, $(s,a) \mapsto \boldsymbol{u}_h(s,a) = (u_{1,h}(s,a), \dots, u_{I,h}(s,a))^T \in [0,1]^I$ refers to the I constraint reward functions, and $\boldsymbol{c} \in [0,H]^I$ are the respective thresholds c_i for the i-th constraint ($i \in [I]$).

The algorithm interacts with the CMDP by playing a policy $\pi \in \Pi$, where

$$\Pi := \left\{ (\pi_1, \dots, \pi_H) \;\middle|\; \forall h \, \forall s \in \mathcal{S} \colon \pi_h(\cdot|s) \in \Delta(\mathcal{A}) \right\}.$$

For any $\pi \in \Pi$, we consider the Markov process given by $a_h \sim \pi_h(\cdot|s_h)$, $s_{h+1} \sim p_h(\cdot|s_h, a_h)$ for $h = 1, \dots, H$. For any function $r' \colon [H] \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $(h,s,a) \mapsto r'_h(s,a)$, every $(s,h) \in \mathcal{S} \times [H]$ and $\pi \in \Pi$, consider the value functions

$$V^{\pi}_{r',h}(s) := \mathbb{E}_{\pi}\left[\sum_{h'=h}^{H} r'_{h'}(s_{h'}, a_{h'}) \;\middle|\; s_h = s \right],$$
$$Q^{\pi}_{r',h}(s,a) := \mathbb{E}_{\pi}\left[\sum_{h'=h}^{H} r'_{h'}(s_{h'}, a_{h'}) \;\middle|\; s_h = s,\ a_h = a \right].$$

For notational convenience, we drop the indices for the step and state if we refer to $h = 1$ and s_1 and write $V^{\pi}_{r'} = V^{\pi}_{r',1}(s_1)$.

In the CMDP setting, we are interested in solving the following optimization problem:

$$\max_{\pi \in \Pi} \quad V^{\pi}_{r} \quad \text{s.t.} \quad V^{\pi}_{u_i} \geq c_i \quad (\forall i \in [I]), \tag{1}$$

and we fix an *optimal solution* $\pi^\star \in \Pi$ for Equation (1). Among all policies that are feasible with respect to the I safety constraints $V^{\pi}_{u_i} \geq c_i$, the goal is to find one that maximizes V^{π}_{r}.

We consider the stochastic reward setting, in which the algorithm observes rewards sampled from random variables $R_h(s,a) \in [0,1]$ and $\boldsymbol{U}_h(s,a) \in [0,1]^I$ such that $\mathbb{E}[R_h(s,a)] = r_h(s,a)$ and $\mathbb{E}[U_{i,h}(s,a)] = u_{i,h}(s,a)$ for all $i \in [I]$ when taking action a in state s at step h. Throughout, we make the following assumption, which is standard in the context of CMDPs (Altman, 1999; Efroni et al., 2020; Li et al., 2021; Ying et al., 2022; Ding et al., 2022c; Paternain et al., 2022; Ding et al., 2023).

Assumption 2.1 (Slater policy)**.** *There exists* $\bar{\pi} \in \Pi$ *and* $\boldsymbol{\xi} \in \mathbb{R}^I_{>0}$ *such that* $V^{\bar{\pi}}_{u_i} \geq c_i + \xi_i$ *for all* $i \in [I]$. *Set the* Slater gap

$$\Xi := \min_{i \in [I]} \xi_i.$$

This assumption asserts that there exists a policy that strictly satisfies the constraints.

Problem Formulation The algorithm interacts with the unknown CMDP over a fixed number of $K > 0$ episodes. Prior to every episode $k \in [K]$, the algorithm selects a policy $\pi_k \in \Pi$ and plays it for one run of the CMDP. The goal is to simultaneously minimize its two *strong regrets*:

$$\mathcal{R}(K; r) := \sum_{k \in [K]} \left[V^{\pi^\star}_{r} - V^{\pi_k}_{r} \right]_+, \tag{Objective}$$
$$\mathcal{R}(K; \boldsymbol{u}) := \max_{i \in [I]} \sum_{k \in [K]} \left[c_i - V^{\pi_k}_{u_i} \right]_+. \tag{Constraints}$$

Only when a policy has a suboptimal objective or violates the constraints, this counts to the respective regret. All existing works on primal-dual (and dual) algorithms (e.g., Liu et al., 2021a; Efroni et al., 2020; Bai et al., 2022; Ding et al., 2022a;c) only prove sublinear guarantees on a *weaker* notion:

$$\mathcal{R}_{\text{weak}}(K; r) := \sum_{k \in [K]} \left(V^{\pi^\star}_{r} - V^{\pi_k}_{r} \right),$$
$$\mathcal{R}_{\text{weak}}(K; \boldsymbol{u}) := \max_{i \in [I]} \sum_{k \in [K]} \left(c_i - V^{\pi_k}_{u_i} \right).$$

The weak regrets allow for the aforementioned *error cancellations* as positive and negative terms count toward each of the regrets. Even if they are sublinear in K (in fact, even if they are *zero*), the algorithm may continue compensating for a constraint violation in one episode with strict constraint satisfaction in another. On the other hand, a sublinear bound on the *stronger* notion of regret guarantees that the algorithm achieves a low constraint violation in most episodes (see Section 5.3). This is crucial for many practical applications where we do not have access to a simulator, but we have to learn our optimal policy in an *online* fashion. In the example of navigating an autonomous vehicle or drone, one would want to avoid crossing the boundaries of a specified track in each episode during learning. It is not helpful to compensate for crashing the vehicle into a wall by driving overly safely in the next episode. However, from a theoretical perspective, it is strictly more challenging to provide a guarantee for the strong regret than for the weaker notion.[3]

[2] It is straightforward to extend this to any initial distribution μ.

[3] For practical purposes, one may consider the strong regret only

3. Primal-Dual Scheme

Vanilla Scheme Primal-dual and dual algorithms arise from the equivalent Lagrangian formulation (Altman, 1999) of Equation (1):

$$\max_{\pi\in\Pi}\min_{\boldsymbol{\lambda}\in\mathbb{R}^I_{\geq 0}} \mathcal{L}(\pi,\boldsymbol{\lambda}), \tag{2}$$

where

$$\mathcal{L}(\pi,\boldsymbol{\lambda}) := V_r^\pi + \sum_{i\in[I]}\lambda_i(V_{u_i}^\pi - c_i) = V^\pi_{r+\boldsymbol{\lambda}^T(\boldsymbol{u}-H^{-1}\boldsymbol{c})}$$

is the *Lagrangian*. Paternain et al. (2019) showed that CMDPs exhibit strong duality, by which Equation (2) is equivalent to finding a saddle point $(\pi^\star,\boldsymbol{\lambda}^\star)$ of the Lagrangian. Primal-dual algorithms solve this saddle point problem via iterated play between two no-regret dynamics for π and $\boldsymbol{\lambda}$. Typically, as considered by Efroni et al. (2020) in the regret minimization setting,

$$\pi_{k+1,h}(a|s) \propto \pi_{k,h}(a|s)\exp\left(\eta Q^{\pi_k}_{r+\boldsymbol{\lambda}_k^T\boldsymbol{u},h}(s,a)\right), \tag{3}$$

$$\boldsymbol{\lambda}_{k+1} = \text{proj}_\Lambda\left(\boldsymbol{\lambda}_k - \eta(\boldsymbol{V}_{\boldsymbol{u}}^{\pi_k} - \boldsymbol{c})\right), \tag{4}$$

where proj_Λ refers to the projection onto a predefined $\Lambda = [0,\lambda_{max}]^I$, which amounts to truncating the coordinates. We refer to Equations (3) and (4) as *vanilla primal-dual scheme*. The mixture policy of the iterates is guaranteed to converge to an optimal solution pair of the min-max problem. However, the last iterate is not guaranteed to converge. Instead, the method oscillates around an optimal solution, which results in the weak regret bounds of previous primal-dual algorithms (Section 6).

Regularized Scheme The key idea of the regularization is to induce strict concavity in the primal variable (to be precise, in the state-action occupancy measure $d_h^\pi(s,a) := P_\pi[s_h = s, a_h = a]$ and not in the policy) and strong convexity in the dual variable $\boldsymbol{\lambda}$. This enables us to establish convergence to the unique solution of the regularized problem. We then show how to retrieve an error bound for the original, unregularized problem by carefully choosing the amount of regularization.

For $\tau > 0$, we define the *regularized Lagrangian* $\mathcal{L}_\tau\colon \Pi\times\mathbb{R}^I\to\mathbb{R}$ as

$$\mathcal{L}_\tau(\pi,\boldsymbol{\lambda}) := \mathcal{L}(\pi,\boldsymbol{\lambda}) + \tau\left(\mathcal{H}(\pi) + \frac{1}{2}\|\boldsymbol{\lambda}\|^2\right),$$

where $\mathcal{H}(\pi) := -\mathbb{E}_\pi[\sum_{h=1}^H \log(\pi_h(a_h|s_h))]$ is the *entropy* of a policy π. Then, consider the following *regularized CMDP problem*:

$$\max_{\pi\in\Pi}\min_{\boldsymbol{\lambda}\in\Lambda} \mathcal{L}_\tau(\pi,\boldsymbol{\lambda}), \tag{5}$$

for the constraint violations and the weak one for the objective. We refer to Appendix G for a discussion of the differences. However, this relaxation does not improve our theoretical results.

The domain of the dual variable $\boldsymbol{\lambda}$ is now a compact set $\Lambda := [0,\lambda_{max}]^I$, with $\lambda_{max} \geq H\Xi^{-1}$ to be specified (crucially, we will choose it depending on the number of episodes K). Thanks to strong duality of the unregularized problem, any saddle point $(\pi^\star,\boldsymbol{\lambda}^\star)$ of $\mathcal{L}$ satisfies $\|\boldsymbol{\lambda}^\star\| \leq H\Xi^{-1}$ (e.g., (Ying et al., 2022) for infinite horizon), which will allow us to constrain the dual variable as above. We denote the regularized primal and dual optimizers as follows:

$$\pi_\tau^\star = \arg\max_{\pi\in\Pi}\min_{\boldsymbol{\lambda}\in\Lambda} \mathcal{L}_\tau(\pi,\boldsymbol{\lambda}),$$

$$\boldsymbol{\lambda}_\tau^\star = \arg\min_{\boldsymbol{\lambda}\in\Lambda}\max_{\pi\in\Pi} \mathcal{L}_\tau(\pi,\boldsymbol{\lambda}).$$

Regularization preserves strong duality (Appendix C), by which we are equivalently looking for a saddle point $(\pi_\tau^\star,\boldsymbol{\lambda}_\tau^\star)$ of the regularized Lagrangian $\mathcal{L}_\tau$. Ding et al. (2023) proposed to perform the ascent-descent scheme in Equations (3) and (4) on $\mathcal{L}_\tau$ rather than $\mathcal{L}$ in the discounted infinite-horizon setting given a value function oracle:

$$\pi_{k+1,h}(a|s) \propto \pi_{k,h}(a|s)\exp\left(\eta Q^{\pi_k}_{r+\boldsymbol{\lambda}_k^T\boldsymbol{u}+\tau\psi_{k,h}}(s,a)\right), \tag{6}$$

$$\boldsymbol{\lambda}_{k+1} = \text{proj}_\Lambda\left((1-\eta\tau)\boldsymbol{\lambda}_k - \eta(\boldsymbol{V}_{\boldsymbol{u}}^{\pi_k} - \boldsymbol{c})\right), \tag{7}$$

where $\psi_{k,h}(s,a) := -\log(\pi_{k,h}(a|s))$. We refer to Equations (6) and (7) as *regularized primal-dual scheme*. In fact, our scheme above is a simplification of Ding et al. (2023)'s algorithm, since their policy update would read $\pi_{k+1,h}(\cdot|s) = \arg\max_{\pi_h(\cdot|s)\in\widehat{\Delta}(\mathcal{A})}\langle \pi_h(\cdot|s), Q^{\pi_k}_{r+\boldsymbol{\lambda}_k^T\boldsymbol{u}+\tau\psi_{k,h}}(s,\cdot)\rangle - \frac{1}{\eta}\text{KL}(\pi_h(\cdot|s)\|\pi_{k,h}(\cdot|s))$, where $\widehat{\Delta}(\mathcal{A}) := \{\pi_h(\cdot|s) \in \Delta(\mathcal{A}) \mid \forall a\in\mathcal{A}\colon \pi_h(a|s) \geq \varepsilon_0/A\}$ for some $\varepsilon_0 > 0$ is a *restricted* probability simplex for technical reasons stemming from the analysis. While this update can be performed via Bregman projections (Orabona, 2019) of the KL divergence onto $\widehat{\Delta}(\mathcal{A})$, this requires solving a convex program at every iteration k of the scheme. In contrast, our scheme admits a closed form of the policy update in Equation (6) due to our modified analysis (see discussion of Lemma 4.1).

4. Last-iterate Convergence

In this section, we prove last-iterate convergence of the regularized primal-dual scheme (Equations (6) and (7)) with an exact value function oracle (e.g., via policy evaluation if the true model is known) for an arbitrary number of constraints. We define last-iterate convergence as follows.

Definition 4.1 (Last-iterate convergence)**.** *A method producing policy iterates $\pi_k \in \Pi$ ($k = 1,2,\dots$) is* last-iterate convergent *if*

$$V_r^{\pi^\star} - V_r^{\pi_k} \to 0 \quad \textit{and} \quad [c_i - V_{u_i}^{\pi_k}]_+ \to 0 \quad (\forall i\in[I])$$

as $k\to\infty$.

The main technical challenges we overcome to show last-iterate convergence are: (a) to prove ascent properties for the primal update (Equation (6)), which optimizes a nonconcave objective with surrogate gradients $Q^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{u}+\tau\psi_{k,h}}(s,a)$ that are unbounded in general; and (b) to bound all *unregularized* constraint violations of the last iterate π_k in the presence of more than one constraint. We provide all proofs for this section in Appendix D.

Regularized Optimizers Our first step is to show that the iterates $(\pi_k, \boldsymbol{\lambda}_k)$ converge to the regularized optimizers $(\pi^\star_\tau, \boldsymbol{\lambda}^\star_\tau)$. Indeed, we formalize this by showing that the *potential function*

$$\Phi_k := \sum_{s,h} P_{\pi^\star_\tau}[s_h = s]\mathrm{KL}_{k,h}(s) + \frac{1}{2}\left\|\boldsymbol{\lambda}^\star_\tau - \boldsymbol{\lambda}_k\right\|^2$$

approaches zero, if we choose the regularization parameter τ and the step size η sufficiently small. Here, $\mathrm{KL}_{k,h}(s) := \mathrm{KL}(\pi^\star_{\tau,h}(\cdot|s), \pi_{k,h}(\cdot|s))$ refers to the Kullback-Leibler divergence between the optimal and the k-th policy, and $P_{\pi^\star_\tau}$ refers to the probability distribution under policy $\pi^\star_\tau$.

Lemma 4.1 (Regularized convergence)**.** *Let $\eta, \tau < 1$ and $\lambda_{max} \geq H\Xi^{-1}$. The iterates in Equations (6) and (7) satisfy*

$$\Phi_{k+1} \leq (1-\eta\tau)^k \Phi_1 + \tilde{O}\left(\eta\tau^{-1} C_{\eta,\tau,\Lambda}\right),$$

where

$$C_{\eta,\tau,\Lambda} = \lambda_{max}^2 H^3 A^{1/2} I^2 \exp\left(\eta H\left(1 + \lambda_{max} I + \log(A)\right)\right) + I\left(H + \tau\lambda_{max}\right)^2.$$

Despite the exponential term, we can control the factor $C_{\eta,\tau,\Lambda}$ to be constant of order $\mathrm{poly}(A, H, I, \Xi^{-1})$ by choosing $\eta < (H\lambda_{max} I \log(A))^{-1}$. For the remaining part, η and τ need to be traded off to have fast linear convergence $(1-\eta\tau)^k\Phi_1$ and a small bias term $\eta\tau^{-1}C_{\eta,\tau,\Lambda}$ simultaneously. Ding et al. (2023) showed a similar result with a different constant $C_{\eta,\tau}$ for their update rule that constrains the policies to the restricted probability simplex $\widehat{\Delta}(\mathcal{A}) := \{\pi_h(\cdot|s) \in \Delta(\mathcal{A}) \mid \forall a \in \mathcal{A}\colon \pi_h(a|s) \geq \varepsilon_0/A\}$ by solving a convex problem in every iteration. They introduce this restriction as their proof requires a uniform bound of $Q^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{u}+\tau\psi_{k,h}}(s,a)$ and thus of $\tau\psi_{k,h} = -\tau\log(\pi_{k,h}(s,a))$, which may be unbounded outside of $\widehat{\Delta}(\mathcal{A})$. Our modified proof overcomes this challenge by leveraging a mirror descent (MD) lemma with local norms rather than the standard online MD lemma (Orabona, 2019). While the standard norm of the regularized Q-values may be unbounded outside of $\widehat{\Delta}(\mathcal{A})$, we are able to bound their local norms to arrive at Lemma 4.1, even though our policy updates (Equation (6)) are not restricted to $\widehat{\Delta}(\mathcal{A})$ and thus closed-form. We refer to Appendix D for the proof.

Unregularized Error Bounds While the bound in Lemma 4.1 depends on the choice of $\eta < 1$, $\tau < 1$ and $\lambda_{max} \geq H\Xi^{-1}$, we show that it is possible though not obvious to choose them (depending on the desired approximation) such that Φ_k decays to zero. Prior to this, we show that this will allow us to upper-bound both the constraint violation and the objective suboptimality in the original problem.

Lemma 4.2 (Error bounds)**.** *For any sequence $(\pi_k)_{k\in[K]}$,*

$$\begin{aligned}\left[V_r^{\pi^\star} - V_r^{\pi_k}\right]_+ &\leq H^{3/2}(2\Phi_k)^{1/2} + \tau H\log(A),\\ \max_{i\in[I]}\left[c_i - V_{u_i}^{\pi_k}\right]_+ &\leq H^{3/2}(2\Phi_k)^{1/2} + \tau\lambda_{max}\\ &\quad + \lambda_{max}^{-1}\left(H^2\Xi^{-1} + \tau H\log(A)\right).\end{aligned}$$

A similar result was provided by Ding et al. (2023) for the case of a single constraint ($I = 1$). Generalizing this is technically challenging as the standard way of showing that approximate saddle points have small constraint violation (Beck, 2017, Theorem 3.60) does not apply in the case of regularized saddle points. Simultaneously, the technique of Ding et al. (2023, Corollary 1) leverages the fact that only one constraint is present. We overcome this by choosing the domain $\Lambda = [0, \lambda_{max}]^I$ larger than standard primal-dual algorithms, making it possible to extract bounds on the individual constraint violations from the approximate saddle points. See Appendix D for the proof. This novel approach yields the rather uncommon inverse dependency on the diameter of Λ in Lemma 4.2, which needs to be chosen such that both terms $\tau\lambda_{max}$ and λ_{max}^{-1} are $O(\varepsilon)$ to obtain an ε-close solution.

Lemma 4.2 tells us that we can bound the objective suboptimality and all constraint violations by controlling the terms $\propto (\Phi_k)^{1/2}$ via Lemma 4.1, and remaining the terms by appropriately choosing the regularization and domain diameter, in terms of ε.

Last-iterate Convergence We are now ready to establish last-iterate convergence to the unregularized optimal policy of the regularized primal-dual scheme.

Theorem 4.1 (Last-iterate convergence)**.** *Let $\varepsilon \in (0,1)$. Then, with appropriate choices of $\eta \propto \varepsilon^6$, $\tau \propto \varepsilon^2$, $\lambda_{max} \propto \varepsilon^{-1}$, we have*

$$\left[V_r^{\pi^\star} - V_r^{\pi_k}\right]_+ \leq \varepsilon, \quad \left[c_i - V_{u_i}^{\pi_k}\right]_+ \leq \varepsilon \quad (\forall i \in [I])$$

for $k = \Omega(poly(A, H, I, \Xi^{-1}) \cdot \varepsilon^{-10})$.

Here, we only highlight the explicit dependency on the desired approximation. The dependency on the CMDP size is (low-degree) polynomial and detailed in Appendix D. The only problem-dependent constant in this bound is the Slater gap Ξ, which is shared by all primal-dual analyses of our

knowledge. While the provided rate is slow and may be improved in the future, all other known rates of primal-dual algorithms for CMDPs with arbitrary constraints ($I > 1$) only hold for the averaged and not the last iterate. More importantly, the technique leading to this result will allow us to achieve sublinear strong regret in the following section.

5. Online Setup

Recall the regularized primal-dual scheme from Equations (6) and (7). In our online learning setup, the true value functions are not known as we are learning the unknown CMDP. Thus, we are required to explore the CMDP and respect safety *during* exploration. Replacing the value functions by optimistic estimates (Shani et al., 2020; Auer et al., 2008) allows us to turn the primal-dual scheme into an online learning algorithm for finite-horizon CMDPs (see Algorithm 1). Importantly, we need to be optimistic with respect to the regularization term $\tau\mathcal{H}(\pi)$ too, rather than just the classical mixture value $V^{\pi}_{r+\boldsymbol{\lambda}_k^T\boldsymbol{u}}$. The main technical challenge is to incorporate the model uncertainty into our primal-dual analysis from Section 4.

5.1. Optimistic Model

For all s, a, h and $k \in [K]$, let $n_{k-1,h}(s,a) := \sum_{l=1}^{k-1} \mathbb{1}_{\{s_h^l = s,\ a_h^l = a\}}$ 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 visited at step h in episode l. First, we compute the empirical averages of the reward and transition probabilities as follows:

$$\begin{aligned}
\bar{r}_{k-1,h}(s,a) :=& \frac{\sum_{l=1}^{k-1} R_h^l(s,a)\mathbb{1}_{\{s_h^l=s,\ a_h^l=a\}}}{n_{k-1,h}(s,a)\vee 1},\\
\bar{u}_{k-1,i,h}(s,a) :=& \frac{\sum_{l=1}^{k-1} U_{i,h}^l(s,a)\mathbb{1}_{\{s_h^l=s,\ a_h^l=a\}}}{n_{k-1,h}(s,a)\vee 1},\\
\bar{p}_{k-1,h}(s'|s,a) :=& \frac{\sum_{l=1}^{k-1} \mathbb{1}_{\{s_h^l=s,\ a_h^l=a,\ s_{h+1}^l=s'\}}}{n_{k-1,h}(s,a)\vee 1},
\end{aligned} \tag{8}$$

where $a \vee b := \max\{a,b\}$ and $\mathbb{1}_A$ is the indicator function of an event A. We consider optimistic estimates

$$\begin{aligned}
\hat{r}_{k,h}(s,a) :=& \bar{r}_{k-1,h}(s,a) + b_{k-1,h}(s,a),\\
\hat{u}_{k,i,h}(s,a) :=& \bar{u}_{k-1,i,h}(s,a) + b_{k-1,h}(s,a),\\
\hat{\psi}_{k,h}(s,a) :=& -\log(\pi_{k,h}(a|s)) + b^p_{k-1,h}(s,a)\log(A),\\
\hat{p}_{k,h}(s'|s,a) :=& \bar{p}_{k-1,h}(s'|s,a),
\end{aligned} \tag{9}$$

where $b_{k-1,h}(s,a) = b^r_{k-1,h}(s,a) + b^p_{k-1,h}(s,a)$, and for any $\delta \in (0,1)$, we specify the correct values for

$$\begin{aligned}
b^r_{k-1,h}(s,a) =& O\left(\sqrt{\frac{\log\left(SAHIK\delta^{-1}\right)}{n_{k-1,h}(s,a)\vee 1}}\right),\\
b^p_{k-1,h}(s,a) =& O\left(H\sqrt{\frac{S+\log\left(SAHK\delta^{-1}\right)}{n_{k-1,h}(s,a)\vee 1}}\right),
\end{aligned}$$

in Appendix E to obtain our regret guarantees with probability at least $1-\delta$. The optimistic model guarantees that, with high probability, the obtained value functions overestimate the true ones and simultaneously allows us to control the estimation error. While optimistic exploration is standard, we here also take the entropy term in the objective into account via $\hat{\psi}_k$. Let

$$\hat{z}_k := \hat{r}_k + \boldsymbol{\lambda}_k^T\hat{\boldsymbol{u}}_k + \tau\hat{\psi}_k \tag{10}$$

be the optimistic reward function mimicking the π-dependency of the regularized Lagrangian at (π_k, λ_k). Consider the *truncated value functions*

$$(h,s,a) \mapsto \hat{Q}^k_{\hat{z}_k,h}(s,a), \qquad \hat{\boldsymbol{V}}^k_{\hat{\boldsymbol{u}}_k} = \hat{\boldsymbol{V}}^k_{\hat{\boldsymbol{u}}_k,1}(s_1)$$

that we compute via truncated policy evaluation (by dynamic programming) of π_k with respect to the optimistic model. We refer to Algorithm 2 in Appendix E, where we also establish the relevant properties of the model.

Algorithm Combining the truncated policy estimation under our learned model with the regularized primal-dual scheme (Equations (6) and (7)) yields Algorithm 1. The computational cost of the algorithm amounts to evaluating a policy $O(I)$ times per episode, which matches the complexity of standard primal-dual algorithms and is more efficient than running dual or LP-based algorithms. Projecting onto Λ is immediate since Λ is a product of intervals.

5.2. Regret Analysis

We now provide the key steps of our regret analysis, showing that Algorithm 1 indeed achieves sublinear strong regret for both the constraint violations and the objective. We defer all proofs for this section to Appendix F.

Lemma 5.1 (Regularized convergence)**.** *Let $\eta, \tau < 1$ and $\lambda_{max} \geq H\Xi^{-1}$. With probability at least $1-\delta$, the iterates of Algorithm 1 satisfy*

$$\begin{aligned}
\Phi_{k+1} \leq& (1-\eta\tau)^k\Phi_1 + \tilde{O}\Big(\eta\tau^{-1}C_{\eta,\tau,\Lambda}\\
&+ \eta\lambda_{max}\left(ISA^{1/2}H^2k^{1/2} + IS^{3/2}AH^2\right)\Big),
\end{aligned}$$

where $C_{\eta,\tau,\Lambda}$ is the same constant as in Lemma 4.1.

