Safe Reinforcement Learning with Instantaneous Constraints: The Role of Aggressive Exploration

Honghao Wei[1], **Xin Liu**[2], **Lei Ying**[3]
[1]Washington State University
[2] ShanghaiTech University
[3]University of Michigan, Ann Arbor
honghao.wei@wsu.edu, liuxin7@shanghaitech.edu.cn, leiying@umich.edu

Abstract

This paper studies safe Reinforcement Learning (safe RL) with linear function approximation and under hard instantaneous constraints where unsafe actions must be avoided at each step. Existing studies have considered safe RL with hard instantaneous constraints, but their approaches rely on several key assumptions: (i) the RL agent knows a safe action set for *every* state or knows a *safe graph* in which all the state-action-state triples are safe, and (ii) the constraint/cost functions are *linear*. In this paper, we consider safe RL with instantaneous hard constraints without assumption (i) and generalize (ii) to Reproducing Kernel Hilbert Space (RKHS). Our proposed algorithm, LSVI-AE, achieves $\tilde{\mathcal{O}}(\sqrt{d^3H^4K})$ regret and $\tilde{\mathcal{O}}(H\sqrt{dK})$ hard constraint violation when the cost function is linear and $\mathcal{O}(H\gamma_K\sqrt{K})$ hard constraint violation when the cost function belongs to RKHS. Here K is the learning horizon, H is the length of each episode, and γ_K is the information gain w.r.t the kernel used to approximate cost functions. Our results achieve the optimal dependency on the learning horizon K, matching the lower bound we provide in this paper and demonstrating the efficiency of LSVI-AE. Notably, the design of our approach encourages aggressive policy exploration, providing a unique perspective on safe RL with general cost functions and no prior knowledge of safe actions, which may be of independent interest.

Introduction

Reinforcement Learning (RL) has shown significant empirical success in improving online decision-making in various applications, including games (Silver et al. 2017), robotic control (Andrychowicz et al. 2020), etc. However, in many real-world scenarios, it is essential to consider more than just maximizing rewards. Safety, ethical considerations, and adherence to predefined constraints are crucial aspects, particularly in critical domains like robotics, finance, and healthcare.

RL with instantaneous constraints addresses this need by introducing constraints that the agent must adhere to at every single time step during the learning process. Unlike constraints imposed on the entire trajectory or episode (Wei, Liu, and Ying 2022a,b; Ghosh, Zhou, and Shroff 2022; Ding et al. 2021; Liu et al. 2021a; Bura et al. 2021; Wei et al. 2023; Singh, Gupta, and Shroff 2020; Ding et al. 2021; Chen, Jain, and Luo 2022; Efroni, Mannor, and Pirotta 2020), instantaneous constraints demand strict compliance with specified limitations at each moment of decision-making so that unsafe actions should be avoided at each step. For instance, in autonomous vehicles, RL agents must consistently adhere to traffic rules and avoid dangerous maneuvers in *any time* to ensure safety. In healthcare, RL algorithms that respect privacy and confidentiality restrictions can recommend personalized treatment plans without violating patient data protection regulations. By enforcing instantaneous hard constraints, RL agents can be trusted and relied upon to operate responsibly in complex and dynamic environments while avoiding unnecessary exploratory actions and adhering to safety guidelines.

Existing literature on safe RL with hard instantaneous constraints has explored various aspects of this complex problem. (Amani, Alizadeh, and Thrampoulidis 2019; Pacchiano et al. 2021) studied the linear bandit problem with instantaneous constraints, which was extended to safe linear MDP with instantaneous constraints in (Amani, Thrampoulidis, and Yang 2021). The most recent work (Shi, Liang, and Shroff 2023) studied designing a safe policy for both unsafe states and actions. However, it is important to note that all of the existing works have made restrictive assumptions. (Amani, Thrampoulidis, and Yang 2021) requires the knowledge of a safe action for every state, while (Shi, Liang, and Shroff 2023) relies on a known safe subgraph where all state-action-state transitions are guaranteed to be safe. Additionally, all of these approaches require the cost function to have a linear structure, which imposes practical limitations on its applicability. In light of these assumptions, their approaches can ensure safe learning during the entire learning process with high probability due to the inherent capability of the algorithm to construct confidence sets along the state feature vector associated with the known safe actions. This, in turn, engenders a more conservative exploration strategy for the agent. A comparison of the theoretical results and the basic assumptions between our paper and the existing results can be found in Table 1.

In this paper, we study safe RL with minimal assumptions, where we neither assume any prior knowledge of cost/constraint functions nor any form of safe guidance (e.g., safe actions or graphs), except the necessary assumption that the cost functions are within Reproducing Kernel Hilbert Space (RKHS) to guarantee the learnability of cost functions. Since the agent is required to explore the environ-

Algorithm	Regret	Cost Function	Assumptions
LSVI-NEW (Shi, Liang, and Shroff 2023)	$\tilde{\mathcal{O}}(\frac{dH^3\sqrt{dK}}{\Delta_c})$	linear	known safe subgraph, star convex sets Lipschitz rewards/transitions
SLUCB-QVI (Amani, Thrampoulidis, and Yang 2021)	$\tilde{\mathcal{O}}(\sqrt{d^3H^4K})$	linear	known safe action for each state star convex sets
This Paper (LSVI-AE)	$\tilde{\mathcal{O}}(\sqrt{d^3H^4K})$	linear / RKHS	✗

Table 1: Regret on linear MDPs for safe RL with instantaneous hard constraints. Here d is the dimension of the feature mapping, H is the duration of each episode, K is the total number of episodes, and Δ_c is a safety-related parameter.

ment from scratch, the constraint violation is *unavoidable*. We consider the strict hard constraint violation, defined as $\sum_{k=1}^{K}\sum_{h=1}^{H} g_h(x_h^k, a_h^k)_+$, which prohibits the cancellation across different steps. Here, K represents the total number of episodes, H is the horizon of the (MDP), g_h is the cost function at step h, (x_h^k, a_h^k) denotes the state-action pair selected at step h during episode k, and $g_h(\cdot)_+ := \max\{g_h(\cdot), 0\}$. The hard constraint violation is much more challenging to minimize than the "soft" constraint violation $[\sum_{k=1}^{K}\sum_{h=1}^{H} g_h(x_h^k, a_h^k)]_+$. For example, if we consider a sequence of decisions such that $g_h(x_h^k, a_h^k) = -1$ if k is odd and $+1$ if k is even. Any positive value of the cost indicates a violation of the constraint. Then, assuming $K = 100$, it becomes evident that the soft violation is 0, but the constraint actually violates half of the K episodes. Therefore, an agent/policy with minimal hard violations can guarantee strong safety. Our main contributions of this paper are summarized below:

- We propose a novel algorithm, LSVI-AE, an acronym for **L**east-squares **V**alue **I**teration with **A**ggressive **E**xploration, which integrates adaptive penalty-based optimization with double optimistic learning. The algorithm guarantees fast learning in an uncertain environment while keeping the hard violation minimal (safe and aggressive exploration). Our design is based on the intuition that aggressive exploration in the initial periods can significantly improve safety and efficiency for the majority of subsequent periods, which is in contrast to the conventional idea of conservative exploration, typically employed in the previous study of safe bandits or RL.

