Can Q-learning be Improved with Advice?

Noah Golowich[*] NZG@MIT.EDU
MIT CSAIL

Ankur Moitra[†] MOITRA@MIT.EDU
MIT Math, SDSC, and CSAIL

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

Despite rapid progress in theoretical reinforcement learning (RL) over the last few years, most of the known guarantees are worst-case in nature, failing to take advantage of structure that may be known a priori about a given RL problem at hand. In this paper we address the question of whether worst-case lower bounds for regret in online learning of Markov decision processes (MDPs) can be circumvented when information about the MDP, in the form of predictions about its optimal Q-value function, is given to the algorithm. We show that when the predictions about the optimal Q-value function satisfy a reasonably weak condition we call *distillation*, then we can improve regret bounds by replacing the set of state-action pairs with the set of state-action pairs on which the predictions are grossly inaccurate. This improvement holds for both uniform regret bounds and gap-based ones. Further, we are able to achieve this property with an algorithm that achieves sublinear regret when given arbitrary predictions (i.e., even those which are not a distillation). Our work extends a recent line of work on *algorithms with predictions*, which has typically focused on simple online problems such as caching and scheduling, to the more complex and general problem of reinforcement learning.

Keywords: Reinforcement learning, Q-learning, Learning-augmented algorithms

1. Introduction

The study of worst-case algorithm design has traditionally been a mainstay of much of computer science, leading to provable and efficient algorithms for various tractable problems. However, many problems encountered in practice are often intractable, in the sense that efficient algorithms for them would violate widely held complexity theoretic hypotheses, or there are strong unconditional lower bounds on the amount of data required to achieve desired error bounds. As such, the study of beyond-worst case algorithm design (Roughgarden, 2021), which aims to use additional information about the structure of problem instances to improve algorithms' guarantees, has attracted significant attention in recent years.

An exciting approach to beyond-worst case algorithm design is to assume that the algorithm has access to certain *predictions* regarding the nature of the problem instance at hand. For example, while millions of samples may be required in order to teach a humanoid robot to walk starting from scratch, physical approximations of the robot's dynamics can be used to furnish an *approximately* optimal policy, viewed as a prediction about the optimal policy. Then, starting from this predicted policy, we can algorithmically fine-tune it using relatively few samples. This general approach,

[*] Supported by a Fannie & John Hertz Foundation Fellowship and an NSF Graduate Fellowship.

[†] This work was supported in part by a Microsoft Trustworthy AI Grant, NSF CAREER Award CCF-1453261, NSF Large CCF1565235, a David and Lucile Packard Fellowship and an ONR Young Investigator Award.

© 2022 N. Golowich & A. Moitra.

known as *algorithm design with predictions* (or *advice*), aims to improve an algorithm's guarantees, such as by reducing sample complexity, when it is given access to accurate predictions. It has been studied from a theoretical perspective in several recent works for online problems such as the ski rental problem, scheduling, caching, and many others (Mitzenmacher and Vassilvitskii, 2020). In this paper we address the design of algorithms with predictions for the much broader problem of reinforcement learning, in particular the setting of no-regret online learning in tabular Markov decision processes (MDPs). In turn, MDPs can be used to model learning problems in a plethora of settings including, for instance, personalized medicine, optimal control, and market design (Sutton and Barto, 2018).

1.1. Model overview

We consider the setting of a tabular finite-horizon episodic MDP with a finite state space $\mathcal{S}$ consisting of S states, a finite action space $\mathcal{A}$ consisting of A actions, and a horizon of length H (Agarwal et al., 2021). In this setting, the rewards and transitions of the MDP are unknown, but the learning algorithm has the ability to simulate trajectories in the MDP corresponding to policies of its choice. In total the learner simulates $K \in \mathbb{N}$ trajectories, each of which consists of H steps; the total number of samples is then $T := KH$. The learner aims to minimize the regret, namely the difference between the reward it would have received had it always followed the optimal policy and its actual aggregate reward; we refer the reader to Section 2 for additional preliminaries. In Definition 1 below, we introduce the setting of *RL with Q-value predictions*, namely where the algorithm is given access to predictions $\widetilde{Q}$ of the optimal Q-value function $Q^\star_h(x, a)$. Recall that for a state $x \in \mathcal{S}$, action $a \in \mathcal{A}$, and step $h \in [H]$, $Q^\star_h(x, a)$ denotes the cumulative expected reward when action a is taken at state x and step h, and thereafter the optimal policy is followed.

Definition 1 (RL with Q-value predictions) *We assume the learning algorithm is given, at the onset of its interaction with the MDP, access to a collection of* predictions $\widetilde{Q} = (\widetilde{Q}_1, \ldots, \widetilde{Q}_H)$, *where each* $\widetilde{Q}_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ *represents a prediction for the optimal Q-value function* $Q^\star_h$.

For many important applications of reinforcement learning (RL), predictions $\widetilde{Q}$ as in Definition 1 may be readily available to a learning algorithm; for instance:

- In robotics, powerful physics-based simulation engines allow easy collection of large amounts of data from simulated environments, but the learned policies from these simulated environments often do not transfer directly to real-world environments due to factors such as measurement error and mis-specification of the simulation parameters. To bridge this gap, information gleaned from learning in the simulated environment can be used as a prediction to be fine-tuned when interacting with the real-world environment. This general approach (sometimes called *sim-to-real*) has attracted much attention recently, such as in Rusu et al. (2018); Chebotar et al. (2019).
- In applications of RL to healthcare, Q-learning based methods (such as deep Q-learning) are commonly employed (Yu et al., 2020). It is therefore reasonable to expect that predictions $\widetilde{Q}$ for a given task can be computed based on data collected for similar tasks, such as from a previous iteration of a clinical trial to treat a particular disease. Such ideas were used in Liao et al. (2020), where various parameters to an algorithm in a mobile health trial, `HeartSteps V2`, were set based on data collected in an earlier iteration, `HeartSteps V1`.[1]

[1] The algorithm used was an ad hoc approach tuned to the particular task, rather than Q-learning.

- More broadly, our framework provides an approach for the related problems of *multi-task learning* and *transfer learning* in RL (Taylor and Stone, 2009; Zhu et al., 2021). These problems consider an RL agent which wishes to perform well on multiple related tasks (e.g., a robot moving in an environment with gradually changing obstacles) throughout its lifetime. An estimate of $Q^\star$ for earlier tasks (e.g., as obtained by Q-learning) may be used as the input predictions $\widetilde{Q}$ for later tasks. As discussed in Section 1.4, there has been a large body of empirical work devoted to improving transfer learning in RL. Some of this work considers the reuse of certain representations of an MDP such as the Q-value function; our results can thus be interpreted as a theoretical justification for such techniques.