Algorithm 1 Regularized Primal-Dual Algorithm with Optimistic Exploration

Require: $\Lambda = [0, \lambda_{max}]^I$, stepsize $\eta > 0$, regularization parameter $\tau > 0$, number of episodes K, initial policy $\pi_{1,h}(a|s) = 1/A \ \ (\forall s, a, h)$, $\boldsymbol{\lambda}_1 := \mathbf{0} \in \mathbb{R}^I$

for $k = 1, \dots, K$ **do**

Update $\hat{r}_k$, $\hat{\boldsymbol{u}}_k$, $\hat{p}_k$, $\hat{\psi}_k$ via Equation (9)

Truncated policy evaluation (Algorithm 2) w.r.t. $\hat{z}_k$ (Equation (10)) and $\hat{\boldsymbol{u}}_k$:

$$\hat{Q}^k_{\hat{z}_k}(\cdot), \hat{\boldsymbol{V}}^k_{\hat{\boldsymbol{u}}_k} := \textsc{Eval}(\pi_k, \lambda_k, \hat{r}_k, \hat{\boldsymbol{u}}_k, \hat{\psi}_k, \hat{p}_k).$$

Update primal variables for all h, s, a:

$$\pi_{k+1,h}(a|s) \propto \pi_{k,h}(a|s) \exp\left(\eta \hat{Q}^k_{h,\hat{z}_k}(s,a)\right)$$

Update dual variables:

$$\boldsymbol{\lambda}_{k+1} = \text{proj}_\Lambda\left((1-\eta\tau)\boldsymbol{\lambda}_k - \eta(\hat{\boldsymbol{V}}^k_{\hat{\boldsymbol{u}}_k} - \boldsymbol{c})\right).$$

Play π_k for one episode, update $\bar{r}_k$, $\bar{\boldsymbol{u}}_k$, $\bar{g}_k$, $\bar{p}_k$ via Equation (8)

end for

Here, we use $\tilde{O}$-notation for asymptotics up to polylogarithmic factors in S, A, H, I, K, Ξ^{-1}, and δ^{-1}. This result is similar to our Lemma 4.1, but now we obtain an additional term corresponding to the model uncertainty (estimation error), which we control when choosing the step size η.

Regret Bound In a final step, we can leverage Lemma 4.2 to turn Lemma 5.1 into a sublinear regret bound for Algorithm 1, when summing up the error terms and choosing η, τ, and $\lambda_{max} \geq H\Xi^{-1}$ optimally depending on K given our bounds. This yields our main result.

Theorem 5.1 (Regret bound)**.** *Let $\tau = K^{-1/7}$, $\eta = (H^2 I)^{-1}\Xi K^{-5/7}$, $\lambda_{max} = H\Xi^{-1}K^{1/14}$. Then with probability at least $1-\delta$, Algorithm 1 obtains a* strong *regret of*

$$\mathcal{R}(K; r) \leq C_r K^{0.93}, \quad \mathcal{R}(K; \boldsymbol{u}) \leq C_u K^{0.93},$$

where C_r, $C_u = poly(S, A, H, I, \Xi^{-1}, \log(1/\delta), \log(K))$ and K is the number of episodes.

Here, we only highlight the leading term in K. The dependency on the CMDP parameters is (low-degree) polynomial and detailed in Appendix F. Again, Ξ is the only problem-dependent constant (and unavoidable).

Remark 5.1. *We remark that our proof of Theorem 5.1, in fact, shows last-iterate convergence in the online setup, which is strictly stronger than a regret bound in general.*

Our strong regret bound of $\tilde{O}(K^{0.93})$ is less tight than the $\tilde{O}(K^{1/2})$ that the vanilla primal-dual algorithm achieves for the *weak* regret, for which there exist well-known lower bounds (Jin et al., 2018; Domingues et al., 2021). Nevertheless, Algorithm 1 is the first primal-dual algorithm for CMDPs provably achieving sublinear strong regret. It is thus the first algorithm of its kind for which we can guarantee that it cannot keep violating constraints indefinitely. While LP-based approaches achieve strong regret of $\tilde{O}(K^{1/2})$, most modern (deep) safe RL algorithms for CMDPs follow primal-dual schemes (Chow et al., 2017; Tessler et al., 2018; Stooke et al., 2020). We believe that it might be possible to tighten our analysis, although this will require novel ideas. Indeed, our numerical evaluations show that the parameter choices in Theorem 5.1 are overly pessimistic.

5.3. Strong vs. Weak Regret and Safety at any Time

We allude to the strong regret several times by saying that a sublinear bound guarantees safety *during learning* or *in most episodes*. As our algorithm does not guarantee safety in *every* episode, one may wonder in which sense safety during learning is formally guaranteed by the strong regret compared to the weak one. Indeed, this is an important theme in CMDPs. In an unknown CMDP, there is no way to explore it without constraint violations unless further assumptions are made (as the constraint rewards u_i and transitions P are unknown, we cannot know that an action is unsafe without trying at least once). However, this is a limitation of the CMDP model with exploration rather than our algorithm. We can thus only argue about safety in *most episodes*.

Approximate Safety in Most Episodes Unlike any previous primal-dual algorithm, our method guarantees that for any fixed $\varepsilon > 0$, the fraction of episodes in which our policy is not ε-safe vanishes to 0 as the number of episodes K grows. Not being ε-safe here means to violate at least one constraint by at least ε. We make this formal in the following remark.

Remark 5.2. *Fix $\varepsilon > 0$ and suppose $\mathcal{R}(K; \mathbf{u}) \leq \tilde{O}(K^\alpha)$ for some $\alpha \in (0, 1)$. Then there exist at most $\tilde{O}(K^\alpha/\varepsilon)$ episodes with a constraint violation of at least ε. In other words, only a small fraction $\tilde{O}(K^{\alpha-1}/\varepsilon) = o(1)$ of the iterates is not ε-safe. In comparison, this is by no means guaranteed by a sublinear bound on $\mathcal{R}_{weak}(K; \mathbf{u})$.*

Hence, our algorithm is approximately safe in *most* episodes, while being safe in *every* episode is not possible by design. This is a remarkable result since previous works on primal-dual algorithms can only guarantee safety of the average policy and not in most of the episodes (such algorithms can be fully unsafe in, e.g., half of the episodes).

Strict Safety in Most Episodes Furthermore, ε-safety can be strengthened by a simple reduction to ensure *strict* safety in most episodes. This is possible by increasing the true

constraint thresholds by a small shift of $\varepsilon = O(K^{(\alpha-1)/2})$ to be more conservative and applying our results. For the formal details of this reduction, see, e.g., Appendix C.4 in Ding et al. (2023). We thus established that the fraction of unsafe episodes is vanishing (in terms of K).

6. Simulation

We perform numerical simulations of our algorithms and compare them to their unregularized counterparts (Efroni et al., 2020). We find that the vanilla primal-dual and dual algorithms can suffer linear strong regret while our regularized counterparts do not, illustrating that *error cancellations* are not merely a hypothetical issue. We provide further details in Appendix G.

6.1. Baselines and Environment

We compare our regularized primal-dual algorithm (Algorithm 1) to the vanilla primal-dual algorithm of Efroni et al. (2020), which corresponds to Equations (3) and (4) with optimistic exploration. We also include the vanilla dual algorithm of Efroni et al. (2020) as a baseline and our regularized dual algorithm (below), which arises from the same regularization framework as Algorithm 1. We test each algorithm for the same total number (6) of hyperparameter configurations and report the best results for each.

Dual Algorithm Leveraging the framework we introduced, it is immediate to also derive a *dual* algorithm for finite-horizon CMDPs. Dual algorithms amount to performing projected dual descent (Beck, 2017; Paternain et al., 2019) on the Lagrangian, where one can again use the optimistic model to estimate the unknown CMDP. Efroni et al. (2020) proved that this algorithm achieves a sublinear *weak* regret. Instead, we perform dual descent on the *regularized* Lagrangian $\mathcal{L}_\tau$. Explicitly,

$$\pi_k = \arg\max_{\pi\in\Pi} \left(V^{\hat{p}_k,\pi}_{\hat{r}_k+\boldsymbol{\lambda}_k^T \hat{\boldsymbol{u}}_k} + \tau \hat{\mathcal{H}}_k(\pi) \right), \tag{11}$$

$$\boldsymbol{\lambda}_{k+1} = \mathrm{proj}_\Lambda \left((1-\eta\tau)\boldsymbol{\lambda}_k - \eta(\boldsymbol{V}^{\hat{p}_k,\pi_k}_{\hat{\boldsymbol{u}}_k} - \boldsymbol{c}) \right), \tag{12}$$

where $V^{\hat{p}_k,\pi}$ refers to the value functions under transition model $\hat{p}_k$. These updates are similar to the ones proposed by Li et al. (2021); Ying et al. (2022), yet both assume a value function oracle. We can compute the first update via regularized dynamic programming, and the second one is the same as before. The dual approach has a higher computational complexity as the primal update requires a planning subroutine rather than just policy evaluation, but shows similar numerical performance. See Appendix G.3 for the full description of the regularized dual algorithm.

Environment We consider a randomly generated CMDP with deterministic rewards and unknown transitions. We draw the reward function r, constraint thresholds $\boldsymbol{c}$, and transitions p uniformly at random. In order for oscillations (and thus error cancellations) to occur, the objective must be conflicting with the constraints (Moskovitz et al., 2023), as they can otherwise easily be satisfied. However, by concentration of measure, two random vectors in high dimension are nearly orthogonal with high probability (Blum et al., 2020). Uniformly sampling the constraints would thus not yield interesting CMDPs, which is why we invoke a negative correlation between reward and constraint function. We sample the constraint function as $(1-r)+\beta\zeta$, where $\zeta \in \mathbb{R}^{HSA}$ is Gaussian with zero mean and identity covariance matrix. We consider $S = A = H = 5$, $\beta = 0.1$, and focus on the case of one constraint for visualization purposes. We refer to Appendix G for further details.

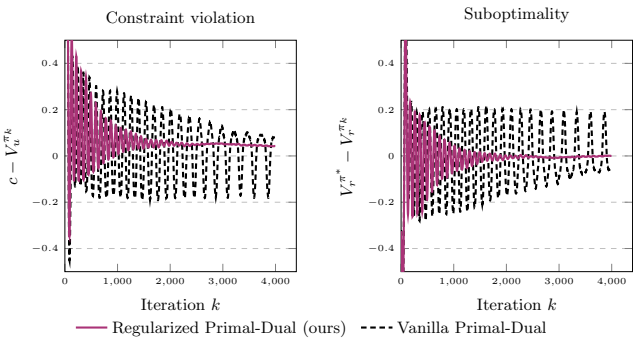

Figure 1. Constraint violation and objective suboptimality of the **vanilla** primal-dual algorithm (Efroni et al., 2020, cf. Equations (3) and (4)) and our **regularized** version (Algorithm 1). We present the values of the individual policies in each episode while learning the CMDP.

6.2. Results

The constraint violation and suboptimality of the iterates in each episode show the oscillatory behavior of the vanilla primal-dual algorithm as opposed to ours (Figure 1). While the on-average errors across episodes are sublinear, the vanilla algorithm keeps violating the constraints indefinitely as the number of episodes grows. In comparison, the oscillations of the regularized method are dampened, thus allowing it to converge to an optimal safe policy.

With respect to the *weak* regret, the vanilla algorithms perform better (Figures 2b and 3b, even constant for the suboptimality). However, with respect to the *strong* regret, the regularized algorithms outperform the unregularized ones, as they achieve sublinear regret *without* allowing for error cancellations (Figures 2a and 3a). While the strong regrets for the vanilla algorithms may look sublinear, a second look at their iterates (Figure 1) reveals that their regret will indeed grow linearly due to the persisting oscillations. This confirms our key point that a sublinear bound on the weak regret is not informative whenever we do not allow compensating for an unsafe episode with a safe one.

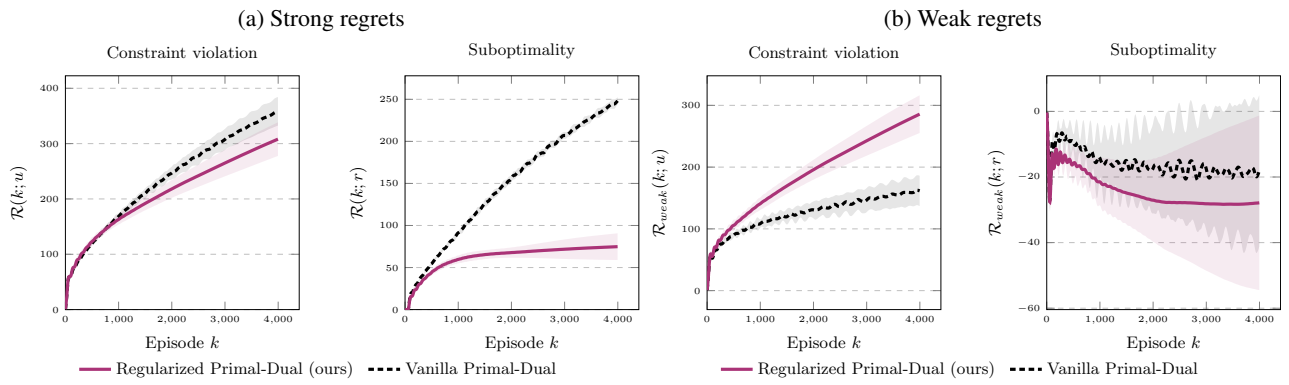

Figure 2. **Vanilla** primal-dual algorithm (Efroni et al., 2020, cf. Equations (3) and (4)) and our **regularized** version (Algorithm 1). Figure 2a shows the strong regret; Figure 2b shows the weak regret. The weak regret regarding the objective can be negative, illustrating that the iterates are superoptimal but unsafe on average. Y-axes differ across plots. All results are averaged over $n = 5$ independent runs, with plotted confidence intervals.

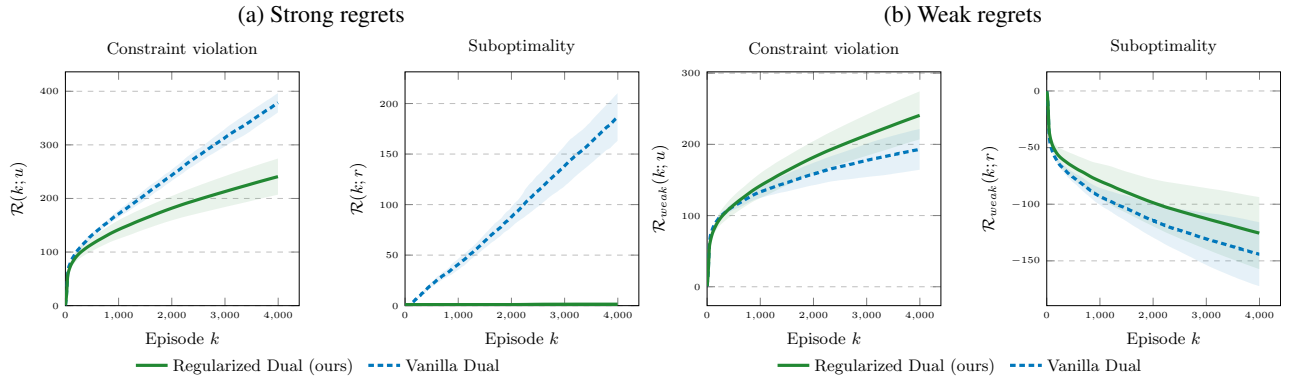

Figure 3. **Vanilla** dual algorithm (Efroni et al., 2020) and our **regularized** version (Equations (11) and (12)). Figure 3a shows the strong regret; Figure 3b shows the weak regret. Y-axes differ across plots. All results are averaged over $n = 5$ independent runs, with plotted confidence intervals.

The vanilla algorithms will suffer linear strong regret even with a potentially better learning rate scheduling. We observed that the learning rate influences the oscillation frequency: With a larger learning rate, the vanilla methods oscillate faster. However, changing the learning rate does not dampen the oscillation magnitude. Hence, the strong regret is still linear. Indeed, we observe a change of magnitude only via the regularization parameter rather than the learning rate.

Comparison with Guarantee With the theoretically derived stepsize η, regularization τ, and exploration from Theorem 5.1, we need many episodes to observe a benefit, due to the slowly vanishing gap between regularized and unregularized problem. Setting hyperparameters empirically, we observe a better regret than the theory suggests. Therefore, the plots in this section refer to the empirical choice.

7. Conclusion

In this paper, we gave the first answer to the open question of Efroni et al. (2020) whether primal-dual algorithms can achieve sublinear strong regret in finite-horizon CMDPs. While our answer is affirmative, it remains open in how far it is possible to lower the gap to the desired $\tilde{O}(K^{1/2})$ regret bound. We hope that our first analysis inspires further research on truly no-regret learning in CMDPs, including improvements in the analysis of our algorithm, incorporating function approximation, algorithms for the infinite-horizon average reward setup, and showing provable benefits of related approaches such as optimistic gradients.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback. P.A. is supported by the AI Center and ETH Foundations of Data Science (ETH-FDS) initiative. V.C. is supported by the Hasler Foundation Program: Hasler Responsible AI (project number 21043). V.C. is supported by the Swiss National Science Foundation (SNSF) under grant number 200021_205011. N.H. is supported by Swiss National Science Foundation Project Funding No. 200021-207343 and SNSF Starting Grant.

References

Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained policy optimization. In *International conference on machine learning*, pp. 22–31. PMLR, 2017.

Agrawal, S. and Devanur, N. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29, 2016.

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

Amani, S., Alizadeh, M., and Thrampoulidis, C. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.

Bai, Q., Bedi, A. S., Agarwal, M., Koppel, A., and Aggarwal, V. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3682–3689, 2022.

Beck, A. *First-order methods in optimization*. SIAM, 2017.

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

Bertsekas, D., Nedic, A., and Ozdaglar, A. *Convex analysis and optimization*, volume 1. Athena Scientific, 2003.

Blum, A., Hopcroft, J., and Kannan, R. *Foundations of data science*. Cambridge University Press, 2020.

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

Brunke, L., Greeff, M., Hall, A. W., Yuan, Z., Zhou, S., Panerati, J., and Schoellig, A. P. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:411–444, 2022.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Calvo-Fullana, M., Paternain, S., Chamon, L. F., and Ribeiro, A. State augmented constrained reinforcement learning: Overcoming the limitations of learning with rewards. *IEEE Transactions on Automatic Control*, 2023.

Chow, Y., Ghavamzadeh, M., Janson, L., and Pavone, M. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

Cover, T. M. *Elements of information theory*. John Wiley & Sons, 1999.

Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Ding, D. and Jovanović, M. R. Policy gradient primal-dual mirror descent for constrained mdps with large state spaces. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 4892–4897. IEEE, 2022.

Ding, D., Zhang, K., Basar, T., and Jovanovic, M. Natural policy gradient primal-dual method for constrained markov decision processes. *Advances in Neural Information Processing Systems*, 33:8378–8390, 2020.

Ding, D., Zhang, K., Başar, T., and Jovanović, M. R. Convergence and optimality of policy gradient primal-dual method for constrained markov decision processes. In *2022 American Control Conference (ACC)*, pp. 2851–2856. IEEE, 2022a.

Ding, D., Zhang, K., Duan, J., Başar, T., and Jovanović, M. R. Convergence and sample complexity of natural policy gradient primal-dual methods for constrained mdps. *arXiv preprint arXiv:2206.02346*, 2022b.

Ding, D., Zhang, K., Duan, J., Başar, T., and Jovanović, M. R. Convergence and sample complexity of natural policy gradient primal-dual methods for constrained mdps. *arXiv preprint arXiv:2206.02346*, 2022c.

Ding, D., Wei, C.-Y., Zhang, K., and Ribeiro, A. Last-iterate convergent policy gradient primal-dual methods for constrained mdps. *arXiv preprint arXiv:2306.11700*, 2023.

Ding, Y. and Lavaei, J. Provably efficient primal-dual reinforcement learning for cmdps with non-stationary objectives and constraints. *arXiv preprint arXiv:2201.11965*, 2022.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pp. 578–598. PMLR, 2021.

Efroni, Y., Merlis, N., Ghavamzadeh, M., and Mannor, S. Tight regret bounds for model-based reinforcement learning with greedy policies. *Advances in Neural Information Processing Systems*, 32, 2019.

Efroni, Y., Mannor, S., and Pirotta, M. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

Ghosh, A., Zhou, X., and Shroff, N. Provably efficient model-free constrained rl with linear function approximation. *Advances in Neural Information Processing Systems*, 35:13303–13315, 2022.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Kazerouni, A., Ghavamzadeh, M., Abbasi Yadkori, Y., and Van Roy, B. Conservative contextual linear bandits. *Advances in Neural Information Processing Systems*, 30, 2017.

Li, T., Guan, Z., Zou, S., Xu, T., Liang, Y., and Lan, G. Faster algorithm and sharper analysis for constrained markov decision process. *arXiv preprint arXiv:2110.10351*, 2021.

Liu, T., Zhou, R., Kalathil, D., Kumar, P., and Tian, C. Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, 34:17183–17193, 2021a.

Liu, T., Zhou, R., Kalathil, D., Kumar, P. R., and Tian, C. Policy optimization for constrained mdps with provable fast global convergence, 2021b. URL https://arxiv.org/abs/2111.00552.

Moskovitz, T., O'Donoghue, B., Veeriah, V., Flennerhag, S., Singh, S., and Zahavy, T. Reload: Reinforcement learning with optimistic ascent-descent for last-iterate convergence in constrained mdps. *arXiv preprint arXiv:2302.01275*, 2023.

Müller, A., Alatur, P., Ramponi, G., and He, N. Cancellation-free regret bounds for lagrangian approaches in constrained markov decision processes. In *Sixteenth European Workshop on Reinforcement Learning*, 2023.

Orabona, F. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Pacchiano, A., Ghavamzadeh, M., Bartlett, P., and Jiang, H. Stochastic bandits with linear constraints. In *International conference on artificial intelligence and statistics*, pp. 2827–2835. PMLR, 2021.

Paternain, S., Chamon, L. F., Calvo-Fullana, M., and Ribeiro, A. Constrained reinforcement learning has zero duality gap. *arXiv preprint arXiv:1910.13393*, 2019.

Paternain, S., Calvo-Fullana, M., Chamon, L. F., and Ribeiro, A. Safe policies for reinforcement learning via primal-dual methods. *IEEE Transactions on Automatic Control*, 2022.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Qiu, S., Wei, X., Yang, Z., Ye, J., and Wang, Z. Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. *Advances in Neural Information Processing Systems*, 33:15277–15287, 2020.

Shani, L., Efroni, Y., Rosenberg, A., and Mannor, S. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pp. 8604–8613. PMLR, 2020.

Sion, M. On general minimax theorems. *Pacific J. Math*, 1958.

Stooke, A., Achiam, J., and Abbeel, P. Responsive safety in reinforcement learning by pid lagrangian methods. In *International Conference on Machine Learning*, pp. 9133–9143. PMLR, 2020.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Tessler, C., Mankowitz, D. J., and Mannor, S. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*, 2018.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Wu, Y., Shariff, R., Lattimore, T., and Szepesvári, C. Conservative bandits. In *International Conference on Machine Learning*, pp. 1254–1262. PMLR, 2016.

Ying, D., Ding, Y., and Lavaei, J. A dual approach to constrained markov decision processes with entropy regularization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1887–1909. PMLR, 2022.

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

A. Summary of Notation

The following table summarizes our general CMDP notation.

State space	$\mathcal{S}$, with cardinality S
Action space	$\mathcal{A}$, with cardinality A
# of constraints	I
Time horizon	H
Transition probability	$p_h(s'\|s,a) = P[s_{h+1} = s' \mid s_h = s, a_h = a]$
Initial state	$s_1 \in \mathcal{S}$
Slater gap of $\bar{\pi}$	$\Xi = \min_{i\in[I]}(V^{\bar{\pi}}_{u_i} - c_i)$
Number of episodes	K
Objective reward	Random variable $R_h(s,a) \in [0,1]$, with $\mathbb{E}[R_h(s,a)] = r_h(s,a)$
Constraint rewards	Random variable $U_{i,h}(s,a)$ with $\mathbb{E}[U_{i,h}(s,a)] = u_{i,h}(s,a)$
Constraint thresholds	$\boldsymbol{c} \in \mathbb{R}^I$, with $c_i \in [0,H]$
Constraint functions	$g_{i,h}(s,a) = u_{i,h}(s,a) - \frac{1}{H}c_i$
Policy	$\pi \in \Pi$ with $(h,s,a) \mapsto \pi_h(a\|s)$ (non-stationary)
Value functions	$V^{\pi}_{r',h}(s) = \mathbb{E}_{\pi}[\sum_{h'=h}^{H} r'_{h'}(s_{h'}, a_{h'}) \mid s_h = s]$
(shorthand)	$V^{\pi}_{r'} = V^{\pi}_{r',1}(s_1)$
(vector-valued)	$\boldsymbol{V}^{\pi}_{\boldsymbol{u'}} = (V^{\pi}_{u'_1}, \ldots, V^{\pi}_{u'_I})^T \in \mathbb{R}^I$
Q-values	$Q^{\pi}_{r',h}(s,a) = \mathbb{E}_{\pi}[\sum_{h'=h}^{H} r'_{h'}(s_{h'}, a_{h'}) \mid s_h = s,\ a_h = a]$
Occupancy measures	$d^{\pi}_h(s,a) = P_{\pi}[s_h = s, a_h = a]$
	$d^{\pi}_h(s) = P_{\pi}[s_h = s]$
Lagrangian	$\mathcal{L}(\pi,\boldsymbol{\lambda}) = V^{\pi}_r + \sum_{i\in[I]} \lambda_i (V^{\pi}_{u_i} - c_i) = V^{\pi}_{r+\boldsymbol{\lambda}^T\boldsymbol{g}}$
Optimal policy	$\pi^\star \in \arg\max_{\pi\in\Pi} \min_{\boldsymbol{\lambda}\in\mathbb{R}^I_{\geq 0}} \mathcal{L}(\pi,\boldsymbol{\lambda})$
Dual optimizer	$\boldsymbol{\lambda}^\star \in \arg\min_{\boldsymbol{\lambda}\in\mathbb{R}^I_{\geq 0}} \max_{\pi\in\Pi} \mathcal{L}(\pi,\boldsymbol{\lambda})$
Confidence level	$1-\delta$
Objective regret	$\mathcal{R}(K;r) = \sum_{k\in[K]} \left[V^{\pi^\star}_r - V^{\pi_k}_r\right]_+$
Constraint regret	$\mathcal{R}(K;\boldsymbol{u}) = \max_{i\in[I]} \sum_{k\in[K]} \left[c_i - V^{\pi_k}_{u_i}\right]_+$

The following table summarizes the notation specific to the algorithm.

