Doubly Pessimistic Algorithms for Strictly Safe Off-Policy Optimization

Sanae Amani
University of California, Los Angeles
samani@ucla.edu

Lin F. Yang
University of California, Los Angeles
linyang@ee.ucla.edu

Abstract

We study offline reinforcement learning (RL) in the presence of safety requirements: from a dataset collected a priori and without direct access to the true environment, learn an optimal policy that is guaranteed to respect the safety constraints. We address this problem by modeling the safety requirement as an unknown cost function of states and actions, whose expected value with respect to the policy must fall below a certain threshold. We then present an algorithm in the context of finite-horizon Markov decision processes (MDPs), termed Safe-DPVI that performs in a doubly pessimistic manner when 1) it constructs a conservative set of safe policies; and 2) when it selects a good policy from that conservative set. Without assuming the sufficient coverage of the dataset or any structure for the underlying MDPs, we establish a data-dependent upper bound on the suboptimality gap of the *safe* policy Safe-DPVI returns. We then specialize our results to linear MDPs with appropriate assumptions on dataset being well-explored. Both data-dependent and specialized bounds nearly match that of state-of-the-art unsafe offline RL algorithms, with an additional multiplicative factor $\frac{\sum_{h=1}^{H}\alpha_h}{H}$, where α_h characterizes the safety constraint at time-step h. We further present numerical simulations that corroborate our theoretical findings.

1 Introduction

Offline/Batch reinforcement learning (RL) as a method that uses previously collected datasets in many real-world decision-making applications where obtaining new experiences is costly has received significant attention Lange et al. (2012). For example, the outcome of a treatment in clinical trials can be evaluated only after several years and thus, a bad decision can cause long-term damages. The main focus of offline RL has been on two directions: 1) offline policy evaluation, which aims at estimating value functions of a target policy, and 2) offline policy optimization, which aims to find an optimal policy that maximizes the expected cumulative reward. A key challenge in offline RL is to address the issue of insufficient coverage in the dataset Wang et al. (2020) due to the lack of exploration in data collecting process. There has been a surge of research activities investigating appropriate conditions on the data collecting process to guarantee an efficient and successful learning either in policy evaluation or policy optimization regions. For example, see Duan et al. (2020); Yang et al. (2020); Zhang et al. (2020); Yin and Wang (2020); Yin et al. (2021).

Most of the existing offline RL methods in the more challenging category of offline policy optimization find a policy that under certain coverage assumptions perform well or at least as well as the behavior policy based on which the available dataset has been collected Lange et al. (2012); Fujimoto et al. (2019); Kumar et al. (2019); Yu et al. (2020); Rafailov et al. (2020); Jin et al. (2020b); Kumar et al. (2020). However, the learned policy in the above-mentioned works explores all possible actions, while freely exploring all actions may be harmful in many real-world systems where playing even one unsafe action may lead to catastrophic results. Safety in RL has become increasingly important in

Offline Reinforcement Learning Workshop at Neural Information Processing Systems, 2021.

recent years. Yet, many of existing solutions fail to strictly avoid choosing unsafe policies, which may lead to catastrophic results in safety-critical systems. Thus, safety in offline RL has become a serious issue that has remained unexplored and restricted the applicability of offline RL algorithms to many risk-sensitive real-world systems. For example, in a self-driving car, it is critical to only explore those policies that avoid crash and damage to the car, people and property. Switching cost limitations in medical applications Bai et al. (2019) and legal restrictions in financial managements Abe et al. (2010) are other examples of safety-critical applications. All the aforementioned safety-critical environments introduce the new challenge of balancing the goal of reward maximization with the restriction of playing safe actions and studying the influence of safety constraints in the sample complexity of finding an optimal safe policy.

In this paper, we focus on a strong notion of safety requirement which is modeled as an unknown cost function of states and actions, whose expected value with respect to the learned policy must fall below a certain threshold at each time-step an action is played with high probability. We then propose a Safe Doubly Pessimistic Value Iteration (Safe-DPVI) algorithm that performs pessimistically when 1) it constructs a conservative set of safe policies; and 2) when it selects a good policy from that conservative set in the value iteration step. Without assuming the sufficient coverage of the dataset or any structure for the underlying Markov decision processes (MDPs), we establish a data-dependent upper bound on the suboptimality gap of the *safe* policy Safe-DPVI returns. We then specialize our results to linear MDPs with appropriate assumptions on dataset being well-explored. We prove that both data-dependent and specialized upper bounds are order-wise comparable to those of their unsafe counter-parts.

Notation. We start by introducing a set of notations that are used throughout the paper. We use lower-case letters for scalars, lower-case bold letters for vectors, and upper-case bold letters for matrices. The Euclidean-norm of $\mathbf{x}$ is denoted by $\|\mathbf{x}\|_2$. We denote the transpose of any column vector $\mathbf{x}$ by $\mathbf{x}^\top$. For any vectors $\mathbf{x}$ and $\mathbf{y}$, we use $\langle \mathbf{x}, \mathbf{y} \rangle$ to denote their inner product. Let $\mathbf{A}$ be a positive definite $d \times d$ matrix and $\boldsymbol{\nu} \in \mathbb{R}^d$. The weighted 2-norm of $\boldsymbol{\nu}$ with respect to $\mathbf{A}$ is defined by $\|\boldsymbol{\nu}\|_{\mathbf{A}} = \sqrt{\boldsymbol{\nu}^\top \mathbf{A} \boldsymbol{\nu}}$. We denote the minimum and maximum eigenvalue of $\mathbf{A}$ by $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$. For positive integer n, $[n]$ denotes the $\{1, 2, \ldots, n\}$. For a real number α, we denote by α^+ the maximum of α and zero.

1.1 Problem Statement

In this section, we first introduce the standard episodic Markov decision process (MDP) which is augmented by an extra safety/cost function and describe the data collecting process based on a *behavior* policy in the underlying MDP. Then, we introduce safety constraint which must be satisfied at all time-steps that actions are played with high probability. Finally, we introduce the performance metric.

Episodic Markov decision process. We consider an episodic Markov decision process (MDP) denoted by $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, R, C)$, where $\mathcal{S}$ is the state set, $\mathcal{A}$ is the action set, H is the length of each episode (horizon), $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ are the transition probabilities, $R = \{R_h\}_{h=1}^H$ are the reward functions, and $C = \{C_h\}_{h=1}^H$ are the safety/cost functions, where $R_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ and $C_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$. For each time-step $h \in [H]$, $\mathbb{P}_h(s'|s, a)$ denotes the probability of transitioning to state s' upon playing action a at state s. At each time-step $h \in [H]$, the agent observes the state s_h, plays an action $a_h \in \mathcal{A}$, and observes the next state $s_{h+1} \sim \mathbb{P}_h(.|s_h, a_h)$, a reward $r_h := R_h(s_h, a_h) + \eta_h$, and a cost $c_h := C_h(s_h, a_h) + \epsilon_h$, where η_h and ϵ_h are random additive noise. We consider a learning problem, where $\mathcal{S}$ and $\mathcal{A}$ are known, while the transition probabilities $\mathbb{P}_h$, rewards R_h and costs C_h are *unknown* to the agent and must be learned from a given dataset $\mathcal{D}$. The dataset $\mathcal{D} := \left\{s_h^k, a_h^k, r_h^k, c_h^k\right\}_{h,k=1}^{H,K}$ is collected from K i.i.d. trajectories under a *behavior* policy denoted as $\bar{\pi}$.

Safety Constraint. We assume that the underlying system is safety-critical and the environment is subject to a side constraint that restricts the choice of policies. A policy $\pi := \{\pi_h\}_{h=1}^H$, where $\pi_h : \mathcal{S} \to \Delta_{\mathcal{A}}$ maps $\mathcal{S}$ to distributions over $\mathcal{A}$, is called *safe* if

$$\mathbb{E}_{a \sim \pi_h(.|s)} \left[C_h(s, a)\right] \leq \tau, \; \forall (s, h) \in \mathcal{S} \times [H] \tag{1}$$

with high probability. We accordingly define the *unknown* set of safe policies by $\Pi^{\text{safe}} := \left\{\pi : \pi_h(.|s) \in \Gamma_h^{\text{safe}}(s),\ \forall (s,h) \in \mathcal{S} \times [H]\right\}$, where

$$\Gamma_h^{\text{safe}}(s) := \left\{\theta(.|s) \in \Delta_{\mathcal{A}} : \mathbb{E}_{a\sim\theta(.|s)}\left[C_h(s,a)\right] \leq \tau\right\}. \tag{2}$$

Thus, after observing state s_h at time-step $h \in [H]$, the agent's choice of policy must belong to $\Gamma_h^{\text{safe}}(s_h)$ with high probability. As a motivating example, consider a self-driving car. On the one hand, the agent (car) is rewarded for getting from point one to point two as fast as possible. On the other hand, the driving behavior must be constrained to respect traffic safety standards.

