Provably Efficient Primal-Dual Reinforcement Learning for CMDPs with Non-stationary Objectives and Constraints

Yuhao Ding, Javad Lavaei

UC Berkeley, Department of Industrial Engineering and Operations Research
{yuhao_ding, lavaei}@berkeley.edu

Abstract

We consider primal-dual-based reinforcement learning (RL) in episodic constrained Markov decision processes (CMDPs) with non-stationary objectives and constraints, which plays a central role in ensuring the safety of RL in time-varying environments. In this problem, the reward/utility functions and the state transition functions are both allowed to vary arbitrarily over time as long as their cumulative variations do not exceed certain known variation budgets. Designing safe RL algorithms in time-varying environments is particularly challenging because of the need to integrate the constraint violation reduction, safe exploration, and adaptation to the non-stationarity. To this end, we identify two alternative conditions on the time-varying constraints under which we can guarantee the safety in the long run. We also propose the Periodically Restarted Optimistic Primal-Dual Proximal Policy Optimization (PROPD-PPO) algorithm that can coordinate with both two conditions. Furthermore, a dynamic regret bound and a constraint violation bound are established for the proposed algorithm in both the linear kernel CMDP function approximation setting and the tabular CMDP setting under two alternative conditions. This paper provides the first provably efficient algorithm for non-stationary CMDPs with safe exploration.

Introduction

Safe reinforcement learning (RL) studies how an agent learns to maximize its expected total reward by interacting with an unknown environment over time while dealing with restrictions/constraints arising from real-world problems (Amodei et al. 2016; Dulac-Arnold, Mankowitz, and Hester 2019; Garcıa and Fernández 2015). A standard approach for modeling the safe RL is based on Constrained Markov Decision Processes (CMDPs) (Altman 1999), where one seeks to maximize the expected total reward under a safety-related constraint on the expected total utility.

While classical safe RL and CMDPs assume that an agent interacts with a time-invariant (stationary) environment, both the reward/utility functions and transition kernels can be time-varying for many real-world safety-critical applications. For example, in autonomous driving (Sallab et al. 2017) or power grid control (Ding, Lavaei, and Arcak 2021), it is essential to guarantee safety, such as collision-avoidance and contingency, while handling time-varying conditions related to traffic and load demands. Similarly, in most safety-critical human-computer interaction applications, e.g., automated medical care, human behavior changes over time. In such scenarios, if the automated system is not adapted to take such changes into account, then the system could quickly violate the safety constraint and incur a severe loss (Chandak et al. 2020; Moore et al. 2014). Despite the importance of non-stationary safe RL problems, the literature lacks provably efficient algorithms and theoretical results.

In this work, we formulate a general non-stationary safe exploration problem as an episodic CMDP in which the transition model is unknown and non-stationary, the reward/utility feedback after each episode is bandit and non-stationary, and the variation budget is known. The goal is to design an algorithm that can perform a non-stationary safe exploration, that is, to adaptively explore the unknown and time-varying environment and learn to satisfy time-varying constraints in the long run.

The safe exploration in non-stationary CMDPs is more challenging since the utilities and dynamics are time-varying and unknown a priori. Thus, it is difficult/impossible to guarantee a small/zero constraint violation without knowing how CMDPs will change. Previous constraint violation analyses (Ding et al. 2021; Liu et al. 2021) strongly rely on the conditions of having the same transition dynamics and rewards over all episodes, which are not applicable to non-stationary CMDPs. In view of the aforementioned challenges, we propose a new primal-dual method and develop novel techniques to decouple the optimality gap and the constraint violation. Our main contributions are summarized below:

- We identify two alternative conditions on the time-varying constraints under which we can guarantee the safety in the long run. The first assumption requires the knowledge of the local variation budgets of the constraint for each epoch, while the second assumption needs the strict feasibility of the constraint at each episode and the knowledge of a uniform strict feasibility threshold.
- We develop a new periodically restarted policy-based primal-dual method, which can coordinate with both two conditions, for general non-stationary CMDP problems.
- We study the proposed algorithm under two alternative

conditions that require different amounts of knowledge on the constraints. Our results are summarized in Table 1 and our method is the first provably efficient algorithm for non-stationary CMDPs with safe exploration.

Related Work

Non-stationary RL. Non-stationary RL has been mostly studied in the unconstrained setting (Jaksch, Ortner, and Auer 2010; Auer, Gajane, and Ortner 2019; Ortner, Gajane, and Auer 2020; Domingues et al. 2021; Mao et al. 2020; Zhou et al. 2020; Touati and Vincent 2020; Fei et al. 2020; Zhong, Yang, and Szepesvári 2021; Cheung, Simchi-Levi, and Zhu 2020; Wei and Luo 2021). Our work is related to policy-based methods for non-stationary RL since the optimal solution of CMDP is usually a stochastic policy (Altman 1999) and thus a policy-based method is preferred. When the variation budget is known a prior, (Fei et al. 2020) propose the first policy-based method for non-stationary RL, but they assume stationary transitions and adversarial full-information rewards in the tabular setting. (Zhong, Yang, and Szepesvári 2021) extends the above results to a more general setting where both the transitions and rewards can vary over episodes. To eliminate the assumption of having prior knowledge on variation budgets, (Wei and Luo 2021) recently outline that an adaptive restart approach can be used to convert any upper-confidence-bound-type stationary RL algorithm to a dynamic-regret-minimizing algorithm. Beyond the non-stationary unconstrained RL, (Qiu et al. 2020) consider the online CMDPs where the reward is adversarial but the transition model is fixed and the constraints are stochastic over episodes. In summary, the above papers only consider the non-stationarity in the objective and may not work for the more general safe RL problems where there is also time-varying constraints.

CMDP. The study of RL algorithms for CMDPs has received considerable attention due to the safety requirement (Altman 1999; Paternain et al. 2019; Yu et al. 2019; Dulac-Arnold, Mankowitz, and Hester 2019; Garcıa and Fernández 2015; Gu et al. 2021, 2022). Our work is closely related to Lagrangian-based CMDP algorithms with optimistic policy evaluations (Efroni, Mannor, and Pirotta 2020; Singh, Gupta, and Shroff 2020; Ding et al. 2021; Liu et al. 2021; Qiu et al. 2020). In particular, (Efroni, Mannor, and Pirotta 2020; Singh, Gupta, and Shroff 2020) leverage upper confidence bound (UCB) bonus on fixed reward/utility and transition probability to propose sample efficient algorithms for tabular CMDPs. (Ding et al. 2021) generalize the above results to the linear kernel CMDPs. Under some mild conditions and additional computation cost, (Liu et al. 2021) propose two algorithms to learn policies with a zero or bounded constraint violation for CMDPs. Beyond the stationary CMDP, (Qiu et al. 2020) consider the online CMDPs where only the rewards in objective can vary over episodes. In contrast, our work focuses on a more general and realistic safe RL setting where the dynamics and rewards/utilities can all change over episodes, and thus we significantly extend the existing results.

Due to space restrictions, we introduce the notations in Appendix.