Step size	$\eta > 0$ (hyperparameter)
Regularization parameter	$\tau > 0$ (hyperparameter)
Dual threshold	$\lambda_{max} > 0$ (hyperparameter)
Dual domain	$\Lambda = [0, \lambda_{max}]^I$
Entropy	$\mathcal{H}(\pi) = -\mathbb{E}_\pi \left[\sum_{h=1}^H \log(\pi_h(a_h\|s_h))\right]$
Regularized Lagrangian	$\mathcal{L}_\tau(\pi, \boldsymbol{\lambda}) = V_r^\pi + \sum_{i\in[I]} \lambda_i(V_{u_i}^\pi - c_i) + \tau\left(\mathcal{H}(\pi) + \frac{1}{2}\|\boldsymbol{\lambda}\|^2\right)$
Regularized optimal policy	$\pi_\tau^\star \in \arg\max_{\pi\in\Pi} \min_{\boldsymbol{\lambda}\in\Lambda} \mathcal{L}_\tau(\pi, \boldsymbol{\lambda})$
Regularized dual optimizer	$\boldsymbol{\lambda}_\tau^\star \in \arg\min_{\boldsymbol{\lambda}\in\Lambda} \max_{\pi\in\Pi} \mathcal{L}_\tau(\pi, \boldsymbol{\lambda})$
Auxiliary function	$\psi_{k,h}(s,a) = -\log(\pi_{k,h}(a\|s))$
KL divergence	$\mathrm{KL}(q, q') = \sum_{a\in\mathcal{A}} q(a) \log\left(\frac{q(a)}{q'(a)}\right)$ $(q, q' \in \Delta(\mathcal{A}))$
	$\mathrm{KL}_{k,h}(s) = \mathrm{KL}(\pi_{\tau,h}^\star(\cdot\|s), \pi_{k,h}(\cdot\|s))$
	$\mathrm{KL}_k = \sum_h \sum_s d_h^{\pi_\tau^\star}(s)\mathrm{KL}_{k,h}(s)$
Potential function	$\Phi_k = \mathrm{KL}_k + \frac{1}{2}\|\boldsymbol{\lambda}_\tau^\star - \boldsymbol{\lambda}_k\|^2 \quad (k \geq 1)$
Visitation counter	$n_{k-1,h}(s,a) = \sum_{l=1}^{k-1} \mathbb{1}_{\{s_h^l = s,\ a_h^l = a\}}$
Averages	$\bar{r}_{k-1,h}(s,a)$, $\bar{u}_{k-1,h}(s,a)$, $\bar{p}_{k-1,h}(s'\|s,a)$
Exploration bonuses	$b_{k-1,h}(s,a) = b_{k-1,h}^r(s,a) + b_{k-1,h}^p(s,a)$
Optimistic estimates	$\hat{r}_k$, $\hat{\boldsymbol{u}}_k$, $\hat{\boldsymbol{g}}_k = \hat{\boldsymbol{u}}_k - \frac{1}{H}\boldsymbol{c}$, $\hat{\psi}_k$, $\hat{p}_k$
Regularized reward function	$z_k = r + \boldsymbol{\lambda}_k^T \boldsymbol{u} + \tau\psi_k$, $\hat{z}_k = \hat{r}_k + \boldsymbol{\lambda}_k^T \hat{\boldsymbol{u}}_k + \tau\hat{\psi}_k$
Success event	G
Truncated value functions	$\hat{Q}_{\hat{z}_k,h}^k(s,a) = \hat{Q}_{\hat{r}_k,h}^k(s,a) + \sum_i \lambda_{k,i}\hat{Q}_{\hat{u}_{k,i},h}^k(s,a)$
	$+\tau\hat{Q}_{\hat{\psi}_k,h}^k(s,a)$
	$\hat{V}_{\hat{z}_k,h}^k(s) = \left\langle \pi_{k,h}(\cdot\|s), \hat{Q}_{\hat{z}_k,h}^k(s,\cdot)\right\rangle$

B. Extended Related Work

In this section, we review further related work and provide a technical comparison with prior works.

Constrained MDPs Efroni et al. (2020) provided the first regret analysis for LP-based (OPTLP), primal-dual (OPTPRIMALDUAL), and dual algorithms (OPTDUAL). OPTLP achieves the optimal strong regret of $\tilde{O}(K^{1/2})$, yet most modern CMDP algorithms are based on primal-dual schemes rather than LP. OPTPRIMALDUAL is akin to our Algorithm 1 but without regularization. It guarantees a weak regret of $\tilde{O}(K^{1/2})$ but no bound on the strong regret, which is left as an open question that we addressed in Section 5. The same holds regarding the guarantees for OPTDUAL, for which the question about strong regret bounds is still unanswered.

Since Efroni et al. (2020) analyzed the vanilla primal-dual (and dual) algorithm, their analysis has been extended in various works, both for the case of an unknown or known CMDP. Specifically, the algorithms have been extended to natural policy gradient methods with policy parameterization (Ding et al., 2020; 2022b; Liu et al., 2021b), function approximation in the linear MDP setup (Ding & Jovanović, 2022; Ghosh et al., 2022), CMDPs with time-varying characteristics (Ding & Lavaei, 2022; Qiu et al., 2020), and have even been shown to achieve bounded on-average constraint violation (Liu et al., 2021a; Bai et al., 2022). However, all these works only established convergence of the *averaged* iterates or a sublinear *weak* regret. In practice, recent works do show empirical success (using optimistic gradients (Moskovitz et al., 2023) and PID control (Stooke et al., 2020)) but without the desired theoretical guarantees.

Comparison with Prior Results Ding et al. (2023) analyzed two algorithms, RPG-PD and OPG-PD, for last-iterate convergence assuming a value function oracle. RPG-PD follows the same scheme as our Equations (6) and (7). However, the analysis is tailored for a single constraint. It is not straightforward (as far as we know) how Corollary 1 in (Ding et al., 2023) can be extended to multiple constraints (which we achieve in Section 4). Our Lemma 4.2 generalizes the analysis to deal with multiple constraints. However this extensions leads to a worse iteration complexity in Theorem 4.1 of $O(\varepsilon^{-10})$ rather than $\tilde{O}(\varepsilon^{-6})$. In addition to this, Ding et al. (2023)'s policy update differs from ours in that it does not allow a closed-form solution but requires projection onto a *restricted* probability simplex for technical reasons (see discussion of Lemma 4.1, providing an analysis for our closed-form updates).

Moreover, Ding et al. (2023) left it as an open question whether the algorithm can be generalized to achieve last-iterate convergence in the online setup, when the CMDP is unknown; we addressed this point in Section 5.

The other algorithm of Ding et al. (2023), OPG-PD, is based on optimistic gradient updates and requires the restrictive assumption that the optimal state-visitation distribution (i.e., occupancy measure) is unique and introduces an extra problem-dependent constant. Moreover, it assumes a uniform lower bound on the state-visitation frequency in the discounted infinite-horizon setting, an assumption that cannot be guaranteed in the finite-horizon setting.

Moskovitz et al. (2023) showed last-iterate convergence of a primal-dual scheme using optimistic gradient updates given a known CMDP, but their analysis concerns an algorithm operating over occupancy measures rather than policies (different from the practical implementation). Its implicit updates are constrained over the set of occupancy measures (i.e., the Bellman flow polytope), making them at least as computationally expensive as solving the CMDP directly via an LP in the first place.

Calvo-Fullana et al. (2023) considered a rather different approach to overcome the problem that CMDPs cannot be modeled by a single (mixture) reward weighted by Lagrange multipliers (sometimes referred to as *scalarization fallacy*). They proposed a state-augmentation technique that addresses this related problem without guaranteeing last-iterate convergence.

Dual Algorithms Li et al. (2021) provided a dual (not primal-dual) algorithm based on the same regularization scheme as ours but considered an accelerated dual update and only proved convergence for a history-weighted mixture policy in a known CMDP. Similarly, Ying et al. (2022) derived a dual (not primal-dual) algorithm with last-iterate convergence but left it open if a sample-based version is possible. Moreover, their analysis covers the discounted infinite-horizon setting and requires a uniform lower bound on the state-visitation frequency, an assumption that cannot be guaranteed in the finite-horizon setting.

Constrained Bandits In the simpler bandit setup where there is only a single state, there are mainly three setups in the literature: Knapsack bandits (Agrawal & Devanur, 2016; Badanidiyuru et al., 2018) consider reward maximization over time as long a some global budget is not used up yet. Conservative bandits (Wu et al., 2016; Kazerouni et al., 2017) concern algorithms whose cumulative reward performs sufficiently well relative to some pre-defined baseline policy. Finally, there is a line of research on stage-wise constrained bandits (Amani et al., 2019; Pacchiano et al., 2021), which require algorithms that obtain a reward and a cost associated with an action, where the latter should stay below a threshold in each round. While these settings may inspire related research in CMDPs, they are rather different from ours: They consider hard thresholds in the single-state setup, while exploration in CMDPs is generally stateful and commonly aims at simultaneous minimization of reward and constraint regrets.

C. Properties of the Lagrangian Formulation

The results in this section are not novel by themselves, but we re-establish them here for finite-horizon CMDPs for completeness. We refer to Appendix I for the relevant convex optimization background. To view the CMDP as a convex optimization problem, we will express it via the common notion of occupancy measures (Borkar, 1988).

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

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

for $s \in \mathcal{S}$, $a \in \mathcal{A}$, $h \in [H]$. *We denote the stacked vector of these values as* $\boldsymbol{d}^\pi \in \mathbb{R}^{HSA}$, *with the element at index* (h,s,a) *being* $d_h^\pi(s,a)$. *Similarly, we define*

$$d_h^\pi(s) := P[s_h = s \mid s_1; p, \pi] = \sum_a d_h^\pi(s,a)$$

for $s \in \mathcal{S}$.

We can now define

$$Q(p) := \left\{ \boldsymbol{d}^\pi \in \mathbb{R}^{HSA} \mid \pi \in \Pi \right\}$$

as the *state-action occupancy measure polytope*. Note that $Q(p)$ is indeed a polytope (Puterman, 2014). Moreover, we have a surjective map $\pi \mapsto \boldsymbol{d}^\pi$ between Π and $Q(p)$, for which we can explicitly compute an element in the pre-image of $\boldsymbol{d} \in Q(p)$ via $\pi_h(a|s) = d_h(s,a)/(\sum_{a'} d_h(s,a'))$.

We can stack the expected rewards $r_h(s,a)$ and constraint rewards $u_{i,h}(s,a)$ in the same way as $d_h^\pi(s,a)$ to obtain vectors $\boldsymbol{r} \in \mathbb{R}^{HSA}$ and $\boldsymbol{u}_i \in \mathbb{R}^{HSA}$. Note that we then have $V_r^\pi = \sum_{h,s,a} d_h^\pi(s,a) r_h(s,a) = \boldsymbol{r}^T \boldsymbol{d}^\pi$ by linearity of expectation. Similarly, for all $i \in [I]$, we have $V_{u_i}^\pi = \boldsymbol{u}_i^T \boldsymbol{d}^\pi$. Moreover, if we stack $\boldsymbol{U} = (\boldsymbol{u}_i)_{i \in [I]} \in \mathbb{R}^{I \times HSA}$ and $\boldsymbol{c} = (c_i)_{i \in [I]} \in \mathbb{R}^I$ as

$$\boldsymbol{U} := \begin{pmatrix} \boldsymbol{u}_1^T \\ \vdots \\ \boldsymbol{u}_I^T \end{pmatrix}, \qquad \boldsymbol{c} := \begin{pmatrix} c_1 \\ \vdots \\ c_I \end{pmatrix},$$

we obtain $\boldsymbol{V}_{\boldsymbol{u}}^\pi = \boldsymbol{U}\boldsymbol{d}^\pi \in [0,H]^I$ for the vector of the constraint value functions. We can thus write

$$\pi^* \in \arg\max_{\pi \in \Pi} \quad V_r^\pi \quad \text{s.t.} \quad V_{u_i}^\pi \geq c_i \quad (\forall i \in [I])$$

equivalently as

$$\boldsymbol{d}^{\pi^\star} \in \arg\max_{\boldsymbol{d}^\pi \in Q(p)} \quad \boldsymbol{r}^T \boldsymbol{d}^\pi \quad \text{s.t.} \quad \boldsymbol{U}\boldsymbol{d}^\pi \geq \boldsymbol{c}, \tag{13}$$

where $\geq$ is understood element-wise. This is a linear program (LP). In particular, by compactness of the state-action occupancy polytope, there exists an optimal solution π^* as we assume feasibility.

Lemma C.1 (Strong duality CMDP (Paternain et al., 2019))**.** *We have*

$$\max_{\pi \in \Pi} \min_{\boldsymbol{\lambda} \in \mathbb{R}^I_{\geq 0}} \mathcal{L}(\pi, \boldsymbol{\lambda}) = \min_{\boldsymbol{\lambda} \in \mathbb{R}^I_{\geq 0}} \max_{\pi \in \Pi} \mathcal{L}(\pi, \boldsymbol{\lambda}),$$

and both optima are attained.

Proof. Note that, under Assumption 2.1, we can view Equation (1) as the convex optimization problem in Equation (13) over $Q(p)$ that satisfies all parts of Assumption I.1 from Appendix I.2. Indeed,

(a) $X := Q(p)$ is a polytope and thus convex

(b) the objective $f(\cdot) := -\boldsymbol{r}^T(\cdot)$ is affine and thus convex

(c) the constraints $g_i(\cdot) := c_i - \boldsymbol{u}_i^T(\cdot)$ are affine and thus convex

(d) by Assumption 2.1, Equation (13) is feasible, and thus its optimum is attained (since the domain is compact and the objective continuous)

(e) a Slater point exists by Assumption 2.1, namely $\boldsymbol{d}^{\bar{\pi}}$

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

where $Q(p) \subset \mathbb{R}^{HSA}$, $\boldsymbol{r} \in \mathbb{R}^{HSA}$ and $\boldsymbol{u}_i \in \mathbb{R}^{SAH}$ are defined as above. The claim now readily follows from Theorem I.1. □

Lemma C.2 (e.g., Ying et al. (2022))**.** *We have* $\|\boldsymbol{\lambda}^\star\|_1 \leq \frac{H}{\Xi}$,

Proof. As in the proof of Lemma C.1, under Assumption 2.1, we can view the CMDP problem as a convex optimization problem in the occupancy measure, in the same setup as Appendix I.2. Specifically, we have $V_r^\pi = \boldsymbol{r}^T \boldsymbol{d}^\pi$ and $\boldsymbol{V}_{\boldsymbol{u}}^\pi = \boldsymbol{U}\boldsymbol{d}^\pi$. Then, set $X = Q(p)$, $\bar{x} = \boldsymbol{d}^{\bar{\pi}}$, $f(\cdot) = -\boldsymbol{r}^T(\cdot)$ and $g_i(\cdot) = c_i - \boldsymbol{u}_i^T(\cdot)$. Plugging this into Theorem I.3 indeed yields

$$\|\boldsymbol{\lambda}^\star\|_1 \leq \frac{V_r^{\pi^\star} - V_r^{\bar{\pi}}}{\min_{i \in [I]}(V_{u_i}^{\bar{\pi}} - c_i)} \leq \frac{H}{\Xi}.$$

□

Lemma C.3 (Saddle point CMDP)**.** *Let $\pi \in \Pi$ and $\boldsymbol{\lambda} \in \mathbb{R}^I_{\geq 0}$. Then*

$$\mathcal{L}(\pi, \boldsymbol{\lambda}^\star) \leq \mathcal{L}(\pi^\star, \boldsymbol{\lambda}^\star) \leq \mathcal{L}(\pi^\star, \boldsymbol{\lambda}).$$

Proof. By Lemma C.1, this immediately follows from Lemma I.2 in Appendix I.1. □

Lemma C.4 (Strong duality regularized CMDP (Ding et al., 2023))**.** *We have*

$$\max_{\pi \in \Pi} \min_{\boldsymbol{\lambda} \in \Lambda} \mathcal{L}_\tau(\pi, \boldsymbol{\lambda}) = \min_{\boldsymbol{\lambda} \in \Lambda} \max_{\pi \in \Pi} \mathcal{L}_\tau(\pi, \boldsymbol{\lambda}),$$

and both primal and dual optimum are attained.

Proof. For all $\pi \in \Pi$, $\boldsymbol{\lambda} \in \Lambda$, we have

$$\begin{aligned}
\mathcal{L}_\tau(\pi, \boldsymbol{\lambda}) =& V^\pi_{r+\boldsymbol{\lambda}^T \boldsymbol{g}} + \tau \left(\mathcal{H}(\pi) + \frac{1}{2} \|\boldsymbol{\lambda}\|^2 \right) \\
=& \sum_{s,a,h} (r_h(s,a) + \sum_i \lambda_i g_{i,h}(s,a)) d^\pi_h(s,a) \\
&+ \tau \left(- \sum_{s,a,h} d^\pi_h(s,a) \log \left(\frac{d^\pi_h(s,a)}{\sum_{a'} d^\pi_h(s,a')} \right) + \frac{1}{2} \|\boldsymbol{\lambda}\|^2 \right) \\
=:& \mathcal{L}^{occ}_\tau(\boldsymbol{d}^\pi, \boldsymbol{\lambda}),
\end{aligned}$$

where $g_{i,h}(s,a) = u_{i,h}(s,a) - \frac{1}{H} c_i$ and where we used the definition of the occupancy measures and the polytope $Q(p)$. Consider the problem

$$\max_{\boldsymbol{d} \in Q(p)} \min_{\boldsymbol{\lambda} \in \Lambda} \mathcal{L}^{occ}_\tau(\boldsymbol{d}, \boldsymbol{\lambda}). \tag{14}$$

For any $\pi \in \Pi$ that is optimal for Equation (5), $\boldsymbol{d}^\pi$ is also optimal for Equation (14). Conversely, for every $\boldsymbol{d} \in Q(p)$ that is optimal for Equation (14), we have that any π given by $\pi_h(a|s) := \frac{d_h(s,a)}{\sum_{a' \in \mathcal{A}} d_h(s,a')}$ for s with $\sum_{a' \in \mathcal{A}} d_h(s,a) > 0$, and arbitrary otherwise, is optimal for Equation (2).

Note that $\mathcal{L}^{occ}_\tau$ is continuous. We further claim that $\mathcal{L}^{occ}_\tau$ is 1-strongly convex in $\boldsymbol{\lambda} \in \Lambda$ and concave in $\boldsymbol{d} \in Q(p)$. Indeed, while the former claim is immediate, we can see the latter via the log-sum inequality (e.g., Cover (1999, Theorem 2.7.1)) with $n = 2$: For non-negative a_i, b_i,

$$- \left(\sum_{i=1} a_i \right) \log \left(\frac{\sum_{i=1} a_i}{\sum_{i=1} b_i} \right) \geq - \sum_{i=1} a_i \log \left(\frac{a_i}{b_i} \right)$$

and equality if and only if a_i / b_i is the same for all i. Only considering the nonlinear term in $\mathcal{L}_\tau(\boldsymbol{\lambda}, \cdot)$, for $\boldsymbol{d}_1, \boldsymbol{d}_2 \in Q(p)$

and $\alpha \in (0,1)$ we have

$$
\begin{aligned}
&-\sum_{s,a,h} (\alpha d_{1,h}(s,a) + (1-\alpha) d_{2,h}(s,a)) \log\left(\frac{\alpha d_{1,h}(s,a) + (1-\alpha) d_{2,h}(s,a)}{\sum_{a'} (\alpha d_{1,h}(s,a') + (1-\alpha) d_{2,h}(s,a'))}\right) \\
=&\sum_{s,a,h} -(\alpha d_{1,h}(s,a) + (1-\alpha) d_{2,h}(s,a)) \log\left(\frac{\alpha d_{1,h}(s,a) + (1-\alpha) d_{2,h}(s,a)}{\sum_{a'} \alpha d_{1,h}(s,a') + \sum_{a'} (1-\alpha) d_{2,h}(s,a')}\right) \\
\geq&\sum_{s,a,h} -\alpha d_{1,h}(s,a) \log\left(\frac{\alpha d_{1,h}(s,a)}{\sum_{a'} \alpha d_{1,h}(s,a')}\right) \\
&+\sum_{s,a,h} -(1-\alpha) d_{2,h}(s,a) \log\left(\frac{(1-\alpha) d_{2,h}(s,a)}{\sum_{a'} (1-\alpha) d_{2,h}(s,a')}\right) \\
=&-\alpha \sum_{s,a,h} d_{1,h}(s,a) \log\left(\frac{d_{1,h}(s,a)}{\sum_{a'} d_{1,h}(s,a')}\right) \\
&-(1-\alpha) \sum_{s,a,h} d_{2,h}(s,a) \log\left(\frac{d_{2,h}(s,a)}{\sum_{a'} d_{2,h}(s,a')}\right).
\end{aligned}
$$

with equality if and only if $\frac{d_{1,h}(s,a)}{\sum_{a'} d_{1,h}(s,a')} = \frac{d_{2,h}(s,a)}{\sum_{a'} d_{2,h}(s,a')}$ for all s, h. By Lemma I.1, we thus have

$$
\max_{\boldsymbol{d} \in Q(p)} \min_{\boldsymbol{\lambda} \in \Lambda} \mathcal{L}_\tau^{occ}(\boldsymbol{d}, \boldsymbol{\lambda}) = \min_{\boldsymbol{\lambda} \in \Lambda} \max_{\boldsymbol{d} \in Q(p)} \mathcal{L}_\tau^{occ}(\boldsymbol{d}, \boldsymbol{\lambda}),
$$

and primal and dual optimizers exist. This implies the same for the original problem Equation (2) by converting the occupancy measures back into policies via $\pi_h(a|s) = d_h(s,a)/(\sum_{a'} d_h(s,a'))$. □

Lemma C.5 (Saddle point regularized CMDP)**.** *Let $\pi \in \Pi$ and $\boldsymbol{\lambda} \in \Lambda$. Then*

$$
\mathcal{L}_\tau(\pi, \boldsymbol{\lambda}_\tau^\star) \leq \mathcal{L}_\tau(\pi_\tau^\star, \boldsymbol{\lambda}_\tau^\star) \leq \mathcal{L}_\tau(\pi_\tau^\star, \boldsymbol{\lambda}).
$$

Proof. By Lemma C.4, this follows from Lemma I.2. □

Lemma C.6. *Let $\pi \in \Pi$ and $\boldsymbol{\lambda} \in \Lambda$. Then*

$$
V^{\pi}_{r+(\boldsymbol{\lambda}_\tau^\star)^T \boldsymbol{g}} - \tau \mathcal{H}(\pi_\tau^\star) \leq V^{\pi_\tau^\star}_{r+(\boldsymbol{\lambda}_\tau^\star)^T \boldsymbol{g}} \leq V^{\pi_\tau^\star}_{r+\boldsymbol{\lambda}^T \boldsymbol{g}} + \frac{\tau}{2} \|\boldsymbol{\lambda}\|^2,
$$

where $\boldsymbol{g} = \boldsymbol{u} - \frac{1}{H}\boldsymbol{c}$.

Proof. Plugging the definition of $\mathcal{L}_\tau$ into Lemma C.5 proves the claim, after using that $\mathcal{H}(\pi) \geq 0$ and $\|\boldsymbol{\lambda}\|^2 \geq 0$. □

D. Last-Iterate Convergence

In this section, we provide the proofs for all results in Section 4, resulting in the proof of last-iterate convergence of the regularized primal-dual scheme (Equations (6) and (7)).

We first establish the convergence of the aforementioned potential function Φ_k.