Performance Metric. We define the state-action and state value function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $V_h^\pi : \mathcal{S} \to \mathbb{R}$ for a policy π at time-step $h \in [H]$ by

$$Q_h^\pi(s,a) := \mathbb{E}\left[\sum_{h'=h+1}^{H} r_{h'}\left(s_{h'}, a_{h'}\right)\middle| s_h = s, a_h = a, \pi\right],$$

$$V_h^\pi(s) := \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}\left(s_{h'}, a_{h'}\right)\middle| s_h = s, \pi\right],$$

where the expectation is over the environment and the randomness of policy π. To simplify the notation, for any function f, we denote $[\mathbb{P}_h f](s,a) := \mathbb{E}_{s'\sim\mathbb{P}_h(.|s,a)} f(s')$ and $[\mathbb{B}_h f](s,a) := R_h(s,a) + [\mathbb{P}_h f](s,a)$. Let π_* be the optimal *safe* policy such that $V_h^{\pi_*}(s) := V_h^*(s) = \sup_{\pi\in\Pi^{\text{safe}}} V_h^\pi(s)$ for all $(s,h) \in \mathcal{S} \times [H]$. Thus, for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, the Bellman equations for the optimal safe policy and an arbitrary policy $\pi \in \Pi^{\text{safe}}$ are:

$$Q_h^*(s,a) = [\mathbb{B}_h V_{h+1}^*](s,a), \quad V_h^*(s) = \max_{\theta(.|s)\in\Gamma_h^{\text{safe}}(s)} \mathbb{E}_{a\sim\theta(.|s)}\left[Q_h^*(s,a)\right] \tag{3}$$

$$Q_h^\pi(s,a) = [\mathbb{B}_h V_{h+1}^\pi](s,a), \quad V_h^\pi(s) = \mathbb{E}_{a\sim\pi_h(.|s)}\left[Q_h^\pi(s,a)\right], \tag{4}$$

where $V_{H+1}^\pi(s) = V_{H+1}^*(s) = 0$. Our goal is to learn a *safe* policy that maximizes the cumulative expected reward given the collected dataset $\mathcal{D}$. To this end, we define the following suboptimality gap of a safe policy π given by the initial state $s_1 = s$ as

$$\Delta(\pi; s) := V_1^*(s) - V_1^\pi(s). \tag{5}$$

1.2 Related Works

Safety in Online RL: In online setting, the problem of Safe RL formulated with Constrained Markov Decision Process (CMDP) is studied in Efroni et al. (2020); Turchetta et al. (2020); Garcelon et al. (2020); Zheng and Ratliff (2020); Ding et al. (2020a); Qiu et al. (2020); Ding et al. (2020b); Xu et al. (2020); Kalagarla et al. (2020); Liu et al. (2021). In the above-mentioned papers, the goal is to find the optimal policy in an online manner that maximizes the reward value function $V_r^\pi(s)$ (expected total reward) over the safe policies that satisfy $V_c^\pi(s) \leq b$, where $V_c^\pi(s)$ is the cumulative expected cost over an entire episode with duration H and b is a threshold. This safety requirement is defined over an *entire* episode, and consequently is less strict than the safety requirement considered in this paper, which must be satisfied at each time-step an action is played.

Safety in Offline RL: The notion of safety has been used in several existing offline RL works. However, they fundamentally differ from the definition of safety considered in our paper. For example, safety in Laroche et al. (2019); Thomas et al. (2015); Ghavamzadeh et al. (2016) means the algorithm returns a policy with performance at least as good as that of behavior/baseline policy, based on which the dataset has been collected. In another line of work, Srinivasan et al. (2020); Thananjeyan et al. (2021) empirically study safety-constrained RL problem and propose algorithms that consist of two distinct offline and online phases and aim to find a an optimal policy for which the expected value of the number of unsafe states visits is less than some threshold $\epsilon \in (0,1)$ in the context of discounted MDPs with discount factor γ. In the offline phase, they estimate the safe set of policies form an available dataset, and then in the online phase, they seek to find the best policy based on the estimated safe set of policies.

```
Algorithm 1: Safe Doubly Pessimistic Value Iteration
Input: Dataset $\mathcal{D} = \left\{s_h^k, a_h^k, r_h^k, c_h^k\right\}_{h,k=1}^{H,K}$
Initialization: $\hat{V}_{H+1}(s) = 0,\ \forall s \in \mathcal{S}$
for time-steps $h = H, \dots, 1$ do
    Compute $[\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a)$, $B'_h(s,a)$ and $\hat{\Gamma}_h(s), \forall (s,a) \in \mathcal{S} \times \mathcal{A}$.    // as defined in Section 3.1 for underlying linear MDP.

    Set $\hat{Q}_h(s,a) = \left\{[\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a) - B'_h(s,a)\right\}^+,\ \forall (s,a) \in \mathcal{S} \times \mathcal{A}$.
    Set $\hat{\pi}_h(.|s) = \arg\max_{\theta(.|s) \in \hat{\Gamma}_h(s)} \mathbb{E}_{a \sim \theta(.|s)}\left[\hat{Q}_h(s,a)\right],\ \forall s \in \mathcal{S}$.
    Set $\bar{V}_h(s) = \mathbb{E}_{a \sim \hat{\pi}_h(.|s)}\left[\hat{Q}_h(s,a)\right],\ \forall s \in \mathcal{S}$.
    Set $\hat{V}_h(s) = \min\left\{\bar{V}_h(s), H\right\},\ \forall s \in \mathcal{S}$.
Output: $\hat{\pi} = \{\hat{\pi}_h\}_{h=1}^H$
```

Per time-step vs Per Episode Safety Constraints: The definition of safety constraints studied in all the above-stated papers is a special case of the notion of safety considered in our paper. For example, if $C(s,a)$ and τ in (1) are set to be the probability of transitioning to an unsafe state by playing action a at state s and $\epsilon(1-\gamma)$ would recover the safety constraint in Thananjeyan et al. (2021) for infinite-horizon discounted MDPs. Furthermore, in all the online safe papers, having $\tau = b/H$ in the definition of safety constraint considered in our paper in (1) would recover $V_c^\pi(s) \leq b$. Therefore, the safety requirement considered in our paper is much stricter than those in the existing literature, and naturally covers a wider range of applications.

2 Safe-DPVI: A General Framework for Safe Offline Policy Optimization

In this section, we formally present Safe Doubly Pessimistic Value Iteration (Safe-DPVI), summarized in Algorithm 1, that employs the dataset and returns a *safe* policy $\hat{\pi}$. We then introduce two uncertainty quantifiers based on which, we are able to control $\Delta(\hat{\pi}; s)$ and state our main results on the suboptimality gap's bound in Theorem 1.

First, we introduce the following assumption, which is necessary to ensure that the safety constraint in (1) is satisfied from the very first time-step.

Assumption 1 (Non-empty safe sets)**.** *There exists a known safe policy π^0 with known costs $\tau_h(s) := \mathbb{E}_{a \sim \pi_h^0(.|s)}\left[C_h(s,a)\right] < \tau$. Thus, the sets $\Gamma_h^{\text{safe}}(s)$ are non-empty, as $\pi_h^0(s) \in \Gamma_h^{\text{safe}}(s)$.*

This assumption is rather standard and has been widely used in the literature of safe online RL Amani et al. (2021); Liu et al. (2021) and safe bandits Amani et al. (2019); Pacchiano et al. (2021). This assumption is also realistic in many practical examples, where the known safe policy could be the one suggested by the current strategy of the company or a very cost-neutral policy that does not necessarily have high reward but its cost is far from the threshold. Note that the known safe policy π^0 is not necessarily the same as the behavior policy $\bar{\pi}$. If the behavior policy $\bar{\pi}$ is also safe, we can simply treat it as the known safe policy, i.e., $\bar{\pi} = \pi^0$. In Appendix C, we show that it is possible to relax the assumption of knowing the costs of the safe policy $\tau_h(s)$ and when $\pi^0 = \bar{\pi}$, this relaxation naturally goes through.

2.1 Overview

From a high-level point of view, based on dataset $\mathcal{D}$, Safe-DPVI constructs estimated cost functions $\hat{C}_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, Q-functions $\hat{Q}_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, value functions $\hat{V}_h : \mathcal{S} \to \mathbb{R}$, and Bellman operator $\hat{\mathbb{B}}_h$ such that $[\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a)$ approximates $[\mathbb{B}_h \hat{V}_{h+1}](s,a)$. Note that the algorithm only relies on construction of $[\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a)$ not $\hat{\mathbb{B}}_h$ itself. The algorithm constructs an estimated set of safe policies $\hat{\Pi}$ based on estimated cost functions $\hat{C}_h$. To see how this happens, we first define the following δ-safety uncertainty quantifier and δ-bellman uncertainty quantifier for $\delta \in (0,1)$

that quantify the uncertainty arising from approximating the cost function C and $[\mathbb{B}_h\hat{V}_{h+1}](s,a)$, respectively.

Definition 1 (Uncertainty quantifiers)**.** *For a fixed $\delta \in (0,1)$, we call $B = \{B_h\}_{h=1}^H$, where $B_h : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$, a δ-safety uncertainty quantifier if $\mathbb{P}\left(\left|C_h(s,a)-\hat{C}_h(s,a)\right| \leq B_h(s,a), \forall(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]\right) \geq 1-\delta$. We also call $B' = \{B'_h\}_{h=1}^H$, where $B'_h : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$, a δ-bellman uncertainty quantifier if $\mathbb{P}\left(\left|[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a)-[\mathbb{B}_h\hat{V}_{h+1}](s,a)\right| \leq B'_h(s,a), \forall(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]\right) \geq 1-\delta$.*

Thus, if the agent can compute a δ-safety uncertainty quantifier B based on the dataset $\mathcal{D}$ and $u_h^c(s,a) = C_h(s,a)+B_h(s,a)$, then, a natural approximation for Π^{safe} is $\hat{\Pi} := \left\{\pi : \pi_h(.|s)\in\hat{\Gamma}_h(s),\ \forall(s,h)\in\mathcal{S}\times[H]\right\}$, where

$$\hat{\Gamma}_h(s) := \pi_h^0(s) \cup \left\{\theta(.|s)\in\Delta_{\mathcal{A}} : \mathbb{E}_{a\sim\theta(.|s)}\left[u_h^c(s,a)\right] \leq \tau\right\}. \tag{6}$$

Thus, Safe-DPVI constructs $\hat{\Gamma}_h(s)$ pessimistically as it relies on $u_h^c(s,a)$, which is an upper confidence bound on $C_h(s,a)$. Note that $\hat{\Gamma}_h(s)$ is not empty as it includes $\pi_h^0(s)$.