Preliminaries

Model. In this paper, we study safe RL in non-stationary environments via episodic CMDPs with adversarial bandit-information reward/utility feedback and unknown adversarial transition kernels. At each episode m, a CMDP is defined by the state space $\mathcal{S}$, the action space $\mathcal{A}$, the fixed length of each episode H, a collection of transition probability measure $\{\mathbb{P}_h^m\}_{h=1}^H$, a collection of reward functions $\{r_h^m\}_{h=1}^H$, a collection of utility functions $\{g_h^m\}_{h=1}^H$ and the constraint offset b_m. We assume that $\mathcal{S}$ is a measurable space with a possibly infinite number of elements, and that $\mathcal{A}$ is a finite set. In addition, we assume $r_h^m : \mathcal{S}\times\mathcal{A} \to [0,1]$ and $g_h^m : \mathcal{S}\times\mathcal{A} \to [0,1]$ are deterministic reward and utility functions. Our analysis readily generalizes to the setting where the reward/utility functions are random. In this paper, we focus on a bandit setting where the agent only observes the values of reward and utility functions, $r_h^m(x_h^m, a_h^m)$ and $g_h^m(x_h^m, a_h^m)$ at the visited state-action pair (x_h^m, a_h^m). To avoid triviality, we take $b_m \in (0, H]$ and assume that it is known to the agent.

Let the policy space $\Delta(\mathcal{A}|\mathcal{S}, H)$ be $\{\{\pi_h(\cdot|\cdot)\}_{h=1}^H : \pi_h(\cdot|s) \in \Delta(\mathcal{A}), \forall x \in \mathcal{S}, h \in [H]\}$, where $\Delta(\mathcal{A})$ denotes a probability simplex over the action space. Let $\pi^m \in \Delta(\mathcal{A}|\mathcal{S}, H)$ be a policy taken by the agent at episode m, where $\pi_h^m(\cdot|x_h^m) : \mathcal{S} \to \mathcal{A}$ is the action that the agent takes at state x_h^m. For simplicity, we assume the initial state x_1^m to be fixed as x_1 in different episodes. The episode terminates at state x_H^m in which no control action is needed and both reward and utility functions are equal to zero.

Given a policy $\pi \in \Delta(\mathcal{A}|\mathcal{S}, H)$ and the episode m, the value function $V_{r,h}^{\pi,m}$ associated with the reward function r at step h in episode m is the expected value of the total reward, $V_{r,h}^{\pi,m}(x) = \mathbb{E}_{\pi,\mathbb{P}^m}\left[\sum_{i=h}^H r_i^m(x_i, a_i)|x_h = x\right]$, for all $x \in \mathcal{S}$ and $h \in [H]$, where the expectation $\mathbb{E}_{\pi,\mathbb{P}^m}$ is taken over the random state-action sequence $\{(x_i^m, a_i^m)\}_{i=h}^H$, the action a_h^m follows the policy $\pi_h^m(\cdot|x_h^m)$, and the next state x_{h+1} follows the transition dynamics $\mathbb{P}_h^m(\cdot|x_h^m, a_h^m)$.

The action-value function is defined as $Q_{r,h}^{\pi,m}(x,a) = \mathbb{E}_{\pi,\mathbb{P}^m}\left[\sum_{i=h}^H r_i^m(x_i^m, a_i^m)|x_h^m = x, a_h^m = a\right]$, for all $x \in \mathcal{S}, a \in \mathcal{A}$ and $h \in [H]$. Similarly, we define the value function $V_{g,h}^{\pi,m} : \mathcal{S} \to \mathbb{R}$ and the action-value function $Q_{g,h}^{\pi,m} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ associated with the utility function g. For brevity, we use the symbol $\diamond$ to denote r or g. and take the shorthand $\mathbb{P}_h^m V_{\diamond,h}^{\pi,m}(x,a) \coloneqq \mathbb{E}_{x'\sim\mathbb{P}_h^m(\cdot|x,a)}\left[V_{\diamond,h+1}^{\pi,m}(x')\right]$. The Bellman equation associated with a policy π is given by

$$Q_{\diamond,h}^{\pi,m}(x,a) = (\diamond_h^m + \mathbb{P}_h^m V_{\diamond,h+1}^{\pi,m})(x,a), \tag{1a}$$

$$V_{\diamond,h}^{\pi,m}(x) = \left\langle Q_{\diamond,h}^{\pi,m}(x,\cdot), \pi_h(\cdot|x)\right\rangle_{\mathcal{A}}, \tag{1b}$$

for all $(x,a) \in \mathcal{S}\times\mathcal{A}$, where $\langle\cdot,\cdot\rangle_{\mathcal{A}}$ denotes the inner product over $\mathcal{A}$ and we will omit the subscript $\mathcal{A}$ in the sequel when it is clear from the context.

Constrained MDP. In constrained MDPs, the agent aims to approximate the optimal non-stationary policy by interacting with the environment. In each episode m, the agent aims to maximize the expected total reward while satisfying the

Setting	Assumption	Dynamic regret	Constraint violation
Tabular	$B_{g,\mathcal{E}}$, $B_{\mathbb{P},\mathcal{E}}$	$\widetilde{\mathcal{O}}\left(\lvert\mathcal{S}\rvert^{\frac{2}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}H^{\frac{5}{3}}M^{\frac{1+\rho}{2}}(B_\Delta+B_*)^{\frac{1}{3}}\right)$	$\widetilde{\mathcal{O}}\left(\lvert\mathcal{S}\rvert^{\frac{2}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}H^{\frac{5}{3}}M^{\frac{2-\rho}{2}}(B_\Delta+B_*)^{\frac{1}{3}}\right)$
Tabular	γ	$\widetilde{\mathcal{O}}\left(\gamma^{-1}\lvert\mathcal{S}\rvert^{\frac{2}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}H^{\frac{5}{2}}M^{\frac{2}{3}}(B_\Delta+B_\star)^{\frac{1}{3}}\right)$	$\widetilde{\mathcal{O}}\left(\gamma^{-1}\lvert\mathcal{S}\rvert^{\frac{2}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}H^{\frac{5}{2}}M^{\frac{2}{3}}(B_\Delta+B_\star)^{\frac{1}{3}}\right)$
Linear kernel	$B_{g,\mathcal{E}}$, $B_{\mathbb{P},\mathcal{E}}$	$\widetilde{\mathcal{O}}\left(d^{\frac{9}{8}}H^{\frac{5}{2}}M^{\frac{3}{4}}(\sqrt{d}B_\Delta+B_*)^{\frac{1}{3}}\right)$	$\widetilde{\mathcal{O}}\left(d^{\frac{9}{8}}H^{\frac{5}{2}}M^{\frac{3}{4}}(\sqrt{d}B_\Delta+B_*)^{\frac{1}{3}}\right)$
Linear kernel	γ	$\widetilde{\mathcal{O}}\left(\gamma^{-1}d^{\frac{9}{8}}H^{\frac{5}{2}}M^{\frac{3}{4}}(\sqrt{d}B_\Delta+B_*)^{\frac{1}{3}}\right)$	$\widetilde{\mathcal{O}}\left(\gamma^{-1}d^{\frac{9}{8}}H^{\frac{5}{2}}M^{\frac{3}{4}}(\sqrt{d}B_\Delta+B_*)^{\frac{1}{3}}\right)$