- We prove that LSVI-AE achieves a regret of $\mathcal{O}(\sqrt{d^3H^4K})$, and a hard constraint violation of $\mathcal{O}(H\gamma_K\sqrt{K})$ (the violation becomes $\tilde{\mathcal{O}}(H\sqrt{dK})$ when the cost functions are linear). To show the sharpness of these results, we provide lower bounds on regret, which is $\Omega(Hd\sqrt{HK})$, and on the violation which is $\Omega(\sqrt{HK})$. The lower bounds show that LSVI-AE achieves the order-optimal regret and violation w.r.t. the episode length K, while the dependencies on d and H can be further improved to match the lower bound using the technique of the "rare-switching" idea (Hu, Chen, and Huang 2022; He et al. 2022). To the best of our knowledge, these are the first results in safe RL with instantaneous hard constraints. Further, the numerical experiments verify the "safe learning" of our algorithm.

Related Work

Safe RL, especially those with expected cumulative constraints, has been extensively studied under model-free approaches (Wei, Liu, and Ying 2022b,a; Wei et al. 2023; Ghosh, Zhou, and Shroff 2022), and model-based approaches(Ding et al. 2021; Liu et al. 2021a; Bura et al. 2021; Singh, Gupta, and Shroff 2020; Ding et al. 2021; Chen, Jain, and Luo 2022). There are also many works (Liu, Jiang, and Li 2022; Wu et al. 2018; Caramanis, Dimitrov, and Morton 2014) that have studied the knapsack constraints, wherein the learning process stops whenever the budget has run out. (Amani, Alizadeh, and Thrampoulidis 2019; Pacchiano et al. 2021) studied safe linear bandits which require a linear safety value for each step to be bounded. (Turchetta, Berkenkamp, and Krause 2016; Wachi et al. 2018) investigated instantaneous hard constraints with unsafe states under deterministic transitions. (Amani, Thrampoulidis, and Yang 2021; Shi, Liang, and Shroff 2023) studied linear MDPs with instantaneous hard constraints but with known safe actions or a safe subgraph, and only for the case with linear cost functions.

Problem Formulation

We consider an episodic Markov decision process (MDP) denoted by $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, g)$, where $\mathcal{S}$ is the state set, $\mathcal{A}$ is the action set, H is the length of each episode, $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ are the transition kernels at step h, $r = \{r_h\}_{h=1}^H$ are the reward functions, and $g = \{r_h\}_{h=1}^H$ are the cost functions. We assume that $\mathcal{S}$ is a measurable space with a possibly infinite number of elements, $\mathcal{A}$ is a finite action set. For any $h \in [H]$, the reward function $r_h : \mathcal{S}\times\mathcal{A} \to [0, 1]$, is assumed to be deterministic. However, it can be readily extended to settings where r_h is random. The unknown safety measures for taking an action a at state x is a random variable $G_h(x, a)$ with expectation $\mathbb{E}[G_h(x, a)] = g_h(x, a)$. Without loss of generality, we assume $g_h(x, a) : \mathcal{S} \times \mathcal{A} \to [-1, 1]$.

A policy $\pi = \{\pi_h\}_{h=1}^H$ for an agent is a set of functions with $\pi_h : \mathcal{S} \to \mathcal{A}$. In an episodic MDP, every episode starts by arbitrarily selecting an initial state x_1. In each subsequent step, an agent observes the state $x_h \in \mathcal{S}$, takes an action $a_h \in \mathcal{A}$ according to policy π_h, and receives a reward $r_h(x_h, a_h)$ and incurs a cost $g_h(x_h, a_h)$. The MDP then moves to the next state x_{h+1} based on the transition kernel $\mathbb{P}_h(\cdot|x_h, a_h)$. The episode ends after the action a_H is taken at the step H.

Given a policy π, let $V_h^\pi(x) : \mathcal{S} \to \mathbb{R}$ denote the expected value of the cumulative reward function starting from step h and state x, when the agent selects action using the policy

$\pi = \{\pi_h\}_{h=1}^H$, which is defined as

$$V_h^\pi(x) = \mathbb{E}\left[\sum_{i=h}^{H} r_i(x_i, a_i) | x_h = x, \pi\right], \forall x \in \mathcal{S}, h \in [H],$$

where $\mathbb{E}$ is taken with respect to the policy π and the transition kernels $\mathbb{P}$. Accordingly, we also let $Q_h^\pi(x,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denote the expected value of the cumulative reward starting from step h and the state-action pair (x,a) and follows the policy π as

$$Q_h^\pi(x,a) = \mathbb{E}\left[\sum_{i=h}^{H} r_i(x_i,a_i)|x_h = x, a_h = a, \pi\right], \quad \forall (x,a) \in \mathcal{S}\times\mathcal{A}, \forall h \in [H]. \tag{1}$$

To simplify the notation, we define

$$[\mathbb{P}_h V_{h+1}](x,a) := \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a)} V_{h+1}(x'). \tag{2}$$

Then we can express the Bellman equation for a given policy π as follows:

$$Q_h^\pi(x,a) = (r_h + \mathbb{P}_h V_{h+1}^\pi)(x,a), \tag{3}$$
$$V_h^\pi(x) = Q_h^\pi(x, \pi_h(x)), \tag{4}$$
$$V_{H+1}^\pi(x) = 0. \tag{5}$$

For an episodic MDP with instantaneous hard constraints, the agent needs to learn the optimal policy while satisfying the constraints at each step of any episode by interacting with the environment. The objective of the agent is to find a safe and optimal policy to solve the following problem:

$$\max_\pi \quad V_1^\pi(x_1) \tag{6}$$
$$\text{s.t.} \quad g_h(x_h, \pi(x_h)) \leqslant 0, \forall h \in [H]. \tag{7}$$

Assumption 1. *(Feasibility) There exists at least a "safe" action for each state $x_h \in \mathcal{S}, \forall h \in [H]$.*

Assumption 1 is necessary to ensure the feasibility of the problem. We remark that the safe actions are unknown to the learner.