Next time Safe-DPVI applies pessimism is when it computes $\hat{Q}_h$ by incorporating δ-bellman uncertainty quantifier B' into the value iteration step as follows

$$\hat{Q}_h(s,a) = \left\{[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B'_h(s,a)\right\}^+. \tag{7}$$

After the construction of $\hat{\Gamma}_h$ and $\hat{Q}_h$, the algorithm is ready to return the *safe* policy $\hat{\pi} = \{\hat{\pi}_h\}_{h=1}^H$, where $\hat{\pi}_h(.|s) = \arg\max_{\theta(.|s)\in\hat{\Gamma}_h(s)} \mathbb{E}_{a\sim\theta(.|s)}\left[\hat{Q}_h(s,a)\right]$, as its output. In the following theorem, we characterize the safeness and suboptimality gap of Safe-DPVI.

Theorem 1. *Fix $\delta\in(0,0.5)$. Let B and B' be δ-safety uncertainty quantifier and δ-bellman uncertainty quantifier, respectively, $\hat{\pi}$ be the output of Algorithm 1, $\alpha_h = 2+\frac{2H}{\tau-\max_{s\in\mathcal{S}}\tau_h(s)}$ and $\bar{B}_h(s,a) := \max\left\{B_h(s,a), B'_h(s,a)\right\}$. Then, under Assumption 1, with probability at least $1-2\delta$, it holds that:*

1. *$\hat{\pi}$ is safe;*
2. *$\Delta(\hat{\pi};s) \leq \max\left\{\sum_{h=1}^H \alpha_h\mathbb{E}\left[\bar{B}_h(s_h,a_h)|s_1=s,\pi^*\right], \sum_{h=1}^H \alpha_h\mathbb{E}\left[\bar{B}_h(s_h,a_h)|s_1=s,\pi^0\right]\right\}$.*

While point 1 is directly proven from the definition of δ-safety uncertainty quantifier B in Definition 1, respectively, the proof of point 2 is more intricate and challenging. The complete proof is given in Appendix A.3. In the following section, we give a proof sketch.

Before that, we comment on the suboptimality gap of Algorithm 1 and how it compares to that of its *unsafe* counterpart in Jin et al. (2020b). The bound on PEVI's suboptimality gap in Jin et al. (2020b) is $2\sum_{h=1}^H \mathbb{E}\left[B'_h(s_h,a_h)|s_1=s,\pi^*\right]$. We observe that our bound is comparable with that of PEVI with the following differences: 1) Instead of B'_h, our bound includes $\alpha_h\bar{B}_h$ to account for the uncertainty regarding the additional unknown safety constraints we have to deal with in our setting; 2) Moreover, we take the maximum of the expected value of the uncertainty of trajectories induced by both the optimal safe policy π^* and the known safe policy π^0, which once again highlights the role of the known safe policy in Safe-DPVI's performance.

2.2 Proof Sketch of Theorem 1

For the proof of point 1, recall the definition of $\hat{\Gamma}_h(s)$ in (6). Since B is a δ-safety uncertainty quantifier, for all $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$, $u_h^c(s,a)$ is an upper bound on $C_h(s,a)$ with probability at least $1-\delta$. Thus, one of the following two cases occurs: 1) $\hat{\pi}_h(.|s) = \pi_h^0(.|s)$, which implies that $\mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[C_h(s,a)\right] = \tau_h(s) < \tau$; 2) $\hat{\pi}_h(.|s) \in \left\{\theta(.|s)\in\Delta_{\mathcal{A}} : \mathbb{E}_{a\sim\theta(.|s)}\left[u_h^c(s,a)\right]\leq\tau\right\}$, which implies that with probability at least $1-\delta$, it holds that $\mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[C_h(s,a)\right] \leq \mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[u_h^c(s,a)\right] \leq \tau$. This concludes point 1 of Theorem 1.

For the proof of point 2, we employ the definitions of uncertainty quantifiers in Definition 1. First, consider a meta-algorithm that employs the dataset to construct an estimated Q-function $\hat{Q}_h : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ and an estimated value function $\hat{V}_h : \mathcal{S}\to\mathbb{R}$. We let $\iota_h(s,a) := [\mathbb{B}_h\hat{V}_{h+1}](s,a) - \hat{Q}_h(s,a)$ be the model evaluation error. In the following lemma, we decompose the upper bound of the suboptimality gap of a policy $\hat{\pi}$ into two key terms.

Lemma 1 (Suboptimality Gap's Upper Bound Decomposition)**.** *Let $\hat{\pi}$ be a policy such that $\hat{V}_h(s) = \min\left\{\mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[\hat{Q}_h(s,a)\right], H\right\}$ for all $(s,h)\in\mathcal{S}\times[H]$. Then, it holds that*

$$\Delta(\hat{\pi};s) \leq \underbrace{V_1^*(s) - \hat{V}_1(s)}_{\text{Term I}} + \underbrace{\sum_{h=1}^{H}\mathbb{E}\left[-\iota_h(s_h,a_h)\Big|s_1=s,\hat{\pi}\right]}_{\text{Term II}}.$$

Note that $\Delta(\hat{\pi};s) = V_1^*(s) - \hat{V}_1(s) + \hat{V}_1(s) - V_1^{\hat{\pi}}(s)$. Thus, the proof is complete once we show $\hat{V}_1(s) - V_1^{\hat{\pi}}(s) \leq$ Term II. We report the complete proof in Appendix A.1.

Now, recall that $\iota_h(s,a) := [\mathbb{B}_h\hat{V}_{h+1}](s,a) - \hat{Q}_h(s,a)$. Thus, ι_h and $\hat{\pi}$ are correlated as they both depend on the dataset $\mathcal{D}$ and thus, the expectation in Term II can be rather large. The definition of δ-bellman uncertainty quantifier B' and the pessimism in computation of $\hat{Q}_h(s,a)$ helps us eliminate Term II. Note that if $[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B_h'(s,a) < 0$, then $\hat{Q}_h(s,a) = 0$ and therefore $-\iota_h(s,a) = -[\mathbb{B}_h\hat{V}_{h+1}](s,a) \leq 0$ as $\hat{V}_h(s)\geq 0$ for all $(s,h)\in\mathcal{S}\times[H]$. Now, suppose $[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B_h'(s,a) \geq 0$. Since B' is a δ-bellman uncertainty quantifier, we have

$$\begin{aligned}
-\iota_h(s,a) &= \hat{Q}_h(s,a) - [\mathbb{B}_h\hat{V}_{h+1}](s,a)\\
&= [\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B_h'(s,a) - [\mathbb{B}_h\hat{V}_{h+1}](s,a)\\
&\leq 0.
\end{aligned}$$

This concludes that for all $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$, with probability at least $1-\delta$, it holds that $-\iota_h(s,a)\leq 0$, and therefore Term II $= \sum_{h=1}^{H}\mathbb{E}\left[-\iota_h(s_h,a_h)\Big|s_1=s,\hat{\pi}\right] \leq 0$.

Our main technical contribution towards bounding $\Delta(\hat{\pi};s)$ is given in the following lemma that is used to prove the more challenging point 2 of Theorem 1.

Lemma 2. *Fix $\delta\in(0,0.5)$. Let B and B' be δ-safety uncertainty quantifier and δ-bellman uncertainty quantifier, respectively. Also, let $\bar{B}_h(s,a) = \max\left\{B_h(s,a), B_h'(s,a)\right\}$ and*

$$F_h(s) := \max\left\{\sum_{h'=h}^{H}\alpha_{h'}\mathbb{E}\left[\bar{B}_{h'}(s_{h'},a_{h'})|s_h=s,\pi^*\right], \sum_{h'=h}^{H}\alpha_{h'}\mathbb{E}\left[\bar{B}_{h'}(s_{h'},a_{h'})|s_h=s,\pi^0\right]\right\}.$$

Then, under Assumption 1 and provided that $\alpha_h = 2+\frac{2H}{\tau-\max_{s\in\mathcal{S}}\tau_h(s)}$, with probability at least $1-2\delta$, it holds that

$$V_h^*(s) - \hat{V}_h(s) \leq F_h(s),\ \forall(s,h)\in\mathcal{S}\times[H]. \tag{8}$$

In safe off-policy optimization, the safe set Π^{safe} is not known. Therefore, at each time-step, the agent's policy must be chosen from a conservative inner approximation of Π^{safe}. Intuitively, the better this approximation is, the more likely that the output policy of Safe-DPVI leads to small suboptimality gap, ideally of the same order as that of PEVI proposed by Jin et al. (2020b) in the classical offline RL setting.

In order to better highlight the challenging part of our analysis compared to classical setting without safety constraint, we observe that for all $(s,h) \in \mathcal{S} \times [H]$, with probability at least $1-\delta$, it holds that $V_h^*(s) - \hat{V}_h(s) \leq \text{Term i} + \text{Term ii}$, where $\text{Term i} = \min\left\{\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[\hat{Q}_h(s,a)\right], H\right\} - \min\left\{\mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[\hat{Q}_h(s,a)\right], H\right\}$ and $\text{Term ii} = \mathbb{E}_{a\sim\pi_h^*(.|s)}\left[2B'_h(s,a) - \left[\mathbb{P}_h\left(\hat{V}_{h+1} - V^*_{h+1}\right)\right](s,a)\right]$.

A key difference in the analysis of Safe-DPVI compared to the classical offline RL without safety constraint is that $\pi_h^*(.|s_h)$ may not lie within the estimated safe set $\hat{\Gamma}_h(s_h)$, which makes controlling Term i and Term ii more delicate. This complication lies at the heart of the new formulation with additional safety constraints. When safety constraints are absent, classical pessimistic offline RL algorithms such as PEVI in Jin et al. (2020b) guarantee that Term i is non-positive and by induction it can be shown that Term ii$\leq 2\sum_{h'=h}^{H}\mathbb{E}\left[B'_{h'}(s_{h'},a_{h'})|s_h = s, \pi^*\right]$. Unfortunately, this is not the case here as $\pi_h^*(.|s_h)$ does not necessarily belong to $\hat{\Gamma}_h(s_h)$, thus Term i can be positive, which also affects the bound on Term ii. This extra positive term in the suboptimality gap is the price paid by Safe-DPVI for choosing safe policies at each time-step $h \in [H]$. The proof of Lemma 2 is given in Appendix A.2.