Table 1: We summarize the dynamic regrets and constraint violations obtained in this paper for tabular and linear kernel CMDPs under different assumptions. Here, $B_{g,\mathcal{E}}$ and $B_{\mathbb{P},\mathcal{E}}$ are the local variation budgets for the constraints and are defined in Assumption 2, γ is the strict feasibility threshold of the constraints and is defined in Assumption 3, H is the horizon of each episode, M is the total number of episodes, d is the dimension of the feature mapping, $|\mathcal{S}|$ and $|\mathcal{A}|$ are the cardinalities of the state and action spaces, and B_Δ, B_* are the variation budgets defined in (6) and (7). There is a trade-off controlled by $\rho \in [\frac{1}{3}, \frac{1}{2}]$ between the dynamic regret and constraint violation for the tabular CMDP under Assumption 2.

constraints on the expected total utility

$$\max_{\pi\in\Delta(\mathcal{A}|\mathcal{S},H)} V_{r,1}^{\pi,m} \text{ subject to } V_{g,1}^{\pi,m} \geq b_m \tag{2}$$

for all $m = 1, 2, \ldots$, where the reward/utility functions and the transition kernels are potentially different across the episodes. The associated Lagrangian of problem (2) is given by

$$\mathcal{L}^m(\pi,\mu) \coloneqq V_{r,1}^{\pi,m} + \mu\left(V_{g,1}^{\pi,m} - b_m\right) \tag{3}$$

where the policy π is the primal variable and $\mu \geq 0$ is the dual variable. We can reformulate the constrained optimization problem (2) as the saddle-point problem $\max_{\pi\in\Delta(\mathcal{A}|\mathcal{S},H)} \min_{\mu\geq 0} \mathcal{L}^m(\pi,\mu)$. Let $\mathcal{D}^m(Y) :=$ maximize $_\pi \mathcal{L}^m(\pi,\mu)$ be the dual function, $\mu^{\star,m} := \operatorname{argmin}_{\mu\geq 0} \mathcal{D}^m(\mu)$ be an optimal dual variable and $\pi^{\star,m}$ be a globally optimal solution of (2) at episode m.

Unlike the unconstrained MDP, the optimal solution of CMDP is usually a stochastic policy and the best deterministic policy can lose as much as the difference between the respective values of the best and the worst policies (Altman 1999). As a consequence, RL methods that implicitly rely on the existence of a deterministic optimal policy (e.g., Q learning) may not be suitable for this type of problem. This further inspires the study of randomized policies and take on a policy gradient approach for non-stationary CMDP.

Performance metrics. Suppose that the agent executes policy π^m in episode m. We now define the dynamic regret and the constraint violation in the long run as:

$$\mathrm{DR}(M) \coloneqq \sum_{m=1}^{M}\left(V_{r,1}^{\pi^{\star,m},m} - V_{r,1}^{\pi^m,m}\right), \tag{4}$$

$$\mathrm{CV}(M) \coloneqq \left[\sum_{m=1}^{M}\left(b_m - V_{g,1}^{\pi^m,m}\right)\right]_+. \tag{5}$$

There are two main reasons for considering the constraint violation in the long run. Firstly, in many applications such as supply chain and energy systems, the requirements of balancing the time-varying and unknown demands with the supply are formulated as some time-varying constraints. As long as the supply and the demand can be balanced in the long run, the policy is considered safe. Secondly, since the utility function g_h^m is unknown *a priori* and time-varying, the constraint $V_{g,1}^{\pi,m} \geq b_m$ may not be satisfied in every episode m. Rather, the agent strives to satisfy the constraints in the long run. In other words, the agent aims to ensure the long-term constraint $\sum_{m=1}^{M}(V_{g,1}^{\pi,m} - b_m) \geq 0$ over some given period of episodes M.

Linear function approximation We focus on a class of CMDPs, where transition kernels and reward/utility functions are linear in feature maps.

Assumption 1 (Linear Kernel CMDP) *For every $m \in [M]$, the CMDP$(\mathcal{S},\mathcal{A},H,\mathbb{P}^m,r^m,g^m)$ satisfies the following conditions: (1) there exist a kernel feature map $\psi : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}^{d_1}$ and a vector $\theta_h^m \in \mathbb{R}^{d_1}$ with $\|\theta_h^m\|_2 \leq \sqrt{d_1}$ such that*

$$\mathbb{P}_h^m(x' \mid x,a) = \langle\psi(x,a,x'),\theta_h^m\rangle$$

for all $(x,a,x') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$ and $h \in [H]$; (2) there exist a feature map $\varphi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^{d_2}$ and vectors $\theta_{r,h}^m, \theta_{g,h}^m \in \mathbb{R}^{d_2}$ such that

$$r_h^m(x,a) = \left\langle\varphi(x,a),\theta_{r,h}^m\right\rangle \text{ and } g_h^m(x,a) = \left\langle\varphi(x,a),\theta_{g,h}^m\right\rangle$$

for all $(x,a) \in \mathcal{S}\times\mathcal{A}$ and $h \in [H]$, where $\max\left(\left\|\theta_{r,h}^m\right\|_2, \left\|\theta_{g,h}^m\right\|_2\right) \leq \sqrt{d_2}$; (3) for every function $V : \mathcal{S} \to [0,H]$, $\left\|\int_{\mathcal{S}}\psi(x,a,x')V(x')dx'\right\| \leq \sqrt{d_1}H$ for all $(x,a) \in \mathcal{S}\times\mathcal{A}$ and $\max(d_1,d_2) \leq d$.

This assumption adapts the definition of linear kernel MDP (Ayoub et al. 2020; Cai et al. 2020; Zhou, He, and Gu 2021) to CMDP and has also been used in (Ding et al. 2021) for stationary constrained MDP problems. We refer the reader to Appendix for more discussions on this assumption.

Variation budget. Note that $\mathbb{P}_h^m$ and r_h^m, g_h^m are determined by the unknown measures $\{\theta_h^m\}_{h\in[H],m\in[M]}$ and the latent vectors $\{\theta_{\diamond,h}^m\}_{h\in[H],m\in[M]}$ for $\diamond = r$ or g which can vary across the indexes $(m,h) \in [M]\times[H]$ in general. We measure the non-stationarity of the CMDP in terms of its variation in $\theta_h^m, \theta_{r,h}^m$ and $\theta_{g,h}^m$:

$$B_{\mathbb{P}} \coloneqq \textstyle\sum_{m=2}^{M}\sum_{h=1}^{H}\left\|\theta_h^m - \theta_h^{m-1}\right\|_2, \tag{6a}$$

$$B_\diamond \coloneqq \textstyle\sum_{m=2}^{M}\sum_{h=1}^{H}\left\|\theta_{\diamond,h}^m - \theta_{\diamond,h}^{m-1}\right\|_2, \text{ for } \diamond = r \text{ or } g, \tag{6b}$$

and denote $B_\Delta = B_\mathbb{P} + B_r + B_g$. Note that our definition of variation only imposes restrictions on the summation of non-stationarity across two different episodes, and it does not put any restriction on the difference between two consecutive steps in the same episode. In addition to the variations defined above, we introduce the total variation in the optimal policies of adjacent episodes:

$$B_\star \coloneqq \sum_{m=2}^{M} \sum_{h=1}^{H} \max_{x \in \mathcal{S}} \left\| \pi_h^{\star,m}(\cdot \mid x) - \pi_h^{\star,m-1}(\cdot \mid x) \right\|_1 . \quad (7)$$

The notion of $B_\star$ is also used for online convex optimization with a dynamic regret criterion (Besbes, Gur, and Zeevi 2015; Hall and Willett 2013, 2015; Cao and Liu 2018) and for policy-based methods in non-stationary unconstrained MDPs (Fei et al. 2020; Zhong, Yang, and Szepesvári 2021). It is worth noting that the variations $(B_\mathbb{P}, B_\diamond)$ and $B_\star$ do not imply each other.