Note that given complete knowledge of reward functions r_h, cost functions g_h, and the transition kernel $\mathbb{P}_h$, one could use dynamic (constrained) programming to determine the optimal policy π^* to (6)-(7) (thought dynamic programming might suffer from high computational overhead). However, this knowledge is not available in advance, and we have to learn this information while interacting with the environment.

To measure the performance of an agent in an online learning setting, we consider two metrics w.r.t. rewards and constraints. Let a policy selected by the agent at episode k be $\pi^k = \{\pi_h^k\}_{h=1}^H$. We define the performance metrics:

$$\text{Regret}(K) = \sum_{k=1}^{K} V_1^{\pi^*}(x_1^k) - V_1^{\pi^k}(x_1^k), \tag{8}$$

$$\text{Violation}(K) = \sum_{k=1}^{K}\sum_{h=1}^{H} \left[g_h(x_h^k, a_h^k)\right]_+, \tag{9}$$

where $[\cdot]_+ = \max\{\cdot, 0\}$. The regret is defined as the gap between the total rewards returned by the optimal policy π^*, and that obtained by following the agent's policy π^k over K episodes. The constraint violation captures the total constraint violation **without** cancellation over all the episodes K. Note that the violation is unavoidable for an online policy because we do not have the knowledge of the environment (e.g., the cost functions g_h). Moreover, the "hard" violation is much stricter than the "soft" violation $[\sum_{k=1}^K \sum_{h=1}^H g_h(x_h^k, a_h^k)]_+$, which is especially important for safety-critical applications.

Linear Constrained Markov Decision Processes

In order to handle a large number or even an infinite number of states, we consider the following linear MDPs.

Assumption 2. *The MDP is a linear MDP with feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, if for any h, there exists d unknown measures $\mu_h = \{\mu_h^1, \ldots, \mu_h^d\}$ over $\mathcal{S}$ such that for any $(x, a, x') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,*

$$\mathbb{P}_h(x'|x,a) = \langle \phi(x,a), \mu_h(x') \rangle, \tag{10}$$

and there exists vector $\theta_{r,h} \in \mathbb{R}^d$ such that for any $(x,a) \in \mathcal{S} \times \mathcal{A}$,

$$r_h(x,a) = \langle \phi(x,a), \theta_{r,h} \rangle.$$

With loss of generality, we assume $\|\phi(x,a)\| \leqslant 1$, for all $(x,a) \in \mathcal{S} \times \mathcal{A}$, and $\max\{\|\mu_h(\mathcal{S})\|, \|\theta_{r,h}\|\} \leqslant \sqrt{d}$ for all $h \in [H]$.

Under Assumption 2, we know that (Jin et al. 2020a) for a linear MDP and any policy π, there exists $\{w_h^\pi\}_{h=1}^H$ such that

$$Q_h^\pi(x,a) = \langle w_h^\pi, \phi(x,a) \rangle, \forall (x,a,h) \in \mathcal{S} \times \mathcal{A} \times [H].$$

Algorithm

In this section, we propose our algorithm, called Least-Squares Value Iteration with Aggressive Exploration (LSVI-AE), in Algorithm 1. The design of our algorithm is based on an *adaptive penalty-based optimization with double optimistic learning framework* to minimize the cumulative hard constraint violation by encouraging aggressive exploration. In this framework, in episode k, at step h, our algorithm learns both the $Q-$value function (Q_h^k) and the cost function $(\hat{g}_h(x,a))$ optimistically. By imposing an adaptive rectified operator on the estimated cost, actions are selected at each step h to maximize a surrogate function:

$$a_h^k = \arg\max_a \{Q_h^k(x_h^k, a) - Z_h^k(\hat{g}_h^k(x_h^k, a)_+)\}. \tag{11}$$

The agent's decision-making process encourages aggressive exploration throughout the learning, in contrast to the conservative policies commonly employed in addressing safe RL with episode constraints or budget limitations. This insight highlights a crucial observation: in the context of safe RL with instantaneous hard constraints and no prior knowledge of safe actions, finding a safe policy requires the agent's prompt exploration of actions that might initially appear unsafe. This strategic emphasis on early exploration of potentially risky actions stands as a foundational principle in our approach.

The nonnegative value Z_h^k is an adaptive penalty factor to control cumulative constraint violation. Note that a standard approach to solving a constrained optimization problem is to

optimize the Lagrange function instead, that is, to select an action to maximize:

$$L(x_h^k, \nu) := Q_h^k(x_h^k, a) - \nu \hat{g}_h(x_h^k, a), \tag{12}$$

where ν is the dual variable related to the cost $g_h(x, a) \leqslant 0$. We approximate the dual variable ν with an adaptive penalty factor Z_h^k which is updated according to the observed cost function: $Z_h^{k+1} := Z_h^k + g_h(x_h^k, a_h^k)_+$ to track the constraint violation during learning. The idea behind the adaptive factor Z_h^k lies in two folds. First the operator $\hat{g}_h^k(x, a)_+$ only penalizes the "unsafe" actions that do not satisfy the constraints. Secondly, a minimum penalty price η_h^k is established as a lower bound for Z_h^k to prevent aggressive decisions when the constraint is not satisfied. Therefore the adaptive rectified factor Z_h^k is updated as

$$Z_h^{k+1} := \max\{Z_h^k + g_h(x_h^k, a_h^k)_+, \eta_h^k\}. \tag{13}$$

This design is inspired by constrained online convex optimization (Guo et al. 2022) and constrained bandit optimization (Guo, Zhu, and Liu 2022). However, reinforcement learning with instantaneous constraints is much more complicated due to its stateful nature where the states/actions and rewards/-costs are all coupled. For example, if a dangerous/unfavorable action has been taken at the initial step in an episode, it might result in cascade effects to the sequential steps. The setting in (Guo et al. 2022; Guo, Zhu, and Liu 2022) can be regarded as a special case of $H = 1$ in this paper.

We remark here that another classical method to track constraint violation is using a virtual queue update approach such that the dual variable is updated as

$$Z_h^{k+1} := \max\{Z_h^k + g_h(x_h^k, a_h^k), 0\}. \tag{14}$$

This approach is usually referred to as the primal-dual approach or the drift-plus-penalty method, which is the most commonly used method for dealing with constraint RL/bandits (Efroni, Mannor, and Pirotta 2020; Ding et al. 2020, 2022; Bai et al. 2022; Liu et al. 2021b) or online convex optimization (Yi et al. 2022, 2021; Yu and Neely 2020). However, this approach or its variants usually require an assumption of Slater's condition or the knowledge of the Slater/slackness constant to achieve a safe policy. The design is primarily due to their target on "soft violation", where the virtual queues/dual variables are the proxy for "soft violation" and the Slater's condition is to guarantee the bounded violation. Apparently, this design cannot handle the RL setting with instantaneous hard constraints. This observation also has been justified in the simulation results.