Employing the results in Lemmas 1 and 2, and under the setting of Theorem 1, we are now ready to bound the suboptimality gap $\Delta(\hat{\pi}; s)$ by $F_1(s)$, which concludes the proof of point 2 of Theorem 1.

3 Safe-DPVI: Linear MDP

In this section, we specialize Safe-DPVI and its theoretical guarantees to the case where the underlying MDP is linear Bradtke and Barto (1996); Yang and Wang (2019); Jin et al. (2020a). We further determine sufficient conditions that allow us to derive finite sample complexity for Safe-DPVI with an underlying linear MDP.

Definition 2 (Linear MDP). *$M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, R, C)$ is a linear MDP with feature map $\boldsymbol{\phi} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$, if for any $h \in [H]$, there exist d unknown measures $\boldsymbol{\mu}_h^* := [\mu_h^{*(1)}, \dots, \mu_h^{*(d)}]^\top$ over $\mathcal{S}$, and unknown vectors $\boldsymbol{\theta}_h^*, \boldsymbol{\zeta}_h^* \in \mathbb{R}^d$ such that $\mathbb{P}_h(.|s,a) = \langle \boldsymbol{\mu}_h^*(.), \boldsymbol{\phi}(s,a)\rangle$, $R_h(s,a) = \langle \boldsymbol{\theta}_h^*, \boldsymbol{\phi}(s,a)\rangle$, and $C_h(s,a) = \langle \boldsymbol{\zeta}_h^*, \boldsymbol{\phi}(s,a)\rangle$.*

3.1 Overview

We introduce the quantities that Safe-DPVI constructs based on the dataset $\mathcal{D}$ when the underlying MDP is linear. Recall that $\hat{\Gamma}_h(s)$ in (6) depends on $\hat{C}_h(s,a)$, an approximation of $C_h(s,a)$, and $B_h(s,a)$. In particular, Safe-DPVI constructs

$$\hat{C}_h(s,a) = \left\langle \hat{\boldsymbol{\zeta}}_h, \boldsymbol{\phi}(s,a)\right\rangle, \tag{9}$$

where $\hat{\boldsymbol{\zeta}}_h := \arg\min_{\boldsymbol{\nu}\in\mathbb{R}^d} \sum_{k=1}^{K}\left(\left\langle \boldsymbol{\nu}, \boldsymbol{\phi}(s_h^k, a_h^k)\right\rangle - c_h^k\right)^2 + \lambda\|\nu\|_2^2$ is the least square estimator of $\boldsymbol{\zeta}_h^*$ with regularization parameter $\lambda > 1$ and has the closed form

$$\hat{\boldsymbol{\zeta}}_h := \Lambda_h^{-1}\left(\sum_{k=1}^{K}\boldsymbol{\phi}(s_h^k, a_h^k).\ c_h^k\right), \tag{10}$$

where $\Lambda_h = \lambda I + \sum_{k=1}^{K}\boldsymbol{\phi}(s_h^k, a_h^k)\boldsymbol{\phi}(s_h^k, a_h^k)^\top$. Moreover, Safe-DPVI computes

$$[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) = \left\langle \hat{\mathbf{w}}_h, \boldsymbol{\phi}(s,a)\right\rangle,\ B_h(s,a) = \beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\Lambda_h^{-1}},\ B'_h(s,a) = \beta'\left\|\boldsymbol{\phi}(s,a)\right\|_{\Lambda_h^{-1}}, \tag{11}$$

where $\hat{\mathbf{w}}_h$ is the minimizer of the empirical mean squared Bellman error (MSBE), with closed form

$$\hat{\mathbf{w}}_h := \Lambda_h^{-1}\left(\sum_{k=1}^{K}\boldsymbol{\phi}(s_h^k, a_h^k).\left[r_h^k + \hat{V}_{h+1}(s_{h+1}^k)\right]\right), \tag{12}$$

and $\beta, \beta' > 0$ are scaling parameters that will be defined shortly in Theorem 2.

3.2 Theoretical guarantees

Now, we specialize our results in Theorem 1 to the case of linear MDP and provide a sample complexity for Safe-DPVI when the underlying MDP is linear and certain conditions hold. First, we make the remaining necessary assumptions under which our proposed algorithm operates and achieves small suboptimality gap.

Assumption 2 (Subgaussian Noise). *For all $(h,k) \in [H] \times [K]$, η_h^k and ϵ_h^k are zero-mean σ-subGaussian random variables.*

Assumption 3 (Boundedness). *Without loss of generality, $\left\|\phi(s,a)\right\|_2 \leq 1$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $\max\left(\left\|\boldsymbol{\mu}_h^*(\mathcal{S})\right\|_2, \left\|\boldsymbol{\theta}_h^*\right\|_2, \left\|\boldsymbol{\zeta}_h^*\right\|_2\right) \leq \sqrt{d}$ for all $h \in [H]$.*

Assumption 4 (Well-Explored Dataset). *There exists an absolute constant $\bar{c} > 0$ such that $\lambda_{\min}(\Sigma_h) \geq \bar{c}$, $\forall h \in [H]$, where $\Sigma_h = \mathbb{E}_{\bar{\pi}}\left[\phi(s_h,a_h)\phi(s_h,a_h)^\top\right]$ and $\mathbb{E}_{\bar{\pi}}$ is the expectation taken with respect to the trajectory induced by behavior policy $\bar{\pi}$.*

Assumptions 2 and 3 are standard in linear MDP and bandit literature Jin et al. (2020a); Pacchiano et al. (2021); Amani et al. (2019). Assumption 4 is necessary to ensure that the data collecting process has sufficiently explored $\mathcal{A}$ and $\mathcal{S}$. This assumption is standard in the literature of offline policy optimization/evaluation; e.g., see Jin et al. (2020b); Duan et al. (2020).

Given these assumptions, we are now ready to present the formal theoretical guarantees of Safe-DPVI, with underlying linear MDP defined in Definition 2, in the following theorem.

Theorem 2 (Suboptimality gap of Safe-DPVI: Linear MDP). *Let the underlying MDP of Safe-DPVI be a linear MDP as stated in Definition 2, $\hat{\pi}$ be the output of Safe-DPVI and $\alpha_h = 2 + \frac{2H}{\tau - \max_{s\in\mathcal{S}} \tau_h(s)}$. Under Assumptions 1, 2, 3, and 4, if we set $\beta = \sigma\sqrt{d\log\left(\frac{2+\frac{2T}{\lambda}}{\delta}\right)} + \sqrt{\lambda d}$, $\beta' = cdH\sqrt{\log(\frac{dT}{\delta})}$ for an absolute constant $c > 0$, $\bar{\beta} = \max\{\beta, \beta'\}$, then for any fixed $\delta \in (0, 1/3)$, for all $s \in \mathcal{S}$, with probaility at least $1 - 3\delta$, $\hat{\pi} \in \Pi^{\text{safe}}$ and it holds that*

$$\Delta(\hat{\pi}; s) \leq \frac{\sqrt{2}\bar{\beta}\sum_{h=1}^{H}\alpha_h}{\sqrt{2\lambda + \bar{c}K}} \tag{13}$$

We observe that under similar wide-coverage assumption (Assumption 4), Safe-DPVI with underlying linear MDP achieves an upper bound on suboptimality gap of the safe policy $\hat{\pi}$, which is nearly of the same order as that of state-of-the-art unsafe algorithm PEVI in Jin et al. (2020b). The complete proof is reported in the Appendix B. In the following section, we give a sketch of the proof.

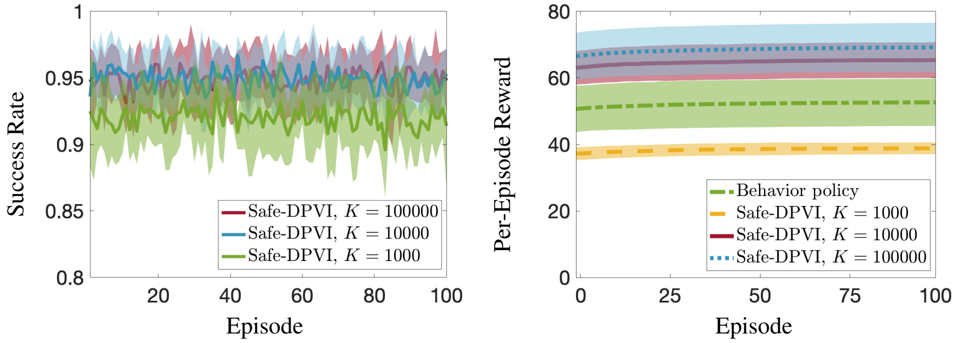
3.3 Proof sketch of Theorem 2

We make use of Theorem 1, which is stated for general MDPs, to prove Theorem 2 in two steps: 1) We first state Lemma 3, in which we specify B and B' such that they are δ-safety uncertainty quantifier and δ-bellman uncertainty quantifier as in Definition 1 for the corresponding to the linear MDP in Definition 2; 2) Next, we lower bound $\lambda_{\min}(\Lambda_h)$ for each $h \in [H]$ in Lemma 4, which is followed by a high probability upper bound on $\bar{B}_h(s,a)$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

Lemma 3 (Theorem 2 in Abbasi-Yadkori et al. (2011) and Lemma 5.2 in Jin et al. (2020b)). *Let the underlying MDP of Safe-DPVI be a linear MDP as in Definition 2. Under Assumptions 2 and 3, if we set $B_h(s,a) = \beta\left\|\phi(s,a)\right\|_{\Lambda_h^{-1}}$ and $B'_h(s,a) = \beta'\left\|\phi(s,a)\right\|_{\Lambda_h^{-1}}$, where $\beta = \sigma\sqrt{d\log\left(\frac{2+\frac{2K}{\lambda}}{\delta}\right)} + \sqrt{\lambda d}$ and $\beta' = cdH\sqrt{\log(\frac{dK}{\delta})}$ for an absolute constant $c > 0$, then B and B' are δ-safety uncertainty quantifier and δ-bellman uncertainty quantifier as in Definition 1.*

Lemma 4. *Let $\delta \in (0,1)$ and Assumption 4 holds. If $K \geq \frac{8}{\bar{c}}\log(\frac{dH}{\delta})$, then $\mathbb{P}\left(\lambda_{\min}(\Lambda_h) \geq \lambda + \frac{\bar{c}K}{2}, \forall h \in [H]\right) \geq 1 - \delta$.*

See Appendix B.1 for the proof. As a direct conclusion of Lemma 4, we upper bound $\bar{B}_h(s,a)$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$. In particular, Assumption 3 and Lemma 4 imply that for all



(a) Success Rate of $\hat{\pi}$

(b) Per-Episode Reward of $\hat{\pi}$ and $\bar{\pi}$

Figure 1: Performance of Safe-DPVI with an underlying linear MDP on Inverted Pendulum. The shaded regions show standard deviation around the average over 100 realizations.