Lemma 4.1 (Regularized convergence)**.** *Let $\eta, \tau < 1$ and $\lambda_{max} \geq H\Xi^{-1}$. The iterates in Equations (6) and (7) satisfy*

$$
\Phi_{k+1} \leq (1-\eta\tau)^k \Phi_1 + \tilde{O}\left(\eta \tau^{-1} C_{\eta,\tau,\Lambda}\right),
$$

where

$$
\begin{aligned}
C_{\eta,\tau,\Lambda} =& \lambda_{max}^2 H^3 A^{1/2} I^2 \exp\left(\eta H \left(1 + \lambda_{max} I + \log(A)\right)\right) \\
&+ I \left(H + \tau \lambda_{max}\right)^2.
\end{aligned}
$$

Proof. We first decompose the k-th *primal-dual gap* as follows:

$$\mathcal{L}_\tau(\pi_\tau^\star, \boldsymbol{\lambda}_k) - \mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}_\tau^\star) = \underbrace{\mathcal{L}_\tau(\pi_\tau^\star, \boldsymbol{\lambda}_k) - \mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}_k)}_{(i)} + \underbrace{\mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}_k) - \mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}_\tau^\star)}_{(ii)}. \tag{15}$$

We first bound term (i):

$$\begin{aligned}
(i) =& \mathcal{L}_\tau(\pi_\tau^\star, \boldsymbol{\lambda}_k) - \mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}_k) \\
=& V^{\pi_\tau^\star}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}} - V^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}} && \text{(as } d_h^\pi(s,a) = d_h^\pi(s)\pi_h(a|s)\text{, and cancel } \|\boldsymbol{\lambda}_k\|^2\text{)} \\
& - \tau \sum_{s,a,h} d_h^{\pi_\tau^\star}(s)\pi_{\tau,h}^\star(a|s)\log(\pi_{\tau,h}^\star(a|s)) + \tau \sum_{s,a,h} d_h^{\pi_k}(s)\pi_{k,h}(a|s)\log(\pi_{k,h}(a|s)) \\
=& V^{\pi_\tau^\star}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} - V^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} && \text{(since } \psi_{k,h}(s,a) = -\log(\pi_{k,h}(a|s))\text{)} \\
& + \tau \sum_{s,a,h} d_h^{\pi_\tau^\star}(s)\pi_{\tau,h}^\star(a|s)\log(\pi_{k,h}(a|s)) - \tau \sum_{s,a,h} d_h^{\pi_k}(s)\pi_{k,h}(a|s)\log(\pi_{k,h}(a|s)) \\
& - \tau \sum_{s,a,h} d_h^{\pi_\tau^\star}(s)\pi_{\tau,h}^\star(a|s)\log(\pi_{\tau,h}^\star(a|s)) + \tau \sum_{s,a,h} d_h^{\pi_k}(s)\pi_{k,h}(a|s)\log(\pi_{k,h}(a|s)) \\
=& V^{\pi_\tau^\star}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} - V^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} \\
& + \tau \sum_{s,a,h} d_h^{\pi_\tau^\star}(s)\pi_{\tau,h}^\star(a|s)\log(\pi_{k,h}(a|s)) \\
& - \tau \sum_{s,a,h} d_h^{\pi_\tau^\star}(s)\pi_{\tau,h}^\star(a|s)\log(\pi_{\tau,h}^\star(a|s)) \\
=& V^{\pi_\tau^\star}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} - V^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} \\
& - \tau \sum_{s,h} d_h^{\pi_\tau^\star}(s) \sum_a \pi_{\tau,h}^\star(a|s) \log\left(\frac{\pi_{\tau,h}^\star(a|s)}{\pi_{k,h}(a|s)}\right) \\
=& V^{\pi_\tau^\star}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} - V^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} - \tau \sum_{s,h} d_h^{\pi_\tau^\star}(s)\mathrm{KL}_{k,h}(s) \\
=& V^{\pi_\tau^\star}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} - V^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} - \tau\mathrm{KL}_k \\
=& V^{\pi_\tau^\star}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{u}+\tau\psi_k} - V^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{u}+\tau\psi_k} - \tau\mathrm{KL}_k && \text{(as } \boldsymbol{g} = \boldsymbol{u} - \tfrac{1}{H}\boldsymbol{c}\text{)} \\
=& V_{z_k}^{\pi_\tau^\star} - V_{z_k}^{\pi_k} - \tau\mathrm{KL}_k \\
=& \sum_{s,h} d_h^{\pi_\tau^\star}(s) \left\langle Q^{\pi_k}_{z_k,h}(s,\cdot), \pi_{\tau,h}^\star(\cdot|s) - \pi_{k,h}(\cdot|s) \right\rangle - \tau\mathrm{KL}_k && \text{(by Lemma H.1)}
\end{aligned}$$

Note that for all s, h,

$$\begin{aligned}
& \left\langle Q^{\pi_k}_{z_k,h}(s,\cdot), \pi_{\tau,h}^\star(\cdot|s) - \pi_{k,h}(\cdot|s) \right\rangle \\
\leq& \frac{\mathrm{KL}_{k,h}(s) - \mathrm{KL}_{k+1,h}(s)}{\eta} + \frac{\eta}{2}\sum_a \pi_{k,h}(a|s)\exp\left(Q^{\pi_k}_{z_k,h}(s,a)\right) Q^{\pi_k}_{z_k,h}(s,a)^2 && \text{(Lemma I.6)} \\
\leq& \frac{\mathrm{KL}_{k,h}(s) - \mathrm{KL}_{k+1,h}(s)}{\eta} && \text{(Lemma E.11)} \\
& + \frac{\eta}{2} A^{1/2} \exp\left(\eta H\left(1 + \lambda_{max} I + \tau\log(A)\right)\right)\left(2H^2\left(1 + I\lambda_{max} + \tau\log(A)\right)^2 + 2\tau^2(64/e^2)\right) \\
=& \frac{\mathrm{KL}_{k,h}(s) - \mathrm{KL}_{k+1,h}(s)}{\eta} + \frac{\eta}{2}\frac{1}{H} D_{\eta,\tau,\Lambda},
\end{aligned}$$

with

$$D_{\eta,\tau,\Lambda} = HA^{1/2}\exp\left(\eta H\left(1 + \lambda_{max} I + \tau\log(A)\right)\right)\left(2H^2\left(1 + I\lambda_{max} + \tau\log(A)\right)^2 + 2\tau^2(64/e^2)\right)$$

and where we were able to apply Lemma I.6 by Lemma I.9 and since $Q^{\pi_k}_{z,h}(s,a) \geq 0$. Hence,

$$\sum_{s,h} d^{\pi^\star_\tau}_h(s) \left\langle Q^{\pi_k}_{z_k,h}(s,\cdot), \pi^\star_{\tau,h}(\cdot|s) - \pi_{k,h}(\cdot|s) \right\rangle \leq \sum_{s,h} d^{\pi^\star_\tau}_h(s) \left(\frac{\mathrm{KL}_{k,h}(s) - \mathrm{KL}_{k+1,h}(s)}{\eta} + \frac{\eta}{2}\frac{1}{H} D_{\eta,\tau,\Lambda} \right)$$
$$= \frac{\mathrm{KL}_k - \mathrm{KL}_{k+1}}{\eta} + \frac{\eta}{2} D_{\eta,\tau,\Lambda}.$$

Plugging in, we thus find

$$(i) = V^{\pi^\star_\tau}_{z_k} - V^{\pi_k}_{z_k} - \tau \mathrm{KL}_k \leq \frac{\mathrm{KL}_k - \mathrm{KL}_{k+1}}{\eta} + \frac{\eta}{2} D_{\eta,\tau,\Lambda} - \tau \mathrm{KL}_k = \frac{(1-\eta\tau)\mathrm{KL}_k - \mathrm{KL}_{k+1}}{\eta} + \frac{\eta}{2} D_{\eta,\tau,\Lambda}. \tag{16}$$

We now bound term (ii):

$$\begin{aligned}
(ii) =& \mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}_k) - \mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}^\star_\tau) \\
=& V^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}} - V^{\pi_k}_{r+(\boldsymbol{\lambda}^\star_\tau)^T \boldsymbol{g}} + \frac{\tau}{2}\|\boldsymbol{\lambda}_k\|^2 - \frac{\tau}{2}\|\boldsymbol{\lambda}^\star_\tau\|^2 && \text{(cancel } \mathcal{H}(\pi_k)\text{)} \\
=& \sum_i (\lambda_{k,i} - \lambda^\star_{\tau,i}) V^{\pi_k}_{g_i} + \frac{\tau}{2}\|\boldsymbol{\lambda}_k\|^2 - \frac{\tau}{2}\|\boldsymbol{\lambda}^\star_\tau\|^2 \\
=& \sum_i (\lambda_{k,i} - \lambda^\star_{\tau,i})(V^{\pi_k}_{u_i} - c_i + \tau\lambda_{k,i}) - \frac{\tau}{2}\|\boldsymbol{\lambda}_k - \boldsymbol{\lambda}^\star_\tau\|^2 \\
\leq& \frac{\|\boldsymbol{\lambda}^\star_\tau - \boldsymbol{\lambda}_k\|^2 - \|\boldsymbol{\lambda}^\star_\tau - \boldsymbol{\lambda}_{k+1}\|^2}{2\eta} + \frac{\eta}{2}\left\|\boldsymbol{V}^{\pi_k}_{u_k} - \boldsymbol{c} + \tau\boldsymbol{\lambda}_k\right\|^2 && \text{(Lemma I.8)} \\
\leq& \frac{\|\boldsymbol{\lambda}^\star_\tau - \boldsymbol{\lambda}_k\|^2 - \|\boldsymbol{\lambda}^\star_\tau - \boldsymbol{\lambda}_{k+1}\|^2}{2\eta} + \frac{\eta}{2} D'_{\tau,\Lambda}, && \text{(Lemma E.11)}
\end{aligned}$$

with $D'_{\tau,\Lambda} = I(H + \tau\lambda_{max})^2$ and where we were able to apply Lemma I.8 by Lemma I.9. Plugging in, we find

$$\begin{aligned}
(ii) =& \sum_i (\lambda_{k,i} - \lambda^\star_{\tau,i})(V^{\pi_k}_{u_i} - c_i + \tau\lambda_{k,i}) - \frac{\tau}{2}\|\boldsymbol{\lambda}_k - \boldsymbol{\lambda}^\star_\tau\|^2 \\
\leq& \frac{\|\boldsymbol{\lambda}^\star_\tau - \boldsymbol{\lambda}_k\|^2 - \|\boldsymbol{\lambda}^\star_\tau - \boldsymbol{\lambda}_{k+1}\|^2}{2\eta} + \frac{\eta}{2} D'_{\tau,\Lambda} - \frac{\tau}{2}\|\boldsymbol{\lambda}_k - \boldsymbol{\lambda}^\star_\tau\|^2 \\
=& \frac{(1-\eta\tau)\|\boldsymbol{\lambda}^\star_\tau - \boldsymbol{\lambda}_k\|^2 - \|\boldsymbol{\lambda}^\star_\tau - \boldsymbol{\lambda}_{k+1}\|^2}{2\eta} + \frac{\eta}{2} D'_{\tau,\Lambda}.
\end{aligned} \tag{17}$$

From Lemma C.3 (with $\pi = \pi_k$, $\boldsymbol{\lambda} = \boldsymbol{\lambda}_k$), we have $0 \leq \mathcal{L}_\tau(\pi^\star_\tau, \boldsymbol{\lambda}_k) - \mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}^\star_\tau)$. Moreover, recall $\Phi_k = \mathrm{KL}_k + \frac{1}{2}\|\boldsymbol{\lambda}_k - \boldsymbol{\lambda}^\star_\tau\|^2$, thus by Equations (16) and (17),

$$\begin{aligned}
\Phi_{k+1} =& \mathrm{KL}_{k+1} + \frac{1}{2}\|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}^\star_\tau\|^2 \\
\leq& (1-\eta\tau)\mathrm{KL}_k + \frac{\eta^2}{2} D_{\eta,\tau,\Lambda} - \eta(i) + (1-\eta\tau)\frac{\|\boldsymbol{\lambda}_k - \boldsymbol{\lambda}^\star_\tau\|^2}{2} + \frac{\eta^2}{2} D'_{\tau,\Lambda} - \eta(ii) && \text{(Equations (16) and (17))} \\
\leq& (1-\eta\tau)\Phi_k + \eta^2(D_{\eta,\tau,\Lambda} + D'_{\tau,\Lambda}) - \eta\left((i) + (ii)\right) && \text{(Def. } \Phi_k\text{)} \\
\leq& (1-\eta\tau)\Phi_k + \eta^2(D_{\eta,\tau,\Lambda} + D'_{\tau,\Lambda}) - \eta\left(\mathcal{L}_\tau(\pi^\star_\tau, \boldsymbol{\lambda}_k) - \mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}^\star_\tau)\right) && \text{(Equation (15))} \\
\leq& (1-\eta\tau)\Phi_k + \eta^2(D_{\eta,\tau,\Lambda} + D'_{\tau,\Lambda}). && \text{(as } \mathcal{L}_\tau(\pi^\star_\tau, \boldsymbol{\lambda}_k) - \mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}^\star_\tau) \geq 0\text{)}
\end{aligned}$$

Finally, the claimed bound follows by noting that

$$\begin{aligned}
& D_{\eta,\tau,\Lambda} + D'_{\tau,\Lambda} \\
=& H A^{1/2} \exp\left(\eta H\left(1 + \lambda_{max} I + \tau\log(A)\right)\right)\left(2H^2\left(1 + I\lambda_{max} + \tau\log(A)\right)^2 + 2\tau^2(64/e^2)\right) + I(H + \tau\lambda_{max})^2 \\
\leq& \tilde{O}\left(\lambda^2_{max} H^3 A^{1/2} I^2 \exp\left(\eta H\left(1 + \lambda_{max} I + \log(A)\right)\right) + I(H + \tau\lambda_{max})^2\right),
\end{aligned}$$

as $\tau \leq 1$ and $\lambda_{max} \geq H\Xi^{-1} \geq 1$. □

We can use the following result to turn the convergence of the potential function into an error bound. We will then choose the optimal values for λ_{max}, τ, and η.

Lemma 4.2 (Error bounds)**.** *For any sequence* $(\pi_k)_{k\in[K]}$,

$$\begin{aligned}\left[V_r^{\pi^\star} - V_r^{\pi_k}\right]_+ \leq& H^{3/2}(2\Phi_k)^{1/2} + \tau H\log(A),\\ \max_{i\in[I]}\left[c_i - V_{u_i}^{\pi_k}\right]_+ \leq& H^{3/2}(2\Phi_k)^{1/2} + \tau\lambda_{max}\\ &+ \lambda_{max}^{-1}\left(H^2\Xi^{-1} + \tau H\log(A)\right).\end{aligned}$$

Proof. (1) We bound the objective optimality gap. First, decompose it as

$$V_r^{\pi^\star} - V_r^{\pi_k} = \underbrace{V_r^{\pi^\star} - V_r^{\pi_\tau^\star}}_{(i)} + \underbrace{V_r^{\pi_\tau^\star} - V_r^{\pi_k}}_{(ii)}. \tag{18}$$

We bound (ii) as follows:

$$\begin{aligned}(ii) =& V_r^{\pi_\tau^\star} - V_r^{\pi_k}\\ =& \sum_{s,a,h} d_h^{\pi_\tau^\star}(s)\left(\pi_{\tau,h}^\star(a|s) - \pi_{k,h}(a|s)\right)Q_{r,h}^{\pi_k}(s,a) && \text{(Lemma H.1)}\\ \leq& H\sum_{s,h} d_h^{\pi_\tau^\star}(s)\left\|\pi_{\tau,h}^\star(\cdot|s) - \pi_{k,h}(\cdot|s)\right\|_1\\ \leq& H\sum_{s,h} d_h^{\pi_\tau^\star}(s)\sqrt{2\mathrm{KL}_{k,h}(s)} && \text{(by Pinsker's)}\\ \leq& H^2\sqrt{2\sum_{s,h}\frac{1}{H}d_h^{\pi_\tau^\star}(s)\mathrm{KL}_{k,h}(s)} && \text{(by Jensen's)}\\ =& H^{3/2}\sqrt{2\mathrm{KL}_k}.\end{aligned}$$

We next bound term (i). By Lemma C.6 with $\pi = \pi^\star$ we have

$$V_r^{\pi^\star} - \tau\mathcal{H}(\pi_\tau^\star) \leq V_r^{\pi_\tau^\star} + \sum_i \lambda_{\tau,i}^\star\left(V_{g_i}^{\pi_\tau^\star} - V_{g_i}^{\pi^\star}\right).$$

By Lemma C.6 with $\boldsymbol{\lambda} = \mathbf{0}$ we have

$$\sum_i \lambda_{\tau,i}^\star V_{g_i}^{\pi_\tau^\star} \leq 0.$$

Moreover, $V_{g_i}^{\pi^\star} \geq 0$ by feasibility and $\lambda_{\tau,i}^\star \geq 0$. Combing these inequalities, we find

$$(i) = V_r^{\pi^\star} - V_r^{\pi_\tau^\star} \leq \tau\mathcal{H}(\pi_\tau^\star) \leq \tau H\log(A), \tag{19}$$

which concludes the proof for the objective optimality gap.

(2) Let $i \in [I]$. We now bound the i-th constraint violation. First, decompose it as

$$c_i - V_{u_i}^{\pi_k} = -V_{g_i}^{\pi_k} = \underbrace{-V_{g_i}^{\pi_\tau^\star}}_{(iii)} + \underbrace{V_{g_i}^{\pi_\tau^\star} - V_{g_i}^{\pi_k}}_{(iv)} \tag{20}$$

We first bound (iv). The same calculation as for the objective optimality gap (1) shows

$$(iv) = V_{g_i}^{\pi_\tau^\star} - V_{g_i}^{\pi_k} \leq H^{3/2}\sqrt{2\mathrm{KL}_k}. \tag{21}$$

We next bound term (iii). Recall $\Lambda = [0, \lambda_{max}]^I$. Lemma C.6 with $\pi = \pi^\star$ and $\boldsymbol{\lambda} \in \Lambda$ as

$$\lambda_j := \begin{cases} 0 & (j \neq i) \\ \lambda_{max} & (j = i) \end{cases}$$

yields

$$V_r^{\pi^\star} + \sum_j \lambda^\star_{\tau,j} V_{g_j}^{\pi^\star} \leq V_r^{\pi^\star_\tau} + \lambda_{max} V_{g_i}^{\pi^\star_\tau} + \frac{\tau}{2}\lambda_{max}^2 + \tau \mathcal{H}(\pi^\star_\tau)$$

From Lemma C.3 (with $\pi = \pi^\star_\tau$) we get

$$V_r^{\pi^\star_\tau} - V_r^{\pi^\star} \leq \sum_j \lambda^\star_j \left(V_{g_j}^{\pi^\star} - V_{g_j}^{\pi^\star_\tau} \right).$$

Adding the two previous inequalities and canceling terms, we get

$$0 \leq \sum_j \lambda^\star_{\tau,j} V_{g_j}^{\pi^\star} \leq \lambda_{max} V_{g_i}^{\pi^\star_\tau} + \frac{\tau}{2}\lambda_{max}^2 + \sum_j \lambda^\star_j \left(V_{g_j}^{\pi^\star} - V_{g_j}^{\pi^\star_\tau} \right) + \tau \mathcal{H}(\pi^\star_\tau),$$

where the first inequality holds since $0 \leq V_{g_j}^{\pi^\star}$ by feasibility and $\boldsymbol{\lambda}^\star_\tau \geq \mathbf{0}$. Rearranging this shows

$$\begin{aligned}
-V_{g_i}^{\pi^\star_\tau} \leq & \frac{\tau}{2}\lambda_{max} + \frac{1}{\lambda_{max}} \sum_j \lambda^\star_j \left(V_{g_j}^{\pi^\star} - V_{g_j}^{\pi^\star_\tau} \right) + \frac{1}{\lambda_{max}} \tau \mathcal{H}(\pi^\star_\tau) \\
= & \frac{\tau}{2}\lambda_{max} + \frac{1}{\lambda_{max}} \sum_j \lambda^\star_j \left(V_{u_j}^{\pi^\star} - V_{u_j}^{\pi^\star_\tau} \right) + \frac{1}{\lambda_{max}} \tau \mathcal{H}(\pi^\star_\tau) && (\boldsymbol{g} = \boldsymbol{u} - \tfrac{1}{H}\boldsymbol{c}) \\
\leq & \frac{\tau}{2}\lambda_{max} + \frac{1}{\lambda_{max}} \|\boldsymbol{\lambda}^\star\|_1 H + \frac{1}{\lambda_{max}} \tau \mathcal{H}(\pi^\star_\tau) && \text{(Hölder's)} \\
\leq & \frac{\tau}{2}\lambda_{max} + \frac{1}{\lambda_{max}} \left(\frac{H^2}{\Xi} + \tau \mathcal{H}(\pi^\star_\tau) \right) && \text{(Lemma C.2)} \\
\leq & \frac{\tau}{2}\lambda_{max} + \frac{1}{\lambda_{max}} \left(\frac{H^2}{\Xi} + \tau H \log(A) \right).
\end{aligned}$$

□

Finally, we are ready to prove last-iterate convergence by combining the previous two lemmas.

Theorem 4.1 (Last-iterate convergence)**.** *Let $\varepsilon \in (0,1)$. Then, with appropriate choices of $\eta \propto \varepsilon^6$, $\tau \propto \varepsilon^2$, $\lambda_{max} \propto \varepsilon^{-1}$, we have*

$$\left[V_r^{\pi^\star} - V_r^{\pi_k} \right]_+ \leq \varepsilon, \quad \left[c_i - V_{u_i}^{\pi_k} \right]_+ \leq \varepsilon \quad (\forall i \in [I])$$

for $k = \Omega(poly(A, H, I, \Xi^{-1}) \cdot \varepsilon^{-10})$.

Proof. The bound follows from Lemma 4.1 and Lemma 4.2. We choose $\tau = \varepsilon^2$, $\eta = (H^2 I \log(A))^{-1} \Xi \varepsilon^6$, $\lambda_{max} = \frac{H}{\Xi}\varepsilon^{-1} \geq \frac{H}{\Xi}$. Set $\Delta_r(k) := \left[V_r^{\pi^\star} - V_r^{\pi_k} \right]_+$ and $\Delta_{g_i}(k) := \left[-V_{g_i}^{\pi_k} \right]_+$.

We first consider the suboptimality for the reward. Plugging Lemma 4.1 into Lemma 4.2 we find, using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $1 + x \leq \exp(x)$,

$$\begin{aligned}
\Delta_r(k) \leq & H^{3/2} \Phi_1^{1/2} \exp\left(-\eta\tau k/2\right) && \text{(a)} \\
& + H^{3/2} \left(\frac{\eta}{\tau}\right)^{1/2} \tilde{O}(C_{\eta,\tau,\Lambda}^{1/2}) && \text{(b)} \\
& + \tau H \log(A). && \text{(c)}
\end{aligned}$$

For (b), note that, using the definitions of η, τ, λ_{max} (and taking $\sqrt{\cdot}$, and $\tau < 1$)

$$\begin{aligned}C_{\eta,\tau,\Lambda}^{1/2} \leq& \lambda_{max} H^{3/2} A^{1/4} I \exp\left(\eta H \left(1 + \lambda_{max} I + \log(A)\right)/2\right) + I^{1/2}(H + \tau\lambda_{max}) \\ \leq& \lambda_{max} H^{3/2} A^{1/4} I \exp(2) + I^{1/2}(H + \tau\lambda_{max}) \\ =& \varepsilon^{-1} \cdot H^{5/2} A^{1/4} I \Xi^{-1} \exp(2) + I^{1/2} H + I^{1/2} \varepsilon^2 H \Xi^{-1} \varepsilon^{-1}) \\ \lesssim& \varepsilon^{-1} \cdot H^{5/2} A^{1/4} I \Xi^{-1}.\end{aligned}$$

Since

$$\left(\frac{\eta}{\tau}\right)^{1/2} = (H^2 I \log(A))^{-1/2} \Xi^{1/2} \varepsilon^{(6-2)/2} = (H^2 I \log(A))^{-1/2} \Xi^{1/2} \varepsilon^2,$$

we thus have

$$(b) = H^{3/2} \left(\frac{\eta}{\tau}\right)^{1/2} C_{\eta,\tau,\Lambda} \lesssim H^3 I^{1/2} A^{1/4} \Xi^{-1/2} \varepsilon = \text{poly}(A, H, I, \Xi^{-1}) \cdot \varepsilon.$$

Similarly,

$$(c) = \tau H \log(A) = H \log(A) \varepsilon^2.$$

For (a), using the standard inequality $e^{-x} \leq 1 - x/2$ (if $0 \leq x \leq 1$) with $x := \eta\tau/2$, we first find

$$\exp(-\eta\tau l/2) \leq (1 - \eta\tau/4)^l$$

and hence,

$$\begin{aligned}(a) = H^{3/2} \Phi_1^{1/2} \exp(-\eta\tau k/2) \leq& H^{3/2} \Phi_1^{1/2} \cdot \frac{1}{k} \sum_{l=1}^{k} \exp(-\eta\tau l/2) \\ \leq& H^{3/2} \Phi_1^{1/2} \cdot \frac{1}{k} \sum_{l=1}^{k} (1 - \eta\tau/4)^l \\ \leq& H^{3/2} \Phi_1^{1/2} \cdot \frac{1}{k} \sum_{l=1}^{\infty} (1 - \eta\tau/4)^l \\ =& H^{3/2} \Phi_1^{1/2} \cdot \frac{1}{k} \frac{4}{\eta\tau} \\ =& H^{3/2} \Phi_1^{1/2} \frac{1}{k} \frac{4}{(H^2 I \log(A))^{-1} \Xi \varepsilon^6 \varepsilon^2} \\ =& 4 H^{7/2} I \Xi^{-1} \log(A) \frac{1}{k} \Phi_1^{1/2} \varepsilon^{-8}.\end{aligned}$$

Furthermore, since π_1 plays actions uniformly at random and $\boldsymbol{\lambda_1} = \mathbf{0}$, we have $\Phi_1^{1/2} \leq (H \log(A) + \frac{1}{2} I \lambda_{max}^2)^{1/2} \leq H^{1/2} \log(A)^{1/2} + I^{1/2} \lambda_{max} = H^{1/2} \log(A)^{1/2} + I^{1/2} H \Xi^{-1} \varepsilon^{-1}$. Hence, the calculation above shows

$$(a) \leq 4 H^{7/2} I \Xi^{-1} \log(A) \frac{1}{k} (H^{1/2} \log(A)^{1/2} + I^{1/2} H \Xi^{-1} \varepsilon^{-1}) \varepsilon^{-8} \leq \text{poly}(A, H, I, \Xi^{-1}) \varepsilon$$

for $k = \Omega(\varepsilon^{-10})$. Hence, summing up terms (a) to (c) and choosing $k = \Omega(\text{poly}(A, H, I, \Xi^{-1}) \varepsilon^{-10})$ yields the bound for the objective.

Next, we consider the regret for the constraints. Plugging Lemma 4.1 into Lemma 4.2 we find, using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$,

$$\begin{aligned}\Delta_{u_i}(k) \leq& H^{3/2} \Phi_1^{1/2} \exp(-\eta\tau k/2) && \text{(a')} \\ &+ H^{3/2} \left(\frac{\eta}{\tau}\right)^{1/2} \tilde{O}(C_{\eta,\tau,\Lambda}^{1/2}) && \text{(b')} \\ &+ \tau\lambda_{max} + \frac{1}{\lambda_{max}} H^2 \Xi^{-1} && \text{(c')} \\ &+ \frac{1}{\lambda_{max}} \tau H \log(A). && \text{(d')}\end{aligned}$$

Note that terms (a'), (b') are identical to (a), (b). Moreover, for (d') we have

$$(d') = \frac{1}{\lambda_{max}}\tau \log(A) = \Xi H^{-1}\log(A)\varepsilon^3.$$

Finally, for (c'), we have

$$\begin{aligned}(c') =& \tau\lambda_{max} + \frac{1}{\lambda_{max}}H^2\Xi^{-1}\\ =& \varepsilon^2 \cdot H\Xi^{-1}\varepsilon^{-1} + H^{-1}\Xi\varepsilon \cdot H^2\Xi^{-1}\\ =& H(1+\Xi^{-1})\varepsilon.\end{aligned}$$

Thus, summing up (a') to (d') and choosing $k = \Omega(\text{poly}(A, H, I, \Xi^{-1})\varepsilon^{-10})$ yields the bound for the constraints. □

E. Properties of the Optimistic Model

In this section, we establish important properties of the model Algorithm 1 builds.

E.1. Building the Model

First, we describe the exact model and how we perform policy evaluation.

We follow Shani et al. (2020) for the optimistic exploration, but we also take the I constraint functions $\boldsymbol{u}$ into account rather than just the reward function r. We also need to pay special attention to the auxiliary term ψ_k.

For all s, a, h and $k \in [K]$, let $n_{k-1,h}(s,a) := \sum_{l=1}^{k-1} \mathbb{1}_{\{s_h^l = s,\ a_h^l = a\}}$ 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 visited at step h in episode l. First, we compute the empirical averages of the reward and transition probabilities as follows:

$$\begin{aligned}\bar{r}_{k-1,h}(s,a) :=& \frac{\sum_{l=1}^{k-1} R_h^l(s,a)\mathbb{1}_{\{s_h^l=s,\ a_h^l=a\}}}{n_{k-1,h}(s,a)\vee 1},\\ \bar{u}_{k-1,i,h}(s,a) :=& \frac{\sum_{l=1}^{k-1} U_{i,h}^l(s,a)\mathbb{1}_{\{s_h^l=s,\ a_h^l=a\}}}{n_{k-1,h}(s,a)\vee 1} \quad (\forall i \in [I]),\\ \bar{p}_{k-1,h}(s'|s,a) :=& \frac{\sum_{l=1}^{k-1} \mathbb{1}_{\{s_h^l=s,\ a_h^l=a,\ s_{h+1}^l=s'\}}}{n_{k-1,h}(s,a)\vee 1},\end{aligned}$$

where $a \vee b := \max\{a, b\}$. We consider optimistic estimates $\hat{r}_k$, $\hat{\boldsymbol{u}}_k$, $\hat{p}_k$:

$$\begin{aligned}\hat{r}_{k,h}(s,a) :=& \bar{r}_{k-1,h}(s,a) + b_{k-1,h}(s,a),\\ \hat{u}_{k,i,h}(s,a) :=& \bar{u}_{k-1,i,h}(s,a) + b_{k-1,h}(s,a) \quad (\forall i \in [I]),\\ \hat{p}_{k,h}(s'|s,a) :=& \bar{p}_{k-1,h}(s'|s,a),\end{aligned}$$

with the bonuses $b_{k-1,h}(s,a) = b^r_{k-1,h}(s,a) + b^p_{k-1,h}(s,a)$ specified below. For ψ_k, we take[4]

$$\hat{\psi}_{k,h}(s,a) := \psi_{k,h}(s,a) + b^p_{k-1,h}(s,a)\log(A).$$

For notational convenience, we write

$$z_k := r + \boldsymbol{\lambda}_k^T\boldsymbol{u} + \tau\psi_k, \qquad \hat{z}_k := \hat{r}_k + \boldsymbol{\lambda}_k^T\hat{\boldsymbol{u}}_k + \tau\hat{\psi}_k$$

for the reward function mimicking the π-dependency of the regularized Lagrangian at $(\pi_k, \boldsymbol{\lambda}_k)$.