In this paper we address the following question(s): *Is it possible to leverage prior knowledge, in the form of access to predictions $\widetilde{Q}$ as in Definition 1, to show a regret bound that beats the worst-case? Moreover, can we achieve such a result with an algorithm that still obtains sublinear regret when the predictions $\widetilde{Q}$ are arbitrary?*

1.2. Overview of results

We begin by reviewing known results regarding worst-case regret in online learning for tabular MDPs. In this paper we prove both uniform regret bounds, which depend only on the parameters S, A, H of the MDP as well as the number of samples T, as well as instance-dependent gap-based bounds, which we proceed to explain. The *gap* for action a at state x and step h is defined to be $\Delta_h(x,a) = V_h^\star(x) - Q_h^\star(x,a)$, where $V_h^\star$ and $Q_h^\star$ are the optimal value function and Q-value function, respectively.[2] $\Delta_h(x,a)$ denotes the marginal loss incurred, relative to the optimal policy, when action a is taken at state x and step h. In this section (including in Theorems 2 and 3) we make the simplifying assumption that for each (x,h) there is a unique optimal action a (this assumption is relaxed in the full statements of our results in Section 3). In this case Xu et al. (2021) exhibit an algorithm, AMB, which satisfies the following regret guarantee:

$$\mathrm{Regret}_T \leq \widetilde{O}\left(\min\left\{\sqrt{H^5SAT}, \sum_{\substack{(x,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]:\\ \Delta_h(x,a)>0}} \frac{H^5}{\Delta_h(x,a)}\right\}\right), \tag{1}$$

The regret bound (1) is optimal up to lower-order terms in the following sense: the uniform regret bound of $\widetilde{O}(\sqrt{H^5SAT})$ matches the minimax regret in tabular MDPs, $\widetilde{O}(\sqrt{H^2SAT})$ (Jin et al., 2018; Zhang et al., 2020), up to a factor of $\sqrt{H^3}$.[3] Moreover, (Simchowitz and Jamieson, 2019, Proposition 2.2) shows that a term of the form $\sum_{\substack{(x,a,h):\mathcal{S}\times\mathcal{A}\times[H]:\\ \Delta_h(x,a)>0}} \frac{\log K}{\Delta_h(x,a)}$ must appear in the gap-based regret bound in (1).

Our first main result addresses the following question: *Suppose the learning algorithm has access to predictions $\widetilde{Q}$ which are accurate on an unknown set of many state-action pairs. Then can we improve upon the worst-case regret bound (1), as if we had fewer state-action pairs to begin with?* We show an affirmative answer to this question, replacing the set of all state-action pairs with the set of those for which $\widetilde{Q}$ is inaccurate. In order for our improved bounds to "kick in",

[2] $Q_h^\star$ was defined previously, and $V_h^\star(x) = \max_{a\in\mathcal{A}}\{Q_h^\star(x,a)\}$; see Section 2 for further details.

[3] In this paper we generally disregard factors polynomial in H and $\log(SAT)$, and do not attempt to optimize the dependence of our own bounds on H and $\log(SAT)$.

though, it is necessary that the predictions $\widetilde{Q}$ satisfy an additional property, which we formalize as being a *(approximate) distillation* of $Q^\star$ (Definition 5). We focus on the case of exact distillation here, which corresponds to $\epsilon = 0$ in Definition 5: we say that $\widetilde{Q}$ is a distillation of $Q^\star$ if for each state $x \in \mathcal{S}$ and step $h \in [H]$, letting $\pi^\star_h(x)$ denote the optimal action at (x, h), it holds that $\widetilde{Q}_h(x, \pi^\star_h(x)) \geq Q^\star_h(x, \pi^\star_h(x))$. Intuitively, $\widetilde{Q}$ "picks out" the good action $\pi^\star_h(x)$ at (x, h).[4] Of course, there could be additional actions a' for which $\widetilde{Q}_h(x, a')$ is equally large, but for which a' is very much suboptimal at (x, h). Theorem 2 shows that when the predictions are a distillation of $Q^\star$, the regret bound (1) can be improved by replacing the set $\mathcal{S} \times \mathcal{A} \times [H]$ with a much smaller set consisting of tuples (x, a, h) for which $\widetilde{Q}_h(x, a)$ is grossly inaccurate:

Theorem 2 (Simplified/informal version of Theorem 8, item 2) *Suppose that $\widetilde{Q}$ is guaranteed to be a distillation of $Q^\star$. Then there is an algorithm (`QLearningPreds`, Algorithm 1[5]) which achieves regret*

$$\widetilde{O}\left(\min\left\{\sqrt{H^5 T \cdot |\mathcal{F}|}, \sum_{(x,a,h)\in\mathcal{F}} \frac{H^4}{\Delta_h(x,a)}\right\}\right), \tag{2}$$

where

$$\mathcal{F} := \left\{(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H] : \widetilde{Q}_h(x, a) > V^\star_h(x) \text{ or } \left(a \neq \pi^\star_h(x) \text{ and } \widetilde{Q}_h(x, a) \geq V^\star_h(x)\right)\right\}. \tag{3}$$

The regret bound (2) of Theorem 2 depends on the set $\mathcal{F}$ of action-state pairs (x, a, h) for which $\widetilde{Q}_h(x, a)$ is larger than the optimal value at (x, h), namely $V^\star_h(x) = \max_{a \in \mathcal{A}} Q^\star_h(x, a)$. Recall that since the distillation property requires that $\widetilde{Q}_h(x, \pi^\star_h(x)) \geq Q^\star_h(x, \pi^\star_h(x))$ for all (x, h), if $\widetilde{Q}_h(x, \pi^\star_h(x)) \neq Q^\star_h(x, \pi^\star_h(x))$, then $(x, \pi^\star_h(x), h) \in \mathcal{F}$. In the full version of Theorem 8 we relax the condition of exact equality by allowing for an approximate version of distillation (Definition 5) and an approximate version of the set $\mathcal{F}$ which set call the *fooling set* (Definition 6).

To complement Theorem 2, it is desirable to have a single algorithm which obtains nontrivial regret bounds (i.e., sublinear regret) for arbitrary predictions $\widetilde{Q}$, i.e., even those which are not an (approximate) distillation of $Q^\star$, which *also* obtains improved regret bounds (such as (2)) when $\widetilde{Q}$ is an approximate distillation. The former guarantee is often known as *robustness* in the literature on algorithms with predictions (Lykouris and Vassilvitskii, 2020; Mitzenmacher and Vassilvitskii, 2020). Robustness in this context is well-motivated since the predictions are often generated by an ad hoc procedure with few provable guarantees (such as the use of deep RL techniques on a simulated environment to estimate $\widetilde{Q}$ for use in the real-world environment), making them liable to be grossly inaccurate. Theorem 3 below gives such a guarantee for the case of uniform regret bounds; Theorems 8 and 9 provide more general robustness bounds that cover the gap-based case as well.