Next, we present the idea of double optimism in estimating $Q-$value functions and cost functions g.

Optimistic Estimates of Q: Estimating $Q-$value functions need to solve a regularized least-squares problem (Jin et al. 2020a); however, we should use a SARSA-type update instead of $Q-$learning because Bellman optimally is no longer hold in RL with constraints, i.e., the $V_{h+1}(\cdot)$ in Line 8 in Algorithm 1 is not a maximize of the Q_{h+1} functions but from the Q function under the current policy.

To encourage exploration, an additional UCB bonus term $\beta(\phi^\top \Lambda_h^{-1} \phi)^{1/2}$ (Line 6 in Algorithm 1) is added when estimating the $Q-$value functions, where Λ_h is the Gram matrix

Algorithm 1: Least-Squares Value Iteration with Aggressive Exploration (LSVI-AE)

1 **Initialization:** $Z_h^1 = 1, \forall h \in [H], \eta_h^k = k, \forall k \in [K]$;
2 **for** *episode* $k = 1, \ldots, K$ **do**
3 Receive the initial state $x_1^k = x_1$. **for** $h = H, H-1 \ldots, 1$ **do**
4 $\Lambda_h^k = \sum_{\tau=1}^{k-1} \phi(x_h^\tau, x_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top + \lambda I$;
5 $w_h^k \leftarrow (\Lambda_h^k)^{-1}[\sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)[r_h(x_h^\tau, a_h^\tau) + V_{h+1}(x_{h+1}^\tau)]$;
6 $Q_h^k(\cdot, \cdot) \leftarrow \min\{\langle w_h^k, \phi(\cdot, \cdot)\rangle + \beta(\phi(\cdot, \cdot)^\top (\Lambda_h^k)^{-1} \phi(\cdot, \cdot))^{1/2}, H\}$;
7 $a_x = \arg\max_a \{Q_h^k(x, a) - Z_h^k(\hat{g}_h^k(x, a)_+)\}$.
8 $V_h^k(x) = Q_h^k(x, a_x)$.
9 **end**
10 **for** *h=1,…,H* **do**
11 Take action a_h^k according to Eq. (11) and observe the next state x_{h+1}^k, and cost $g_h(x_h^k, a_h^k)$;
12 Update estimates of the cost $\hat{g}_h^k(x, a)$;
13 **end**
14 **for** *h=1,…,H* **do**
15 $Z_h^{k+1} = \max\{Z_h^k + (g_h(x_h^k, a_h^k))_+, \eta_h^k\}$.
16 **end**
17 **end**

of the regularized least-square problem, and β is a scalar. The term $(\phi^\top \Lambda_h^{-1} \phi)^{-1}$ basic represents the effective number of samples that the agent has observed so far along the ϕ direction, and the bonus term represents the uncertainty along the ϕ direction. Therefore, we can prove that the estimate $Q-$value function Q_h^k is always an upper bound of Q_h^* for all state-action pairs (see Lemma 3). Proving this property also leverages the design of the adaptive penalty operator on the cost function.

Optimistic Estimates of g: Assuming the cost functions belongs to RKHS, we present the optimistic estimation when g_h are approximated by GP and also illustrate a special case when g_h are approximated by linear functions.

- **Gaussian Process approximation of cost functions:** When $G_h(x, a)$ is a Gaussian process. We let $y = (x, a)$ denote a state-action pair and denote $\mathcal{Y} = \mathcal{S} \times \mathcal{A}$ to simplify the notation. Gaussian process $GP(\mu(y), ker(y, y'))$ over a state space $y \in \mathcal{Y}$ is specified by its mean $\mu(y)$ and covariance $ker(y, y')$. If we assume that for any $h \in [H]$ the cost function $G_h(y)$ is a Gaussian process such that $g_h(y) = \mathbb{E}[G_h(y)]$, and $ker_h(y, y') = \mathbb{E}[(g_h(y) - \mu_h(y))(g_h(y') - \mu_h(y'))]$, where ker_h is the kernel function associated with the Reproducing Kernel Hilbert Space (RKHS) with a bounded norm. Then given a collection of states and actions $\mathcal{B}_h^k = \{y_h^1, \ldots, y_h^{k-1}\}$, we use the GP-LCB (Chowdhury and Gopalan 2017) to optimistically estimate the cost function for $h \in [H], k \in [K], y \in \mathcal{Y}$ in particular, $\hat{g}_h^k(y) = g_h^k(y) - \beta_h^k(p/H)\sigma_h^k(y)$, where

$\beta_h^k(p) = 1 + \sqrt{2(\gamma_h^k + 1 + \ln(2/p))}$ with $p \in (0,1)$. The information gain $\gamma_h^k := \max_{y \in \mathcal{Y}:} \frac{1}{2} \ln |I + \lambda^{-1} KER_h^k|$. The estimate model includes parameters $\{\mu_h^k, \sigma_h^k\}_{h=1}^H$ and for $h \in [H]$ they are updated as:

$$g_h^k(y) = ker_h^k(y)(V_h^k(\lambda))^{-1} g_h^{1:k}, ker_h^k(y, y') = ker_h(y, y') - ker_h^k(y)^\top (V_h^k(\lambda))^{-1} ker_h^k(y')$$

$$\sigma_h^k(y) = \sqrt{ker_h^k(y, y)},$$

where $V_h^k(\lambda) = KER_h^k + \lambda I, \lambda = 1 + 2/K, KER_h^k = [ker_h(y, y')]_{y,y' \in \mathcal{B}_h^k}, g_h^{1:k} = \{g_h^1(y_h^1), \dots g_h^{k-1}(y_h^{k-1})\}$, and $ker_h^k(y) = [ker_h(y_h^1, y), \dots, ker_h(y_h^{k-1}, y)]^\top$. Without loss of generality, we assume that the RKHS norm of the cost function is bounded, i.e., $\|f\|_{ker} = \sqrt{\langle f, f \rangle_{ker}} \leqslant 1$.