For any $\delta \in (0,1)$, we specify the correct bonuses to obtain our regret guarantees with probability at least $1-\delta$:

$$b_{k-1,h}(s,a) := b^r_{k-1,h}(s,a) + b^p_{k-1,h}(s,a),$$

[4] In other words, there is no bonus for the function, only for the transitions. This is because ψ_k is known in episode k, and so the only uncertainty in the corresponding value function is due to estimating the transitions p. Note that the extra $\log(A)$ factor corrects for the fact that ψ_k is not a function to $[0,1]$.

where

$$b^r_{k-1,h}(s,a) := \sqrt{\frac{\frac{1}{2}\log\left(\frac{2SAH(I+1)K}{\delta'}\right)}{n_{k-1,h}(s,a)\vee 1}}, \quad b^p_{k-1,h}(s,a) := H\sqrt{\frac{2S+2\log\left(\frac{SAHK}{\delta'}\right)}{n_{k-1,h}(s,a)\vee 1}}.$$

For ψ_k, recall

$$\hat{\psi}_{k,h}(s,a) := \psi_{k,h}(s,a) + b^p_{k-1,h}(s,a)\log(A).$$

We define the *truncated value functions*[5]

$$\hat{Q}^k_{\hat{z}_k,h}(s,a) := \hat{Q}^k_{\hat{r}_k,h}(s,a) + \sum_i \lambda_{k,i}\hat{Q}^k_{\hat{u}_{k,i},h}(s,a) + \tau\hat{Q}^k_{\hat{\psi}_k,h}(s,a), \tag{22}$$

$$\hat{V}^k_{\hat{z}_k,h}(s) := \left\langle \pi_{k,h}(\cdot|s), \hat{Q}^k_{\hat{z}_k,h}(s,\cdot)\right\rangle, \tag{23}$$

where we compute $\hat{Q}^k_{\hat{r}_k,h}(s,a)$, $\hat{Q}^k_{\hat{u}_{k,i},h}(s,a)$, $\hat{Q}^k_{\hat{\psi}_k,h}(s,a)$ via truncated policy evaluation with respect to the estimated model, see Algorithm 2.

Algorithm 2 EVAL (Truncated Policy Evaluation)

Initialize $\hat{V}^k_{\hat{r}_k,H+1}(s) = \hat{V}^k_{\hat{u}_{k,i},H+1}(s) = \hat{V}^k_{\hat{\psi}_k,H+1}(s) = 0$ (for $s \in \mathcal{S}$)
for $h = H, \dots, 1$ **do**
  for $(s,a) \in \mathcal{S}\times\mathcal{A}$ **do**
    Truncated DP step:

$$\hat{Q}^k_{\hat{r}_k,h}(s,a) := \min\left\{H-h+1,\ \hat{r}_{k,h}(s,a) + \left\langle \hat{p}_{k,h}(\cdot|s,a), \hat{V}^k_{\hat{r}_k,h+1}(\cdot)\right\rangle\right\}$$
$$\hat{Q}^k_{\hat{u}_{k,i},h}(s,a) := \min\left\{H-h+1,\ \hat{u}_{k,i,h}(s,a) + \left\langle \hat{p}_{k,h}(\cdot|s,a), \hat{V}^k_{\hat{u}_{k,i},h+1}(\cdot)\right\rangle\right\} \quad (\forall i \in [I])$$
$$\hat{Q}^k_{\hat{\psi}_k,h}(s,a) := \min\left\{\psi_{k,h}(s,a) + (H-h+1)\log(A),\ \hat{\psi}_{k,h}(s,a) + \left\langle \hat{p}_{k,h}(\cdot|s,a), \hat{V}^k_{\hat{\psi}_k,h+1}(\cdot)\right\rangle\right\}$$

    Retrieve V-function:

$$\hat{V}^k_{\hat{r}_k,h}(s) := \left\langle \pi_{k,h}(\cdot|s), \hat{Q}^k_{\hat{r}_k,h}(s,\cdot)\right\rangle$$
$$\hat{V}^k_{\hat{u}_{k,i},h}(s) := \left\langle \pi_{k,h}(\cdot|s), \hat{Q}^k_{\hat{u}_{k,i},h}(s,\cdot)\right\rangle \quad (\forall i \in [I])$$
$$\hat{V}^k_{\hat{\psi}_k,h}(s) := \left\langle \pi_{k,h}(\cdot|s), \hat{Q}^k_{\hat{\psi}_k,h}(s,\cdot)\right\rangle$$

  end for
end for
output $\hat{Q}^k_{\hat{z}_k}(\cdot) := \hat{Q}^k_{\hat{r}_k}(\cdot) + \sum_i \lambda_{k,i}\hat{Q}^k_{\hat{u}_{k,i}}(\cdot) + \tau\hat{Q}^k_{\hat{\psi}_k}(\cdot)$, and $(\hat{V}^k_{\hat{u}_{k,i}}(s_1,1))_i$

The main reason for truncating during the otherwise standard policy evaluation algorithm is the need for a bonus-independent upper bound on the surrogate value functions so that Lemma E.10 holds.[6] Clearly, the truncated value functions are all lower bounded by zero and upper bounded by the actual value functions under the estimated model. Finally, note that these truncated value functions need not correspond to the true value function of a policy in some MDP.

Recall the truncated value functions from Algorithm 2 in Appendix E.1. Note that due to the separate definition of $\hat{Q}^k_{\hat{z}_k,h}(s,a)$

[5] Importantly, note that this is the definition of $\hat{Q}^k_{\hat{z}_k,h}(s,a)$, rather than running truncated policy evaluation on z_k.

[6] In fact, truncation is only required for the update of π, and for the update of $\boldsymbol{\lambda}$, we can use either truncated or exact values.

and the updates in Algorithm 2, for all $s \in \mathcal{S}$, $h \in [H]$,

$$\begin{aligned}
\hat{V}^k_{\hat{z}_k,h}(s) &= \left\langle \pi_{k,h}(\cdot|s), \hat{Q}^k_{\hat{z}_k,h}(s,\cdot) \right\rangle && \text{(by def.)} \\
&= \left\langle \pi_{k,h}(\cdot|s), \hat{Q}^k_{\hat{r}_k,h}(s,a) + \sum_i \lambda_{k,i} \hat{Q}^k_{\hat{u}_{k,i},h}(s,a) + \tau \hat{Q}^k_{\hat{\psi}_k,h}(s,a) \right\rangle && \text{(by def.)} \\
&= \left\langle \pi_{k,h}(\cdot|s), \hat{Q}^k_{\hat{r}_k,h}(s,a) \right\rangle + \sum_i \lambda_{k,i} \left\langle \pi_{k,h}(\cdot|s), \hat{Q}^k_{\hat{u}_{k,i},h}(s,a) \right\rangle \\
&\quad + \tau \left\langle \pi_{k,h}(\cdot|s), \hat{Q}^k_{\hat{\psi}_k,h}(s,a) \right\rangle \\
&= \hat{V}^k_{\hat{r}_k,h}(s) + \sum_i \lambda_{k,i} \hat{V}^k_{\hat{u}_{k,i},h}(s) + \tau \hat{V}^k_{\hat{\psi}_k,h}(s). && (24)
\end{aligned}$$

Similarly, by linearity of expectation

$$V^{\pi_k}_{z_k,h}(s) = V^{\pi_k}_{r_k,h}(s) + \sum_i \lambda_{k,i} V^{\pi_k}_{u_{k,i},h}(s) + \tau V^{\pi_k}_{\psi_k,h}(s) \tag{25}$$

for the true value functions.

E.2. Properties of the Model

We are now ready to establish the properties of the model. In particular, we will show that it is optimistic with respect to the value function and prove bounds on the estimation error during the learning procedure.

Success Event We will condition our regret analysis on a *success event* G, which we formally define below. Fix $\delta > 0$, and construct the estimated model as in Appendix E.1. G ensures that (a) the optimistic reward estimates are in fact optimistic and (b) the true transitions are close to the estimated ones, i.e.:

$$\begin{aligned}
r &\leq \hat{r}_k, \\
u_i &\leq \hat{u}_{k,i} \quad (\forall i \in [I]), \\
\|p_h(\cdot|s,a) - \bar{p}_{k-1,h}(\cdot|s,a)\|_1 H &\leq b^p_{k-1,h}(s,a) \quad (\forall s,a,h),
\end{aligned}$$

for every episode $k \in [K]$. Formally, with $\delta' := \delta/3$, define the *failure events*

$$\begin{aligned}
F^r_k &:= \left\{ \exists s,a,h\colon\ |r_h(s,a) - \bar{r}_{k-1,h}(s,a)| \geq b^r_{k-1,h}(s,a) \right\}, \\
F^u_k &:= \left\{ \exists i,s,a,h\colon\ |u_{i,h}(s,a) - \bar{u}_{k-1,i,h}(s,a)| \geq b^r_{k-1,h}(s,a) \right\}, \\
F^p_k &:= \left\{ \exists s,a,h\colon\ \|p_h(\cdot|s,a) - \bar{p}_{k-1,h}(\cdot|s,a)\|_1 H \geq b^p_{k-1,h}(s,a) \right\}, \\
F^n_k &:= \left\{ \exists s,a,h\colon n_{k-1,h}(s,a) \leq \frac{1}{2} \sum_{j<k} d^{\pi_j}_h(s,a) - H \log\left(\frac{SAH}{\delta'}\right) \right\},
\end{aligned}$$

where $d^{\pi_j}_h(s,a)$ refers to the occupancy measure (Appendix C), and

$$\begin{aligned}
F^r &:= \left(\bigcup_{k=1}^K F^r_k \right) \bigcup \left(\bigcup_{k=1}^K F^u_k \right), \\
F^p &:= \bigcup_{k=1}^K F^p_k, \\
F^n &:= \bigcup_{k=1}^K F^n_k.
\end{aligned}$$

We define the *success event* G as the complement of $F^r \cup F^p \cup F^n$, i.e.

$$G := \overline{F^r \cup F^p \cup F^n}.$$

We now show that this event holds with high probability. The proof of this theorem relies on standard concentration bounds (specifically, Hoeffding for the rewards and $L1$-concentration for the transitions) and a union bound over all involved indices.

Lemma E.1 (Success event)**.** *Let $\delta > 0$ and define the bonuses accordingly. Suppose that for all $k \in [K]$, in episode k, policy π_k is played. Then $P[G] \geq 1 - \delta$.*

Proof. By Hoeffding's for any possible realization of $n_{k-1,h}(s,a)$ (and total probability), we have $P[F^r] \leq \delta'$ after union bound over all indices s, a, h and all episodes k. For $n_{k-1,h}(s,a) = 0$ the bound holds trivially.

By the $L1$ concentration bound of Weissman et al. (2003, Theorem 2.1), for any possible realization of $n_{k-1,h}(s,a)$ (and total probability), we have $P[F^p] \leq \delta'$ after union bound over all indices s, a, h and all episodes k. For $n_{k-1,h}(s,a) = 0$ the bound holds trivially.

By Dann et al. (2017, Corollary E.4), we also have $P[F^n] \leq \delta'$.

We conclude by union bound over the three events. □

Decomposition via Extended Value Difference Lemma The following lemma allows us to decompose the instantaneous regret into three terms that we will bound separately.

Lemma E.2 (Decomposition via simulation lemma)**.** *We have the following decomposition:*

$$\begin{aligned}
&V_{z_k}^{\pi_\tau^\star} - V_{z_k}^{\pi_k} \\
=&\hat{V}_{\hat{z}_k}^k - V_{z_k}^{\pi_k} && \text{(a)} \\
+&\sum_h \mathbb{E}\left[\left\langle \hat{Q}_{\hat{z}_k,h}^k(s_h,\cdot), \pi_{\tau,h}^\star(\cdot|s_h) - \pi_{k,h}(\cdot|s_h)\right\rangle \;\middle|\; s_1, \pi_\tau^\star, p\right] && \text{(b)} \\
+&\sum_h \mathbb{E}\left[-\hat{Q}_{\hat{z}_k,h}^k(s_h,a_h) + z_{k,h}(s_h,a_h) + \left\langle p_h(\cdot|s_h,a_h), \hat{V}_{z_k,h+1}^k(\cdot)\right\rangle \;\middle|\; s_1, \pi_\tau^\star, p\right]. && \text{(c)}
\end{aligned}$$

Proof. First, expand

$$V_{z_k}^{\pi_\tau^\star} - V_{z_k}^{\pi_k} = \left(\hat{V}_{\hat{z}_k}^k - V_{z_k}^{\pi_k}\right) + \left(V_{z_k}^{\pi_\tau^\star} - \hat{V}_{\hat{z}_k}^k\right).$$

Then apply Lemma H.2 to $\pi = \pi_k$, $\pi' = \pi_\tau^\star$ and $M = (\mathcal{S}, \mathcal{A}, \hat{p}_k, \hat{z}_k)$, $M' = (\mathcal{S}, \mathcal{A}, p, z_k)$ to the second term (after multiplying both sides by -1). □

General On-Policy Bounds The following two results are standard and will allow us to bound the estimation errors during learning. Consider the setup in which policy π_k is derived based on the previous episodes $1, \ldots, k-1$, and then played in episode k. Recall that for all s, a, h and $k \in [K]$, $n_{k-1,h}(s,a) := \sum_{l=1}^{k-1} \mathbb{1}_{\{s_h^l = s,\, a_h^l = a\}}$ counts the visits of state-action pair (s,a) at step h before episode k. We write $\lesssim$ for asymptotic inequality up to polylogarithmic terms.

Note that in the following two lemmas, the exponent of H is different from the one in the referenced proofs. This is because the referenced works consider the case of stationary transition probabilities, whereas we consider non-stationary dynamics. See Shani et al. (2020, Lemmas 18, 19).

Lemma E.3 (Lemma 36, Efroni et al. (2020))**.** *Suppose for all s, a, h, $k \in [K]$, we have*

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

Then for all $K' \leq K$

$$\sum_{k'=1}^{K'} \sum_{h=1}^{H} \mathbb{E}\left[\frac{1}{\sqrt{n_{k'-1,h}(s_h^{k'}, a_h^{k'})}} \mid \mathcal{F}_{k'-1}\right] \leq \tilde{O}(\sqrt{SAH^2K'} + SAH),$$

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 E.4 (Lemma 37, Efroni et al. (2020))**.** *Suppose for all s, a, h, $k \in [K]$, we have*

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

Then for all $K' \leq K$

$$\sum_{k'=1}^{K'} \sum_{h=1}^{H} \mathbb{E}\left[\frac{1}{n_{k'-1,h}(s_h^{k'}, a_h^{k'})} \mid \mathcal{F}_{k'-1}\right] \leq \tilde{O}(SAH^2),$$

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 & Brunskill (2019, Lemma 13) for a proof of the statement. □

Estimation Error (On-Policy Error Bounds) We next prove bounds on the estimation error obtained while learning the model.

Lemma E.5 (Estimation error $\hat{r}$, $\hat{\boldsymbol{u}}$)**.** *Conditioned on G, for every $K' \in [K]$, we have*

$$\sum_{k=1}^{K'} \left(\hat{V}_{\hat{r}_k}^k - V_r^{\pi_k}\right) \leq \tilde{O}\left(\sqrt{S^2AH^4K'} + S^{3/2}AH^2\right),$$
$$\sum_{k=1}^{K'} \left(\hat{V}_{\hat{u}_{k,i}}^k - V_{u_i}^{\pi_k}\right) \leq \tilde{O}\left(\sqrt{S^2AH^4K'} + S^{3/2}AH^2\right) \quad (\forall i \in [I]).$$

Proof. By Lemma H.2 (with $\pi = \pi' = \pi_k$ and $M = (\mathcal{S}, \mathcal{A}, \hat{p}_k, \hat{z}_k)$, $M' = (\mathcal{S}, \mathcal{A}, p, z_k)$), we have according to the truncated policy evaluation (Algorithm 2),

$$\begin{aligned}
&\hat{V}_{\hat{r}_k}^k - V_r^{\pi_k} \\
=&\sum_{h=1}^{H} \mathbb{E}\left[\hat{Q}_{\hat{r}_k,h}^k(s_h,a_h) - r_h(s_h,a_h) - \left\langle p_h(\cdot|s_h,a_h), \hat{V}_{\hat{r}_k,h+1}^k(\cdot)\right\rangle \,\middle|\, s_1, \pi_k, p\right] \\
\leq&\sum_{h=1}^{H} \mathbb{E}\left[\hat{r}_{k,h}(s_h,a_h) - r_h(s_h,a_h) \,\middle|\, s_1, \pi_k, p\right] \\
&+ \sum_{h=1}^{H} \mathbb{E}\left[\left\langle \hat{p}_{k,h}(\cdot|s_h,a_h) - p_h(\cdot|s_h,a_h), \hat{V}_{\hat{r}_k,h+1}^k(\cdot)\right\rangle \,\middle|\, s_1, \pi_k, p\right] \\
=&\sum_{h=1}^{H} \mathbb{E}\left[\bar{r}_{k-1,h}(s_h,a_h) + b_{k-1,h}^r(s_h,a_h) - r_h(s_h,a_h) \,\middle|\, s_1, \pi_k, p\right] \\
&+ \sum_{h=1}^{H} \mathbb{E}\left[b_{k-1,h}^p(s_h,a_h) + \left\langle \hat{p}_{k,h}(\cdot|s_h,a_h) - p_h(\cdot|s_h,a_h), \hat{V}_{\hat{r}_k,h+1}^k(\cdot)\right\rangle \,\middle|\, s_1, \pi_k, p\right].
\end{aligned}$$

Since G occurs, we have $\bar{r}_{k-1,h}(s_h,a_h) - r_h(s_h,a_h) \leq b^r_{k-1,h}(s_h,a_h)$. Moreover,

$$\begin{aligned}
&\left\langle \hat{p}_{k,h}(\cdot|s_h,a_h) - p_h(\cdot|s_h,a_h), \hat{V}^k_{\hat{r}_k,h+1}(\cdot)\right\rangle \\
\leq& \left\|\hat{p}_{k,h}(\cdot|s_h,a_h) - p_h(\cdot|s_h,a_h)\right\|_1 \left\|\hat{V}^k_{\hat{r}_k,h+1}(\cdot)\right\|_\infty \\
\leq& \left\|\hat{p}_{k,h}(\cdot|s_h,a_h) - p_h(\cdot|s_h,a_h)\right\|_1 H \\
\leq& b^p_{k-1,h}(s_h,a_h)
\end{aligned}$$

by Hölder's, the truncation in the policy evaluation, and since G occurs. Plugging this into the inequality above,

$$\begin{aligned}
\hat{V}^k_{\hat{r}_k} - V^{\pi_k}_r \leq& \sum_{h=1}^H \mathbb{E}\left[2b^r_{k-1,h}(s_h,a_h) \,\middle|\, s_1,\pi_k,p\right] \\
&+ \sum_{h=1}^H \mathbb{E}\left[2b^p_{k-1,h}(s_h,a_h) \,\middle|\, s_1,\pi_k,p\right].
\end{aligned}$$

Recalling the definition of $b^r_{k-1,h}(s,a) \lesssim \frac{1}{\sqrt{n_{k-1,h}(s,a)\vee 1}}$, $b^p_{k-1,h}(s,a) \lesssim \frac{H\sqrt{S}}{\sqrt{n_{k-1,h}(s,a)\vee 1}}$ and summing up, we thus find

$$\begin{aligned}
\sum_{k=1}^{K'}\left(\hat{V}^k_{\hat{r}_k} - V^{\pi_k}_r\right) \lesssim& H\sqrt{S}\sum_{k=1}^{K'}\sum_{h=1}^H \mathbb{E}\left[\frac{1}{\sqrt{n_{k-1,h}(s_h,a_h)\vee 1}} \,\middle|\, s_1,\pi_k,p\right] \\
=& H\sqrt{S}\sum_{k=1}^{K'}\sum_{h=1}^H \mathbb{E}\left[\frac{1}{\sqrt{n_{k-1,h}(s^k_h,a^k_h)\vee 1}} \,\middle|\, \mathcal{F}_{k-1}\right] && \text{(play } \pi_k\text{)} \\
\lesssim& H\sqrt{S}\left(\sqrt{SAH^2K'} + SAH\right) && \text{(Lemma E.3)} \\
=& \sqrt{S^2AH^4K'} + S^{3/2}AH^2
\end{aligned}$$

where we used that we play π_k in episode k and Lemma E.3, which applies since G occurs.

The proof for u_i ($i \in [I]$) is identical. □

Lemma E.6 (Estimation error $\hat{\psi}$)**.** *Conditioned on G, for every $K' \in [K]$, we have*

$$\sum_{k=1}^{K'}\left(\hat{V}^k_{\hat{\psi}_k} - V^{\pi_k}_{\psi_k}\right) \leq \tilde{O}\left(\sqrt{S^2AH^4K'} + S^{3/2}AH^2\right).$$

Proof. By Lemma H.2 (with $\pi = \pi' = \pi_k$ and $M = (\mathcal{S},\mathcal{A},\hat{p}_k,\hat{z}_k)$, $M' = (\mathcal{S},\mathcal{A},p,z_k)$), we have according to the

truncated policy evaluation,

$$
\begin{aligned}
&\hat{V}^k_{\hat{r}_k} - V^{\pi_k}_r \\
=&\sum_{h=1}^{H} \mathbb{E}\left[\hat{Q}^k_{\hat{\psi}_k,h}(s_h,a_h) - \psi_{k,h}(s_h,a_h) - \left\langle p_h(\cdot|s_h,a_h), \hat{V}^k_{\hat{\psi}_k,h+1}(\cdot)\right\rangle \;\middle|\; s_1,\pi_k,p\right] \\
\leq&\sum_{h=1}^{H} \mathbb{E}\left[\hat{\psi}_{k,h}(s_h,a_h) - \psi_{k,h}(s_h,a_h) \;\middle|\; s_1,\pi_k,p\right] \\
&+\sum_{h=1}^{H} \mathbb{E}\left[\left\langle \hat{p}_{k,h}(\cdot|s_h,a_h) - p_h(\cdot|s_h,a_h), \hat{V}^k_{\hat{\psi}_k,h+1}(\cdot)\right\rangle \;\middle|\; s_1,\pi_k,p\right] \\
=&\sum_{h=1}^{H} \mathbb{E}\left[\psi_{k,h}(s_h,a_h) - \psi_{k,h}(s_h,a_h) \;\middle|\; s_1,\pi_k,p\right] \\
+&\sum_{h=1}^{H} \mathbb{E}\left[b^p_{k-1,h}(s_h,a_h)\log(A) + \left\langle \hat{p}_{k,h}(\cdot|s_h,a_h) - p_h(\cdot|s_h,a_h), \hat{V}^k_{\hat{\psi}_k,h+1}(\cdot)\right\rangle \;\middle|\; s_1,\pi_k,p\right].
\end{aligned}
$$

Since G occurs, we have

$$
\begin{aligned}
&\left\langle \hat{p}_{k,h}(\cdot|s_h,a_h) - p_h(\cdot|s_h,a_h), \hat{V}^k_{\hat{\psi}_k,h+1}(\cdot)\right\rangle \\
\leq& \left\|\hat{p}_{k,h}(\cdot|s_h,a_h) - p_h(\cdot|s_h,a_h)\right\|_1 \left\|\hat{V}^k_{\hat{\psi}_k,h+1}(\cdot)\right\|_\infty \\
\leq& \left\|\hat{p}_{k,h}(\cdot|s_h,a_h) - p_h(\cdot|s_h,a_h)\right\|_1 H\log(A) \\
\leq& b^p_{k-1,h}(s_h,a_h)\log(A)
\end{aligned}
$$

by Hölder's, the truncation in the policy evaluation, and since G occurs. Plugging this into the inequality above,

$$
\hat{V}^k_{\hat{\psi}_k} - V^{\pi_k}_{\psi_k} \leq \sum_{h=1}^{H} \mathbb{E}\left[2b^p_{k-1,h}(s_h,a_h)\log(A) \;\middle|\; s_1,\pi_k,p\right]
$$

and the rest of the proof follows exactly as in the proof of Lemma E.5, with an extra $\log(A)$ factor. □

The following lemma allows us to control the total estimation error (a) from the optimistic model. Roughly speaking, it guarantees that the estimates are not too optimistic.

Lemma E.7 (Estimation error with regularization)**.** *Conditioned on G, for every $K' \in [K]$, we have (if $\tau \leq 1$)*

$$
\sum_{k=1}^{K'} \left(\hat{V}^k_{\hat{z}_k} - V^{\pi_k}_{z_k}\right) \lesssim (2 + I\lambda_{max})\left(\sqrt{S^2AH^4K'} + S^{3/2}AH^2\right).
$$

Proof. By Equations (22), (24) and (25), conditioned on G,

$$
\begin{aligned}
&\sum_{k=1}^{K'} \left(\hat{V}^k_{\hat{z}_k} - V^{\pi_k}_{z_k}\right) \\
=&\sum_{k=1}^{K'} \left(\hat{V}^k_{\hat{r}_k} - V^{\pi_k}_{r}\right) + \sum_{i=1}^{I}\lambda_k(i)\sum_{k=1}^{K'}\left(\hat{V}^k_{\hat{u}_{k,i}} - V^{\pi_k}_{u_i}\right) + \tau\sum_{k=1}^{K'}\left(\hat{V}^k_{\hat{\psi}_k} - V^{\pi_k}_{\psi_k}\right) \\
\lesssim&(1 + I\lambda_{max} + \tau)\left(\sqrt{S^2AH^4K'} + S^{3/2}AH^2\right),
\end{aligned}
$$

and $\tau \leq 1$. The last inequality holds due to Lemmas E.5 and E.6. □

Per-State Optimism In the following, we show per-state optimism for the optimistic model.