Theorem 3 (Simplified/informal version of Theorem 8, uniform version) *There is an algorithm (`QLearningPreds`, Algorithm 1) which satisfies the following two guarantees, when given as input a parameter $\lambda \in \left(\frac{SAH^4}{T}, 1\right)$ and predictions $\widetilde{Q}$:*

1. *For an arbitrary choice of $\widetilde{Q}$, the regret is $\widetilde{O}\left(\sqrt{\frac{TSAH^{10}}{\lambda}}\right)$.*

[4] We give an example in Section 3.1 showing that this assumption is necessary to beat minimax lower bounds.

[5] The parameter λ is set to 0 for the purposes of Theorem 2.

2. *If the predictions $\widetilde{Q}$ are a distillation of $Q^\star$, then the regret is*

$$\widetilde{O}\left(\sqrt{\lambda \cdot SATH^{10}} + \sqrt{|\mathcal{F}| \cdot TH^5}\right), \tag{4}$$

where $\mathcal{F}$ was defined in (3).

As was the case for Theorem 2, the notion of distillation and the set $\mathcal{F}$ are relaxed to their approximate analogues in Theorems 8 and 9. Notice that there is a tradeoff (mediated by the parameter λ) in Theorem 3 between the regret for arbitrary $\widetilde{Q}$ (i.e., the robustness) and the improved regret bound for $\widetilde{Q}$ that is an ϵ-approximate distillation of $Q^\star$. It follows from (Lattimore, 2015, Theorem 1) that even in the simpler multi-armed bandit setting, the dependence of this tradeoff on λ cannot be improved. This is a common occurrence in the study of algorithms with predictions, occurring, for instance, in the ski rental problem (Purohit et al., 2018; Wei and Zhang, 2020) and the problem of non-clairvoyent scheduling (Purohit et al., 2018; Wei and Zhang, 2020). We remark that in many such problems where the tradeoff occurs (including the ski rental problem), there is only a single episode of play. In contrast, in episodic RL one could allow the algorithm designer to tune λ after multiple epochs; we leave a thorough investigation of such a possibility to future work.

1.3. Warm up: Stochastic multi-armed bandits

In this section we give a brief overview of the techniques used to prove our regret upper bounds; a detailed description of the algorithm is given in Section A, and a more in-depth overview of the proof is given in Section B. As a warm-up, we begin by addressing the easier case of multi-armed bandits: in this case there is a single state, and for each action (also known as an *arm*) $a \in \mathcal{A}$, we denote the expected reward of taking a as $Q^\star(a)$.[6] Again we assume that there is a unique optimal action, denoted by $a^\star$. The algorithm receives at onset a function $\widetilde{Q} : \mathcal{A} \to [0,1]$ denoting predictions for the mean reward of each arm. Moreover, $\widetilde{Q}$ is a distillation of $Q^\star$ if $\widetilde{Q}(a^\star) \geq Q^\star(a^\star)$. The below proposition specializes Theorem 3 to the multi-armed bandit setting.

Proposition 4 (Bandit case) *There is an algorithm (`BanditPreds`, Algorithm 5) which satisfies the following two guarantees, when given as input a parameter $\lambda \in \left(\frac{A}{T}, 1\right)$ and predictions $\widetilde{Q}$:*

1. *If the predictions $\widetilde{Q}$ are a distillation of $Q^\star$, then the regret is $\widetilde{O}(\sqrt{|\mathcal{G}| \cdot T} + \sqrt{\lambda \cdot AT})$, where $\mathcal{G} := \left\{a \in \mathcal{A} \backslash \{a^\star\} : \widetilde{Q}(a) \geq Q^\star(a^\star)\right\}$.*
2. *For an arbitrary choice of $\widetilde{Q}$, the regret is $\widetilde{O}\left(\sqrt{\frac{TA}{\lambda}}\right)$.*

For simplicity we have only stated uniform regret bounds in Proposition 4, but gap-based regret bounds specializing Theorems 8 and 9 to the bandit case may readily be derived using similar techniques. Algorithm 5, which establishes Proposition 4, generally speaking aims to choose the action a which maximizes $\widetilde{Q}(a)$. Of course, when $\widetilde{Q}$ is not accurate (e.g., because it is not a distillation or $\mathcal{G}$ is nonempty) Algorithm 5 must make the following modifications:

[6] In the bandit setting, the reward received upon taking action a is a random variable; this is in contrast to the full RL setting where we assume the immediate rewards $r(x, a)$ are deterministic. This discrepancy does not lead to any significant differences in the algorithm or analysis, though.

- To handle non-optimal actions a which are in the set $\mathcal{G}$ (i.e., actions for which $\widetilde{Q}$ predicts them as having higher reward than $a^\star$), we maintain both upper and lower confidence bounds for the mean reward of each action a. For each t we then project $\widetilde{Q}(a)$ onto the interval $[\underline{Q}^t(a), \overline{Q}^t(a)]$ and use the resulting projected value instead of $\widetilde{Q}(a)$.
- Even with the use of upper and lower confidence bounds, we could run into the following difficulty: if $\widetilde{Q}(a) = Q^\star(a)$ for all $a \neq a^\star$, but $\widetilde{Q}(a^\star) \ll Q^\star(a^\star)$ (which can happen when $\widetilde{Q}$ is not a distillation), then choosing the action a to maximize $\widetilde{Q}(a)$ would simply choose the second-best action at all time steps, thus incurring linear regret. To deal with this situation, we insert an initial *exploration* phase consisting of λT time steps, in which we choose an action with the highest upper confidence bound (as per the UCB algorithm (Lattimore and Szepesvari, 2020, Chapter 7)). If $\widetilde{Q}(a^\star)$ is significantly sub-optimal, this initial exploration phase will discover that and subsequently learn to ignore the prediction $\widetilde{Q}(a^\star)$ (i.e., round it up to $\underline{Q}^t(a^\star)$, which, over time, will approach $Q^\star(a^\star)$).

The case of full RL (i.e., learning in MDPs) requires significant innovation beyond the above techniques for the multi-armed bandit case. At a high level, this occurs because errors in $\widetilde{Q}$ can compound over multiple steps h in the standard Q-learning updates. To handle such challenges, we have to use a more sophistocated rule to modify $\widetilde{Q}(a)$ over time than simply projecting it onto $[\underline{Q}^t(a), \overline{Q}^t(a)]$. Additionally, the initial exploration phase described above must be made *state-specific*, meaning that different states may leave the exploration phase at different times, according to the current value estimates at each respective state. We refer the reader to Sections A and B for further details.