- **Linear function approximation for cost functions:** For any $h \in [H], (x, a) \in \mathcal{S} \times \mathcal{A}$, the cost function $g_h(x, a) : \mathcal{S} \times \mathcal{A} \to [-1, 1]$ is assumed to be linear such that there exists vector $\theta_{g,h} \in \mathbb{R}^d$ and $g_h(x, a) = \langle \phi(x, a), \theta_{g,h} \rangle$. Recall that at the kth episode, we have the Gram matrix $\Lambda_h^k = \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau) \phi(x_h^\tau, a_h^\tau)^\top + \lambda I$ and then we can have an optimization for any (x, a) at the step h with high probability according to:

$$\hat{\theta}_h^k(x, a) = (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau) g_h(x_h^\tau, a_h^\tau)$$

$$\tilde{\beta}_h^k(p) = \sqrt{\lambda d} + \sqrt{d \log((1 + k/\lambda)/p)}$$

$$\hat{g}_h^k(x, a) = \langle \phi(x, a), \hat{\theta}_h^k(x, a) \rangle - \tilde{\beta}_h^k(\frac{p}{H}) \|\phi(x, a)\|_{(\Lambda_h^k)^{-1}},$$

where $\|x\|_\Sigma = \sqrt{x^\top \Sigma x}$.

Note that $\hat{g}_h^k(x, a)$ is called an optimistic estimation of $g_h(x, a)$ because we are optimistic about $g_h(x, a) \leqslant 0$, which would imply $\hat{g}_h^k(x, a) \leqslant 0$ with high probability. Next, we introduce an important condition on the estimation error, which is the key to quantify regret and violation.

Condition 1. *There exist nonnegative values $e_h^k(p, x, a)$, we have for all $x \in \mathcal{S}, a \in \mathcal{A}$, and $h \in [H], k \in [K]$, for any $p \in (0, 1)$ we have with probability at least $1 - p$:*

$$0 \leqslant g_h(x, a) - \hat{g}_h^k(x, a) \leqslant e_h^k(p, x, a), \tag{15}$$

where $e_h^k(p, x, a) = 2\tilde{\beta}_h^k(p/H) \|\phi(x, a)\|_{(\Lambda_h^k)^{-1}}$ for the linear case and $2\beta_h^k(p/H)\sigma_h^k(x, a)$ for the Gaussian approximation case.

We will show that Condition 1 is satisfied by our optimistic learning in Lemma 6, and we defer the proof to the appendix due to the page limit.

Main Results

In this section, we present the main theoretical result of our algorithm (LSVI-AE), which includes a double optimistic estimation and an adaptive penalty-based rectified factor to encourage aggressive exploration. We also present a theorem that establishes an information-theoretic lower bound for episodic MDP with instantaneous hard constraints to show the tightness of our results.

Performance Guarantee

Our results are shown as follows:

Theorem 1. *Under Condition 1 and Assumptions 1 and 2, there exists an absolute constant $c > 0$ that for any fixed $p \in (0, 1/2)$, if we set $\lambda = 1, \beta = cdH\sqrt{\iota}$ in Algorithm 1 with $\iota = \log(2dHK/p)$, then with probability at least $1 - 2p$, the total regret and violation of Algorithm 1 satisfy:*

$$\begin{aligned} \textit{Regret}(K) &= \sum_{k=1}^K V_1^{\pi^*}(x_1^k) - V_1^{\pi^k}(x_1^k) \\ &= \mathcal{O}(\sqrt{d^3 H^4 K \iota^2}), \\ \textit{Violation}(K) &= \sum_{k=1}^K \sum_{h=1}^H g_h(x_h^k, a_h^k)_+ \\ &\leqslant \sum_{k=1}^K \sum_{h=1}^H e_h^k(p, x, a) + 2H^2 \log(K). \end{aligned}$$

We can observe that the dominant term for constraint violation comes from the error in the estimation of cost functions. For the different types of cost functions mentioned above, we have the following results.

- **Gaussian Processes:**

  Lemma 1. *Considering the cost function is a Gaussian process, the cumulative estimation error can be bounded as follows:*

$$\sum_{k=1}^K \sum_{h=1}^H e_h^k(p, x, a) \leqslant \mathcal{O}(H\gamma_K \sqrt{K}), \tag{16}$$

  where $\gamma_K = \max_h \{\gamma_h^K\}$.

- **Linear cost function:**

  Lemma 2. *Considering the cost function in a linear structure, the cumulative estimation error can be bounded as:*

$$\sum_{k=1}^K \sum_{h=1}^H e_h^k(p, x, a) \leqslant \tilde{\mathcal{O}}(H\sqrt{dK}) \tag{17}$$

Lower Bound

To demonstrate the sharpness of our results, We construct a hard-to-learn linear CMDP with the same state space $\mathcal{S}$, action space $\mathcal{A}$, episode length H, reward function $\{r_h\}_{h=1}^H$, cost function $\{g_h\}_{h=1}^H$ and the transition kernel $\{\mathbb{P}_h\}_{h=1}^H$ as in (Zhou, Gu, and Szepesvari 2021; Hu, Chen, and Huang 2022). This is the first result in safe RL under instantaneous hard constraints. The information-theoretic lower bound for the episodic CMDP with hard instantaneous constraints setting studied in this paper is shown is the following theorem.

Theorem 2. *Let $d \geqslant 4, H \geqslant 3$, and suppose that $K \geqslant \max\{(d-1)^2 H/2, (d-1)/(32H(d-1))\}$. Then there exists an episodic linear CMDP parameterized by μ_h, θ and satisfies the norm assumption given in Assumption 2, such that the expected regret and violation of constraints are lower bounded as follows by using any algorithm:*

$$\mathbb{E}[\textit{Regret}(K)] = \Omega(Hd\sqrt{HK}), \tag{18}$$

$$\mathbb{E}[\textit{Violation}(K)] = \Omega(\sqrt{HK}). \tag{19}$$

We can observe that both our regret and violation have the optimal dependencies on the episode length K when $p \leqslant 1/\sqrt{K}$. The dependencies on d and H can be further improved to match the lower bound using the technique of the "rare-switching" idea in (Hu, Chen, and Huang 2022; He et al. 2022).

Discussion on Extension to More General MDPs

Our adaptive penalty-based optimization with a double optimistic learning framework can be generalized to general function approximation beyond linear MDPs when the cost function belongs to RKHS. The more general LSVI-AE with function approximation is meant to solve a least-squares regression problem:

$$\hat{Q}_h^k \leftarrow \min_{f\in\mathcal{F}} \Bigg\{ \sum_{\tau=1}^{k-1} [r_h(x_h^\tau, a_h^\tau) + V_{h+1}^k(x_{h+1}^\tau) - f(x_h^\tau, a_h^\tau)]^2 + pen(f) \Bigg\}, \tag{20}$$

where $pen(f)$ is a regularization term, $\mathcal{F}$ is a function class. Then to ensure an overestimation, we can update Q function by adding a bonus term $b_h^k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$:

$$Q_h^k(x,a) := \min\left\{\hat{Q}_h^k + \beta \cdot b_h^k(x,a), H\right\}, \tag{21}$$

and $V_h^k(x) = Q_h^k(x,a)$, where

$$a = \arg\max_{a'\in\mathcal{A}} \{Q_h^k(x,a') - Z_h^k(\hat{g}_h^k(x_h^k, a'))_+\}.$$

The function $\mathcal{F}$ can be chosen as a RKHS as in (Yang et al. 2020), which covers the linear MDP discussed in this paper, or a more general function $\mathcal{F}$ with low Bellman Eluder Dimension (Jin, Liu, and Miryoosefi 2021).