A special but important example of the non-stationarity is the system with piece-wise constant dynamics and rewards/utilities where the number of switches is S. In this case, all variation budgets $(B_\mathbb{P}, B_\diamond)$ and $B_\star$ can be upper bounded by $\mathcal{O}(SH)$. As one of the first works to investigate the non-stationary CMDP, we assume that we have access to quantities B_Δ and $B_\star$ or some upper bounds on them via an oracle.

Assumptions on Time-Varying Constraints

In this paper, we consider two scenarios for the non-stationary CMDPs, each requiring some specific knowledge to enable safe exploration under the non-stationarity.

The first scenario assumes the knowledge of local variation budgets of constraints. We first define local variation budgets of constraints. To adapt the non-stationarity, the restart estimation of the value function is used (see Section Periodically Restarted Optimistic Policy Evaluation), which breaks the M episodes into $\lceil \frac{M}{L} \rceil$ epochs. For every $\mathcal{E} \in \left[\lceil \frac{M}{L} \rceil\right]$, define $B_{g,\mathcal{E}}$ and $B_{\mathbb{P},\mathcal{E}}$ to be the local variation budgets of the utility function and transitions within epoch $\mathcal{E}$. By definition, we have $\sum_{\mathcal{E}=1}^{\lceil \frac{M}{L} \rceil} B_{g,\mathcal{E}} \le B_g$ and $\sum_{\mathcal{E}=1}^{\lceil \frac{M}{L} \rceil} B_{\mathbb{P},\mathcal{E}} \le B_\mathbb{P}$.

Assumption 2 (Local variation budgets of constraints) *We have access to the local variation budget $B_{g,\mathcal{E}}$ and $B_{\mathbb{P},\mathcal{E}}$ for every $\mathcal{E} \in \left[\lceil \frac{M}{L} \rceil\right]$, and also the constrained optimization problems given in* (2) *are uniformly feasible.*

The second scenario extends the strict feasibility (also known as Slater condition) for problem (2) to non-stationary constrained optimization problems.

Assumption 3 (Uniformly strict feasibility) *We have access to a sequence of constraint thresholds $\{b_m\}_{m=1}^M$ and a constant γ such that the constrained optimization problems in* (2) *are γ−uniformly strictly feasible, i.e., there exist $\gamma > 0$ and $\bar{\pi}^m \in \Delta(\mathcal{A} \mid \mathcal{S}, H)$ such that $V_{g,1}^{\bar{\pi}^m,m}(x_1) \ge b_m + \gamma$ for all $m = 1, \ldots, M$.*

Under this assumption, one can establish the strong duality and the boundedness of the optimal dual variable.

Lemma 4 (Lemma 1 in (Ding et al. 2021)) *Under Assumption 3, it holds that $V_{r,1}^{\pi^{\star,m},m}(x_1) = \mathcal{D}^m(\mu^{\star,m})$ and $0 \le \mu^{\star,m} \le H/\gamma$ for all $m = 1, \ldots, M$.*

Remark 5 *We require either Assumption 2 or Assumption* (3)*, and both of them need not hold simultaneously. Assumption 2 requires the local variation budgets of constraints, but does not enforce every instance problem* (2) *to be strictly feasible. It is suitable for the case with a forecasting oracle for the constraints. For example, in supply chain or energy systems, the supply is desired to match the time-varying and unknown demands where a forecasting oracle for the demands is usually available. In addition, it is also suitable for the case with only non-stationary rewards such as collision avoidance in a maze with a moving target. On the other hand, Assumption 3 needs the knowledge of strict feasible constraint thresholds, but does not require the local variation budgets of constraints. It is suitable for the case with a relatively large feasibility threshold γ.*

Safe Exploration under The Non-Stationarity

In Algorithm 1, we develop a new efficient method named Periodically Restarted Optimistic Primal-Dual Proximal Policy Optimization (PROPD-PPO) algorithm. In each episode, our algorithm consists of three main stages: periodically restarted policy improvement, dual update, and periodically restarted policy evaluation. We first present the high-level idea behind our method.

High-Level Idea

Safe exploration in non-stationary CMDPs is more challenging in that we need to reduce the constraint violation even when the constraints vary over the episodes. To overcome this issue, we develop our method based on some assumed knowledge on the constraints. Under Assumption 2, since the optimal dual variables may not be well-bounded, we need to add a dual regularization to stabilize the dual updates and fully utilize the convexity of the dual function. In addition, the knowledge of local variation of the constraints is needed to obtain an optimistic estimator of constraint functions, so that a large dual variable cannot amplify the estimation error of the constraint functions. This is different from the dual update that has been used in Lagrangian-based stationary CMDPs under the strict feasible condition (Ding et al. 2020, 2021; Ying, Ding, and Lavaei 2021; Efroni, Mannor, and Pirotta 2020; Liu et al. 2021; Qiu et al. 2020). On the other hand, under Assumption 3, the optimal dual variables can be bounded by Lemma 4. Then, the dual regularization and an optimistic estimator for the constraint functions are not necessary. Thus, a standard dual update will be enough.

Periodically Restarted Policy Improvement

One way to update the policy π^m is to solve the Lagrangian-based policy optimization problem $\max_{\pi \in \Delta(\mathcal{A}|\mathcal{S},H)} \mathcal{L}_\xi^m(\pi, \mu^{m-1})$, where $\mathcal{L}_\xi^m(\pi, \mu^{m-1})$ is defined in (9) and the dual variable μ^{m-1} is from episode $m-1$. Motivated by the policy improvement step in NPG (Kakade 2001), TRPO (Schulman et al. 2015), and PPO (Schulman et al. 2017), we perform a simple policy update

Algorithm 1: Periodically Restarted Optimistic Primal-Dual Proximal Policy Optimization

Inputs: Time horizon M, restart period W, L, $\{Q^0_{r,h}, Q^0_{g,h}\}_{h=1}^H$ and $V^0_{g,1}$ being zero functions, initial policy $\{\pi^0_h\}_{h\in[H]}$ being uniform distributions on $\mathcal{A}$, initial dual variable $\mu^0 = 0$, dual regularization parameter ξ, learning rates $\alpha, \eta > 0$, χ.
for $m = 1, \ldots, M$ **do**
  Set the initial state $x_1^m = x_1$, $\ell_\pi^m = (\lceil \frac{m}{L} \rceil - 1)L + 1$, $\ell_Q^m = (\lceil \frac{m}{W} \rceil - 1)W + 1$.
  if $m = \ell_\pi^m$ **then**
    Set $\{Q^{m-1}_{r,h}, Q^{m-1}_{g,h}\}_{h=1}^H$ as zero functions and set $\{\pi^{m-1}_h\}_{h=1}^H$ as uniform distributions on $\mathcal{A}$.
  end if
  for $h = 1, 2, \ldots, H$ **do**
    Update the policy $\pi_h^m(\cdot \mid \cdot) \propto \pi_h^{m-1}(\cdot \mid \cdot) \exp\left(\alpha \left(Q^{m-1}_{r,h} + \mu^{m-1} Q^{m-1}_{g,h}\right)(\cdot, \cdot)\right)$.
    Take an action $a_h^m \sim \pi_h^m(\cdot \mid x_h^m)$ and receive reward/utility $r_h(x_h^m, a_h^m), g_h(x_h^m, a_h^m)$.
    Observe the next state x_{h+1}^m.
  end for
  Update the dual variable by $\mu^m = \mathrm{Proj}_{[0,\chi]}\left(\mu^{m-1} + \eta\left(b_m - V^{m-1}_{g,1}(x_1) - \xi\mu^{m-1}\right)\right)$.
  Estimate $\{Q^m_{r,h}, Q^m_{g,h}\}_{h=1}^H$ and $V^m_{g,1}$ via $\mathrm{LSTD}\left(\{x_h^\tau, a_h^\tau, r_h^\tau(x_h^\tau, a_h^\tau), g_h^\tau(x_h^\tau, a_h^\tau)\}_{h=1,\tau=\ell_Q^m}^{H,m}\right)$.
end for