1.4. Related work

Algorithms with predictions Many recent works studying algorithms with predictions have primarily focused on relatively specific online problems including the ski rental problem (Purohit et al., 2018; Wei and Zhang, 2020), scheduling (Purohit et al., 2018; Wei and Zhang, 2020; Mitzenmacher, 2019b; Lattanzi et al., 2019), caching (Lykouris and Vassilvitskii, 2020; Rohatgi, 2019), design of bloom filters (Kraska et al., 2018; Mitzenmacher, 2019a), and revenue optimization (Medina and Vassilvitskii, 2017); see also (Mitzenmacher and Vassilvitskii, 2020) for an overview of the above papers. In many of these problems, the performance parameter optimized by the algorithm (and improved with access to predictions) is the *competitive ratio*, namely the ratio between a cost measure specific to the problem and the optimal cost in hindsight, rather than the regret. There has also been a fruitful line of work showing that by using predictions it is possible, using variants of *optimistic mirror descent*, to significantly decrease the *regret* in settings including online linear optimization (Hazan and Kale, 2010; Rakhlin and Sridharan, 2012, 2013; Steinhardt and Liang, 2014; Mohri and Yang, 2015; Bhaskara et al., 2020) and contextual bandits (Wei et al., 2020). As some of these works include the case of bandit feedback, they might seem to generalize Proposition 4; however, this is not the case, since they face the significant limitation that a prediction is required by the algorithm at each time step, and the regret bound depends on the aggregate distance between the predictions and the realized values of the cost vectors or rewards over all time steps. In the setting of stochastic multi-armed bandits, this would require the predictions to track the noise of the realized reward over all T time steps in order to achieve sublinear regret; in contrast, Proposition 4 only requires a single set of predictions which must be close to the mean reward vector.

Transfer learning in RL More closely related to our results is a collection of work which proposes to solve the problem of *transfer learning in RL* (Taylor and Stone, 2009; Zhu et al., 2021) by reusing information (such as Q-values) from certain RL tasks in order to solve related RL tasks. In the particular case of Q-value reuse, the Q-values for each successive task may be initialized as some function (e.g., the mean) of the Q-values from the previous tasks; these Q-values are then updated over the course of the learning procedure for the current task (Singh, 1992; Asada et al., 1994; Tanaka and Yamamura, 2003; Torrey et al., 2005; Taylor et al., 2009). Many of these papers show that doing so outperforms an initialization of Q which is agnostic to previous tasks. These works, however, are purely empirical in nature, with no supporting theory.

Very recently there has been some effort to perform theoretical analyses for such transfer learning techniques; Tkachuk et al. (2021) shows that if the algorithm is given at onset predictions $\widetilde{Q}$ which are known to be equal to $Q^\star$ at all state-action pairs except a *single known* state-action pair at step $h = 1$, then it is possible to achieve regret $\widetilde{O}(\sqrt{H^2T})$ using Q-learning (thus eliminating the dependence on SA). Additionally, the recent work (Zhang and Wang, 2021) shows that if M agents are interacting with separate MDPs whose optimal Q-value functions are known to be ϵ-close in ℓ_∞ distance, then by sharing information about their respective MDPs, they can decrease their aggregate regret by a factor of $\sqrt{M}$ (in the uniform case) or M (in the gap-based case). Unlike our work, Zhang and Wang (2021) does not show that minimax regret bounds can be beaten for a single MDP; moreover, both Tkachuk et al. (2021); Zhang and Wang (2021) do not consider any notion of robustness nor do they allow relaxations to the ℓ_∞-closeness of the Q-value functions.

Theoretical RL background The minimax optimal regret for online learning in tabular MDPs is (up to polylogarithmic factors) $\Theta(\sqrt{SATH^2})$; the lower bound is shown by Jin et al. (2018), and the upper bound is known for both model-based algorithms such as `UCBVI` (Azar et al., 2017), as well as the model-free algorithm `UCB-Advantage` (Zhang et al., 2020) (which is a variant of the Q-learning algorithm). Non-asymptotic gap-based upper bounds for tabular MDPs were shown in Simchowitz and Jamieson (2019) using the model-based `StrongEuler` algorithm, a variant of `EULER` (Zanette and Brunskill, 2019); additional algorithms achieving gap-based bounds were shown in Lykouris et al. (2020); Yang et al. (2021); Xu et al. (2021). Our gap-based bounds are based on the techniques in Xu et al. (2021). Very recently some works (Dann et al., 2021; Tirinzoni et al., 2021; Wagenmaker et al., 2021) have derived new instance-dependent bounds in RL, such as by making alternative definitions of gaps; using insights from these works to improve our gap-based bounds is an interesting direction left for future work. The books Agarwal et al. (2021); Lattimore and Szepesvari (2020) contain a more comprehensive overview of the flurry of recent work in theoretical RL.

2. Preliminaries

We consider the setting of a tabular finite-horizon episodic Markov decision process (MDP) $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, where $\mathcal{S}$ denotes the (finite) state space, $\mathcal{A}$ denotes the (finite) action space, $H \in \mathbb{N}$ denotes the horizon, $\mathbb{P} = (\mathbb{P}_1, \dots, \mathbb{P}_H)$ denotes the transitions, and $r = (r_1, \dots, r_H)$ denotes the reward functions. In particular, for each $h \in [H]$, $\mathbb{P}_h(x'|x, a)$ (for $x, x' \in \mathcal{S}$, $a \in \mathcal{A}$) denotes the probability of transitioning to x' from x at step h when action a is taken; and $r_h(x, a)$ denotes the reward received when at state x and step h when action a is taken. We assume each reward lies in

$[0, 1]$, i.e., $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$. We write $S := |\mathcal{S}|$ and $A := |\mathcal{A}|$. A *policy* π is a collection of mappings $\pi_h : \mathcal{S} \to \mathcal{A}$, for each $h \in [H]$.[7]

In each episode, a state x_1 is picked by an adversary. For each $h \in [H]$, the agent observes the state x_h, picks an action $a_h \in \mathcal{A}$ (usually given according to some policy π, i.e., $a_h = \pi_h(x_h)$), receives reward $r_h(x_h, a_h)$, and transitions to a new state x_{h+1}, drawn according to $\mathbb{P}_h(\cdot|x_h, a_h)$. Upon receiving the reward $r_H(x_H, a_H)$ at the final step H, the episode ends. For a policy π, we let $V_h^\pi : \mathcal{S} \to \mathbb{R}$ denote the *V-value function* at step h; in particular, $V_h^\pi(x)$ gives the expected total reward received by the agent when it starts in state x at step h and thereafter follows policy π. In a similar manner, we let $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denote the *Q-value function* at step h; $Q_h^\pi(x, a)$ gives the expected total reward received by the agent when it starts in state x at step h, takes action a, and thereafter follows policy π. Formally, V_h^π and Q_h^π are defined as follows:

$$V_h^\pi(x) := \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_h(x_{h'}, a_{h'})|x_h = x\right], \qquad Q_h^\pi(x, a) = \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(x_{h'}, a_{h'})|x_h = x, a_h = a\right],$$

where $\mathbb{E}_\pi[\cdot]$ denotes that π is used to choose the action at each state.