Proof of Theorem 1

Next, we briefly review the key intuitions behind the main results in Theorem 1. We first introduce two lemmas that are useful to prove the theorem. The first lemma shows that Q_h^k is always an upper bound on Q_h^* at any episode k.

Lemma 3. *Given the event $\mathcal{E}$ defined in Lemma 15 and condition 1, the following inequality holds simultaneously for all (x,a), step h and episode k,*

$$Q_h^k(x,a) \geqslant Q_h^*(x,a). \tag{22}$$

In the lemma, we bound the difference between the value function maintained in Algorithm 1 and the true value function under policy π^k used in each episode k.

Lemma 4. *Under the event $\mathcal{E}$ defined in Lemma 15, for any fixed $p \in (0,1)$, if we set $\lambda = 1, \beta = c \cdot dH\sqrt{\iota}$ in Algorithm 1 with $\iota = \log(2dHK/p)$, then with probability at least $1 - p/2$, we have :*

$$\sum_{k=1}^{K} V_h^k(x_1^k) - V_1^{\pi^k}(x_1^k) = \mathcal{O}(\sqrt{d^3H^4K\iota^2}). \tag{23}$$

In the next lemma, we show an upper bound on the entire "regret plus violation" term over K episodes using the results from Lemma 3 and Lemma 4.

Lemma 5. *Under the event $\mathcal{E}$ defined in Lemma 15 and condition 1, for any fixed $p \in (0,1)$, we set the parameters in our algorithm as indicated in Lemma 4, then with probability at least $1 - p/2$, we have:*

$$\begin{aligned}&\sum_{k=1}^{K} V_h^*(x_h^k) - V_h^{\pi^k}(x_h^k) + Z_h^k(\hat{g}_h^k(x_h^k, a_h^k)_+) \\ =&\mathcal{O}(\sqrt{d^3H^4K\iota^2})\end{aligned} \tag{24}$$

Proof. For any $h \in [H], k \in [K]$, according to the action selection (Eq.(11)) in our algorithm we have

$$\begin{aligned}&Q_h^k(x_h^k, a_h^k) - Z_h^k(\hat{g}_h^k(x_h^k, a_h^k)_+) \\ \geqslant& Q_h^k(x_h^k, a_h^*) - Z_h^k(\hat{g}_h^k(x_h^k, a_h^*)_+) \\ &+ Q_h^*(x_h^k, a_h^*) - Q_h^*(x_h^k, a_h^*),\end{aligned} \tag{25}$$

where a_h^* is the optimal action selected by the optimal policy π^*. Therefore rearranging the equation and subtracting $Q_h^{\pi^k}(x_h^k, a_h^k)$ at both sides we have:

$$\begin{aligned}&Q_h^*(x_h^k, a_h^*) - Q_h^{\pi^k}(x_h^k, a_h^k) + Z_h^k(\hat{g}_h^k(x_h^k, a_h^k)_+) \\ \leqslant& Q_h^*(x_h^k, a_h^*) - Q_h^k(x_h^k, a_h^*) &(26)\\ &+ Z_h^k(\hat{g}_h^k(x_h^k, a_h^*)_+) &(27)\\ &+ Q_h^k(x_h^k, a_h^k) - Q_h^{\pi^k}(x_h^k, a_h^k). &(28)\end{aligned}$$

Eq. (26) is nonpositive due to the overestimation Lemma 3. Eq. (27) is also nonpositive because the optimistic estimation of the cost function ensures that $Z_h^k(\hat{g}_h^k(x_h^k, a_h^*)_+) \leqslant Z_h^k(g_h(x_h^k, a_h^*)_+) \leqslant 0$. Bounding the last term (28) with Lemma 4 we prove the lemma. □

Using the results from Lemma 5, we are ready to prove the main results:

Regret: According to the results from Lemma 5, we have:

$$\begin{aligned}\text{Regret}(K) =& \sum_{k=1}^{K} V_1^{\pi^*}(x_1^k) - V_1^{\pi^k}(x_1^k) \\ =&\mathcal{O}(\sqrt{d^3H^4K\iota^2}) - Z_1^k(\hat{g}_1^k(x_1^k, a_1^*)_+ \\ =&\mathcal{O}(\sqrt{d^3H^4K\iota^2})\end{aligned} \tag{29}$$

Violation: Using the intermediate results in Lemma 5 (Eq. (26)-(28)) we have that:

$$\begin{aligned}\hat{g}_h^k(x_h^k, a_h^k)_+ \leqslant& \frac{1}{Z_h^k}\Big([Q_h^k(x_h^k, a_h^k) - Q_h^{\pi^k}(x_h^k, a_h^k)]) \\ &- [Q_h^*(x_h^k, a_h^*) - Q_h^{\pi^k}(x_h^k, a_h^k)]\Big) \\ \leqslant& \frac{1}{k}\Big|[Q_h^k(x_h^k, a_h^k) - Q_h^{\pi^k}(x_h^k, a_h^k)] \\ &- [Q_h^*(x_h^k, a_h^*) - Q_h^{\pi^k}(x_h^k, a_h^k)]\Big|.\end{aligned} \tag{30}$$