in the online mirror descent fashion by

$$\underset{\pi\in\Delta(\mathcal{A}|\mathcal{S},H)}{\operatorname{argmax}} \sum_{h=1}^H \left\langle \left(Q^{m-1}_{r,h} + \mu^{m-1} Q^{m-1}_{g,h}\right)(x_h, \cdot), \pi_h - \pi_h^{m-1} \right\rangle - \frac{1}{\alpha} \sum_{h=1}^H D\left(\pi_h(\cdot|x_h) \mid \pi_h^{m-1}(\cdot|x_h)\right). \quad (8)$$

Since the above update is separable over H steps, we can update the policy π^m as line 8 in Algorithm 1, leading to a closed-form solution for each step $h \in [H]$. Furthermore, in order to guarantee the policy to be exploratory enough in new environments, our policy improvement step also features a periodic restart mechanism, which resets its policy to a uniform distribution over the action space $\mathcal{A}$ every L episodes.

Remark 6 *Although policy improvement step* (8) *has been used in stationary CMDPs (Ding et al. 2021), our method differs in the sense that we remove the requirement to mix the policy with a uniform policy at every iteration. This is due to a technical improvement in the analysis by replacing the "pushback property of KL-divergence lemma" (Lemma 14 in (Ding et al. 2021)) with the "one-step descent lemma" for the KL-regularized optimization.*

Dual Update

We first define the modified Lagrangian of (3) to be

$$\mathcal{L}_\xi^m(\pi, \mu) \coloneqq V_{r,1}^{\pi,m} + \mu\left(V_{g,1}^{\pi,m} - b_m\right) + \frac{\xi}{2}\|\mu\|_2^2 \quad (9)$$

where $\xi \geq 0$ is the dual regularization parameter to be determined later. Since the value function $V_{g,1}^{\pi,m}$ is unknown, in order to infer the constraint violation for the dual update, we estimate $V_{g,1}^{\pi^m,m}(x_1)$ via an optimistic policy evaluation. We update the Lagrange multiplier μ by moving μ^m to the direction of minimizing the estimated Lagrangian $\mathcal{L}(\pi, \mu)$:

$$\widetilde{\mathcal{L}}_\xi^m(\pi, \mu) \coloneqq V_{r,1}^m + \mu\left(V_{g,1}^m - b_m\right) + \frac{\xi}{2}\|\mu\|_2^2. \quad (10)$$

over $\mu \geq 0$ in line 14 of Algorithm 1, where $\eta > 0$ is a stepsize and $\mathrm{Prof}_{[0,\chi]}$ is a projection onto $[0, \chi]$ with an upper bound χ on μ^m. The choices of the parameters χ and ξ depend on the assumption:

$$\xi > 0,\ \chi = \infty, \text{ under Assumption 2,}$$
$$\xi = 0,\ \chi = \frac{2H}{\gamma}, \text{under Assumption 3.}$$

Under Assumption 2, since the strictly feasibility may not hold for all episodes (corresponding to $\gamma = 0$), we may not have a finite upper bound on the dual variable μ. Thus, a dual regularization with $\xi > 0$ is needed to stabilize the dual updates under the non-stationarity. The value of ξ depends on the number of episodes M and the variation budgets $B_\mathbb{P}, B_g$. On the other hand, under Assumption 3, we choose $\chi = \frac{2H}{\gamma} \geq 2\mu^{*,m}$ similarly as (Ding et al. 2021; Efroni, Mannor, and Pirotta 2020), so that the projection interval $[0, \chi]$ includes all optimal dual variables $\{\mu^{*,m}\}_{m=1}^M$ in light of Lemma 4.

Periodically Restarted Optimistic Policy Evaluation

To evaluate the policy under the unknown nonstationarity, we take the Least-Squares Temporal Difference (LSTD) (Bradtke and Barto 1996; Lazaric, Ghavamzadeh, and Munos 2010) with UCB to properly handle the exploration-exploitation trade-off and apply the restart strategy to adapt to the unknown nonstationarity. In particular, we apply the restart strategy and evaluate the policy π^m only based on the previous historical trajectories from the episode ℓ_Q^m to the episode m instead of the all previous historical trajectories. The method is standard and summarized in Appendix.

After obtaining the estimates of $\mathbb{P}_h^m V_{\diamond,h+1}^m$ and $\diamond_h^m(\cdot,\cdot)$ for $\diamond = r$ or g, we update the estimated action-value function $\left\{Q^m_{\diamond,h}\right\}_{h=1}^H$ iteratively and add UCB bonus terms $\Gamma_h^m(\cdot,\cdot)$, $\Gamma^m_{\diamond,h}(\cdot,\cdot) : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^+$ so that

$$\Omega_{1,\diamond} \coloneqq (\varphi^m)^\top u^m_{\diamond,h} + \Gamma_h^m \text{ and } \Omega_{2,\diamond} \coloneqq \left(\phi^m_{\diamond,h}\right)^\top w^m_{\diamond,h} + \Gamma^m_{\diamond,h}$$

all become upper bounds on $\mathbb{P}_h^m V^m_{\diamond,h+1}$ and $\diamond_h^m(\cdot,\cdot)$ (up to some errors due to the non-stationarity). Here, the weights $u^m_{\diamond,h}, w^m_{\diamond,h}$ and the bonus terms $\Gamma_h^m, \Gamma^m_{\diamond,h}$ are defined in Appendix. Moreover,

$$Q^m_{r,h}(\cdot,\cdot) = \min\left(H - h + 1, \Omega_{1,r}(\cdot,\cdot) + \Omega_{2,r}(\cdot,\cdot)\right)_+,$$
$$Q^m_{g,h}(\cdot,\cdot) = \min\left(H - h + 1, \Omega_{1,g}(\cdot,\cdot) + \Omega_{2,g}(\cdot,\cdot) + LV\right)_+$$

where $LV > 0$ depends on the local variation budgets of the constraint $B_{\mathbb{P},\mathcal{E}}$, $B_{g,\mathcal{E}}$ under Assumption 2, $LV = 0$ under Assumption 3, and $(x)_+$ denotes the maximum between x and 0. The reason for introducing a positive LV term under Assumption 2 is to guarantee that the model prediction error in $Q^m_{g,h}$ is non-positive when the dual variable μ is very large.

Main Results

We now present the dynamic regret and the constraint violation bounds for Algorithm 1 under the two alternative assumptions introduced in Section Assumptions on Time-Varying Constraints. The choices of the algorithm parameters will depend on the assumption used for the analysis. When both assumptions are satisfied, one can check which one yields a tighter bound, and this depends on the value of the strict feasibility threshold γ (and the values of H, M if in the tabular CMDP setting).