We let $\pi^\star$ denote the *optimal policy*, namely the policy which maximizes $V_h^{\pi^\star}(x)$ for all $(x, h) \in \mathcal{S} \times [H]$. We write $V_h^\star(x) := V_h^{\pi^\star}(x)$ for all x, h. With slight abuse of notation, we let $\mathbb{P}_h$ denote the Markov operator $\mathbb{P}_h : \mathbb{R}^S \to \mathbb{R}^{S \times A}$, defined by, for any value function $V_{h+1} : \mathcal{S} \to \mathbb{R}$, $(\mathbb{P}_h V_{h+1})(x, a) := \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a)}[V_{h+1}(x')]$. The following relations (*Bellman equation* and *Bellman optimality equation*) are standard and follow easily from the definitions: for all $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$,

$$\begin{cases} V_h^\pi(x) = & Q_h^\pi(x, \pi_h(x)) \\ Q_h^\pi(x, a) = & (r_h + \mathbb{P}_h V_{h+1}^\pi)(x, a) \\ V_{H+1}^\pi(x) = & 0 \end{cases} \quad \text{and} \quad \begin{cases} V_h^\star(x) = & \max_{a \in \mathcal{A}} Q_h^\star(x, a) \\ Q_h^\star(x, a) = & (r_h + \mathbb{P}_h V_{h+1}^\star(x, a) \\ V_{H+1}^\star(x) = & 0. \end{cases}$$

For some $K \in \mathbb{N}$, over a series of K *episodes*, the RL agent interacts with the MDP M as follows: for each $k \in K$, the agent chooses a policy π^k, and applies the policy π^k in the MDP to obtain a *trajectory* $(x_1^k, a_1^k, r_1^k), \ldots, (x_H^k, a_H^k, r_H^k)$, as explained above; here $r_h^k := r_h(x_h^k, a_h^k) \in [0, 1]$ denotes the reward received at step h. We measure the agent's performance with the *regret*:

$$\mathrm{Regret}_K := \sum_{k=1}^{K} \mathbb{E}\left[V_1^\star(x_1^k) - V_1^{\pi^k}(x_1^k)\right],$$

where the expectation is taken over the randomness of the environment (in particular, the policies π^k are random variables since they depend on trajectories from previous episodes).

Notation for gap-based bounds In this paper we will derive gap-dependent regret bounds; for $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, the *gap at* (x, a, h) is defined as:

$$\Delta_h(x, a) := V_h^\star(x) - Q_h^\star(x, a).$$

[7] Note that our setting of finite-horizon MDPs is equivalent to the setting of *layered* MDPs in the literature (e.g., Xu et al. (2021)), where a different copy of the state space $\mathcal{S}$ is created for each step h, and transitions from states in layer h always to go states in layer $h + 1$.

The gap denotes the marginal sub-optimality in reward the agent suffers as a result of taking action a at state x and step h. For $\epsilon > 0$, we write, for $(x, h) \in \mathcal{S} \times [H]$,

$$\mathcal{A}^{\mathrm{opt}}_{h,\epsilon}(x) := \{a \in \mathcal{A} : \Delta_h(x) \leq \epsilon\}$$

to denote the set of actions with gap at most ϵ at (x, h). For $x \in \mathcal{S}$ and $h \in [H]$, define $\Delta_{\min,h}(x) := \min_{a \notin \mathcal{A}^{\mathrm{opt}}_{h,0}(x)} \{\Delta_h(x, a)\}$ to be the minimum positive gap at (x, h). Also define the minimum positive gap in the entire MDP to be $\Delta_{\min} := \min_{x,a,h:\Delta_h(x,a)>0}\{\Delta_h(x, a)\}$. Following Xu et al. (2021), our gap-based bounds will have a term depending on the number of state-action pairs which are optimal and for which there is *not* a unique optimal action at that state, i.e., the size of the set:

$$\mathcal{A}^{\mathrm{mul}} := \{(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H] : \ \Delta_h(x, a) = 0 \quad \text{and} \quad |\mathcal{A}^{\mathrm{opt}}_{h,0}(x)| > 1\}.$$

Prior worst-case regret bound In the special case that each state has a unique optimal action we discussed the worst-case regret bound (1) and prior work showing its optimality. In the general case, (Xu et al., 2021, Theorem B.1 & Corollary B.10) showed the following regret bound:

$$\mathrm{Regret}_T \leq O\left(H^2SA + \log(SAT) \cdot \min\left\{\frac{H^5|\mathcal{A}^{\mathrm{mul}}|}{\Delta_{\min}} + \sum_{\substack{(x,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]:\\ \Delta_h(x,a)>0}} \frac{H^5}{\Delta_h(x, a)}, \sqrt{H^5SAT}\right\}\right). \tag{5}$$

(Xu et al., 2021, Theorem 5.1) shows that a term of the form $\log K \cdot \frac{|\mathcal{A}^{\mathrm{mul}}|}{\Delta_{\min}}$ is necessary, even in the presence of the term $\sum_{\substack{(x,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]:\\ \Delta_h(x,a)>0}} \frac{1}{\Delta_h(x,a)}$ of the regret bound. Thus, in general, the bound (5) cannot be improved by more than $\mathrm{poly}(H, \log(SAT))$ factors.

Additional notation Given a real number x, let $[x]_+$ denote x if $x > 0$, and 0 otherwise. We will write $T = HK$ to denote the total number of samples over K episodes; note that $\mathrm{Regret}_K \leq T$ always holds. We also set $\iota := \log(SAT)$. For $(x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, $N^k_h(x, a)$ denotes the number of episodes before episode k in which (x, a, h) is visited, i.e., action a was taken at state x and step h ($N^k_h(x, a)$ is also defined in step 2(b)i of Algorithm 1). For integers $i \geq 1$ and $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, we let $k^i_h(x, a)$ denote the episode k which is the ith episode that (x, a, h) was visited. If no such episode exists, we set $k^i_h(x, a) = K + 1$ as a matter of convention.

3. Learning in MDPs with predictions: main results

3.1. Properties of the predictions $\widetilde{Q}$

Our main result shows that in the presence of arbitrary predictions $\widetilde{Q}$, we are able to obtain a sublinear regret bound for our algorithm `QLearningPreds`, and moreover, if $\widetilde{Q}$ satisfies an additional property, then we can obtain an improved regret bound that can *beat* the minimax regret bounds for learning in MDPs (i.e., (5)), replacing the space $\mathcal{S} \times \mathcal{A} \times [H]$ with a smaller space representing the set of state-action pairs where $\widetilde{Q}$ is inaccurate (consistency). Definition 5 below captures the additional property (referred to as being an *approximate distillation* of $Q^\star$) that $\widetilde{Q}$ needs to satisfy in order to obtain improved regret bounds.