$(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ with probability at least $1 - \delta$, it holds that

$$\left\|\phi(s,a)\right\|_{\Lambda_h^{-1}} \leq \left\|\phi(s,a)\right\|_2 \sqrt{\lambda_{\max}\left(\Lambda_h^{-1}\right)}$$
$$= \left\|\phi(s,a)\right\|_2 \sqrt{\frac{1}{\lambda_{\min}\left(\Lambda_h\right)}} \leq \sqrt{\frac{2}{2\lambda + \bar{c}K}}.$$

Now that we have established B and B', and therefore $\bar{B}$ as defined in Theorem 1, and obtained an upper bound on $\left\|\phi(s,a)\right\|_{\Lambda_h^{-1}}$ under Assumption 4, for when the underlying MDP of Safe-DPVI is linear, we are able to exploit the results stated in Theorem 1 to establish the final bound on $\Delta(\hat{\pi}; s)$ while $\hat{\pi} \in \Pi^{\text{safe}}$ with high probability in Theorem 2.

4 Experiments

In this section, we present numerical simulations to complement and confirm our theoretical findings. We apply Safe-DPVI to the control of a simulated *Inverted Pendulum* environment from OpenAI Gym Brockman et al. (2016). We consider a pendulum with mass $m = 1$, length $l = 1$, which is actuated by torque $u \in [-15, 15]$. The environment's state is described by the pendulum's angular position $\theta \in [-\pi, \pi]$ and its angular rate $\dot{\theta} \in [-5, 5]$. The system dynamics are defined as follows

$$\theta_{h+1} = \theta_h + \dot{\theta}_h \delta h + \frac{3g}{2l}\sin(\theta_t)\delta h^2 + \frac{3}{ml^2}u\delta h^2,$$
$$\dot{\theta}_{h+1} = \dot{\theta}_h + \frac{3g}{2l}\sin(\theta_h)\delta h + \frac{3}{ml^2}u\delta h, \tag{14}$$

where $g = 9.8$ is the gravity constant and δh is the simulation step and we set it to 1.

For real numbers a and b and positive integer number n, let $\text{Disc}([a, b], n)$ be a discretized set formed of uniformly dividing $[a, b]$ into n intervals. We discretize the continuous state and action spaces and consider that $\mathcal{S} = \text{Disc}([-\pi, \pi], 10) \times \text{Disc}([-5, 5], 5)$ and $\mathcal{A} = \text{Disc}([-15, 15], 15)$. Thus, $|\mathcal{S}| = 50$ and $|\mathcal{A}| = 15$. For any $s \in \mathcal{S}$, let $s(1)$ and $s(2)$ be the corresponding pendulum's angular position and pendulum's angular rate.

We consider that the transition probability $\mathbb{P}$, the reward R, and the cost C do not vary during an episode. In order to induce stochasticity and parametrize $\mathbb{P}(s'|s, a)$, we assumed that when a torque a is chosen, an additive random torque affects it. In particular, we considered that $\mathbb{P}(s'|s, a) = 0.8$ for s' being the closest element of $\mathcal{S}$ to the next state of playing torque a at pendulum's angular position $s(1)$ and pendulum's angular rate $s(2)$ according to system's dynamics in (14). Moreover, $\mathbb{P}(s'|s, a) = 0.2/4$ for s' being the closest element of $\mathcal{S}$ to the next state of playing torque $a + i,\ i \in \{-6, -3, 3, 6\}$ at pendulum's angular position $s(1)$ and pendulum's angular rate $s(2)$ according to (14). We also let $R(s, a) = c - s(1)^2 + 0.1s(2)^2 + 0.001a^2$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, where c

is a constant that makes the rewards positive, and divided them by $\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} R(s,a)$. This definition for the reward function encourages learning a controller that keeps the pendulum upright. We further defined the set of unsafe states as $\mathcal{S}^{\text{unsafe}} = \left\{s \in \mathcal{S} : s(1) \notin [-\pi/3, \pi/3]\right\}$ and specified $C(s,a) = \sum_{s'\in\mathcal{S}^{\text{unsafe}}} \mathbb{P}(s'|s,a)$, and $\tau = 0.01$. Therefore a safe policy ensures that the expected value of the probability of moving to an unsafe state is a small value ($\tau = 0.01$). We consider an underlying linear MDP with random feature maps $\boldsymbol{\phi}(s,a)$ of dimension $d = |\mathcal{S}||\mathcal{A}|$, and episode length $H = 100$.

The performance of Batch RL algorithms can vary greatly from one dataset to another. To properly assess Safe-DPVI, we repeated the following for 100 times: 1) fixed a randomly selected safe behavior policy, $\bar{\pi}$, used in the data collecting process, and created datasets with size $K = 1000$, $K = 10000$, and $K = 100000$, on Inverted Pendulum environment discussed above; 2) implemented Safe-DPVI on each of these three datasets, and employed the output policies for 100 episodes with randomly selected initial state; 3) reported the per-episode reward and success rate, number of time-steps the pendulum was in safe states during an episode divided by the duration of each episode $H = 100$, for each of the output policies. The results shown in Figure 1 depict averages over these 100 realizations, for which we have chosen $\delta = 0.01$, $\sigma = 0.05$, $\lambda = 1$. In this figure, we have numerically confirmed the result of Theorem 2. Figure 1a showcases that the output policy $\hat{\pi}$ is safe with high probability, and therefore the rate of unsafe states visits is low (success rate is high) and Figure 1b confirms that $\hat{\pi}$, for sufficiently large datasets that satisfy wide-coverage assumption (see Assumption 4), performs near-optimally and better than the behavior policy $\bar{\pi}$.

5 Conclusion

In this paper, we developed Safe-DPVI, a safe offline RL algorithm in the setting of episodic MDPs, that performs in a pessimistic manner when 1) it constructs a conservative set of safe policies; and 2) when it selects a good policy from that conservative set in the value iteration step. We guaranteed that Safe-DPVI outputs a policy $\hat{\pi}$ which is strictly safe in the sense that it respects the safety constraint at each time-step that it suggests an action to be played with high probability. Without assuming the sufficient coverage of the dataset or any structure for the underlying MDPs, we first established a data-dependent upper bound on the suboptimality gap of the *safe* policy Safe-DPVI returns. Then, we specialized our results to linear MDPs with appropriate assumptions on dataset being well-explored and proved a high probability upper bound on the suboptimality gap of $\hat{\pi}$, i.e., $\Delta(\hat{\pi}; s) \leq \frac{\sqrt{2}\bar{\beta}\sum_{h=1}^{H}\alpha_h}{\sqrt{2\lambda+\bar{c}K}}$, $\forall s \in \mathcal{S}$, which is order-wise comparable to those of its unsafe counter-parts. Finally, we implemented Safe-DPVI on Inverted Pendulum environment to empirically confirm our theoretical findings. It is an exciting future direction to find sample complexity for safe offline RL using richer function approximations with milder assumptions on *realizability* (that the underlying MDP is linear) and under appropriate assumptions on the exploratoriness of the dataset that hold in practice.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320.

Abe, N., Melville, P., Pendus, C., Reddy, C. K., Jensen, D. L., Thomas, V. P., Bennett, J. J., Anderson, G. F., Cooley, B. R., Kowalczyk, M., et al. (2010). Optimizing debt collections using constrained reinforcement learning. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84.

Amani, S., Alizadeh, M., and Thrampoulidis, C. (2019). Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems*, pages 9252–9262.

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

Bai, Y., Xie, T., Jiang, N., and Wang, Y.-X. (2019). Provably efficient q-learning with low switching cost. *arXiv preprint arXiv:1905.12849*.

Bradtke, S. J. and Barto, A. G. (1996). Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1-3):33–57.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.

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

Ding, D., Wei, X., Yang, Z., Wang, Z., and Jovanović, M. R. (2020a). Provably efficient safe exploration via primal-dual policy optimization. *arXiv preprint arXiv:2003.00534*.

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

Duan, Y., Jia, Z., and Wang, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR.

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

Fujimoto, S., Meger, D., and Precup, D. (2019). Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062. PMLR.

Garcelon, E., Ghavamzadeh, M., Lazaric, A., and Pirotta, M. (2020). Conservative exploration in reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1431–1441. PMLR.

Ghavamzadeh, M., Petrik, M., and Chow, Y. (2016). Safe policy improvement by minimizing robust baseline regret. *Advances in Neural Information Processing Systems*, 29:2298–2306.

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

Jin, Y., Yang, Z., and Wang, Z. (2020b). Is pessimism provably efficient for offline rl? *arXiv preprint arXiv:2012.15085*.

Kalagarla, K. C., Jain, R., and Nuzzo, P. (2020). A sample-efficient algorithm for episodic finite-horizon mdp with constraints. *arXiv preprint arXiv:2009.11348*.

Kumar, A., Fu, J., Tucker, G., and Levine, S. (2019). Stabilizing off-policy q-learning via bootstrapping error reduction. *arXiv preprint arXiv:1906.00949*.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020). Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*.