Linear Kernal CMDP

We first present the results for linear Kernal CMDP under each of Assumptions 2 and 3.

Theorem 7 (Linear Kernal CMDP + Assumption 2) *Let Assumptions 1 and 2 hold. Given $p \in (0,1)$, we set $\alpha = H^{-1}M^{-\frac{1}{2}}(\sqrt{d}B_\Delta + B_\star)^{\frac{1}{3}}$, $L = M^{\frac{3}{4}}(\sqrt{d}B_\Delta + B_\star)^{-\frac{2}{3}}$, $\eta = M^{-\frac{1}{2}}$, $\xi = 2H(\sqrt{d}B_\Delta + B_\star)^{\frac{1}{3}}M^{-\frac{1}{2}}$, $W = d^{-\frac{1}{4}}H^{-1}M^{\frac{1}{2}}B_\Delta^{-\frac{1}{2}}$, in Algorithm 1 and set set $\beta = C_1\sqrt{dH^2\log(dW/p)}$, $LV = B_{\mathbb{P},\mathcal{E}}H^2 d_1\sqrt{d_1 W} + B_{g,\mathcal{E}}\sqrt{d_2 W}$. Then, with probability $1-p$, the dynamic regret and the constraint violation satisfy*

$$\mathrm{DR}(M) \le \widetilde{\mathcal{O}}\left(d^{\frac{9}{8}}H^{\frac{5}{2}}M^{\frac{3}{4}}(\sqrt{d}B_\Delta + B_*)^{\frac{1}{3}}\right),$$
$$\mathrm{CV}(M) \le \widetilde{\mathcal{O}}\left(d^{\frac{9}{8}}H^{\frac{5}{2}}M^{\frac{3}{4}}(\sqrt{d}B_\Delta + B_*)^{\frac{1}{3}}\right).$$

Theorem 8 (Linear Kernal CMDP + Assumption 3) *Let Assumptions 1 and 3 hold. Given $p \in (0,1)$, we set $\alpha = \gamma H^{-\frac{3}{2}}M^{-\frac{1}{3}}(\sqrt{d}B_\Delta + B_\star)^{\frac{1}{3}}$, $L = M^{\frac{2}{3}}(\sqrt{d}B_\Delta + B_\star)^{-\frac{2}{3}}$, $\eta = M^{-\frac{1}{2}}$, $\xi = 0$, $W = d^{-\frac{1}{4}}H^{-1}M^{\frac{1}{2}}B_\Delta^{-\frac{1}{2}}$ in Algorithm 1 and set set $\beta = C_1\sqrt{dH^2\log(dW/p)}$, $LV = 0$. Then, with probability $1-p$, the dynamic regret and the constraint violation satisfy*

$$\mathrm{DR}(M) \le \widetilde{\mathcal{O}}\left(\gamma^{-1}d^{\frac{9}{8}}H^{\frac{5}{2}}M^{\frac{3}{4}}(\sqrt{d}B_\Delta + B_*)^{\frac{1}{3}}\right),$$
$$\mathrm{CV}(M) \le \widetilde{\mathcal{O}}\left(\gamma^{-1}d^{\frac{9}{8}}H^{\frac{5}{2}}M^{\frac{3}{4}}(\sqrt{d}B_\Delta + B_*)^{\frac{1}{3}}\right).$$

The proofs for Theorems 7 and 8 can be found in Appendix. Our dynamic regret bounds in Theorems 7 and 8 have the optimal dependence on the total number of episodes M. This matches the existing bounds in the general non-stationary linear kernel MDP setting without any constraints (Zhong, Yang, and Szepesvári 2021; Zhou et al. 2020; Touati and Vincent 2020). The dependence on the variation budgets $(B_\Delta, B_\star)$ also matches the existing bound in policy-based method for the non-stationary linear kernel MDP setting (Zhong, Yang, and Szepesvári 2021). Regarding the long-term safe exploration, we provide the first finite-time constraint violation result in the non-stationary CMDP setting.

In the linear kernel CMDP setting, the same dynamic regret and constraint violation bounds are obtained under either of Assumptions 2 and 3, except that the dynamic regret and constraint violation under Assumption 3 also depend on the strict feasibility threshold γ. When γ is small, i.e., there exist some episodes for which the CMDP problem (2) does not have a large enough strict feasibility threshold, the dynamic regret and constraint violation bounds in Theorem 8 may be large.

Tabular CMDP

A special case of the linear kernel CMDP in Assumption 1 is the tabular CMDP with $|\mathcal{S}| < \infty$ and $|\mathcal{A}| < \infty$. In the tabular case, improved results can be obtained by incorporating Algorithm 1 with a variant of the optimistic policy evaluation method. We refer the reads to Appendix for such procedures and state the result below:

Theorem 9 (Tabular CMDP + Assumption 2) *Let Assumption 2 hold and consider a tabular CMDP. Given $p \in (0,1)$ and $\rho \in [\frac{1}{3}, \frac{1}{2}]$, we set $\alpha = H^{-\frac{1}{3}}M^{-\rho}(B_\Delta + B_\star)^{\frac{1}{3}}$, $L = H^{-\frac{1}{3}}M^{\frac{1+\rho}{2}}(B_\Delta + B_\star)^{-\frac{2}{3}}$, $\eta = H^{-\frac{1}{3}}M^{-\frac{1}{2}}$, $\xi = 2H^{\frac{5}{3}}(B_\Delta + B_\star)^{\frac{1}{3}}M^{-\rho}$, $W = H^{\frac{2}{3}}|\mathcal{S}|^{\frac{2}{3}}|\mathcal{A}|^{\frac{1}{3}}\left(\frac{M}{B_\Delta}\right)^{\frac{2}{3}}$ in Algorithm 1 and $\beta = C_4 H\sqrt{|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|W/p)}$, $LV = B_{\mathbb{P},\mathcal{E}}H + B_{g,\mathcal{E}}$. Then, with probability $1-p$, the dynamic regret and the constraint violation satisfy*

$$\mathrm{DR}(M) \le \widetilde{\mathcal{O}}\left(|\mathcal{S}|^{\frac{2}{3}}|\mathcal{A}|^{\frac{1}{3}}H^{\frac{5}{3}}M^{\frac{1+\rho}{2}}(B_\Delta + B_*)^{\frac{1}{3}}\right),$$
$$\mathrm{CV}(M) \le \widetilde{\mathcal{O}}\left(|\mathcal{S}|^{\frac{2}{3}}|\mathcal{A}|^{\frac{1}{3}}H^{\frac{5}{3}}M^{\frac{2-\rho}{2}}(B_\Delta + B_*)^{\frac{1}{3}}\right).$$