To motivate the definition, consider the setting where there is a single state x_0 and $H = 1$ (which is equivalent to the stochastic multi-armed bandit problem). Moreover suppose there is a unique optimal action $a^\star$ with reward 1 and all other $A - 1$ actions have reward $1 - \Delta$ for some positive $\Delta < 1/A$. If we are given the predictions $\widetilde{Q}_1^{\mathrm{F}}$, where $\widetilde{Q}_1^{\mathrm{F}}(x_0, a) := 1 - \Delta$ for all a, then $\widetilde{Q}_1^{\mathrm{F}}$ is only incorrect at a single action (namely, $a^\star$), but it provides no information about what $a^\star$ is, and it is straightforward to show that, even given $\widetilde{Q}_1^{\mathrm{F}}$, the regret of any algorithm must be $\Omega(A/\Delta)$, giving no improvement over the setting without predictions (Lattimore and Szepesvari, 2020, Chapter 16). On the other hand, consider the predictions $\widetilde{Q}_1^{\mathrm{T}}$ defined as equal to $Q_1^\star$ except at a single (unknown) *non-optimal action* a'.[8] Though both $\widetilde{Q}_1^{\mathrm{F}}, \widetilde{Q}_1^{\mathrm{T}}$ both differ from $\widetilde{Q}_1^\star$ at a single action, it will follow from Theorem 8 (with $\lambda = A/T$) that given $\widetilde{Q}_1^{\mathrm{T}}$, `QLearningPreds` obtains the much smaller regret bound of $\widetilde{O}(1/\Delta)$. As this example shows, a set of accurate predictions $\widetilde{Q}$ cannot entirely mitigate the exploration problem: even if the predictions are accurate at nearly all states and actions, if they do not provide any information as to the identity of the optimal action at a given state (e.g., as for $\widetilde{Q}_1^{\mathrm{F}}$), then we cannot hope to beat existing regret bounds. The notion of *approximate distillation*, defined below, formalizes the notion that $\widetilde{Q}$ must provide information about the optimal action at each state:

Definition 5 (Approximate distillation) *Consider a predicted Q-value function $\widetilde{Q} \in \mathbb{R}^{[H]\times\mathcal{S}\times\mathcal{A}}$. For $\epsilon > 0$, we say that $\widetilde{Q}$ is an ϵ-*approximate distillation *of the optimal value function $Q^\star$ if the following holds: for each $(x, h) \in \mathcal{S} \times [H]$, there is some $a \in \mathcal{A}$ so that*

$$\Delta_h(x, a) + [Q_h^\star(x, a) - \widetilde{Q}_h(x, a)]_+ \leq \epsilon.$$

In words, Definition 5 requires that for each (x, h), there is some action a which is nearly optimal and for which $\widetilde{Q}_h(x, a)$ does not greatly underestimate the value of $Q_h^\star(x, a)$. It follows from Definition 5 that $\max_{a\in\mathcal{A}} \widetilde{Q}_h(x, a) \geq V_h^\star(x) - \epsilon$ for all x, h; thus a trivial way to utilize the predictions $\widetilde{Q}$ is to follow the greedy policy with respect to $\widetilde{Q}$ for $O((H/\epsilon)^2)$ iterations, which suffices to check if the greedy policy's value is ϵ-close to $\max_{a\in\mathcal{A}} \widetilde{Q}_h(x, a)$, and if not, then apply a standard worst-case RL algorithm. This approach suffers from two serious shortcomings: first, it may not obtain the optimal dependence of regret on T (e.g., if ϵ decays with T), and second, there may be a few actions whose value is grossly over-estimated by a distillation $\widetilde{Q}$, so that the greedy policy fails yet the predictions can still be effectively used. Our regret bounds, in contrast, can handle the presence of such actions, which we formally define below as the *fooling set*:

Definition 6 (Fooling set) *Given a set of predictions $\widetilde{Q}$ for any $\epsilon_1, \epsilon_2 > 0$, we define the set of* (ϵ_1, ϵ_2)-fooling tuples (x, a, h), *denoted $\mathcal{F}(\epsilon_1, \epsilon_2) \subset \mathcal{S} \times \mathcal{A} \times [H]$, to be the set of tuples (x, a, h) so that*

$$\widetilde{Q}_h(x, a) - Q_h^\star(x, a) \geq \Delta_h(x, a) - \epsilon_1 \geq \epsilon_2 - \epsilon_1 \quad \text{or} \quad \widetilde{Q}_h(x, a) > V_h^\star(x) + \epsilon_2. \tag{6}$$

In this context, we will always have $\epsilon_2 > \epsilon_1 > 0$.

Notice that the first condition in (6) subsumes the second when $\Delta_h(x, a) \geq \epsilon_2$ (as it requires $\widetilde{Q}_h(x, a) \geq V_h^\star(x) - \epsilon_1$); the second condition is added to account for near-optimal actions, namely those satisfying $\Delta_h(x, a) < \epsilon_2$. Moreover, we could alternatively define the fooling set as those

[8] $\widetilde{Q}_1^{\mathrm{T}}(x, a')$ can be set to any real number.

(x,a,h) for which $|\widetilde{Q}_h(x,a) - Q^\star_h(x,a)| > \epsilon_2 - \epsilon_1$; this set, however, is in general larger than $\mathcal{F}(\epsilon_1,\epsilon_2)$, and so using $\mathcal{F}(\epsilon_1,\epsilon_2)$ allows us to obtain tighter regret bounds.

One of our results will also make use of the following additional assumption on $\widetilde{Q}$:

Definition 7 (Optimal fooling actions) *For $\epsilon' > 0$, we say that predictions $\widetilde{Q}$* lack ϵ'-fooling optimal actions *if there is no (x,h) with multiple optimal actions (i.e., for which $|\mathcal{A}^{\mathrm{opt}}_{h,0}(x)| > 1$) so that for some $a \in \mathcal{A}^{\mathrm{opt}}_{h,0}(x)$, $\widetilde{Q}_h(x,a) > V^\star_h(x) + \epsilon'$.*

Note that in the context of Definition 7, $\widetilde{Q}_h(x,a) > V^\star_h(x) + \epsilon'$ implies that $(x,a,h) \in \mathcal{F}(\epsilon,\epsilon')$ for any ϵ, explaining the terminology of the definition.