Lange, S., Gabel, T., and Riedmiller, M. (2012). Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer.

Laroche, R., Trichelair, P., and Des Combes, R. T. (2019). Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pages 3652–3661. PMLR.

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*.

Pacchiano, A., Ghavamzadeh, M., Bartlett, P., and Jiang, H. (2021). Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 2827–2835. PMLR.

Qiu, S., Wei, X., Yang, Z., Ye, J., and Wang, Z. (2020). Upper confidence primal-dual optimization: Stochastically constrained markov decision processes with adversarial losses and unknown transitions. *arXiv preprint arXiv:2003.00660*.

Rafailov, R., Yu, T., Rajeswaran, A., and Finn, C. (2020). Offline reinforcement learning from images with latent space models. *arXiv preprint arXiv:2012.11547*.

Srinivasan, K., Eysenbach, B., Ha, S., Tan, J., and Finn, C. (2020). Learning to be safe: Deep rl with a safety critic. *arXiv preprint arXiv:2010.14603*.

Thananjeyan, B., Balakrishna, A., Nair, S., Luo, M., Srinivasan, K., Hwang, M., Gonzalez, J. E., Ibarz, J., Finn, C., and Goldberg, K. (2021). Recovery rl: Safe reinforcement learning with learned recovery zones. *IEEE Robotics and Automation Letters*, 6(3):4915–4922.

Thomas, P., Theocharous, G., and Ghavamzadeh, M. (2015). High confidence policy improvement. In *International Conference on Machine Learning*, pages 2380–2388. PMLR.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*.

Turchetta, M., Kolobov, A., Shah, S., Krause, A., and Agarwal, A. (2020). Safe reinforcement learning via curriculum induction. *arXiv preprint arXiv:2006.12136*.

Wang, R., Foster, D. P., and Kakade, S. M. (2020). What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*.

Xu, T., Liang, Y., and Lan, G. (2020). A primal approach to constrained policy optimization: Global optimality and finite-time analysis. *arXiv preprint arXiv:2011.05869*.

Yang, L. and Wang, M. (2019). Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR.

Yang, M., Nachum, O., Dai, B., Li, L., and Schuurmans, D. (2020). Off-policy evaluation via the regularized lagrangian. *Advances in Neural Information Processing Systems*, 33.

Yin, M., Bai, Y., and Wang, Y.-X. (2021). Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1567–1575. PMLR.

Yin, M. and Wang, Y.-X. (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3948–3958. PMLR.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., and Ma, T. (2020). Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*.

Zhang, R., Dai, B., Li, L., and Schuurmans, D. (2020). Gendice: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*.

Zheng, L. and Ratliff, L. J. (2020). Constrained upper confidence reinforcement learning. *arXiv preprint arXiv:2001.09377*.

A Analysis of Safe-DPVI

In this section, we first prove Lemma 1 and then prove the three points stated in Theorem 1.

A.1 Proof of Lemma 1

First, we summarize Lemma A.1 in Jin et al. (2020b) and Lemma 4.2 in Cai et al. (2020) in the following lemma.

Lemma 5. *Let π and π' be two arbitrary policies and let Q be any given Q-function such that $V_h(s) = \mathbb{E}_{a\sim\pi_h(.|s)}\left[Q_h(s,a)\right]$ for all $(s,h) \in \mathcal{S}\times[H]$. Then*

$$\begin{aligned} V_1(s) - V_1^{\pi'}(s) = &\sum_{h=1}^{H} \mathbb{E}\left[\mathbb{E}_{a\sim\pi_h(.|s_h)}\left[Q_h(s_h,a)\right] - \mathbb{E}_{a\sim\pi'_h(.|s_h)}\left[Q_h(s_h,a)\right]\Big| s_1 = s, \pi'\right] \\ &+ \sum_{h=1}^{H} \mathbb{E}\left[Q_h(s_h,a_h) - [\mathbb{B}_h V_{h+1}](s_h,a_h)\Big| s_1 = s, \pi'\right]. \end{aligned} \tag{15}$$

Recall the definition of suboptimality gap $\Delta(\pi; s)$ in (5). We have

$$\Delta(\hat{\pi}; s) = \underbrace{V_1^*(s) - \hat{V}_1(s)}_{\text{Term I}} + \hat{V}_1(s) - V_1^{\hat{\pi}}(s). \tag{16}$$

Let $\pi = \pi' = \hat{\pi}$. Thus, applying Lemma 5, we have

$$\begin{aligned} \hat{V}_1(s) - V_1^{\hat{\pi}}(s) &= \min\left\{\mathbb{E}_{a\sim\hat{\pi}_1(.|s)}\left[\hat{Q}_1(s,a)\right], H\right\} - V_1^{\hat{\pi}}(s) \\ &\leq \mathbb{E}_{a\sim\hat{\pi}_1(.|s)}\left[\hat{Q}_1(s,a)\right] - V_1^{\hat{\pi}}(s) \\ &= \sum_{h=1}^{H}\mathbb{E}\left[\hat{Q}_h(s_h,a_h) - [\mathbb{B}_h \bar{V}_{h+1}](s_h,a_h)\Big| s_1 = s, \hat{\pi}\right] \\ &\leq \sum_{h=1}^{H}\mathbb{E}\left[\hat{Q}_h(s_h,a_h) - [\mathbb{B}_h \hat{V}_{h+1}](s_h,a_h)\Big| s_1 = s, \hat{\pi}\right] \\ &= \sum_{h=1}^{H}\mathbb{E}\left[-\iota_h(s_h,a_h)\Big| s_1 = s, \hat{\pi}\right] = \text{Term II}. \end{aligned} \tag{17}$$

A.2 Proof of Lemma 2

First note that

$$\begin{aligned} [\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a) - [\mathbb{B}_h \hat{V}_{h+1}](s,a) &= [\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a) - R_h(s,a) - [\mathbb{P}_h \hat{V}_{h+1}](s,a) \\ &= [\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a) - Q_h^{\pi}(s,a) + Q_h^{\pi}(s,a) - R_h(s,a) - [\mathbb{P}_h \hat{V}_{h+1}](s,a) \\ &= [\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a) - Q_h^{\pi}(s,a) + R_h(s,a) + [\mathbb{P}_h V_{h+1}^{\pi}](s,a) - R_h(s,a) - [\mathbb{P}_h \hat{V}_{h+1}](s,a) \\ &= [\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a) - Q_h^{\pi}(s,a) - \left[\mathbb{P}_h\left(\hat{V}_{h+1} - V_{h+1}^{\pi}\right)\right](s,a). \end{aligned} \tag{18}$$

Thus, if B' is a δ-bellman uncertainty quantifier, then for any policy π and $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, with probability at least $1-\delta$, it holds that

$$\left|[\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a) - Q_h^{\pi}(s,a) - \left[\mathbb{P}_h\left(\hat{V}_{h+1} - V_{h+1}^{\pi}\right)\right](s,a)\right| \leq B'_h(s,a). \tag{19}$$

Now, we start the formal proof of the lemma. We prove this lemma by induction. First, we prove the base case at time-step $H+1$. The statement holds for $H+1$ because $F_{H+1}(s) = 0 = V^*_{H+1}(s) = \hat{V}_{H+1}(s) = 0$. Now, suppose the statement holds for time-step $h+1$. We prove it also holds for time-step h. We consider the following two cases:

1) If $\pi^*_h(.|s) \in \hat{\Gamma}_h(s)$, we have

$$
\begin{aligned}
\hat{V}_h(s) + F_h(s) &= \min\left\{\mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[\hat{Q}_h(s,a)\right], H\right\} + F_h(s) \\
&\geq \min\left\{\mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[\hat{Q}_h(s,a)\right], H\right\} + \sum_{h'=h}^{H} \alpha_{h'}\mathbb{E}\left[\bar{B}_{h'}(s_{h'}, a_{h'})|s_h = s, \pi^*\right] \\
&\geq \min\left\{\mathbb{E}_{a\sim\pi^*_h(.|s)}\left[\hat{Q}_h(s,a)\right], H\right\} + \sum_{h'=h}^{H} \alpha_{h'}\mathbb{E}\left[\bar{B}_{h'}(s_{h'}, a_{h'})|s_h = s, \pi^*\right] \\
&\geq \min\left\{\mathbb{E}_{a\sim\pi^*_h(.|s)}\left[Q^*_h(s,a) + \left[\mathbb{P}_h\left(\hat{V}_{h+1} - V^*_{h+1}\right)\right](s,a) - 2B'_h(s,a)\right], H\right\} \\
&\quad + \sum_{h'=h}^{H} \alpha_{h'}\mathbb{E}\left[\bar{B}_{h'}(s_{h'}, a_{h'})|s_h = s, \pi^*\right] && \text{(Eqn. (19))} \\
&\geq \min\left\{\mathbb{E}_{a\sim\pi^*_h(.|s)}\left[Q^*_h(s,a) + \alpha_h\bar{B}_h(s,a) - 2B_h(s,a)\right], H\right\} && \text{(Induction assumption)} \\
&\geq \min\left\{\mathbb{E}_{a\sim\pi^*_h(.|s)}\left[Q^*_h(s,a) + (\alpha_h - 2)\bar{B}_h(s,a)\right], H\right\} \\
&= \min\left\{V^*_h(s), H\right\} && (\bigstar) \\
&= V^*_h(s).
\end{aligned}
$$

$\bigstar$ is true because $\alpha_h \geq 2$.