The inequality holds because our choice of Z_h^k in our algorithm such that $Z_h^k \geqslant \eta_k = k$. Therefore we have:

$$\text{Violation}(K) = \sum_{k=1}^{K}\sum_{h=1}^{H} g_h(x_h^k, a_h^k)_+$$

$$
\begin{aligned}
&= \sum_{k=1}^{K}\sum_{h=1}^{H}\left(g_h(x_h^k,a_h^k)-\hat{g}_h(x_h^k,a_h^k)+\hat{g}_h(x_h^k,a_h^k)\right)_+ \\
&\leqslant \sum_{k=1}^{K}\sum_{h=1}^{H}\left(g_h(x_h^k,a_h^k)-\hat{g}_h(x_h^k,a_h^k)\right)_+ \\
&\qquad + \sum_{k=1}^{K}\sum_{h=1}^{H}\hat{g}_h(x_h^k,a_h^k)_+ \\
&\leqslant \sum_{k=1}^{K}\sum_{h=1}^{H} e_h^k(p,x,a) + \sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{k}\Big|[Q_h^k(x_h^k,a_h^k) \\
&\qquad - Q_h^{\pi^k}(x_h^k,a_h^k)] - [(Q_h^*(x_h^k,a_h^*) - Q_h^{\pi^k}(x_h^k,a_h^k)]\Big| \\
&\leqslant \sum_{k=1}^{K}\sum_{h=1}^{H} e_h^k(p,x,a) + \sum_{k=1}^{K}\frac{2H^2}{k} \\
&\leqslant \sum_{k=1}^{K}\sum_{h=1}^{H} e_h^k(p,x,a) + 2H^2\log(K), \qquad (31)
\end{aligned}
$$

where the first inequality holds because of the fact $(a+b)_+ \leqslant a_+ + b_+$, the second inequality is due to Eq.(30), the third inequality is because of the assumption that reward is bounded by 1, and the last inequality is true by using the fact that $\sum_{k=1}^{K}\frac{1}{k} \leqslant \int_1^K \frac{1}{k}dk \leqslant \log(K)$.

Simulation

In this section, we evaluate the performance of our algorithm in the Frozen Lake environment (Amani, Thrampoulidis, and Yang 2021), as illustrated in Figure 1. The agent's objective is to navigate a 10×10 grid map to reach a goal while avoiding hazards. At each time step, four actions are available, with a 0.9 probability of moving in the intended direction, and a 0.05 probability for each orthogonal direction. For this simulation, we set $H = 15$, $K = 1000$, and $d = |\mathcal{S}| \times |\mathcal{A}|$. The feature vector is defined as $\phi(x,a) = e_{x,a}$, where $e_{x,a}$ is a d-dimensional vector with the element corresponding to the state-action pair (x,a) set to 1 and zero for other values. The agent receives a reward of 6 upon reaching the goal, and 0.01 otherwise. Taking dangerous actions (hitting the hazards) incurs a cost of 1, while safe actions result in a cost of -1. If the agent reaches the goal, it remains there until the end of the episode.

To highlight the benefits of our algorithm and its aggressive exploration strategy in addressing safe RL with instantaneous hard constraints, we compare our approach against two baselines:

- Classical Least-Squares Value Iteration (LSVI) (Jin et al. 2020a) without accounting for any constraints during learning.
- LSVI-Primal, representing the virtual queue (dual variable) update based on Eq. (14) in the traditional primal-dual/drift-plus-penalty approach for dealing with long-term or budget constraints in safe RL.

We present the results of our evaluation in Figure 2, depicting the moving average reward and the cost

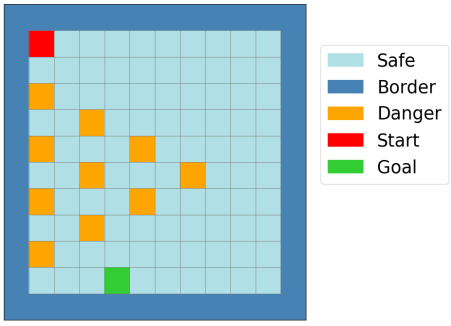

Figure 1: Frozen Lake Environment

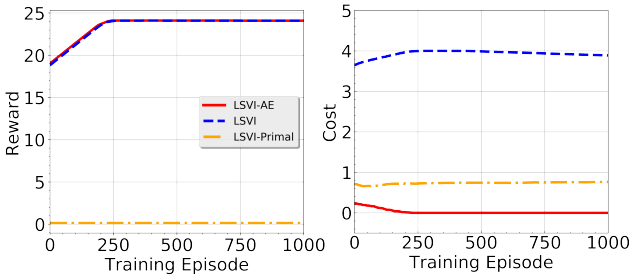

Figure 2: Reward and Cost Performance During Training

($\sum_{h=1}^{H} g_h(x_h,a_h)+$) return. Our LSVI-AE algorithm obtains an optimal reward comparable to that achieved by the LSVI algorithm designed for unconstrained MDPs. However, our approach significantly outperforms in terms of cost. Intriguingly, the LSVI-Primal approach designed for episodic constraint scenarios fails to perform effectively in this environment, exhibiting limited learning progress and only surpassing the unconstrained case in terms of cost. However, this cost improvement still fails to guarantee the desired performance of safe RL with instantaneous hard constraints, where the objective is to ensure $\sum_{h=1}^{H} g_h(x_h,a_h)_+ \leqslant 0$. These observations validate the key principles underlying our approach. A head map illustrating the exploration and exploitation of the agent can be found in the Appendix.

Conclusion

In this paper, we introduce LSVI-AE, an innovative algorithm designed for safe reinforcement learning with instantaneous hard constraints, addressing scenarios where no prior knowledge of safe actions or safe graphs is available. For the first time, we propose an adaptive penalty-based optimization with a double optimistic learning framework for taking care of this setting under a more general cost function. Our approach establishes both sub-linear regret bound and hard constraint violation bound, which both are optimal w.r.t K and match the information-theoretic lower bound. A notable feature of our approach lies in its emphasis on promoting aggressive policy exploration, contributing to the paradigm of algorithm design in this context.

References

Abbasi-yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems 24*.

Amani, S.; Alizadeh, M.; and Thrampoulidis, C. 2019. Linear stochastic bandits under safety constraints. In *Advances Neural Information Processing Systems (NeurIPS)*, 9256–9266.

Amani, S.; Thrampoulidis, C.; and Yang, L. 2021. Safe reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, 243–253. PMLR.

Andrychowicz, O. M.; Baker, B.; Chociej, M.; Jozefowicz, R.; McGrew, B.; Pachocki, J.; Petron, A.; Plappert, M.; Powell, G.; Ray, A.; et al. 2020. Learning dexterous in-hand manipulation. *The Int. Journal of Robotics Research*, 39(1): 3–20.

Bai, Q.; Bedi, A. S.; Agarwal, M.; Koppel, A.; and Aggarwal, V. 2022. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In *AAAI Conf. Artificial Intelligence*, 3682–3689.

Bura, A.; HasanzadeZonuzy, A.; Kalathil, D.; Shakkottai, S.; and Chamberland, J.-F. 2021. Safe exploration for constrained reinforcement learning with provable guarantees. *arXiv preprint arXiv:2112.00885*.

Caramanis, C.; Dimitrov, N. B.; and Morton, D. P. 2014. Efficient algorithms for budget-constrained markov decision processes. *IEEE Transactions on Automatic Control*, 59(10): 2813–2817.