Lemma E.8 (State optimism $\hat{r}$, $\hat{\boldsymbol{u}}$)**.** *Conditioned on the success event G, for all s, a, h, and $k \in [K]$,*

$$-\hat{Q}^k_{\hat{r}_k,h}(s,a) + r_h(s,a) + \left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{r}_k,h+1}(\cdot)\right\rangle \leq 0,$$
$$-\hat{Q}^k_{\hat{u}_{k,i},h}(s,a) + u_{k,i,h}(s,a) + \left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{u}_{k,i},h+1}(\cdot)\right\rangle \leq 0 \quad (i \in [I])$$

Proof. For $\hat{r}_k$, by Algorithm 2 we have

$$\begin{aligned}
&\hat{Q}^k_{\hat{r}_k,h}(s,a)\\
=&\min\left\{H-h+1,\ \hat{r}_{k,h}(s,a) + \left\langle \hat{p}_{k,h}(\cdot|s,a), \hat{V}^k_{\hat{r}_k,h+1}(\cdot)\right\rangle\right\}\\
=&\min\left\{H-h+1,\right.\\
&\left.\bar{r}_{k-1,h}(s,a) + b^r_{k-1,h}(s,a) + \left\langle \hat{p}_{k,h}(\cdot|s,a), \hat{V}^k_{\hat{r}_k,h+1}(\cdot)\right\rangle + b^p_{k-1,h}(s,a)\right\}\\
\geq&\min\left\{1,\ \bar{r}_{k-1,h}(s,a) + b^r_{k-1,h}(s,a)\right\}\\
&+\min\left\{H-h, \left\langle \hat{p}_{k,h}(\cdot|s,a), \hat{V}^k_{\hat{r}_k,h+1}(\cdot)\right\rangle + b^p_{k-1,h}(s,a)\right\},
\end{aligned}$$

where we used $\min\{a+b, c+d\} \geq \min\{a,c\} + \min\{b,d\}$. Thus

$$\begin{aligned}
&-\hat{Q}^k_{\hat{r}_k,h}(s,a) + r_h(s,a) + \left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{r}_k,h+1}(\cdot)\right\rangle\\
\leq&-\min\left\{1,\ \bar{r}_{k-1,h}(s,a) + b^r_{k-1,h}(s,a)\right\} + r_h(s,a)\\
&-\min\left\{H-h, \left\langle \hat{p}_{k,h}(\cdot|s,a), \hat{V}^k_{\hat{r}_k,h+1}(\cdot)\right\rangle + b^p_{k-1,h}(s,a)\right\} + \left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{r}_k,h+1}(\cdot)\right\rangle\\
=&-\min\left\{1 - r_h(s,a),\ \bar{r}_{k-1,h}(s,a) - r_h(s,a) + b^r_{k-1,h}(s,a)\right\}\\
&-\min\left\{H-h-\left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{r}_k,h+1}(\cdot)\right\rangle,\right.\\
&\left.\left\langle \hat{p}_{k,h}(\cdot|s,a) - p_h(\cdot|s,a), \hat{V}^k_{\hat{r}_k,h+1}(\cdot)\right\rangle + b^p_{k-1,h}(s,a)\right\}\\
=&\max\left\{\underbrace{r_h(s,a)-1}_{(a)},\ \underbrace{-\bar{r}_{k-1,h}(s,a) + r_h(s,a) - b^r_{k-1,h}(s,a)}_{(b)}\right\}\\
&+\max\left\{\underbrace{-(H-h) + \left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{r}_k,h+1}(\cdot)\right\rangle}_{(c)},\right.\\
&\left.\underbrace{-\left\langle \hat{p}_{k,h}(\cdot|s,a) - p_h(\cdot|s,a), \hat{V}^k_{\hat{r}_k,h+1}(\cdot)\right\rangle - b^p_{k-1,h}(s,a)}_{(d)}\right\}.
\end{aligned}$$

Now, for each of the four terms appearing in the maxima, conditioned on G, we have

$$(a) = r_h(s,a) - 1 \leq 0$$

by boundedness of the reward functions,

$$(b) = -\bar{r}_{k,h}(s,a) + r_h(s,a) - b^r_{k-1,h}(s,a) \leq 0$$

since G occurs,

$$(c) = -(H-h) + \left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{r}_k,h+1}(\cdot)\right\rangle \leq -(H-h) + 1\cdot(H-(h+1)+1) = 0$$

by Hölder's and the truncation of our evaluation procedure, and finally

$$\begin{aligned}(d) &= -\left\langle \hat{p}_{k,h}(\cdot|s,a) - p_h(\cdot|s,a), \hat{V}^k_{\hat{r}_k,h+1}(\cdot)\right\rangle - b^p_{k-1,h}(s,a)\\ &\le \|\hat{p}_{k,h}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \cdot (H-(h+1)+1) - b^p_{k-1,h}(s,a)\\ &\le \|\hat{p}_{k,h}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \cdot H - b^p_{k-1,h}(s,a) \le 0\end{aligned}$$

by Hölder's, the truncation of our evaluation procedure, since $\hat{p}_k = \bar{p}_{k-1}$, and since G occurs. We are thus taking the minimum of non-positive terms, which shows

$$-\hat{Q}^k_{\hat{r}_k,h}(s,a) + r_h(s,a) + \left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{r}_k,h+1}(\cdot)\right\rangle \le 0.$$

The proof for $\hat{u}_{k,i}$ ($i \in [I]$) is identical. □

Lemma E.9 (State optimism $\hat{\psi}$)**.** *Conditioned on the success event G, for all s, a, h, and $k \in [K]$,*

$$-\hat{Q}^k_{\hat{\psi}_k,h}(s,a) + \psi_{k,h}(s,a) + \left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{\psi}_k,h+1}(\cdot)\right\rangle \le 0.$$

Proof. For $\hat{\psi}_{k,h}(s,a) = \psi_{k,h}(s,a) + b^p_{k-1,h}(s,a)\log(A)$, from Algorithm 2 we have

$$\begin{aligned}&\hat{Q}^k_{\hat{\psi}_k,h}(s,a)\\ =&\min\left\{\psi_{k,h}(s,a) + (H-h+1)\log(A),\ \hat{\psi}_{k,h}(s,a) + \left\langle \hat{p}_{k,h}(\cdot|s,a), \hat{V}^k_{\hat{\psi}_k,h+1}(\cdot)\right\rangle\right\}\\ =&\psi_{k,h}(s,a)\\ &+\min\left\{(H-h+1)\log(A),\ b^p_{k-1,h}(s,a)\log(A) + \left\langle \hat{p}_{k,h}(\cdot|s,a), \hat{V}^k_{\psi_k,h+1}(\cdot)\right\rangle\right\}.\end{aligned}$$

Thus

$$\begin{aligned}&-\hat{Q}^k_{\hat{\psi}_k,h}(s,a) + \psi_{k,h}(s,a) + \left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{\psi}_k,h+1}(\cdot)\right\rangle\\ =&-\min\left\{(H-h+1)\log(A),\ b^p_{k-1,h}(s,a)\log(A) + \left\langle \hat{p}_{k,h}(\cdot|s,a), \hat{V}^k_{\hat{\psi}_k,h+1}(\cdot)\right\rangle\right\}\\ &+\left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{\psi}_k,h+1}(\cdot)\right\rangle\\ =&-\min\Bigg\{(H-h+1)\log(A) - \left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{\psi}_k,h+1}(\cdot)\right\rangle,\\ &b^p_{k-1,h}(s,a)\log(A) + \left\langle \hat{p}_{k,h}(\cdot|s,a) - p_h(\cdot|s,a), \hat{V}^k_{\hat{\psi}_k,h+1}(\cdot)\right\rangle\Bigg\}\\ =&+\max\Bigg\{\underbrace{-(H-h+1)\log(A) + \left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{\psi}_k,h+1}(\cdot)\right\rangle}_{=:(a)},\\ &\underbrace{-b^p_{k-1,h}(s,a)\log(A) - \left\langle \hat{p}_{k,h}(\cdot|s,a) - p_h(\cdot|s,a), \hat{V}^k_{\hat{\psi}_k,h+1}(\cdot)\right\rangle}_{=:(b)}\Bigg\}.\end{aligned}$$

For (a), first note that by the truncation in the policy evaluation (Algorithm 2)

$$\begin{aligned}\hat{V}^k_{\hat{\psi}_k,h+1}(s) &= \left\langle \pi_{k,h+1}(\cdot|s), \hat{Q}^k_{\hat{\psi}_k,h+1}(s,\cdot)\right\rangle && \text{(by definition)}\\ &\le \sum_a \pi_{k,h+1}(a|s)\cdot\psi_{k,h+1}(s,a)\\ &\quad+\sum_a \pi_{k,h+1}(a|s)\cdot(H-(h+1)+1)\log(A) && \text{(truncated update)}\\ &\le \log(A) + \sum_a \pi_{k,h+1}(a|s)\cdot(H-h)\log(A)\\ &\le (H-h+1)\log(A). && (26)\end{aligned}$$

Hence

$$\begin{aligned}(a) =& -(H-h+1)\log(A) + \left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{\psi}_k,h+1}(\cdot)\right\rangle \\ \leq& -(H-h+1)\log(A) + (H-h+1)\log(A) \\ =& 0\end{aligned}$$

by Hölder's and Equation (26). For (b),

$$\begin{aligned}(b) =& -b^p_{k-1,h}(s,a)\log(A) - \left\langle \hat{p}_{k,h}(\cdot|s,a) - p_h(\cdot|s,a), \hat{V}^k_{\hat{\psi}_k,h+1}(\cdot)\right\rangle \\ \leq& -b^p_{k-1,h}(s,a)\log(A) + \|\hat{p}_{k,h}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \cdot (H-h+1)\log(A) \\ \leq& -b^p_{k-1,h}(s,a)\log(A) + \|\hat{p}_{k,h}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \cdot H\log(A) \\ \leq& 0\end{aligned}$$

by Hölder's, and since $\hat{p}_k = \bar{p}_{k-1}$ and G occurs. □

Finally, the following lemma shows that we can discard term (c) in Lemma E.2. It guarantees that for every state, an optimistic *Bellman-type* inequality holds.

Lemma E.10 (State optimism $\hat{z}$)**.** *Conditioned on the success event G,*

$$-\hat{Q}^k_{\hat{z}_k,h}(s,a) + z_{k,h}(s,a) + \left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{z}_k,h+1}(\cdot)\right\rangle \leq 0.$$

Proof. By Equations (22) and (24), conditioned on G,

$$\begin{aligned}& -\hat{Q}^k_{\hat{z}_k,h}(s,a) + z_{k,h}(s,a) + \left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{z}_k,h+1}(\cdot)\right\rangle \\ =& -\hat{Q}^k_{\hat{r}_k,h}(s,a) + r_{k,h}(s,a) + \left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{r}_k,h+1}(\cdot)\right\rangle \\ & \sum_i \lambda_{k,i}\left(-\hat{Q}^k_{\hat{u}_{k,i},h}(s,a) + u_{k,i,h}(s,a) + \left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{u}_{k,i},h+1}(\cdot)\right\rangle\right) \\ & \tau\left(-\hat{Q}^k_{\hat{\psi}_k,h}(s,a) + \psi_{k,h}(s,a) + \left\langle p_h(\cdot|s,a), \hat{V}^k_{\hat{\psi}_k,h+1}(\cdot)\right\rangle\right) \\ \leq& 0,\end{aligned}$$

where the last inequality holds due to Lemmas E.8 and E.9 and since all $\lambda_{k,i} \geq 0$. □

Value Function Bounds In the following, we bound the regularized value functions, which allows us to apply descent properties of the regularized primal-dual algorithm. Recall that $\boldsymbol{g} = \boldsymbol{u} - \frac{1}{H}\boldsymbol{c}$.

Lemma E.11 (Value function bounds)**.** *For any s, a, h, we have*

$$0 \leq Q^{\pi_k}_{r+\boldsymbol{\lambda}_k^T\boldsymbol{u}+\tau\psi_k,h}(s,a) \leq -\tau\log(\pi_{k,h}(a|s)) + H(1 + I\lambda_{max} + \tau\log(A)) \tag{27}$$

Moreover, for any s, a, h,

$$\begin{aligned}& \sum_a \pi_{k,h}(a|s)\exp\left(\eta Q^{\pi_k}_{r+\boldsymbol{\lambda}_k^T\boldsymbol{u}+\tau\psi_k,h}(s,a)\right) Q^{\pi_k}_{r+\boldsymbol{\lambda}_k^T\boldsymbol{u}+\tau\psi_k,h}(s,a)^2 \\ \leq& \sqrt{A}\exp\left(\eta H\left(1 + \lambda_{max}I + \tau\log(A)\right)\right)\left(2H^2\left(1 + I\lambda_{max} + \tau\log(A)\right)^2 + 2\tau^2(64/e^2)\right)\end{aligned}$$

and

$$\left\|\boldsymbol{V}^{\pi_k}_{\boldsymbol{g}} + \tau\boldsymbol{\lambda}_k\right\| \leq \sqrt{I}\left(H + \tau\lambda_{max}\right)$$

Proof. We prove the first inequality. Non-negativity is immediate. Moreover, for all s, a,

$$\begin{aligned}
Q^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{u}+\tau\psi_k,h}(s,a) =& \left| Q^{\pi_k}_{r,h}(s,a) + \sum_i \lambda_{k,i} Q^{\pi_k}_{u_i,h}(s,a) + \tau Q^{\pi_k}_{\psi_k,h}(s,a) \right| \\
\leq& \left| Q^{\pi_k}_{r,h}(s,a) \right| + \sum_i \lambda_{k,i} \left| Q^{\pi_k}_{u_i,h}(s,a) \right| + \tau \left| Q^{\pi_k}_{\psi_k,h}(s,a) \right| \\
\leq& H + I\lambda_{max} H + \tau \mathbb{E}_{\pi_k} \left[\sum_{h'=h}^{H} -\log(\pi_{k,h'}(a_{h'}|s_{h'})) \mid s_h = s, a_h = a \right].
\end{aligned}$$

We finish by bounding

$$\begin{aligned}
&\mathbb{E}_{\pi_k} \left[\sum_{h'=h}^{H} -\log(\pi_{k,h'}(a_{h'}|s_{h'})) \,\middle|\, s_h = s, a_h = a \right] \\
=& -\log(\pi_{k,h}(a|s)) + \mathbb{E}_{\pi_k} \left[\sum_{h'=h+1}^{H} -\log(\pi_{k,h'}(a_{h'}|s_{h'})) \,\middle|\, a_h = a, s_h = s \right] \\
=& -\log(\pi_{k,h}(a|s)) + \mathbb{E}_{\pi_k} \left[\sum_{h'=h+1}^{H} -\log(\pi_{k,h'}(a_{h'}|s_{h'})) \,\middle|\, s_{h+1} \sim p_h(\cdot|s,a) \right] \\
=& -\log(\pi_{k,h}(a|s)) + \sum_{h'=h+1} \sum_{s'} d^{\pi_k}_{s_{h+1}\sim p_h(\cdot|s,a),h'}(s') \sum_{a'} -\pi_{k,h'}(a'|s') \log(\pi_{k,h'}(a'|s')) \\
\leq& -\log(\pi_{k,h}(a|s)) + \sum_{h'=h+1} \sum_{s'} d^{\pi_k}_{s_{h+1}\sim p_h(\cdot|s,a),h'}(s') \log(A) \\
=& -\log(\pi_{k,h}(a|s)) + \sum_{h'=h+1} \log(A) \\
\leq& -\log(\pi_{k,h}(a|s)) + H\log(A),
\end{aligned}$$

where we used the standard bound on the entropy in the third to last step, and we are considering unnormalized occupancy measures throughout.

For the second inequality, first note that (using $(a+b)^2 \leq 2a^2 + 2b^2$)

$$Q^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{u}+\tau\psi_k,h}(s,a)^2 \leq \underbrace{2H^2 \left(1 + I\lambda_{max} + \tau\log(A)\right)^2}_{=:C_1} + 2\tau^2 \log^2\left(\frac{1}{\pi_{k,h}(a|s)}\right).$$

Moreover, using Equation (27) we have

$$\begin{aligned}
&\pi_{k,h}(a|s) \exp\left(\eta Q^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{u}+\tau\psi_k,h}(s,a)\right) \\
\leq& \pi_{k,h}(a|s)^{1-\eta\tau} \underbrace{\exp\left(\eta H \left(1 + \lambda_{max} I + \tau\log(A)\right)\right)}_{=:C_2}
\end{aligned}$$

We thus find, if $\eta\tau \leq 1/4 < 1/2$ (which is easily satisfied since we choose η, τ small),

$$\begin{aligned}
&\sum_a \pi_{k,h}(a|s) \exp\left(\eta Q^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{u}+\tau\psi_k,h}(s,a)\right) \cdot C_1 \\
\leq& \sum_a \pi_{k,h}(a|s)^{1-\eta\tau} \cdot C_2 \cdot C_1 \\
\leq& \sum_a \pi_{k,h}(a|s)^{1/2} \cdot C_2 \cdot C_1 \\
\leq& \left(\sum_a \pi_{k,h}(a|s)\right)^{1/2} \left(AC_2^2 \cdot C_1^2\right)^{1/2} \\
=& \sqrt{A} C_1 C_2,
\end{aligned}$$

where the last inequality is Cauchy-Schwarz, and

$$\begin{aligned}
&\sum_a \pi_{k,h}(a|s) \exp\left(\eta Q^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{u}+\tau\psi_k,h}(s,a)\right) \cdot 2\tau^2 \log^2\left(\frac{1}{\pi_{k,h}(a|s)}\right) \\
\leq &\sum_a \pi_{k,h}(a|s)^{1-\eta\tau} \cdot C_2 \cdot 2\tau^2 \log^2\left(\frac{1}{\pi_{k,h}(a|s)}\right) \\
\leq &\sum_a \pi_{k,h}(a|s)^{1/2} \cdot C_2 \cdot 2\tau^2 \pi_k(a|s,h)^{1/4} \log^2\left(\frac{1}{\pi_{k,h}(a|s)}\right) \\
\leq &\sum_a \pi_{k,h}(a|s)^{1/2} \cdot C_2 \cdot 2\tau^2 (64/e^2) \\
\leq &\sqrt{A} C_2 \cdot 2\tau^2 (64/e^2),
\end{aligned}$$

where we used the fact that $q^{1/4}\log^2(1/q) \leq 64/e^2$ for $q \in (0,1)$, and Cauchy-Schwarz in the same manner as before. Adding up the previous two terms and plugging in the definitions of the constants yields the second inequality.

For the third inequality, we find (recalling that $g_h(s,a) \in [-1,1]^I$)

$$\begin{aligned}
\left\|\boldsymbol{V}_{\boldsymbol{g}}^{\pi_k} + \tau\boldsymbol{\lambda}_k\right\| \leq &\left\|\boldsymbol{V}_{\boldsymbol{g}}^{\pi_k}\right\| + \tau\left\|\boldsymbol{\lambda}_k\right\| \\
\leq &\sqrt{I}H + \tau\sqrt{I}\lambda_{max},
\end{aligned}$$

concluding the proof. □

Lemma E.12 (Truncated value function bounds)**.** *For all s, a, h, we have*

$$0 \leq \hat{Q}^k_{\hat{z}_k,h}(s,a) \leq -\tau\log(\pi_{k,h}(a|s)) + H(1 + I\lambda_{max} + \tau\log(A)) \tag{28}$$

and

$$\begin{aligned}
&\sum_a \pi_{k,h}(a|s) \exp\left(\eta \hat{Q}^k_{\hat{z}_k,h}(s,a)\right) \hat{Q}^k_{\hat{z}_k,h}(s,a)^2 \\
\leq &\sqrt{A} \exp\left(\eta H\left(1 + \lambda_{max} I + \tau\log(A)\right)\right)\left(2H^2\left(1 + I\lambda_{max} + \tau\log(A)\right)^2 + 2\tau^2(64/e^2)\right)
\end{aligned}$$

and

$$\left\|\hat{\boldsymbol{V}}^k_{\hat{\boldsymbol{u}}_k} - \boldsymbol{c} + \tau\boldsymbol{\lambda}_k\right\| \leq \sqrt{I}\left(H + \tau\lambda_{max}\right)$$

Proof. Since the truncated value functions are bounded between 0 and the true value functions, the statement follows from Lemma E.11. □

F. Regret Analysis

In this section, we provide all proofs of the result from Section 5, leading to a regret bound of the regularized primal-dual algorithm (Algorithm 3).

We write $\lesssim$ for asymptotic inequality up to polylogarithmic terms. First, we note that the primal-dual updates indeed correspond to mirror descent updates.

Observation F.1. *The closed-form expressions in Algorithm 1 are solutions to*

$$\begin{aligned}
&\max_{\pi_h(\cdot|s)\in\Delta(\mathcal{A})} \left(\sum_{a\in\mathcal{A}} \pi_h(a|s)\hat{Q}^k_{\hat{z}_k,h}(s,a) - \frac{1}{\eta} KL(\pi_h(\cdot|s), \pi_{k,h}(\cdot|s))\right), \\
&\min_{\boldsymbol{\lambda}\in\Lambda} \left(\boldsymbol{\lambda}^T\left(\hat{\boldsymbol{V}}^k_{\hat{\boldsymbol{u}}_k} - \boldsymbol{c} + \tau\boldsymbol{\lambda}_k\right) + \frac{1}{2\eta}\left\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_k\right\|^2\right),
\end{aligned}$$

respectively.

This means that both the primal and dual variables are updated via mirror descent with different regularizers. Hence, we can make use of the descent lemmas for online mirror descent (we refer to Lemmas I.5 and I.8 in Appendix I.3). However, the value functions that serve as (surrogate) gradients are only estimates. Thus, the convergence proof for the potential function Φ_k that measures the distance from $(\pi_k, \boldsymbol{\lambda}_k)$ to $(\pi_\tau^\star, \boldsymbol{\lambda}_\tau^\star)$ needs to take the estimation errors into account.

Lemma 5.1 (Regularized convergence)**.** *Let $\eta, \tau < 1$ and $\lambda_{max} \geq H\Xi^{-1}$. With probability at least $1-\delta$, the iterates of Algorithm 1 satisfy*

$$\Phi_{k+1} \leq (1-\eta\tau)^k \Phi_1 + \tilde{O}\Bigg(\eta\tau^{-1} C_{\eta,\tau,\Lambda} + \eta\lambda_{max}\left(ISA^{1/2}H^2k^{1/2} + IS^{3/2}AH^2\right)\Bigg),$$

where $C_{\eta,\tau,\Lambda}$ is the same constant as in Lemma 4.1.

Proof. Condition on G, which occurs with probability at least $1-\delta$ by Lemma E.1. We first decompose the k-th *primal-dual gap* as follows:

$$\mathcal{L}_\tau(\pi_\tau^\star, \boldsymbol{\lambda}_k) - \mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}_\tau^\star) = \underbrace{\mathcal{L}_\tau(\pi_\tau^\star, \boldsymbol{\lambda}_k) - \mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}_k)}_{(i)} + \underbrace{\mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}_k) - \mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}_\tau^\star)}_{(ii)}. \tag{29}$$

We first bound term (i):

$$\begin{aligned}
(i) =& \mathcal{L}_\tau(\pi_\tau^\star, \boldsymbol{\lambda}_k) - \mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}_k) \\
=& V^{\pi_\tau^\star}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}} - V^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}} && \text{(as } d_h^\pi(s,a) = d_h^\pi(s)\pi_h(a|s)\text{, and cancel } \|\boldsymbol{\lambda}_k\|^2\text{)} \\
& - \tau \sum_{s,a,h} d_h^{\pi_\tau^\star}(s)\pi_{\tau,h}^\star(a|s)\log(\pi_{\tau,h}^\star(a|s)) + \tau \sum_{s,a,h} d_h^{\pi_k}(s)\pi_{k,h}(a|s)\log(\pi_{k,h}(a|s)) \\
=& V^{\pi_\tau^\star}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} - V^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} && \text{(since } \psi_{k,h}(s,a) = -\log(\pi_{k,h}(a|s))\text{)} \\
& + \tau \sum_{s,a,h} d_h^{\pi_\tau^\star}(s)\pi_{\tau,h}^\star(a|s)\log(\pi_{k,h}(a|s)) - \tau \sum_{s,a,h} d_h^{\pi_k}(s)\pi_{k,h}(a|s)\log(\pi_{k,h}(a|s)) \\
& - \tau \sum_{s,a,h} d_h^{\pi_\tau^\star}(s)\pi_{\tau,h}^\star(a|s)\log(\pi_{\tau,h}^\star(a|s)) + \tau \sum_{s,a,h} d_h^{\pi_k}(s)\pi_{k,h}(a|s)\log(\pi_{k,h}(a|s)) \\
=& V^{\pi_\tau^\star}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} - V^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} \\
& + \tau \sum_{s,a,h} d_h^{\pi_\tau^\star}(s)\pi_{\tau,h}^\star(a|s)\log(\pi_{k,h}(a|s)) \\
& - \tau \sum_{s,a,h} d_h^{\pi_\tau^\star}(s)\pi_{\tau,h}^\star(a|s)\log(\pi_{\tau,h}^\star(a|s)) \\
=& V^{\pi_\tau^\star}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} - V^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} \\
& - \tau \sum_{s,h} d_h^{\pi_\tau^\star}(s) \sum_a \pi_{\tau,h}^\star(a|s) \log\left(\frac{\pi_{\tau,h}^\star(a|s)}{\pi_{k,h}(a|s)}\right) \\
=& V^{\pi_\tau^\star}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} - V^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} - \tau \sum_{s,h} d_h^{\pi_\tau^\star}(s)\mathrm{KL}_{k,h}(s) \\
=& V^{\pi_\tau^\star}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} - V^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}+\tau\psi_k} - \tau\mathrm{KL}_k \\
=& V^{\pi_\tau^\star}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{u}+\tau\psi_k} - V^{\pi_k}_{r+\boldsymbol{\lambda}_k^T \boldsymbol{u}+\tau\psi_k} - \tau\mathrm{KL}_k && \text{(as } \boldsymbol{g} = \boldsymbol{u} - \tfrac{1}{H}\boldsymbol{c}\text{)} \\
=& V^{\pi_\tau^\star}_{z_k} - V^{\pi_k}_{z_k} - \tau\mathrm{KL}_k
\end{aligned}$$