2) Now, we focus on the other case when $\pi^*_h(.|s) \notin \hat{\Gamma}_h(s)$, which means

$$
\mathbb{E}_{a\sim\pi^*_h(.|s)}\left[u^c_h(s,a)\right] > \tau. \tag{20}
$$

Let $\tilde{\pi}_h(.|s) := \gamma_h(s)\pi^*_h(.|s) + (1-\gamma_h(s))\pi^0_h(.|s)$, where

$$
\gamma_h(s) := \left\{\max \gamma \in [0,1] : \gamma\pi^*_h(.|s) + (1-\gamma)\pi^0_h(.|s) \in \hat{\Gamma}_h(s)\right\}. \tag{21}
$$

Now, we show that $\gamma_h(s) \geq \frac{\tau - \tau_h(s)}{\tau - \tau_h(s) + 2\mathbb{E}_{a\sim\pi^*_h(.|s)}\left[\bar{B}_h(s,a)\right]}$, which eventually leads to a proper value for α_h that guarantees for all $s \in \mathcal{S}$, with probability at least $1-2\delta$, it holds that $\hat{V}_h(s) + F_h(s) \geq V^*_h(s)$. Definitions of $\gamma_h(s)$ in (21) and the estimated safe set $\hat{\Gamma}_h(s)$ in (6) imply that

$$
\begin{aligned}
\mathbb{E}_{a\sim\tilde{\pi}_h(.|s)}\left[u^c_h(s,a)\right] &= \gamma_h(s)\mathbb{E}_{a\sim\pi^*_h(.|s)}\left[u^c_h(s,a)\right] + (1-\gamma_h(s))\mathbb{E}_{a\sim\pi^0_h(.|s)}\left[u^c_h(s,a)\right] \\
&= \gamma_h(s)\mathbb{E}_{a\sim\pi^*_h(.|s)}\left[u^c_h(s,a)\right] + (1-\gamma_h(s))\tau_h(s) \\
&\leq \tau. && (22)
\end{aligned}
$$

Thus

$$
0 < \gamma_h(s) = \frac{\tau - \tau_h(s)}{\mathbb{E}_{a\sim\pi^*_h(.|s)}\left[u^c_h(s,a)\right] - \tau_h(s)} < 1. \tag{23}
$$

Recall the definition of $\Gamma^{\text{safe}}_h(s)$ in (2) and note that $\pi^*_h(.|s) \in \Gamma^{\text{safe}}_h(s)$. Due to the definition δ-safety uncertainty quantifier B, for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, with probability at least $1-\delta$, it holds that

$$
\begin{aligned}
\mathbb{E}_{a\sim\pi^*_h(.|s)}\left[u^c_h(s,a)\right] &\leq \mathbb{E}_{a\sim\pi^*_h(.|s)}\left[\hat{C}_h(s,a) + B_h(s,a)\right] \\
&\leq \mathbb{E}_{a\sim\pi^*_h(.|s)}\left[C_h(s,a) + 2B_h(s,a)\right] \\
&\leq \tau + 2\mathbb{E}_{a\sim\pi^*_h(.|s)}\left[B_h(s,a)\right] && (\pi^*_h(.|s) \in \Gamma^{\text{safe}}_h(s)) \\
&\leq \tau + 2\mathbb{E}_{a\sim\pi^*_h(.|s)}\left[\bar{B}_h(s,a)\right]. && (24)
\end{aligned}
$$

Combining (23) and (24), we conclude that

$$\gamma_h(s) \geq \frac{\tau - \tau_h(s)}{\tau - \tau_h(s) + 2\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[\bar{B}_h(s,a)\right]}. \tag{25}$$

We have

$$\begin{aligned}
\hat{V}_h(s) + F_h(s) &= \min\left\{\mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[\hat{Q}_h(s,a)\right], H\right\} + F_h(s) \\
&= \min\left\{\mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[\left\{[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B_h'(s,a)\right\}^+\right], H\right\} + F_h(s) \\
&\geq \min\left\{\mathbb{E}_{a\sim\tilde{\pi}_h(.|s)}\left[\left\{[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B_h'(s,a)\right\}^+\right], H\right\} + F_h(s) \\
&\geq \min\left\{\mathbb{E}_{a\sim\tilde{\pi}_h(.|s)}\left[[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B_h'(s,a)\right], H\right\} + F_h(s) \\
&= \min\left\{\gamma_h(s)\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B_h'(s,a)\right]\right. \\
&\left. +(1-\gamma_h(s))\mathbb{E}_{a\sim\pi_h^0(.|s)}\left[[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B_h'(s,a)\right], H\right\} + F_h(s) \\
&\geq \min\left\{\gamma_h(s)\left(\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B_h'(s,a)\right] + F_h(s)\right)\right. \\
&\left. +(1-\gamma_h(s))\left(\mathbb{E}_{a\sim\pi_h^0(.|s)}\left[[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B_h'(s,a)\right] + F_h(s)\right), H\right\} \\
&\geq \min\left\{\gamma_h(s)\left(\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B_h'(s,a)\right] + F_h(s)\right), H\right\} \qquad (\bigstar) \\
&\geq \min\left\{\gamma_h(s)\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[Q_h^*(s,a) + (\alpha_h - 2)\bar{B}_h(s,a)\right], H\right\}. \qquad (\bigstar\bigstar)
\end{aligned}$$

$\bigstar$ is true because $(1-\gamma_h(s)) \geq 0$ and

$$\begin{aligned}
\mathbb{E}_{a\sim\pi_h^0(.|s)}\left[[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B_h'(s,a)\right] + F_h(s) &\geq \mathbb{E}_{a\sim\pi_h^0(.|s)}\left[Q_h^0(s,a) + \left[\mathbb{P}_h\left(\hat{V}_{h+1} - V_{h+1}^0\right)\right](s,a) - 2B_h'(s,a)\right] + F_h(s) \\
&\qquad \text{(Equation (19))} \\
&\geq \mathbb{E}_{a\sim\pi_h^0(.|s)}\left[Q_h^0(s,a) + \left[\mathbb{P}_h\left(\hat{V}_{h+1} - V_{h+1}^0\right)\right](s,a) - 2B_h'(s,a)\right] \\
&+ \sum_{h'=h}^{H}\alpha_{h'}\mathbb{E}\left[\bar{B}_{h'}(s_{h'},a_{h'})|s_h = s, \pi^0\right] \quad \text{(Equation (8))} \\
&\geq \mathbb{E}_{a\sim\pi_h^0(.|s)}\left[Q_h^0(s,a) + \alpha_h\bar{B}_h(s,a) - 2B_h'(s,a)\right] \\
&\qquad \text{(Induction assumption)} \\
&\geq \mathbb{E}_{a\sim\pi_h^0(.|s)}\left[Q_h^0(s,a) + (\alpha_h - 2)\bar{B}_h(s,a)\right] \\
&\geq \mathbb{E}_{a\sim\pi_h^0(.|s)}\left[Q_h^0(s,a)\right] \qquad (\alpha_h \geq 2) \\
&\geq 0.
\end{aligned}$$

$\bigstar\bigstar$ is true because

$$\begin{aligned}
\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B'_h(s,a)\right] + F_h(s) &\geq \mathbb{E}_{a\sim\pi_h^*(.|s)}\left[Q_h^*(s,a) + \left[\mathbb{P}_h\left(\hat{V}_{h+1} - V^*_{h+1}\right)\right](s,a) - 2B'_h(s,a)\right] + F_h(s) \\
&\qquad \text{(Equation (19))} \\
&\geq \mathbb{E}_{a\sim\pi_h^*(.|s)}\left[Q_h^*(s,a) + \left[\mathbb{P}_h\left(\hat{V}_{h+1} - V^*_{h+1}\right)\right](s,a) - 2B'_h(s,a)\right] \\
&+ \sum_{h'=h}^{H} \alpha_{h'}\mathbb{E}\left[\bar{B}_{h'}(s_{h'}, a_{h'})|s_h = s, \pi^*\right] \quad \text{(Equation (8))} \\
&\geq \mathbb{E}_{a\sim\pi_h^*(.|s)}\left[Q_h^*(s,a) + \alpha_h\bar{B}_h(s,a) - 2B'_h(s,a)\right] \\
&\qquad \text{(Induction assumption)} \\
&\geq \mathbb{E}_{a\sim\pi_h^*(.|s)}\left[Q_h^*(s,a) + (\alpha_h - 2)\bar{B}_h(s,a)\right].
\end{aligned}$$

Now, we continue from ★★ and observe that $\hat{V}_h(s) + F_h(s) \geq V_h^*(s)$ if and only if

$$\begin{aligned}
&\gamma_h(s)\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[Q_h^*(s,a) + (\alpha_h - 2)\bar{B}_h(s,a)\right] \geq V_h^*(s) \\
\overset{(25)}{\iff} \; & \frac{\left(\tau - \tau_h(s)\right)V_h^*(s) + \left(\tau - \tau_h(s)\right)\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[(\alpha_h - 2)\bar{B}_h(s,a)\right]}{\tau - \tau_h(s) + 2\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[\bar{B}_h(s,a)\right]} \geq V_h^*(s) \\
\iff \; & (\alpha_h - 2)\left(\tau - \tau_h(s)\right)\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[\bar{B}_h(s,a)\right] \geq 2\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[\bar{B}_h(s,a)\right]V_h^*(s) \\
\overset{H \geq V_h^*(s)}{\iff} & (\alpha_h - 2)\left(\tau - \tau_h(s)\right)\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[\bar{B}_h(s,a)\right] \geq 2\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[\bar{B}_h(s,a)\right]H \\
\iff \; & \alpha_h \geq 2 + \frac{2H}{\tau - \tau_h(s)}
\end{aligned}$$

as desired.

A.3 Proof of Theorem 1

Proof of point 1 of Theorem 1 Recall the definition of $\hat{\Gamma}_h(s)$ in (6). Since B is a δ-safety uncertainty quantifier, thus for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, $u_h^c(s,a)$ is an upper bound on $C_h(s,a)$ with probability at least $1-\delta$. Thus, one of the following two cases occurs:

1. $\hat{\pi}_h(.|s) = \pi_h^0(.|s)$, which implies that

$$\mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[C_h(s,a)\right] = \tau_h(s) < \tau. \tag{26}$$

2. $\hat{\pi}_h(.|s) \in \left\{\theta(.|s) \in \Delta_{\mathcal{A}} : \mathbb{E}_{a\sim\theta(.|s)}\left[u_h^c(s,a)\right] \leq \tau\right\}$, which implies that with probability at least $1-\delta$, it holds that

$$\mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[C_h(s,a)\right] \leq \mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[u_h^c(s,a)\right] \leq \tau, \tag{27}$$

This concludes point 1 of Theorem 1.