Theorem 10 (Tabular CMDP + Assumption 3) *Let Assumption 3 hold and consider a tabular CMDP. Given $p \in (0,1)$, we set $\alpha = \gamma H^{-\frac{3}{2}}M^{-\frac{1}{3}}(B_\Delta + B_\star)^{\frac{1}{3}}$, $L = M^{\frac{2}{3}}(B_\Delta + B_\star)^{-\frac{2}{3}}$, $\eta = M^{-\frac{1}{2}}$, $\xi = 0$, $W = |\mathcal{S}|^{\frac{2}{3}}|\mathcal{A}|^{\frac{1}{3}}\left(\frac{M}{B_\Delta}\right)^{\frac{2}{3}}$ in Algorithm 1 and $\beta = C_4 H\sqrt{|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|W/p)}$, $LV = 0$. Then, with probability $1-p$, the dynamic regret and the constraint violation satisfy*

$$\mathrm{DR}(M) \le \widetilde{\mathcal{O}}\left(\gamma^{-1}|\mathcal{S}|^{\frac{2}{3}}|\mathcal{A}|^{\frac{1}{3}}H^{\frac{5}{2}}M^{\frac{2}{3}}(B_\Delta + B_*)^{\frac{1}{3}}\right),$$
$$\mathrm{CV}(M) \le \widetilde{\mathcal{O}}\left(\gamma^{-1}|\mathcal{S}|^{\frac{2}{3}}|\mathcal{A}|^{\frac{1}{3}}H^{\frac{5}{2}}M^{\frac{2}{3}}(B_\Delta + B_*)^{\frac{1}{3}}\right).$$

The proofs for Theorems 9 and 10 can be found in Appendix. For the tabular CMDP under Assumption 2, there is a trade-off for the dependence on the total number of episodes M between the dynamic regret and the constraint violation. This trade-off is controlled by the primal update parameter α and the dual regularization parameter ξ. Such trade-off does not appear in the linear kernel CMDP setting because the dynamic regret and constraint violation in the linear kernel CMDP are bottlenecked by the error in the non-stationary policy evaluation.

The dynamic regret and constraint violation bounds in Theorem 10 have an improved dependence on the total number of episodes M compared to Theorems 7 and 8. This improvement is due to the improved result of the policy evaluation step in the tabular setting. The dependence on M in Theorem 10 is also better than that of Theorem 9. This is due to a sharper analysis for the constraint violation under Assumption 3 based on (Beck 2017, Proposition 3.60). However, the dynamic regret and constraint violation bounds in Theorem 10 have a worse dependence on the horizon H and are also dependent on the feasibility threshold γ compared to Theorem 9. In addition, the dependence of the dynamic regret on M and $(B_\Delta, B_\star)$ matches the existing bound in the non-stationary tabular MDP setting without any constraints (Mao et al. 2020).

Proof Sketch

The safe exploration in non-stationary CMDP is more challenging since the utilities and dynamics are time-varying and unknown a priori. In this section, we outline some of the key ideas behind the proof, especially how to decouple the dynamic regret and constraint violation under the non-stationarity. We defer the full proof to Appendix.

Dynamic Regret

By combining the primal-dual analysis of stationary CMDPs (Ding et al. 2021) and the analysis for the non-stationary MDP (Zhong, Yang, and Szepesvári 2021; Fei et al. 2020), we can obtain the following bound on the Lagrangian function:

$$\sum_{m=1}^{M}\left(V_{r,1}^{\pi^{\star,m},m}-V_{r,1}^{\pi^m,m}+\mu^m\left(b_m-V_{g,1}^{\pi^m,m}\right)\right)$$
$$\leq \delta_1+\alpha H^2\sum_{m=1}^{M}|\mu^m|^2+\sum_{m=1}^{M}\sum_{h=1}^{H}\mu^m\mathbb{E}_{\pi^{\star,m},\mathbb{P}^m}\left[\iota_{g,h}^m\right] \quad (11)$$

where δ_1 contains all terms irrelated to the dual variables $\{\mu^m\}_{m=1}^M$ and $\iota_{g,h}^m$ is the model prediction error of the constraint. Furthermore, with the dual regularization and the dual update, it holds that

$$-\sum_{m=1}^{M}\mu^m\left(V_{g,1}^{\pi^{\star,m},m}-V_{g,1}^m\right)$$
$$\leq \eta H^2(M+1)+(\eta\xi^2-\xi)\sum_{m=1}^{M}(\mu^m)^2. \quad (12)$$

Combining the inequalities (11) and (12) yields

$$\mathrm{DR}(M)\leq\delta_1+(\alpha H^2+\eta\xi^2-\xi)\sum_{m=1}^{M}(\mu^m)^2$$
$$+\eta H^2(M+1)+\sum_{m=1}^{M}\sum_{h=1}^{H}\mu^m\mathbb{E}_{\pi^{\star,m},\mathbb{P}^m}\left[\iota_{g,h}^m\right].$$

Under Assumption 2, since μ^m is not well-bounded (when CMDP is not strictly feasible), a positive dual regularization ξ is needed to guarantee $\alpha H^2+\eta\xi^2-\xi\leq 0$ and the knowledge of the local variation budgets of the constraint $B_{\mathcal{P},\mathcal{E}}, B_{g,\mathcal{E}}$ are needed to guarantee that $\iota_{g,h}^m$ is non-positive. On the other hand, under Assumption 3, μ^m is bounded by $\frac{2H}{\gamma}$ and the dynamic regret can be well-controlled without any additional requirement on ξ and $\iota_{g,h}^m$.

Constraint Violation

The techniques used for our analyses under Assumptions 2 and 3 are different. We first consider Assumption 2. If we set $\xi\eta\leq\frac{1}{2}$ and $\chi=\infty$ in Algorithm 1, then from the convexity of the Lagrangian function with respect to the dual variable, we have

$$\sum_{m=1}^{M}(\mu-\mu^m)\left(b_m-V_{g,1}^m\right)-(\frac{\xi M}{2}+\frac{1}{2\eta})\mu^2\leq\eta H^2M$$

for every $\mu\geq 0$. By combining the above inequality with the inequality (11) and using the fact that $\left|V_{r,1}^{\pi^{\star,m},m}-V_{r,1}^{\pi^m,m}\right|\leq H$, it holds that

$$\mu\sum_{m=1}^{M}\left(b_m-V_{g,1}^{\pi^m,m}\right)-(\frac{\xi M}{2}+\frac{1}{2\eta})\mu^2\leq\delta_2,$$

where δ_2 is irrelated to μ. Then, by maximizing both sides of the above inequality over $\mu\geq 0$, we can obtain the constraint violation under Assumption 2. On the other hand, the analysis of the constraint violation under Assumption 3 relies on the extension of (Beck 2017, Proposition 3.60) or (Ding et al. 2021, Lemma 10). In particular, it shows that if there exist δ_3 and $\bar{C}^\star\geq 2\max_{m\in[M]}\mu^{\star,m}$ such that

$$\sum_{m=1}^{M}V_{r,1}^{\pi^{\star,m},m}-V_{r,1}^{\pi^m,m}+\bar{C}^\star\sum_{m=1}^{M}\left(b_m-V_{g,1}^{\pi^m,m}\right)\leq\delta_3,$$

then the constraint violation can be bounded by $\sum_{m=1}^{M}\left(b_m-V_{g,1}^{\pi^m,m}\right)\leq\frac{2\delta_3}{\bar{C}^\star}$.

Conclusion

In this paper, we formulate a general non-stationary safe RL problem as a non-stationary episodic CMDP. To solve this problem, we identify two alternative conditions on the time-varying constraints under which we can guarantee the safety in the long run. We also develop a new algorithm named PROPD-PPO, which consists of three main mechanisms: periodic-restart-based policy improvement, dual update with dual regularization, and periodic-restart-based optimistic policy evaluation. We establish the dynamic regret bound and a constraint violation bounds for the proposed algorithm in both the linear kernel CMDP function approximation setting and the tabular CMDP setting under two alternative assumptions. This paper provides the first provably efficient algorithm for non-stationary CMDPs with safe exploration.