Now by Lemma E.2, we have

$$
\begin{aligned}
& V_{z_k}^{\pi_\tau^\star} - V_{z_k}^{\pi_k} \\
=& (\hat{V}_{\hat{z}_k}^k - V_{z_k}^{\pi_k}) && \text{(a)} \\
&+ \sum_h \mathbb{E}\left[\left\langle \hat{Q}_{\hat{z}_k,h}^k(s_h,\cdot), \pi_{\tau,h}^\star(\cdot|s_h) - \pi_{k,h}(\cdot|s_h) \right\rangle \;\middle|\; s_1, \pi_\tau^\star, p \right] && \text{(b)} \\
&+ \sum_h \mathbb{E}\left[-\hat{Q}_{\hat{z}_k,h}^k(s_h,a_h) + z_{k,h}(s_h,a_h) + \left\langle p_h(\cdot|s_h,a_h), \hat{V}_{\hat{z}_k,h+1}^{\pi_k}(\cdot) \right\rangle \;\middle|\; s_1, \pi_\tau^\star, p \right] && \text{(c)}
\end{aligned}
$$

We leave term (a) as is and will sum over k later. For term (b), note that for all s, h,

$$
\begin{aligned}
& \left\langle \hat{Q}_{\hat{z}_k,h}^k(s,\cdot), \pi_{\tau,h}^\star(\cdot|s) - \pi_{k,h}(\cdot|s) \right\rangle \\
\leq& \frac{\mathrm{KL}_{k,h}(s) - \mathrm{KL}_{k+1,h}(s)}{\eta} + \frac{\eta}{2} \sum_a \pi_{k,h}(a|s) \exp\left(\hat{Q}_{\hat{z}_k,h}^k(s,a) \right) \hat{Q}_{\hat{z}_k,h}^k(s,a)^2 && \text{(Lemma I.6)} \\
\leq& \frac{\mathrm{KL}_{k,h}(s) - \mathrm{KL}_{k+1,h}(s)}{\eta} + \frac{\eta}{2} \frac{1}{H} D_{\eta,\tau,\Lambda}, && \text{(Lemma E.12)}
\end{aligned}
$$

with

$$
D_{\eta,\tau,\Lambda} = HA^{1/2} \exp\left(\eta H \left(1 + \lambda_{max} I + \tau \log(A)\right)\right) \left(2H^2 \left(1 + I\lambda_{max} + \tau \log(A)\right)^2 + 2\tau^2 (64/e^2) \right)
$$

and where we were able to apply Lemma I.6 by Observation F.1 and since $\hat{Q}_{\hat{z}_k,h}^k(s,a) \geq 0$. Hence,

$$
\begin{aligned}
\sum_h \mathbb{E}\left[\left\langle \hat{Q}_{\hat{z}_k,h}^k(s_h,\cdot), \pi_{\tau,h}^\star(\cdot|s_h) - \pi_{k,h}(\cdot|s_h) \right\rangle \;\middle|\; s_1, \pi_\tau^\star, p \right] =& \sum_{s,h} d_h^{\pi_\tau^\star}(s) \left\langle \hat{Q}_{\hat{z}_k,h}^k(s,\cdot), \pi_{\tau,h}^\star(\cdot|s) - \pi_{k,h}(\cdot|s) \right\rangle \\
\leq& \sum_{s,h} d_h^{\pi_\tau^\star}(s) \left(\frac{\mathrm{KL}_{k,h}(s) - \mathrm{KL}_{k+1,h}(s)}{\eta} + \frac{\eta}{2} \frac{1}{H} D_{\eta,\tau,\Lambda} \right) \\
=& \frac{\mathrm{KL}_k - \mathrm{KL}_{k+1}}{\eta} + \frac{\eta}{2} D_{\eta,\tau,\Lambda}.
\end{aligned}
$$

Term (c) is ≤ 0 by Lemma E.10, which applies since G occurs. Plugging in, we thus find

$$
\begin{aligned}
(i) =& V_{z_k}^{\pi_\tau^\star} - V_{z_k}^{\pi_k} - \tau \mathrm{KL}_k \\
\leq& (\hat{V}_{\hat{z}_k}^k - V_{z_k}^{\pi_k}) + \frac{\mathrm{KL}_k - \mathrm{KL}_{k+1}}{\eta} + \frac{\eta}{2} D_{\eta,\tau,\Lambda} + 0 - \tau \mathrm{KL}_k \\
=& (\hat{V}_{\hat{z}_k}^k - V_{z_k}^{\pi_k}) + \frac{(1-\eta\tau)\mathrm{KL}_k - \mathrm{KL}_{k+1}}{\eta} + \frac{\eta}{2} D_{\eta,\tau,\Lambda}. && (30)
\end{aligned}
$$

We now bound term (ii):

$$
\begin{aligned}
(ii) =& \mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}_k) - \mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}_\tau^\star) \\
=& V_{r+\boldsymbol{\lambda}_k^T \boldsymbol{g}}^{\pi_k} - V_{r+(\boldsymbol{\lambda}_\tau^\star)^T \boldsymbol{g}}^{\pi_k} + \frac{\tau}{2} \|\boldsymbol{\lambda}_k\|^2 - \frac{\tau}{2} \|\boldsymbol{\lambda}_\tau^\star\|^2 && \text{(cancel } \mathcal{H}(\pi_k)\text{)} \\
=& \sum_i (\lambda_{k,i} - \lambda_{\tau,i}^\star) V_{g_i}^{\pi_k} + \frac{\tau}{2} \|\boldsymbol{\lambda}_k\|^2 - \frac{\tau}{2} \|\boldsymbol{\lambda}_\tau^\star\|^2 \\
=& \sum_i (\lambda_{k,i} - \lambda_{\tau,i}^\star)(V_{u_i}^{\pi_k} - c_i + \tau\lambda_{k,i}) - \frac{\tau}{2} \|\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_\tau^\star\|^2 \\
=& \sum_i (\lambda_{k,i} - \lambda_{\tau,i}^\star)(\hat{V}_{\hat{u}_{k,i}}^k - c_i + \tau\lambda_{k,i}) && \text{(a)} \\
& \sum_i (\lambda_{k,i} - \lambda_{\tau,i}^\star)(V_{u_i}^{\pi_k} - \hat{V}_{\hat{u}_{k,i}}^k) && \text{(b)} \\
& - \frac{\tau}{2} \|\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_\tau^\star\|^2
\end{aligned}
$$

To bound (a), we note that

$$\begin{aligned}
&\sum_i (\lambda_{k,i} - \lambda^\star_{\tau,i})(\hat{V}^k_{\hat{u}_{k,i}} - c_i + \tau\lambda_{k,i}) \\
\leq &\frac{\|\boldsymbol{\lambda}^\star_\tau - \boldsymbol{\lambda}_k\|^2 - \|\boldsymbol{\lambda}^\star_\tau - \boldsymbol{\lambda}_{k+1}\|^2}{2\eta} + \frac{\eta}{2}\left\|\hat{\boldsymbol{V}}^k_{\hat{u}_k} - \boldsymbol{c} + \tau\boldsymbol{\lambda}_k\right\|^2 && \text{(Lemma I.8)} \\
\leq &\frac{\|\boldsymbol{\lambda}^\star_\tau - \boldsymbol{\lambda}_k\|^2 - \|\boldsymbol{\lambda}^\star_\tau - \boldsymbol{\lambda}_{k+1}\|^2}{2\eta} + \frac{\eta}{2}D'_{\tau,\Lambda}, && \text{(Lemma E.12)}
\end{aligned}$$

with $D'_{\tau,\Lambda} = I(H + \tau\lambda_{max})^2$ and where we were able to apply Lemma I.8 by Observation F.1. We can bound (b) via

$$\sum_i (\lambda_{k,i} - \lambda^\star_{\tau,i})(V^{\pi_k}_{u_i} - \hat{V}^k_{\hat{u}_{k,i}}) \leq \sum_i \lambda_{max}\left|\hat{V}^k_{\hat{u}_{k,i}} - V^{\pi_k}_{u_i}\right|.$$

Plugging in, we find

$$\begin{aligned}
(ii) = &\sum_i (\lambda_{k,i} - \lambda^\star_{\tau,i})(V^{\pi_k}_{u_i} - c_i + \tau\lambda_{k,i}) - \frac{\tau}{2}\|\boldsymbol{\lambda}_k - \boldsymbol{\lambda}^\star_\tau\|^2 \\
\leq &\frac{\|\boldsymbol{\lambda}^\star_\tau - \boldsymbol{\lambda}_k\|^2 - \|\boldsymbol{\lambda}^\star_\tau - \boldsymbol{\lambda}_{k+1}\|^2}{2\eta} + \frac{\eta}{2}D'_{\tau,\Lambda} + \sum_i \lambda_{max}\left|\hat{V}^k_{\hat{u}_{k,i}} - V^{\pi_k}_{u_i}\right| - \frac{\tau}{2}\|\boldsymbol{\lambda}_k - \boldsymbol{\lambda}^\star_\tau\|^2 \\
= &\frac{(1-\eta\tau)\|\boldsymbol{\lambda}^\star_\tau - \boldsymbol{\lambda}_k\|^2 - \|\boldsymbol{\lambda}^\star_\tau - \boldsymbol{\lambda}_{k+1}\|^2}{2\eta} + \frac{\eta}{2}D'_{\tau,\Lambda} + \sum_i \lambda_{max}\left|\hat{V}^k_{\hat{u}_{k,i}} - V^{\pi_k}_{u_i}\right|. && (31)
\end{aligned}$$

Before proceeding, note that conditioned on G, we have

$$\begin{aligned}
\hat{V}^k_{\hat{u}_{k,i}} - V^{\pi_k}_{u_i} \geq &0, \\
\hat{V}^k_{\hat{z}_k} - V^{\pi_k}_{z_k} \geq &0,
\end{aligned}$$

by Lemmas E.8 and E.10, respectively. Hence, we can treat these differences and their absolute values interchangeably in what follows.

From Lemma C.3 (with $\pi = \pi_k$, $\boldsymbol{\lambda} = \boldsymbol{\lambda}_k$), we have $0 \leq \mathcal{L}_\tau(\pi^\star_\tau, \boldsymbol{\lambda}_k) - \mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}^\star_\tau)$. Moreover, recall $\Phi_k = \mathrm{KL}_k + \frac{1}{2}\|\boldsymbol{\lambda}_k - \boldsymbol{\lambda}^\star_\tau\|^2$, thus by Equations (30) and (31),

$$\begin{aligned}
\Phi_{k+1} = &\mathrm{KL}_{k+1} + \frac{1}{2}\|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}^\star_\tau\|^2 \\
\leq &(1-\eta\tau)\mathrm{KL}_k + \frac{\eta^2}{2}D_{\eta,\tau,\Lambda} + \eta(\hat{V}^k_{\hat{z}_k} - V^{\pi_k}_{z_k}) - \eta(i) && \text{(Equations (30) and (31))} \\
&+ (1-\eta\tau)\frac{\|\boldsymbol{\lambda}_k - \boldsymbol{\lambda}^\star_\tau\|^2}{2} + \frac{\eta^2}{2}D'_{\tau,\Lambda} + \eta\sum_i \lambda_{max}\left|\hat{V}^k_{\hat{u}_{k,i}} - V^{\pi_k}_{u_i}\right| - \eta(ii) \\
\leq &(1-\eta\tau)\Phi_k + \eta^2(D_{\eta,\tau,\Lambda} + D'_{\tau,\Lambda}) \\
&+ \eta(\hat{V}^k_{\hat{z}_k} - V^{\pi_k}_{z_k}) + \eta\sum_i \lambda_{max}\left|\hat{V}^k_{\hat{u}_{k,i}} - V^{\pi_k}_{u_i}\right| && \text{(Def. } \Phi_k) \\
&- \eta\left((i) + (ii)\right) \\
\leq &(1-\eta\tau)\Phi_k + \eta^2(D_{\eta,\tau,\Lambda} + D'_{\tau,\Lambda}) \\
&+ \eta(\hat{V}^k_{\hat{z}_k} - V^{\pi_k}_{z_k}) + \eta\sum_i \lambda_{max}\left|\hat{V}^{\pi_k}_{\hat{u}_{k,i}} - V^{\pi_k}_{u_i}\right| \\
&- \eta\left(\mathcal{L}_\tau(\pi^\star_\tau, \boldsymbol{\lambda}_k) - \mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}^\star_\tau)\right) && \text{(Equation (29))} \\
\leq &(1-\eta\tau)\Phi_k + \eta^2(D_{\eta,\tau,\Lambda} + D'_{\tau,\Lambda}) && \text{(as } \mathcal{L}_\tau(\pi^\star_\tau, \boldsymbol{\lambda}_k) - \mathcal{L}_\tau(\pi_k, \boldsymbol{\lambda}^\star_\tau) \geq 0) \\
&+ \eta(\hat{V}^k_{\hat{z}_k} - V^{\pi_k}_{z_k}) + \eta\sum_i \lambda_{max}\left|\hat{V}^k_{\hat{u}_{k,i}} - V^{\pi_k}_{u_i}\right|.
\end{aligned}$$

By induction and geometric series bound we find

$$\begin{aligned}
\Phi_{k+1} \leq & (1-\eta\tau)^k \Phi_1 + \sum_{k'=1}^{k} (1-\eta\tau)^{k+1-k'} \eta^2 (D_{\eta,\tau,\Lambda} + D'_{\tau,\Lambda}) \\
& + \sum_{k'=1}^{k} (1-\eta\tau)^{k+1-k'} \left(\eta(\hat{V}^{k'}_{z_{k'}} - V^{\pi_{k'}}_{z_{k'}}) + \eta \sum_i \lambda_{max} \left| \hat{V}^{k'}_{\hat{u}_{k',i}} - V^{\pi_{k'}}_{u_i} \right| \right) \\
\leq & (1-\eta\tau)^k \Phi_1 + \frac{1}{\eta\tau} \eta^2 (D_{\eta,\tau,\Lambda} + D'_{\tau,\Lambda}) \\
& + \eta \sum_{k'=1}^{k} \left((\hat{V}^{k'}_{\hat{z}_{k'}} - V^{\pi_{k'}}_{z_k}) + \sum_i \lambda_{max} \left| \hat{V}^{k'}_{\hat{u}_{k',i}} - V^{\pi_{k'}}_{u_i} \right| \right) \\
\leq & (1-\eta\tau)^k \Phi_1 + \frac{\eta}{\tau} (D_{\eta,\tau,\Lambda} + D'_{\tau,\Lambda}) \\
& + \tilde{O}\left(\eta \left((2 + I\lambda_{max}) + I\lambda_{max} \right) \left(\sqrt{S^2 A H^4 k} + S^{3/2} A H^2 \right) \right),
\end{aligned}$$

where we used Lemmas E.5 and E.7 (which apply since G occurs) in the final step. Finally, noting that $D_{\eta,\tau,\Lambda} + D'_{\tau,\Lambda} \leq \tilde{O}(C_{\eta,\tau,\Lambda})$ (see Theorem 4.1) and invoking $\lambda_{max} \geq 1$ concludes the proof. □

Invoking Lemma 4.2, we can leverage the convergence of the potential function to show a sublinear regret bound for Algorithm 3.

Theorem 5.1 (Regret bound)**.** *Let $\tau = K^{-1/7}$, $\eta = (H^2 I)^{-1} \Xi K^{-5/7}$, $\lambda_{max} = H \Xi^{-1} K^{1/14}$. Then with probability at least $1-\delta$, Algorithm 1 obtains a* strong *regret of*

$$\mathcal{R}(K; r) \leq C_r K^{0.93}, \quad \mathcal{R}(K; \boldsymbol{u}) \leq C_u K^{0.93},$$

where $C_r,\ C_u = poly(S, A, H, I, \Xi^{-1}, \log(1/\delta), \log(K))$ and K is the number of episodes.

Proof. The bound follows from Lemma 5.1 and Lemma 4.2.

Set $\Delta_r(k) := \left[V_r^{\pi^\star} - V_r^{\pi_k} \right]_+$ and $\Delta_{g_i}(k) := \left[-V_{g_i}^{\pi_k} \right]_+$. Condition on the success event G, which happens with probability at least $1-\delta$ by Lemma E.1.

We first consider the regret for the reward. Plugging Lemma 5.1 into Lemma 4.2 we find, using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $1 + x \leq \exp(x)$,

$$\begin{aligned}
\Delta_r(k) \lesssim & H^{3/2} \Phi_1^{1/2} \exp\left(-\eta\tau k/2\right) && \text{(a)} \\
& + H^{3/2} \left(\frac{\eta}{\tau} \right)^{1/2} C_{\eta,\tau,\Lambda}^{1/2} && \text{(b)} \\
& + H^{3/2} \left(\eta \lambda_{max} I S A^{1/2} H^2 k^{1/2} \right)^{1/2} && \text{(c)} \\
& + H^{3/2} \left(\eta \lambda_{max} S^{3/2} A H^2 \right)^{1/2} && \text{(d)} \\
& + \tau H \log(A). && \text{(e)}
\end{aligned}$$

We first show that we can ignore terms (b), (d), and (e) since they are $o(K^{-13/14})$. For (b), note that, using the definitions of η, τ, λ_{max} (and taking $\sqrt{\cdot}$, and $\tau < 1$)

$$\begin{aligned}
C_{\eta,\tau,\Lambda}^{1/2} \leq & \lambda_{max} H^{3/2} A^{1/4} I \cdot \exp\left((\eta H \left(1 + \lambda_{max} I + \log(A)\right))/2 \right) + I^{1/2} \left(H + \tau \lambda_{max} \right) \\
\lesssim & \lambda_{max} H^{3/2} A^{1/4} I \cdot \exp\left((HI)^{-1} \Xi K^{-5/7} \left(1 + \frac{H}{\Xi} K^{1/14} I + \log(A) \right) \right) + I^{1/2} \left(H + H \Xi^{-1} K^{-1/14} \right) \\
\lesssim & \lambda_{max} H^{3/2} A^{5/4} I \cdot \exp(2) + I^{1/2} \left(H + H \Xi^{-1} K^{-1/14} \right) \\
\lesssim & H^{5/2} A^{5/4} I \Xi^{-1} K^{1/14}.
\end{aligned}$$

Since

$$\left(\frac{\eta}{\tau}\right)^{1/2} = (H^2 I)^{-1/2}\Xi^{1/2}K^{(-5/7+1/7)/2} = (H^2 I)^{-1/2}\Xi^{1/2}K^{-2/7},$$

we thus have

$$(b) = H^{3/2}\left(\frac{\eta}{\tau}\right)^{1/2} C_{\eta,\tau,\Lambda}^{1/2} \leq \text{poly}(A, H, I, \Xi^{-1})K^{-3/14}.$$

Hence, when summing (b) over $k = 1, \ldots, K$, it only contributes $o(K^{-13/14})$ to the regret. Similarly,

$$\begin{aligned}(d) =& H^{3/2}\left(\eta\lambda_{max}S^{3/2}AH^2\right)^{1/2} \leq H^{3/2}\left(S^{3/2}AH^2\right)^{1/2} K^{-(9/2)/14},\\ (e) =& \tau H\log(A) = K^{-1/7}H\log(A).\end{aligned}$$

Hence, when summing (d) and (e) over $k = 1, \ldots, K$, they only contribute $o(K^{-13/14})$ to the regret. We now turn to terms (a) and (c). For (a), using the standard inequality $e^{-x} \leq 1 - x/2$ (if $0 \leq x \leq 1$) with $x := \eta\tau/2$, we first find

$$\exp(-\eta\tau k/2) \leq (1 - \eta\tau/4)^k$$

and hence, (after summing (a) over $k = 1, \ldots, K$)

$$\begin{aligned}\sum_{k=1}^{K} H^{3/2}\Phi_1^{1/2}\exp\left(-\eta\tau k/2\right) \leq& H^{3/2}\Phi_1^{1/2}\sum_{k=1}^{K}(1-\eta\tau/4)^k\\ \leq& H^{3/2}\Phi_1^{1/2}\sum_{k=1}^{\infty}(1-\eta\tau/4)^k\\ =& H^{3/2}\Phi_1^{1/2}\frac{4}{\eta\tau}\\ =& H^{3/2}\Phi_1^{1/2}\frac{4}{(H^2I)^{-1}\Xi K^{-5/7}K^{-1/7}}\\ =& 4H^{7/2}I\Xi^{-1}\Phi_1^{1/2}K^{12/14}.\end{aligned}$$

Furthermore, since π_1 plays actions uniformly at random and $\boldsymbol{\lambda_1} = \mathbf{0}$, we have $\Phi_1^{1/2} \leq (H\log(A) + \frac{1}{2}I\lambda_{max}^2)^{1/2} \leq H^{1/2}\log(A)^{1/2} + I^{1/2}\lambda_{max} = H^{1/2}\log(A)^{1/2} + I^{1/2}H\Xi^{-1}K^{1/14}$. Hence, the calculation above shows

$$\begin{aligned}(a) \leq& 4H^{7/2}I\Xi^{-1}\left(H^{1/2}\log(A)^{1/2} + I^{1/2}H\Xi^{-1}K^{1/14}\right)K^{12/14}\\ \leq& 4H^4I\Xi^{-1}\log(A)^{1/2}K^{12/14} + 4H^{9/2}I\Xi^{-2}I^{1/2}K^{13/14}.\end{aligned}$$

For (c), we find (after summing over $k = 1, \ldots, K$)

$$\begin{aligned}\sum_{k=1}^{K} H^{3/2}\left(\eta\lambda_{max}ISA^{1/2}H^2k^{1/2}\right)^{1/2} =& \sum_{k=1}^{K} H^{3/2}\left((H^2I)^{-1}\Xi K^{-5/7}H\Xi^{-1}K^{1/14}ISA^{1/2}H^2k^{1/2}\right)^{1/2}\\ \leq& K\cdot H^{3/2}\left((H^2I)^{-1}\Xi K^{-5/7}H\Xi^{-1}K^{1/14}ISA^{1/2}H^2K^{1/2}\right)^{1/2}\\ =& H^{3/2}\left((H^2I)^{-1}\Xi H\Xi^{-1}ISA^{1/2}H^2\right)^{1/2}K^{1+(-5/7+1/14+1/2)/2}\\ =& H^{3/2}\left(SA^{1/2}H\right)^{1/2}K^{13/14}\\ =& H^2S^{1/2}A^{1/4}K^{13/14}.\end{aligned}$$

Hence, summing up terms (a) to (e) over $k = 1, \ldots, K$ indeed yields the bound for the objective (using $13/14 \simeq 0.93$):

$$\mathcal{R}(K; r) \lesssim \left(H^{9/2}I\Xi^{-2}I^{1/2} + H^2S^{1/2}A^{1/4}\right)K^{0.93}.$$

Next, we consider the regret for the constraints. Plugging Lemma 5.1 into Lemma 4.2 we find, using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$,

$$\begin{aligned}
\Delta_{u_i}(k) \lesssim & H^{3/2}\Phi_1^{1/2} \exp\left(-\eta\tau k/2\right) && \text{(a')}\\
& + H^{3/2}\left(\frac{\eta}{\tau}\right)^{1/2} C_{\eta,\tau,\Lambda}^{1/2} && \text{(b')}\\
& + H^{3/2}\left(\eta\lambda_{max} I S A^{1/2} H^2 k^{1/2}\right)^{1/2} && \text{(c')}\\
& + H^{3/2}\left(\eta\lambda_{max} S^{3/2} A H^2\right)^{1/2} && \text{(d')}\\
& + \tau\lambda_{max} + \frac{1}{\lambda_{max}} H^2 \Xi^{-1} && \text{(e')}\\
& + \frac{1}{\lambda_{max}} \tau H \log(A). && \text{(f')}
\end{aligned}$$

Note that terms (a'), (b'), (c'), (d') are identical to (a), (b), (c), (d). Summed up, they thus correspond to the same regret as the one for the reward. Moreover, for (f') we have

$$(f') = \frac{1}{\lambda_{max}} \tau \log(A) = \Xi H^{-1} \log(A) K^{-3/14},$$

and thus, we can ignore (f') when summing from $k = 1, \dots, K$. Finally, for (e'), we have

$$\begin{aligned}
(e') =& \tau\lambda_{max} + \frac{1}{\lambda_{max}} H^2 \Xi^{-1}\\
=& K^{-1/7} \cdot H\Xi^{-1} K^{1/14} + H^{-1}\Xi K^{-1/14} \cdot H^2 \Xi^{-1}\\
=& H(1 + \Xi^{-1}) K^{-1/14}.
\end{aligned}$$

Thus, summing up (a') to (f') from $k = 1$ to K yields the regret for the constraints (using $13/14 \simeq 0.93$):

$$\mathcal{R}(K; \boldsymbol{u}) \lesssim \left(H^{9/2} I \Xi^{-2} I^{1/2} + H^2 S^{1/2} A^{1/4} + H(1 + \Xi^{-1})\right) K^{0.93}.$$

□

G. Examples and Simulation

In this section, we expand on the description provided in Section 6.

G.1. Examples

In this section, we provide examples to highlight crucial differences between the strong regret $\mathcal{R}$ and the weak regret $\mathcal{R}_{\text{weak}}$.

A Minimal Example Consider the case in which an agent repeatedly has the option between three different investment strategies (o_1, o_2, o_3). Each of them yields a respective reward $(r_1, r_2, r_3) = (1/5, 1, 3/5)$. There is an initial cost $(u_1, u_2, u_3) = (9/10, 1/10, 5/10)$ associated with each option, which must not exceed the budget $c = 1/2$ of the agent. If the budget is exceeded, the agent will be in debt, and a larger debt is associated with a higher risk. Clearly, we can model this scenario as an optimization problem $\max_{x \in \Delta([3])} \sum_{i \in [3]} x_i r_i$, subject to $\sum_{i \in [3]} x_i r_i \leq c$, where x describes the distribution of investments in the respective strategies. We can equivalently model this problem as a CMDP with one state s_1, horizon $H = 1$ and three actions $\mathcal{A} = [3]$ (i.e., a constrained bandit), in which $r_1(s_1, a) = r_a$, $\boldsymbol{u}_1(s_1, a) = 1 - u_a$ ($a \in [3]$) and $\boldsymbol{c} = 1 - c$. Here, o_1 is highly profitable but too risky, o_2 is less profitable but safe, and o_3 is a compromise between both. Note that $\mathcal{R}(K, \boldsymbol{u})$ now measures the total amount of debt the algorithm accumulated during K episodes. A strategy A that always plays o_3 will have a total debt of 0 since o_3 does not violate the constraint. On the other hand, a strategy B that plays o_1 and o_2 in an alternating fashion will have a total debt of $K/5$, which is linear in K, despite having a weak regret of 0 due to the aforementioned cancellations. Both strategies, A and B, have the same accumulated objective. This simple example, which does not even require an unknown environment, illustrates why weak regret cannot be the right notion of safety during learning. We point to Calvo-Fullana et al. (2023); Moskovitz et al. (2023); Stooke et al. (2020) for other examples exhibiting similar behavior.