Proof of point 2 of Theorem 1 Note that if $[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B'_h(s,a) < 0$, then $\hat{Q}_h(s,a) = 0$ and therefore $-\iota_h(s,a) = -[\mathbb{B}_h\hat{V}_{h+1}](s,a) \leq 0$. Now, suppose $[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B'_h(s,a) \geq 0$. Since B' is a δ-bellman uncertainty quantifier, we have

$$\begin{aligned}
-\iota_h(s,a) &= \hat{Q}_h(s,a) - [\mathbb{B}_h\hat{V}_{h+1}](s,a) \\
&= [\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B'_h(s,a) - [\mathbb{B}_h\hat{V}_{h+1}](s,a) \\
&\leq 0.
\end{aligned}$$

This concludes that for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, with probability at least $1-\delta$, it holds that $-\iota_h(s,a) \leq 0$, and therefore

$$\text{Term II} = \sum_{h=1}^{H} \mathbb{E}\left[-\iota_h(s_h, a_h) \middle| s_1 = s, \hat{\pi} \right] \leq 0. \tag{28}$$

Now, we are ready to use Lemma 2 and (28) to complete the proof of point 2 as follows

$$V_h^*(s) - \hat{V}_h(s) \leq F_1(s) = \max\left\{ \sum_{h=1}^{H} \alpha_h \mathbb{E}\left[\bar{B}_h(s_h, a_h) | s_1 = s, \pi^* \right], \sum_{h=1}^{H} \alpha_h \mathbb{E}\left[\bar{B}_h(s_h, a_h) | s_1 = s, \pi^0 \right] \right\}, \tag{29}$$

as desired.

B Analysis of Safe-DPVI: Linear MDP

In this section, we prove the technical statements in Section 3.

B.1 Proof of Lemma 4

In order to bound the minimum eigenvalue of the Gram matrix Λ_h, we use the Matrix Chernoff Inequality (Tropp, 2015, Thm. 5.1.1).

Theorem 3 (Matrix Chernoff Inequality, Tropp (2015)). *Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random, symmetric matrices in $\mathbb{R}^{d\times d}$. Assume that $\lambda_{\min}(\mathbf{X}_k) \geq 0$ and $\lambda_{\max}(\mathbf{X}_k) \leq L$ for each index k. Introduce the random matrix $\mathbf{Y} = \sum_k \mathbf{X}_k$. Let μ_{min} denote the minimum eigenvalue of the expectation $\mathbb{E}[\mathbf{Y}]$,*

$$\mu_{\min} = \lambda_{\min}\left(\mathbb{E}[\mathbf{Y}]\right) = \lambda_{\min}\left(\sum_k E[\mathbf{X}_k]\right).$$

Then, for any $\epsilon \in (0,1)$, it holds,

$$\mathbb{P}\left(\lambda_{\min}(\mathbf{Y}) \leq \epsilon\mu_{\min}\right) \leq d \cdot \exp\left(-(1-\epsilon)^2 \frac{\mu_{\min}}{2L}\right).$$

Completing the Proof of Lemma 4. Let $\mathbf{X}_k = \boldsymbol{\phi}(s_h^k, a_h^k)\boldsymbol{\phi}(s_h^k, a_h^k)^\top$, such that each $\mathbf{X}_k$ is a symmetric matrix with $\lambda_{\min}(\mathbf{X}_k) \geq 0$ and $\lambda_{\max}(\mathbf{X}_k) \leq 1$ (see Assumption 3). In this notation, $\Lambda_h = \lambda I + \sum_{k=1}^K \mathbf{X}_k$. In order to apply Theorem 3, we compute

$$\mu_{\min} := \lambda_{\min}\left(\sum_{k=1}^K \mathbb{E}[\mathbf{X}_k]\right) = \lambda_{\min}\left(\sum_{k=1}^K \mathbb{E}[\boldsymbol{\phi}(s_h^k, a_h^k)\boldsymbol{\phi}(s_h^k, a_h^k)^\top]\right) = \lambda_{\min}\left(K\Sigma_h\right) \geq \bar{c}K,$$

where the last inequity follows from Assumption 4. Thus, the theorem implies the following for any $\epsilon \in [0,1)$:

$$\mathbb{P}\left[\lambda_{\min}(\sum_{k=1}^K \mathbf{X}_k) \leq \epsilon\bar{c}K\right] \leq d \cdot \exp\left(-(1-\epsilon)^2 \frac{\bar{c}K}{2}\right). \tag{30}$$

To complete the proof of the lemma, simply choose $\epsilon = 0.5$ (say) and $K \geq \frac{8}{\bar{c}}\log(\frac{dH}{\delta})$ in (30). This gives $\mathbb{P}\left(\lambda_{\min}(\Lambda_h) \geq \lambda + \frac{\bar{c}K}{2}, \ \forall h \in [H]\right) \geq 1-\delta$, as desired.

B.2 Proof of Theorem 2

A direct conclusion of Lemma 4, Assumption 3 and Lemma 4 imply that for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ with probability at least $1-\delta$, it holds that

$$\left\|\boldsymbol{\phi}(s,a)\right\|_{\Lambda_h^{-1}} \leq \left\|\boldsymbol{\phi}(s,a)\right\|_2 \sqrt{\lambda_{\max}\left(\Lambda_h^{-1}\right)} = \left\|\boldsymbol{\phi}(s,a)\right\|_2 \sqrt{\frac{1}{\lambda_{\min}\left(\Lambda_h\right)}} \leq \sqrt{\frac{2}{2\lambda + \bar{c}K}}, \tag{31}$$

and thus

$$\mathbb{P}\left(F \le \frac{\bar{\beta}\sqrt{2}}{\sqrt{2\lambda + \bar{c}K}}\right) \ge 1-\delta. \tag{32}$$

Now that we have established B and B' and obtained an upper bound on $\left\|\phi(s,a)\right\|_{\Lambda_h^{-1}}$ under Assumption 4, for when the underlying MDP of Safe-DPVI is linear, we are able to exploit the results stated in Theorem 1 to establish the final bound on $\Delta(\hat{\pi}; s)$ while $\hat{\pi} \in \Pi^{\text{safe}}$ with high probability in Theorem 2 to conclude that

$$\mathbb{P}\left(\Delta(\hat{\pi}; s) \le \frac{\sqrt{2}\bar{\beta}\sum_{h=1}^{H}\alpha_h}{\sqrt{2\lambda + \bar{c}K}},\ \forall s \in \mathcal{S} \quad \text{and} \quad \hat{\pi} \in \Pi^{\text{safe}}\right) \ge 1-3\delta. \tag{33}$$

C Unknown $\tau_h(s)$

In this section, we relax Assumption 1, and instead assume that we only have the knowledge of a safe policy π^0, and remove the assumption on the knowledge about the costs $\tau_h(s)$.

In this case, we compute a conservative estimation of the gap $\tau - \tau_h(s)$ in an adaptive manner. We show that the agent needs N samples of each tuple $(s, a, C_h(s, a_0(s)) + \epsilon_h)$ in the dataset that are collected by executing policy π^0 in order to be able to construct this conservative estimators of the gap $\tau - \tau_h(s)$, and thereafter rely on these conservative estimates in the computation of estimated safe set of policies (discussed shortly). We show that if $\frac{16\log(K)}{(\tau-\tau_h(s))^2} \le N \le \frac{64\log(K)}{(\tau-\tau_h(s))^2}$, then the agent is able to construct these conservative estimates.

Let k be the number of times policy π^0 has been executed in the dataset, and $\hat{\tau}_h(s)$ be the empirical mean estimator of $\tau_h(s)$. Then, for any $\delta \in (0,1)$, we have

$$\mathbb{P}\left(\tau_h(s) \le \hat{\tau}_h(s) + \sqrt{2\log(1/\delta)/k}\right) \ge 1-\delta. \tag{34}$$

If we let $\delta = 1/K^2$, then we have

$$\mathbb{P}\left(\left|\hat{\tau}_h(s) - \tau_h(s)\right| \le 2\sqrt{\log(K)/k},\ \forall k \in [K]\right) \ge 1-2/K. \tag{35}$$

We start from the first sample of $(s, a, C_h(s,a) + \epsilon_h)$ and continue to update the empirical mean $\hat{\tau}_h(s)$. Let N be the first time that $\hat{\tau}_h(s) + 6\sqrt{\log(K)/N} \le \tau$. Thus, we have

$$\tau_h(s) + 4\sqrt{\log(K)/N} \le \tau \Rightarrow \frac{16\log(K)}{(\tau-\tau_h(s))^2} \le N. \tag{36}$$

Note that in this case $4\sqrt{\log(K)/N}$ is a conservative estimation for $\tau - \tau_h(s)$. Thus, we have

$$\tau_h(s) + 4\sqrt{\log(K)/N} \le \tau \Rightarrow \frac{16\log(K)}{(\tau-\tau_h(s))^2} \le N. \tag{37}$$

Now we show that it will not take much more number of this tuple than $\frac{16\log(K)}{(\tau-\tau_h(s))^2}$ that this first time happens. Conversely, for any $N \ge \frac{64\log(K)}{(\tau-\tau_h(s))^2}$, we observe that

$$\hat{\tau}_h(s) + 6\sqrt{\log(K)/N} \le \tau_h(s) + 8\sqrt{\log(K)/N} \le \tau. \tag{38}$$

Therefore, we conclude that

$$\frac{16\log(K)}{(\tau-\tau_h(s))^2} \le N \le \frac{64\log(K)}{(\tau-\tau_h(s))^2}, \tag{39}$$

and $4\sqrt{\log(K)/N}$ is a conservative estimator for $\tau - \tau_h(s)$.