Cesa-Bianchi, N.; and Lugosi, G. 2006. *Prediction, Learning, and Games*. Cambridge University Press.

Chen, L.; Jain, R.; and Luo, H. 2022. Learning Infinite-horizon Average-reward Markov Decision Process with Constraints. In *Int. Conf. Machine Learning (ICML)*, 3246–3270. PMLR.

Chowdhury, S. R.; and Gopalan, A. 2017. On kernelized multi-armed bandits. In *Int. Conf. Machine Learning (ICML)*, 844–853. PMLR.

Ding, D.; Wei, X.; Yang, Z.; Wang, Z.; and Jovanovic, M. 2021. Provably Efficient Safe Exploration via Primal-Dual Policy Optimization. In *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, volume 130, 3304–3312. PMLR.

Ding, D.; Zhang, K.; Basar, T.; and Jovanovic, M. 2020. Natural Policy Gradient Primal-Dual Method for Constrained Markov Decision Processes. In *Advances Neural Information Processing Systems (NeurIPS)*, volume 33, 8378–8390. Curran Associates, Inc.

Ding, D.; Zhang, K.; Başar, T.; and Jovanović, M. R. 2022. Convergence and optimality of policy gradient primal-dual method for constrained Markov decision processes. In *acc*, 2851–2856. IEEE.

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

Ghosh, A.; Zhou, X.; and Shroff, N. 2022. Provably Efficient Model-Free Constrained RL with Linear Function Approximation. In *NeurIPS*.

Guo, H.; Liu, X.; Wei, H.; and Ying, L. 2022. Online Convex Optimization with Hard Constraints: Towards the Best of Two Worlds and Beyond. In *Advances Neural Information Processing Systems (NeurIPS)*.

Guo, H.; Zhu, Q.; and Liu, X. 2022. Rectified Pessimistic-Optimistic Learning for Stochastic Continuum-armed Bandit with Constraints. *arXiv preprint arXiv:2211.14720*.

He, J.; Zhao, H.; Zhou, D.; and Gu, Q. 2022. Nearly minimax optimal reinforcement learning for linear markov decision processes. In *Int. Conf. Machine Learning (ICML)*, 12790–12822. PMLR.

Hu, P.; Chen, Y.; and Huang, L. 2022. Nearly minimax optimal reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, 8971–9019. PMLR.

Jin, C.; Liu, Q.; and Miryoosefi, S. 2021. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34: 13406–13418.

Jin, C.; Yang, Z.; Wang, Z.; and Jordan, M. I. 2020a. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, 2137–2143. PMLR.

Jin, C.; Yang, Z.; Wang, Z.; and Jordan, M. I. 2020b. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, 2137–2143. PMLR.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit Algorithms*. Cambridge University Press.

Liu, S.; Jiang, J.; and Li, X. 2022. Non-stationary bandits with knapsacks. In *Advances Neural Information Processing Systems (NeurIPS)*, volume 35, 16522–16532.

Liu, T.; Zhou, R.; Kalathil, D.; Kumar, P.; and Tian, C. 2021a. Learning policies with zero or bounded constraint violation for constrained MDPs. In *Advances Neural Information Processing Systems (NeurIPS)*, volume 34.

Liu, X.; Li, B.; Shi, P.; and Ying, L. 2021b. An Efficient Pessimistic-Optimistic Algorithm for Stochastic Linear Bandits with General Constraints. In *Advances Neural Information Processing Systems (NeurIPS)*.

Pacchiano, A.; Ghavamzadeh, M.; Bartlett, P.; and Jiang, H. 2021. Stochastic Bandits with Linear Constraints. In *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*.

Shi, M.; Liang, Y.; and Shroff, N. 2023. A Near-Optimal Algorithm for Safe Reinforcement Learning Under Instantaneous Hard Constraints. *iclm*.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. 2017. Mastering the game of go without human knowledge. *nature*, 550(7676): 354–359.

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

Turchetta, M.; Berkenkamp, F.; and Krause, A. 2016. Safe exploration in finite markov decision processes with gaussian processes. *Advances in neural information processing systems*, 29.

Vershynin, R. 2010. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Wachi, A.; Sui, Y.; Yue, Y.; and Ono, M. 2018. Safe exploration and optimization of constrained MDPs using Gaussian processes. In *AAAI Conf. Artificial Intelligence*, volume 32, 6548–6555. ISBN 978-1-57735-800-8.

Wei, H.; Ghosh, A.; Shroff, N.; Ying, L.; and Zhou, X. 2023. Provably Efficient Model-Free Algorithms for Non-stationary CMDPs. In *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, 6527–6570. PMLR.

Wei, H.; Liu, X.; and Ying, L. 2022a. A Provably-Efficient Model-Free Algorithm for Infinite-Horizon Average-Reward Constrained Markov Decision Processes. In *AAAI Conf. Artificial Intelligence*.

Wei, H.; Liu, X.; and Ying, L. 2022b. Triple-Q: A Model-Free Algorithm for Constrained Reinforcement Learning with Sublinear Regret and Zero Constraint Violation. In *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*.

Wu, D.; Chen, X.; Yang, X.; Wang, H.; Tan, Q.; Zhang, X.; Xu, J.; and Gai, K. 2018. Budget constrained bidding by model-free reinforcement learning in display advertising. In *Proc. ACM Int. Conf. Information and Knowledge Management (CIKM)*, 1443–1451.

Yang, Z.; Jin, C.; Wang, Z.; Wang, M.; and Jordan, M. I. 2020. On function approximation in reinforcement learning: Optimism in the face of large state spaces. *arXiv preprint arXiv:2011.04622*.

Yi, X.; Li, X.; Yang, T.; Xie, L.; Chai, T.; and Johansson, K. 2021. Regret and cumulative constraint violation analysis for online convex optimization with long term constraints. In *Int. Conf. Machine Learning (ICML)*, 11998–12008. PMLR.

Yi, X.; Li, X.; Yang, T.; Xie, L.; Chai, T.; and Karl, H. 2022. Regret and cumulative constraint violation analysis for distributed online constrained convex optimization. *IEEE Transactions on Automatic Control*.

Yu, H.; and Neely, M. J. 2020. A Low Complexity Algorithm with $O(\sqrt{T})$ Regret and $O(1)$ Constraint Violations for Online Convex Optimization with Long Term Constraints. *Journal of Machine Learning Research*, 21(1): 1–24.

Zhou, D.; Gu, Q.; and Szepesvari, C. 2021. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *colt*, 4532–4576. PMLR.