A Slight Relaxation For practical purposes, one may consider the strong regret only for the constraint violations and the weak one for the objective. The reasoning behind this possible relaxation is that we may tolerate superoptimal performance with respect to the reward (when the algorithm violates the constraints) while we still do not tolerate additive negative constraint violations (as discussed). For example, an agent maximizing a wealth function may be allowed to obtain a higher return than any safe method, adding a negative term to the objective regret. In this case, we may want to allow compensating suboptimal returns by superoptimal ones, while we do not want to allow compensating unsafe strategies by strictly safe ones. However, this relaxation does not improve our theoretical results in their current form. It remains open whether, under this relaxation, stronger results are possible to obtain.

We argue that despite the possible relaxation, it is sensible to require the strong regret for both the violation and the objective, as done in this paper, which is also what Efroni et al. (2020) referred to. Indeed, one can think of settings in which one gets paid out the return of an episode only up until the limit that would be attainable subject to the constraints. For instance, if there is an illegal set of options, the controller of the environment may decide to pay out only so much as attainable when they are not being used (but may not know whether the illegal actions have, in fact, been used). In other words, the return is $W_k = \min\{V_r^{\pi_k}, V_r^{\pi^\star}\}$. Hence, per-episode regret would be

$$V_r^{\pi^\star} - W_k = V_r^{\pi^\star} - \min\{V_r^{\pi_k}, V_r^{\pi^\star}\} = [V_r^{\pi^\star} - V_r^{\pi_k}]_+,$$

which is exactly the strong regret of the episode.

G.2. Experiment Details

In this section, we report all details of the parameters, environments, and hardware used for our simulations.

Hyperparameters For the vanilla algorithms, we run for $K = 4000$ episodes for each step size $\eta \in \{0.05, 0.075, 0.1, 0.125, 0.15, 0.2\}$, which we observed to be a reasonable range across CMDPs when fixing the number of episodes. Similarly, for the regularized algorithms, we perform the same parameter search across all pairs of step size $\eta \in \{0.05, 0.1, 0.2\}$ and regularization parameter $\tau \in \{0.01, 0.02\}$, totaling a number of six hyperparameter configurations as well. We always set $\lambda_{max} = 6$, which did not play a role in our simulations as long as it was chosen sufficiently large. We use exploration bonuses $0.08 \cdot n_h(s,a)^{-1/2}$. For each hyperparameter configuration, we sample $n = 5$ independent runs with K episodes, obtain the regret curves and plot their average. For each algorithm, we then report the result for the best hyperparameter configuration in hindsight (with respect to the strong regrets).

Environment As described, we sample the rewards r uniformly at random and the constraints as $c = (1 - r) + \beta\zeta$, for a Gaussian vector $\eta \in \mathbb{R}^{HSA}$ and $\beta = 0.1$. We sample an environment with $S = A = H = 5$ according to the procedure above. Throughout seeds (and other CMDP sizes), this led to CMDPs in which the oscillations of the iterates and error cancellations can be observed. As argued, sampling constraints and rewards fully independently does not provide CMDPs that are interesting test beds. Indeed, unlike in random CMDPs, in real-life situations, we often observe goals that are explicitly conflicting with safety constraints, as otherwise, there is no need to encode them via a CMDP. For instance, consider a vehicle that aims to arrive fast but not go over the speed limit or cross the sidewalk. The latter would be the fastest option, but it conflicts with the constraint. In other words, the constraint and the reward function are negatively correlated.

All simulations were performed on a MacBook Pro 2.8 GHz Quad-Core Intel Core i7. We provide the code in the supplementary material. For all experiments, we set the seed to 123.

G.3. Dual Approach

In this section, we provide the pseudo-code for the regularized dual algorithm. One approach to solving the regularized min-max CMDP problem is to perform dual descent on the regularized Lagrangian. Evaluating the dual function then corresponds to entropy-regularized dynamic programming in an MDP, and the descent step on the dual function corresponds to a mirror descent step of the Lagrange multipliers on the regularized Lagrangian. Replacing the required value functions in this scheme with the optimistically estimated ones yields Algorithm 3. Note that here, we consider the estimated value functions without truncation. Set $\hat{\mathcal{H}}_k(\pi) := V^{\pi}_{\psi_{\pi,k}}$, where $\psi_{\pi,k,h}(s,a) := -\log(\pi_h(a|s)) + b^p_{k-1,h}(s,a)\log(A)$.

Algorithm 3 Regularized Dual Algorithm with Optimistic Exploration

Require: $\Lambda = [0, \lambda_{max}]^I$, stepsize $\eta > 0$, regularization parameter $\tau > 0$, number of episodes K, initial policy $\pi_{1,h}(a|s) = 1/A$ $(\forall s, a, h)$, $\boldsymbol{\lambda}_1 := \mathbf{0} \in \mathbb{R}^I$

for $k = 1, \dots, K$ **do**

Update primal variable via regularized dynamic programming

$$\pi_k = \arg\min_{\pi \in \Pi} \left(V_{\hat{r}_k}^{\hat{p}_k, \pi} + \boldsymbol{\lambda}_k^T \boldsymbol{V}_{\hat{\boldsymbol{u}}_k}^{\hat{p}_k, \pi} + \tau \hat{\mathcal{H}}_k(\pi) \right)$$

Evaluate $\boldsymbol{V}_{\hat{\boldsymbol{u}}_k}^{\hat{p}_k, \pi_k}$

Update dual variables:

$$\boldsymbol{\lambda}_{k+1} = \text{proj}_\Lambda \left(\boldsymbol{\lambda}_k - \eta (\boldsymbol{V}_{\hat{\boldsymbol{u}}_k}^{\hat{p}_k, \pi_k} - \boldsymbol{c} + \tau \boldsymbol{\lambda}_k) \right).$$

Play π_k for one episode, update model estimates: $\hat{r}_{k+1}$, $\hat{\boldsymbol{u}}_{k+1}$, $\hat{p}_{k+1}$, $\hat{\psi}_{k+1}$

end for

H. Difference Lemmas

In the following, we recap the well-known performance difference and value difference lemma.

Lemma H.1 (Performance difference lemma)**.** *For all $\pi, \pi' \in \Pi$ and all $r' : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we have*

$$\begin{aligned} V_{r'}^{\pi} - V_{r'}^{\pi'} =& \mathbb{E}_\pi \left[\sum_{h=1}^{H} \sum_a Q_{r',h}^{\pi'}(s_h, a) \left(\pi_h(a|s_h) - \pi'_h(a|s_h) \right) \right] \\ =& \sum_{h=1}^{H} \sum_s d_h^\pi(s) \sum_a Q_{r',h}^{\pi'}(s, a) \left(\pi_h(a|s) - \pi'_h(a|s) \right) \\ =& \sum_{h=1}^{H} \sum_s d_h^\pi(s) \langle Q_{r',h}^{\pi'}(s, \cdot), \pi_h(\cdot|s) - \pi'_h(\cdot|s) \rangle. \end{aligned}$$

Proof. See Cai et al. (2020, Lemma 3.2) for the first equality. The second equality follows since we consider unnormalized occupancy measures. The third equality holds by definition of the inner product. □

From Shani et al. (2020, Lemma 1):

Lemma H.2 (Extended value difference lemma (aka simulation lemma))**.** *Let π, π' be policies and $M = (\mathcal{S}, \mathcal{A}, p, r)$, $M' = (\mathcal{S}, \mathcal{A}, p', r')$ be MDPs. Let $\hat{Q}_h^M(s, a)$ be an approximation of the value function $Q_{r,h}^{p,\pi}(s, a)$. Let $\hat{V}_h^M(s) := \left\langle \hat{Q}_h^M(s, \cdot), \pi_h(\cdot|s) \right\rangle$. Then*

$$\begin{aligned} & \hat{V}^M(s_1, 1) - V_{r',1}^{p',\pi'}(s_1) \\ =& \sum_{h=1}^{H} \mathbb{E} \left[\left\langle \hat{Q}_h^M(s_h, \cdot), \pi_h(\cdot|s_h) - \pi'_h(\cdot|s_h) \right\rangle \;\middle|\; s_1; p', \pi' \right] \\ &+ \sum_{h=1}^{H} \mathbb{E} \left[\hat{Q}_h^M(s_h, a_h) - r'_h(s, a) - \left\langle p'_h(\cdot|s_h, a_h), \hat{V}_{h+1}^M(\cdot) \right\rangle \;\middle|\; s_1; p', \pi' \right], \end{aligned}$$

where $V_{r',1}^{p',\pi'}(s_1)$ is the value function of π' in M'.

Note that $\hat{Q}$ need not correspond to a true value function under some model.

I. Convex Optimization Background

In this section, we review fundamental results from the optimization literature. All of these results are standard, and we include them for completeness. They are not novel by themselves nor specific to the sections in which we make use of them.

I.1. Convex Min-Max Optimization

While the following results from min-max optimization are commonly used, we establish them here for our setup (both for completeness and due to the lack of a unifying resource for our case).

Setup Let $\mathcal{X} \subset \mathbb{R}^{d_x}$, $\mathcal{Y} \subset \mathbb{R}^{d_y}$ be (nonempty) compact convex sets and let $f\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous and convex-concave function. Set $\bar{f}\colon \mathcal{X} \to \mathbb{R}$, $\bar{f}(x) := \max_{y\in\mathcal{Y}} f(x,y)$, and $\underline{f}\colon \mathcal{Y} \to \mathbb{R}$, $\underline{f}(y) := \min_{x\in\mathcal{X}} f(x,y)$, which both exist by continuity of f on a compact domain.

Lemma I.1 (Existence minimax points)**.** *We have*

$$\inf_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y) = \sup_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x,y), \tag{32}$$

and the maximum and minimum are attained, i.e., there exist $x^\star \in \mathcal{X}$, $y^\star \in \mathcal{Y}$ such that

$$\bar{f}(x^\star) = \inf_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y), \tag{33}$$

$$\underline{f}(y^\star) = \sup_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x,y). \tag{34}$$

Proof. The first equality holds due to Sion's Minimax Theorem (Sion, 1958). Note that the second part of the lemma is not immediate from Sion-like statements.

We shall prove that $\bar{f}$ is continuous. By compactness of $\mathcal{X}$, this implies the existence of $x^\star$. By symmetry, this also settles the existence of $y^\star$ (by repeating the argument for $-f$). Thus, let $x \in \mathcal{X}$ and consider a sequence $(x_k)_k$ in $\mathcal{X}$ such that $x_k \to x$. We aim to show that $\bar{f}(x_k) \to \bar{f}(x)$, which would conclude the proof.

Let $y \in \arg\max_{y'\in\mathcal{Y}} f(x,y)$, which exists by continuity. Then, for every k we have

$$\bar{f}(x_k) = \max_{y'\in\mathcal{Y}} f(x_k, y') \geq f(x_k, y) \to f(x,y).$$

Taking $\liminf_k$ on both sides yields

$$\liminf_k \bar{f}(x_k) \geq f(x,y) = \bar{f}(x). \tag{35}$$

Assume by contradiction that $\limsup_k \bar{f}(x_k) > \bar{f}(x)$. Then we can pick $\delta > 0$ such that $\limsup_k \bar{f}(x_k) \geq \bar{f}(x) + \delta$. Thus we can pick a subsequence $x_{n(k)}$ and $y_{n(k)} \in \arg\max_{y'\in\mathcal{Y}} f(x_{n(k)}, y')$ such that for all k,

$$\bar{f}(x_{n(k)}) \geq \bar{f}(x) + \delta/2 \geq f(x, y_{n(k)}) + \delta/2. \tag{36}$$

Since $\mathcal{Y}$ is compact, by further picking a subsequence if needed, we can WLOG assume that there exists $\tilde{y} \in \mathcal{Y}$ such that $y_{n(k)} \to \tilde{y}$. Then by Equation (36),

$$f(x_{n(k)}, y_{n(k)}) = \bar{f}(x_{n(k)}) \geq f(x, y_{n(k)}) + \delta/2.$$

Taking $k \to \infty$ and using continuity of f yields the contradiction $f(x,\tilde{y}) \geq f(x,\tilde{y}) + \delta/2$. We therefore must have $\limsup_k \bar{f}(x_k) \leq \bar{f}(x) \leq \liminf_k \bar{f}(x_k)$, proving $\bar{f}(x_k) \to \bar{f}(x)$. Thus, $\bar{f}$ is indeed continuous. □

General Setup The statements in this paragraph all concern the following more general setup (dropping convex-concavity and boundedness of the domain). As we showed in the previous paragraph, all assertions made here hold in the continuous, convex-concave compactly constrained setup.

Let $\mathcal{X} \subset \mathbb{R}^{d_x}$, $\mathcal{Y} \subset \mathbb{R}^{d_y}$ be (nonempty) closed sets and let $f\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous function. Consider $\bar{f}\colon \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$, $\bar{f}(x) := \max_{y\in\mathcal{Y}} f(x,y)$ and $\underline{f}\colon \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$, $\underline{f}(y) := \min_{x\in\mathcal{X}} f(x,y)$.

Lemma I.2 (Min-max to saddle point)**.** *Let*

$$x^\star \in \arg\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \; f(x,y),$$
$$y^\star \in \arg\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} \; f(x,y),$$

and assume $\bar{f}(x^\star) = \underline{f}(y^\star)$*. Then* $(x^\star, y^\star)$ *is a* saddle point*, i.e., for all* $x \in \mathcal{X}$*,* $y \in \mathcal{Y}$*, we have*

$$f(x^\star, y) \le f(x^\star, y^\star) \le f(x, y^\star).$$

Proof. We have

$$f(x^\star, y) \le \max_{y'\in\mathcal{Y}} f(x^\star, y') = \bar{f}(x^\star) = \underline{f}(y^\star) = \min_{x'\in\mathcal{X}} f(x', y^\star) \le f(x^\star, y^\star),$$

proving the first inequality, and for the second, we have

$$f(x, y^\star) \ge \min_{x'\in\mathcal{X}} f(x', y^\star) = \underline{f}(y^\star) = \bar{f}(x^\star) = \max_{y'\in\mathcal{Y}} f(x^\star, y') \ge f(x^\star, y^\star).$$

Note that this proof does not require convexity or compactness. □

Lemma I.3 (Saddle point to min-max)**.** *Let* $(x^\star, y^\star)$ *be a saddle point, i.e., for all* $x \in \mathcal{X}$*,* $y \in \mathcal{Y}$*, we have*

$$f(x^\star, y) \le f(x^\star, y^\star) \le f(x, y^\star).$$

Then

$$x^\star \in \arg\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \; f(x,y),$$
$$y^\star \in \arg\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} \; f(x,y).$$

Proof. We first note that the assertion implies $\max_{y'} f(x^\star, y') \le \min_{x'\in\mathcal{X}} f(x', y^\star)$. Hence

$$\bar{f}(x^\star) = \max_{y'} f(x^\star, y') \le \min_{x'\in\mathcal{X}} f(x', y^\star) \le \min_{x'\in\mathcal{X}} \max_{y'\in\mathcal{Y}} f(x', y'),$$

showing the claim for $x^\star$. For $y^\star$, we note that similarly,

$$\underline{f}(y^\star) = \min_{x'\in\mathcal{X}} f(x', y^\star) \ge \max_{y'\in\mathcal{Y}} f(x^\star, y') \ge \max_{y'\in\mathcal{Y}} \min_{x'\in\mathcal{X}} f(x', y'),$$

concluding the proof.

Note that this proof does not require convexity or compactness. □

Lemma I.4 (Invariance of saddle points)**.** *Let*

$$x^\star \in \arg\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \; f(x,y),$$
$$y^\star \in \arg\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} \; f(x,y),$$

and assume $\bar{f}(x^\star) = \underline{f}(y^\star)$*. Consider closed sets* $\mathcal{X}' \subset \mathcal{X}$*,* $\mathcal{Y}' \subset \mathcal{Y}$*. If* $(x^\star, y^\star) \in \mathcal{X}' \times \mathcal{Y}'$*, then*

$$x^\star \in \arg\min_{x\in\mathcal{X}'} \max_{y\in\mathcal{Y}'} \; f(x,y),$$
$$y^\star \in \arg\max_{y\in\mathcal{Y}'} \min_{x\in\mathcal{X}'} \; f(x,y).$$

Proof. By Lemma I.2 (which applies since $\bar{f}(x^\star) = \underline{f}(y^\star)$), $(x^\star, y^\star)$ is a saddle point for the minmax problem with domain $\mathcal{X} \times \mathcal{Y}$. Thus, since $y^\star \in \mathcal{Y}' \subset \mathcal{Y}$, we have $f(x^\star, y^\star) = \max_{y \in \mathcal{Y}} f(x^\star, y) = \max_{y \in \mathcal{Y}'} f(x^\star, y)$. Moreover, since $\mathcal{X}' \subset \mathcal{X}$ and $y^\star \in \mathcal{Y}'$, we have $f(x^\star, y^\star) = \min_{x \in \mathcal{X}} f(x, y^\star) \leq \min_{x \in \mathcal{X}'} f(x, y^\star) \leq \min_{x \in \mathcal{X}'} \max_{y \in \mathcal{Y}'} f(x, y)$. Hence $\max_{y \in \mathcal{Y}'} f(x^\star, y) \leq \min_{x \in \mathcal{X}'} \max_{y \in \mathcal{Y}'} f(x, y)$, proving

$$x^\star \in \arg\min_{x \in \mathcal{X}'} \max_{y \in \mathcal{Y}'} f(x, y).$$

The proof for $y^\star$ follows by repeating the argument for $-f$.

Note that this proof does not require convexity or compactness. □

I.2. Constrained Convex Optimization

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

Consider the (primal) optimization problem

$$\begin{aligned} f^* := \min \quad & f(x) \\ \text{s.t.} \quad & g(x) \leq 0 \\ & x \in X \end{aligned} \tag{37}$$

with the following assumptions.

Assumption I.1 (Assumption 8.41, Beck (2017))**.** *In Equation (37),*

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

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

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

(d) Equation (37) 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 $\lambda \in \mathbb{R}^m_{\geq 0}$, $\min_{x \in X}(f(x) + \lambda^T g(x))$ *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 Equation (37). The *dual problem* is then defined as

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

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

Theorem I.1 (Theorem A.1, Beck (2017))**.** *Under Assumption I.1, 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 I.2. *Suppose Assumption I.1 holds. Let $x^* \in X^*$, $\lambda^* \in \Lambda^*$ and $x \in X$. Then*

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

Proof. We have

$$\begin{aligned} f(x) =& f(x) + (\lambda^*)^T g(x) - (\lambda^*)^T g(x) \\ \geq& q(\lambda^*) - (\lambda^*)^T g(x) && \text{(definition of } q(\cdot)) \\ =& f(x^*) - (\lambda^*)^T g(x) && \text{(since by Theorem I.1, } q^* = f^*) \end{aligned}$$

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 I.3. *Under Assumption I.1, 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

$$\begin{aligned} f(x^*) =& q(\lambda^*) && \text{(Theorem I.1)} \\ \leq& f(\bar{x}) + (\lambda^*)^T g(\bar{x}) && \text{(definition of } q(\cdot)) \\ \leq& f(\bar{x}) + \|\lambda^*\|_1 \max_{i\in[m]} g_i(\bar{x}) && \text{(since } \lambda^* \geq 0) \\ =& f(\bar{x}) - \|\lambda^*\|_1 \min_{i\in[m]}(-g_i(\bar{x})) \end{aligned}$$

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

I.3. Online Mirror Descent

Setup In the following, we consider a convex set $X \subset \mathbb{R}^d$ and a non-empty closed convex set $V \subset X$. Let $\psi\colon X \to \mathbb{R}$ be proper, closed, and strictly convex on V. Let $B_\psi\colon X \times \operatorname{int}(X) \to \mathbb{R}$ be the *Bregman divergence* associated with ψ. Define $\|x\|_A := \sqrt{x^T A x}$. Assume that

$$\lim_{\lambda\to 0} \left(\nabla\psi(x + \lambda(y - x))\right)^T (y - x) = -\infty \quad (\forall x \in \operatorname{bdry}(X),\ y \in \operatorname{int}(X)), \text{ or} \tag{38}$$

$$V \subset \operatorname{int}(X). \tag{39}$$

Consider the following descent lemma using local norms.

Lemma I.5 (MD descent lemma, Orabona (2019) Lemma 6.16)**.** *Suppose ψ is twice differentiable, with positive definite Hessian in the interior of its domain. Assume*

$$\tilde{x} \in \arg\min_{\bar{x}\in X} g^T \bar{x} + \frac{1}{\eta} B_\psi(\bar{x}, x), \tag{40}$$

$$x' \in \arg\min_{\bar{x}\in V} g^T \bar{x} + \frac{1}{\eta} B_\psi(\bar{x}, x) \tag{41}$$

exist. Then, for all $x^\star \in V$, there exist z on the line segment[7] between x and x', and z' on the line segment between x and $\tilde{x}$ such that

$$\eta g^T(x - x^\star) \leq B_\psi(x^\star, x) - B_\psi(x^\star, x') + \frac{\eta^2}{2} \min\left\{\|g\|^2_{(\nabla^2\psi(z))^{-1}}, \|g\|^2_{(\nabla^2\psi(z))^{-1}}\right\}.$$

[7]The line segment between two vectors is the convex hull of the set containing those two vectors.

We get the following descent lemma for exponentiated Q-ascent.

Lemma I.6. *Let $V := \Delta([d])$, and $g \in \mathbb{R}^d_{\geq 0} =: X$. Then $\tilde{x} := \arg\max_{\bar{x}\in X} g^T\bar{x} - \frac{1}{\eta}KL(\bar{x}, x)$ and $\arg\max_{\bar{x}\in V} g^T\bar{x} - \frac{1}{\eta}KL(\bar{x}, x)$ exist and are unique. Moreover, if g only has non-negative entries, then for all $x^\star \in V$ we have*

$$g^T(x^\star - x) \leq \frac{KL(x^\star, x) - KL(x^\star, x')}{\eta} + \frac{\eta}{2}\sum_{i=1}^{d} \tilde{x}_i g_i^2.$$

Proof. Note that the negative entropy $\psi(x) = \sum_i x_i \log(x_i)$ is strictly convex and twice differentiable and satisfies Equation (38), as a short calculation reveals. Moreover, for $p, q \in V$, we have $B_\psi(p, q) = \mathrm{KL}(p, q)$ (Orabona, 2019, Example 6.4). Existence and uniqueness are discussed in Orabona (2019).

Maximizing $g^T\bar{x} - \frac{1}{\eta}\mathrm{KL}(\bar{x}, x)$ is equivalent to minimizing $(-g)^T\bar{x} + \frac{1}{\eta}\mathrm{KL}(\bar{x}, x)$, allowing us to apply Lemma I.5. Note that for $z \in \mathbb{R}^I_{>0}$, $\nabla^2\psi(z) = \mathrm{diag}(1/z_1, \ldots, 1/z_d)$. Thus, $\|-g\|_{(\nabla^2\psi(z))^{-1}} = \sum_i z_i g_i^2$. Moreover, we have $\tilde{x}_i = x\exp(-\eta(-g_i)) \geq x_i$ (Orabona, 2019), since $g_i \geq 0$, and thus by Lemma I.5, we can pick $z' \leq \tilde{x}$ (componentwise) such that

$$\begin{aligned} g^T(x^\star - x) =& (-g)^T(x - x^\star) \\ \leq& \frac{\mathrm{KL}(x^\star, x) - \mathrm{KL}(x^\star, x')}{\eta} + \frac{\eta}{2}\sum_{i=1}^{d} z_i' g_i^2 \\ \leq& \frac{\mathrm{KL}(x^\star, x) - \mathrm{KL}(x^\star, x')}{\eta} + \frac{\eta}{2}\sum_{i=1}^{d} \tilde{x}_i g_i^2. \end{aligned}$$

□

Lemma I.7 (MD descent lemma, cf. Orabona (2019) Lemma 6.9)**.** *Let $x \in V$, $g \in \mathbb{R}^d$, and $\eta > 0$. Assume further that ψ is μ-strongly convex w.r.t. some norm $\|\cdot\|_{\mathbb{R}^d}$ in V. Then,*

$$x' = \arg\min_{\bar{x}\in V} g^T\bar{x} + \frac{1}{\eta}B_\psi(\bar{x}, x) \tag{42}$$

exists and is unique. Moreover, for all $x^\star \in V$, the following inequality holds:

$$\eta g^T(x - x^\star) \leq B_\psi(x^\star, x) - B_\psi(x^\star, x') + \frac{\eta^2}{2\mu}\|g\|_*^2,$$

where $\|\cdot\|_*$ is the dual norm *associated with $\|\cdot\|_{\mathbb{R}^d}$.*

We can deduce the descent lemma for projected gradient descent.

Lemma I.8 (Descent lemma PGD)**.** *Let $x \in V$, $g \in \mathbb{R}^d$, and $\eta > 0$. Then $x' := \arg\min_{\bar{x}\in V} \bar{x}^T g + \frac{1}{2\eta}\|\bar{x} - x\|^2$ exists and is unique. Moreover, for all $x^\star \in V$ we have*

$$g^T(x - x^\star) \leq \frac{\|x^\star - x\|^2 - \|x^\star - x'\|^2}{\eta} + \frac{\eta}{2}\|g\|^2.$$

Proof. Note that $\psi(x) = \frac{1}{2}\|x\|^2$ is 1-strongly convex w.r.t. the L2 norm. Moreover, for $a, b \in V$, we have $B_\psi(a, b) = \frac{1}{2}\|a - b\|^2$ (Orabona, 2019, Example 6.4). Since the L2 norm is its own dual norm, applying Lemma I.7 to the minimization of $g^T\bar{x} + \frac{1}{2\eta}\|\bar{x} - x\|^2$ yields the claim. □

Finally, the following lemma shows that the updates of the algorithms indeed fall into the category of (online) mirror descent.

Lemma I.9 ((Orabona, 2019)). *Consider a compact set $Y \subset \mathbb{R}^I$ with $y \in Y$, and let $x \in \Delta([d])$ for some $d \in \mathbb{Z}_{\geq 1}$. Then, the closed-form expressions*

$$x'_i = x_i = \frac{x_i \exp(\eta g_i)}{\sum_{i' \in [d]} x_{i'} \exp(\eta x_{i'})} \quad (i \in [d]),$$
$$y' = \mathit{proj}_Y (y - \eta g),$$

are the unique solutions to

$$\max_{\bar{x} \in \Delta([d])} \bar{x}^T g - \frac{1}{\eta} \mathit{KL}(\bar{x}, x),$$
$$\min_{\bar{y} \in Y} \bar{y}^T g + \frac{1}{2\eta} \|\bar{y} - y\|^2,$$

respectively.

Proof. For the primal variable, the derivation of exponentiated gradient is standard, see, e.g., Orabona (2019, Section 6.6).

For the dual variable, the derivation of projected gradient descent simply follows from the first-order optimality criterion and convexity of the objective. □