3.2. Main theorems

A common thread in the literature on algorithms with predictions is an inherent tradeoff between an algorithm's robustness and its accuracy when it receives correct predictions (sometimes called *consistency*) (Purohit et al., 2018; Wei and Zhang, 2020). Such a tradeoff occurs in our setting too. To describe this tradeoff, we introduce a parameter $\lambda \in (0,1)$: as λ decreases to 0, the regret in the presence of predictions which are an (approximate) distillation improves but the robustness (i.e., regret in the presence of arbitrary predictions) worsens.

λ-Cost We will be able to obtain both gap-based regret bounds and (instance-independent) uniform ones for both robustness and consistency in the presence of predictions. To simplify the dependence of these bounds on the parameter λ introduced above, we define the *λ-cost* for an MDP M as follows: given an MDP M, a value $T \in \mathbb{N}$ and a value $\lambda \in (0,1)$, the λ-cost of M, denoted $\mathscr{C}_{M,T,\lambda}$, is the following quantity:

$$\mathscr{C}_{M,T,\lambda} := \min\left\{ \sqrt{\lambda \cdot TSAH^8\iota},\ H^8\iota \cdot \left(\sum_{(x,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]:a\notin\mathcal{A}^{\mathrm{opt}}_{h,0}(x)} \frac{1}{\Delta_h(x,a)} + \frac{|\mathcal{A}^{\mathrm{mul}}|}{\Delta_{\min}} \right) \right\} \tag{7}$$

Recall that $\iota = \log(SAT)$. Note that, ignoring $\mathrm{poly}(H,\iota)$ factors, $\mathscr{C}_{M,T,\lambda}$ is in general no greater than the worst-case regret bound (5): moreover, if the first term in the minimum in (7) (i.e., $\sqrt{\lambda \cdot TSAH^8\iota}$) is much smaller than the second, then due to the factor of $\sqrt{\lambda}$ in this term, $\mathscr{C}_{M,T,\lambda}$ will be much smaller than the right-hand side of (5) (again, ignoring $\mathrm{poly}(H,\iota)$ factors).

Explicit-λ guarantee Our first main result is stated below; for simplicity, we present here the result under the additional assumption that each (x,h) has a unique optimal action (i.e., $|\mathcal{A}^{\mathrm{opt}}_{h,0}(x)| = 1$); this assumption has been made previously in Xu et al. (2021). As we show in an extended version of the theorem (see the version in Section E), this assumption may be removed if we assume that $\Delta_{\min}$ is known to the algorithm; further, our second main result (Theorem 9) avoids making either assumption altogether. Theorem 8 states that the regret of `QLearningPreds` (Algorithm 1) under arbitrary predictions $\widetilde{Q}$ is $\widetilde{O}(\frac{H}{\lambda} \cdot \mathscr{C}_{M,T,\lambda})$, whereas the regret under accurate predictions (i.e., predictions which are an approximate distillation) is the sum of $\widetilde{O}(H \cdot \mathscr{C}_{M,T,\lambda})$ plus a quantity that grows as the degree of accuracy of the predictions degrades.

Theorem 8 *Suppose that for each (x,h) there is a unique optimal action (i.e., $|\mathcal{A}^{\mathrm{opt}}_{h,0}(x)| = 1$). The algorithm `QLearningPreds` (Algorithm 1) with the `DeltaIncr` subroutine (Algorithm 4)* *with parameter* $\widetilde{\Delta}_{\min} = 0$ *satisfies the following two guarantees, when given as input a parameter* $\lambda \in [0,1]$ *and predictions* $\widetilde{Q}$*:*

1. *Suppose* $\lambda \geq \frac{SAH^4}{T}$*. Then for an arbitrary choice of input predictions* $\widetilde{Q}$*, the regret of* `QLearningPreds` *is* $O(\frac{H\iota}{\lambda} \cdot \mathscr{C}_{M,T,\lambda})$*.*
2. *Fix any* $\epsilon > 0$*, and set* $\epsilon' = 4\epsilon \cdot (H+1)$*. When the input predictions* $\widetilde{Q}$ *are an* ϵ*-approximate distillation of* $Q^\star$ *(Definition 5), the regret of* `QLearningPreds` *is*

$$O\left(H^2\iota \cdot \mathscr{C}_{M,T,\lambda} + \epsilon' TH + \min\left\{\sqrt{H^5 T\iota \cdot |\mathcal{F}(\epsilon'/2,\epsilon')|}, \sum_{(x,a,h)\in\mathcal{F}(\epsilon'/2,\epsilon')} \frac{H^4\iota}{[\Delta_h(x,a) - \epsilon'/2]_+}\right\}\right). \tag{8}$$

Notice that the gap-based term in $\mathscr{C}_{M,T,\lambda}$ (see (7)) does not depend on λ (i.e., it does not decrease when λ decreases). Nevertheless, Theorem 8 implies that accurate predictions can improve logarithmic regret bounds involving gaps, if additional information about the MDP is known. For instance, suppose the algorithm is promised that one of the following holds: either $\widetilde{Q} = Q^\star$ (and so $\widetilde{Q}$ is a 0-approximate distillation),[9] or it holds that all non-zero gaps are at least a constant, which implies that $\mathscr{C}_{M,T,\lambda} \leq \mathrm{poly}(H,\iota) \cdot O(SA)$; however, which of these possibilities holds is unknown. Then by choosing $\lambda = \sqrt{\frac{SA}{T}}$ in Theorem 8 (with $\epsilon = \epsilon' = 0$), we obtain a regret bound of $\mathrm{poly}(H,\iota) \cdot O(SA)$ (which is independent of K) in the case that $\widetilde{Q} = Q^\star$, and a regret bound of $\mathrm{poly}(H,\iota) \cdot O(\sqrt{SAT})$ in the other case. Thus we always manage to achieve regret at least as small as the minimax bound of $\widetilde{O}(\sqrt{H^2 SAT})$ (up to $\mathrm{poly}(H,\iota)$ factors), and in the case of accurate predictions can get a much-improved regret bound that is polylogarithmic in K. This example shows that it can be beneficial to use Q-value predictions for gap-dependent guarantees: if the algorithm designer has a prior over the space of MDPs indicating that the former possibility (accurate $\widetilde{Q}$, but possibly small gaps) is much more likely than the latter, then it is beneficial to use Theorem 8 and obtain the polylogarithmic regret bound with high probability over the prior (and near-minimax regret in all cases), rather than using a worst-case gap-based result such as Xu et al. (2021).