Acknowledgements

This work was supported by grants from ARO, AFOSR, ONR and NSF. We thank Xingyu Zhou, Yuntian Deng, Kaiqing Zhang, Dongsheng Ding and Donghao Ying for the fruitful discussions. In particular, we are grateful to Xingyu Zhou and Yuntian Deng for helping us improve our work and providing some ideas regarding constraint violations in Appendix.

References

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

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Auer, P.; Gajane, P.; and Ortner, R. 2019. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, 138–158. PMLR.

Ayoub, A.; Jia, Z.; Szepesvari, C.; Wang, M.; and Yang, L. 2020. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, 463–474. PMLR.

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

Besbes, O.; Gur, Y.; and Zeevi, A. 2015. Non-stationary stochastic optimization. *Operations research*, 63(5): 1227–1244.

Bradtke, S. J.; and Barto, A. G. 1996. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1): 33–57.

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

Cao, X.; and Liu, K. R. 2018. Online convex optimization with time-varying constraints and bandit feedback. *IEEE Transactions on automatic control*, 64(7): 2665–2680.

Chandak, Y.; Theocharous, G.; Shankar, S.; White, M.; Mahadevan, S.; and Thomas, P. 2020. Optimizing for the future in non-stationary MDPs. In *International Conference on Machine Learning*, 1414–1425. PMLR.

Cheung, W. C.; Simchi-Levi, D.; and Zhu, R. 2020. Reinforcement learning for non-stationary Markov decision processes: The blessing of (more) optimism. In *International Conference on Machine Learning*, 1843–1854. PMLR.

Ding, D.; Wei, X.; Yang, Z.; Wang, Z.; and Jovanovic, M. 2021. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, 3304–3312. PMLR.

Ding, D.; Zhang, K.; Basar, T.; and Jovanovic, M. R. 2020. Natural Policy Gradient Primal-Dual Method for Constrained Markov Decision Processes. In *NeurIPS*.

Ding, Y.; Lavaei, J.; and Arcak, M. 2021. Time-variation in online nonconvex optimization enables escaping from spurious local minima. *IEEE Transactions on Automatic Control*.

Domingues, O. D.; Ménard, P.; Pirotta, M.; Kaufmann, E.; and Valko, M. 2021. A kernel-based approach to non-stationary reinforcement learning in metric spaces. In *International Conference on Artificial Intelligence and Statistics*, 3538–3546. PMLR.

Dulac-Arnold, G.; Mankowitz, D.; and Hester, T. 2019. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*.

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

Fei, Y.; Yang, Z.; Wang, Z.; and Xie, Q. 2020. Dynamic regret of policy optimization in non-stationary environments. *arXiv preprint arXiv:2007.00148*.

Garcıa, J.; and Fernández, F. 2015. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1): 1437–1480.

Gu, S.; Kuba, J. G.; Wen, M.; Chen, R.; Wang, Z.; Tian, Z.; Wang, J.; Knoll, A.; and Yang, Y. 2021. Multi-agent constrained policy optimisation. *arXiv preprint arXiv:2110.02793*.

Gu, S.; Yang, L.; Du, Y.; Chen, G.; Walter, F.; Wang, J.; Yang, Y.; and Knoll, A. 2022. A Review of Safe Reinforcement Learning: Methods, Theory and Applications. *arXiv preprint arXiv:2205.10330*.

Hall, E.; and Willett, R. 2013. Dynamical models and tracking regret in online convex programming. In *International Conference on Machine Learning*, 579–587. PMLR.

Hall, E. C.; and Willett, R. M. 2015. Online convex optimization in dynamic environments. *IEEE Journal of Selected Topics in Signal Processing*, 9(4): 647–662.

Jaksch, T.; Ortner, R.; and Auer, P. 2010. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11(4).

Kakade, S. M. 2001. A natural policy gradient. *Advances in neural information processing systems*, 14.

Lazaric, A.; Ghavamzadeh, M.; and Munos, R. 2010. Finite-sample analysis of LSTD. In *ICML-27th International Conference on Machine Learning*, 615–622.

Liu, T.; Zhou, R.; Kalathil, D.; Kumar, P.; and Tian, C. 2021. Learning Policies with Zero or Bounded Constraint Violation for Constrained MDPs. *arXiv preprint arXiv:2106.02684*.

Mao, W.; Zhang, K.; Zhu, R.; Simchi-Levi, D.; and Basar, T. 2020. Model-Free Non-Stationary RL: Near-Optimal Regret and Applications in Multi-Agent RL and Inventory Control. *https://arxiv.org/abs/2010.03161*.

Moore, B. L.; Pyeatt, L. D.; Kulkarni, V.; Panousis, P.; Padrez, K.; and Doufas, A. G. 2014. Reinforcement learning for closed-loop propofol anesthesia: a study in human volunteers. *The Journal of Machine Learning Research*, 15(1): 655–696.

Ortner, R.; Gajane, P.; and Auer, P. 2020. Variational regret bounds for reinforcement learning. In *Uncertainty in Artificial Intelligence*, 81–90. PMLR.

Paternain, S.; Calvo-Fullana, M.; Chamon, L. F.; and Ribeiro, A. 2019. Safe policies for reinforcement learning via primal-dual methods. *arXiv preprint arXiv:1911.09101*.

Qiu, S.; Wei, X.; Yang, Z.; Ye, J.; and Wang, Z. 2020. Upper confidence primal-dual reinforcement learning for CMDP with adversarial loss. *arXiv preprint arXiv:2003.00660*.

Sallab, A. E.; Abdou, M.; Perot, E.; and Yogamani, S. 2017. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017(19): 70–76.

Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015. Trust region policy optimization. In *International conference on machine learning*, 1889–1897. PMLR.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Singh, R.; Gupta, A.; and Shroff, N. B. 2020. Learning in Markov decision processes under constraints. *arXiv preprint arXiv:2002.12435*.

Touati, A.; and Vincent, P. 2020. Efficient learning in non-stationary linear Markov decision processes. *arXiv preprint arXiv:2010.12870*.

Wei, C.-Y.; and Luo, H. 2021. Non-stationary Reinforcement Learning without Prior Knowledge: An Optimal Black-box Approach. *arXiv preprint arXiv:2102.05406*.

Ying, D.; Ding, Y.; and Lavaei, J. 2021. A Dual Approach to Constrained Markov Decision Processes with Entropy Regularization. *arXiv preprint arXiv:2110.08923*.

Yu, M.; Yang, Z.; Kolar, M.; and Wang, Z. 2019. Convergent policy optimization for safe reinforcement learning. *Advances in Neural Information Processing Systems*, 32: 3127–3139.

Zhong, H.; Yang, Z.; and Szepesvári, Z. W. C. 2021. Optimistic Policy Optimization is Provably Efficient in Non-stationary MDPs. *arXiv preprint arXiv:2110.08984*.

Zhou, D.; He, J.; and Gu, Q. 2021. Provably efficient reinforcement learning for discounted MDPs with feature mapping. In *International Conference on Machine Learning*, 12793–12802. PMLR.

Zhou, H.; Chen, J.; Varshney, L. R.; and Jagmohan, A. 2020. Nonstationary reinforcement learning with linear function approximation. *arXiv preprint arXiv:2010.04244*.