Implicit-λ guarantee Unlike in much of the literature on algorithms with predictions (Mitzenmacher and Vassilvitskii, 2020), the quantity $\mathscr{C}_{M,T,\lambda}$ which appears in our regret bounds is in general *unknown* to the algorithm, as the quantities $\Delta_h(x,a)$, $|\mathcal{A}^{\mathrm{mul}}|$, $\Delta_{\min}$ are all unknown. Therefore, the standard paradigm in which a user chooses a parameter λ and then runs an algorithm depending on λ is somewhat less well-motivated because the user does not have an explicit formula for how the choice of λ influences the regret bounds in the case when either the predictions are accurate or inaccurate. Therefore, in our next main result, Theorem 9, we adopt the alternative procedure in which the user instead inputs a parameter $\mathscr{R} < T$. Given $\mathscr{R}$, the algorithm's robustness (i.e., performance under arbitrary predictions) is guaranteed to be $O(\mathscr{R})$, while the performance under accurate predictions grows with $\mathscr{C}_{M,T,\widehat{\lambda}}$ for $\widehat{\lambda}$ implicitly chosen optimally so as to still guarantee regret $O(\mathscr{R})$ in the worst case.

Theorem 9 *The algorithm* `QLearningPreds` *with the* `DeltaConst` *subroutine satisfies the following two guarantees, when given as input a parameter* $\mathscr{R} \in [SAH^3, \frac{T}{SA}]$ *and predictions* $\widetilde{Q}$*:*

[9] More generally, we could assume that $\widetilde{Q}$ is an ϵ-approximate distillation and that $\mathcal{F}(\epsilon'/2,\epsilon')$ is small.

1. *If $\mathscr{R} \geq \mathscr{C}_{M,T,1}$, for* any *choice of predictions $\widetilde{Q}$, the regret of* `QLearningPreds` *is $O(\mathscr{R})$.*

2. *Fix any $\epsilon > 0$, and set $\epsilon' = 4\epsilon \cdot (H+1)$. When the input predictions $\widetilde{Q}$ are an ϵ-approximate distillation of $Q^\star$ (Definition 5) and lack ϵ'-fooling optimal actions (Definition 7), the regret of* `QLearningPreds` *is*

$$O\left(H \cdot \mathscr{C}_{M,T,\widehat{\lambda}} + \epsilon' TH + \min\left\{\sqrt{H^5 T\iota \cdot |\mathcal{F}(\epsilon'/2, \epsilon')|}, \sum_{(x,a,h)\in\mathcal{F}(\epsilon'/2,\epsilon')} \frac{H^4 \iota}{[\Delta_h(x,a) - \epsilon'/2]_+}\right\}\right), \tag{9}$$

where $\widehat{\lambda} \in (0,1)$ is chosen so that $\frac{1}{\widehat{\lambda}} \cdot \mathscr{C}_{M,T,\widehat{\lambda}} = \mathscr{R}$.

References

Alekh Agarwal, Nan Jiang, and Sham M Kakade. *Reinforcement Learning: Theory and Algorithms*. 2021. URL `https://rltheorybook.github.io/`.

M. Asada, S. Noda, S. Tawaratsumida, and K. Hosoda. Vision-based behavior acquisition for a shooting robot by using a reinforcement learning. In *Proceedings of the Workshop on Visual Behaviors*, pages 112–118, Seattle, WA, USA, 1994. IEEE Comput. Soc. Press. ISBN 978-0-8186-6660-5. doi: 10.1109/VL.1994.365601. URL `https://ieeexplore.ieee.org/document/365601`.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax Regret Bounds for Reinforcement Learning. *arXiv:1703.05449 [cs, stat]*, July 2017. URL `http://arxiv.org/abs/1703.05449`. arXiv: 1703.05449.

Aditya Bhaskara, Ashok Cutkosky, Ravi Kumar, and Manish Purohit. Online Learning with Imperfect Hints. *arXiv:2002.04726 [cs, math, stat]*, October 2020. URL `http://arxiv.org/abs/2002.04726`. arXiv: 2002.04726.

Yevgen Chebotar, Ankur Handa, Viktor Makoviychuk, Miles Macklin, Jan Issac, Nathan Ratliff, and Dieter Fox. Closing the Sim-to-Real Loop: Adapting Simulation Randomization with Real World Experience. *arXiv:1810.05687 [cs]*, March 2019. URL `http://arxiv.org/abs/1810.05687`. arXiv: 1810.05687.

Christoph Dann, Teodor V. Marinov, Mehryar Mohri, and Julian Zimmert. Beyond Value-Function Gaps: Improved Instance-Dependent Regret Bounds for Episodic Reinforcement Learning. *arXiv:2107.01264 [cs]*, July 2021. URL `http://arxiv.org/abs/2107.01264`. arXiv: 2107.01264.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *Journal of Machine Learning Research*, 7(39):1079–1105, 2006. ISSN 1533-7928. URL `http://jmlr.org/papers/v7/evendar06a.html`.

Elad Hazan and Satyen Kale. Extracting certainty from uncertainty: regret bounded by variation in costs. *Machine Learning*, 80(2-3):165–188, September 2010. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-010-5175-x. URL http://link.springer.com/10.1007/s10994-010-5175-x.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. Is Q-learning Provably Efficient? *arXiv:1807.03765 [cs, math, stat]*, July 2018. URL http://arxiv.org/abs/1807.03765. arXiv: 1807.03765.

Tim Kraska, Alex Beutel, Ed H. Chi, Jeffrey Dean, and Neoklis Polyzotis. The Case for Learned Index Structures. *arXiv:1712.01208 [cs]*, April 2018. URL http://arxiv.org/abs/1712.01208. arXiv: 1712.01208.

Silvio Lattanzi, Thomas Lavastida, Benjamin Moseley, and Sergei Vassilvitskii. Online Scheduling via Learned Weights. In *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, Proceedings, pages 1859–1877. Society for Industrial and Applied Mathematics, December 2019. doi: 10.1137/1.9781611975994.114. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611975994.114.

Tor Lattimore. The pareto regret frontier for bandits. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Tor Lattimore and Csaba Szepesvari. *Bandit Algorithms*. 2020.

Peng Liao, Kristjan Greenewald, Predrag Klasnja, and Susan Murphy. Personalized HeartSteps: A Reinforcement Learning Algorithm for Optimizing Physical Activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22, March 2020. ISSN 2474-9567. doi: 10.1145/3381007. URL https://dl.acm.org/doi/10.1145/3381007.

Thodoris Lykouris and Sergei Vassilvitskii. Competitive caching with machine learned advice. *arXiv:1802.05399 [cs]*, August 2020. URL http://arxiv.org/abs/1802.05399. arXiv: 1802.05399.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. *arXiv:1803.09353 [cs, stat]*, March 2018. URL http://arxiv.org/abs/1803.09353. arXiv: 1803.09353.

Thodoris Lykouris, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Corruption robust exploration in episodic reinforcement learning. *arXiv:1911.08689 [cs, stat]*, April 2020. URL http://arxiv.org/abs/1911.08689. arXiv: 1911.08689.

Andrés Muñoz Medina and Sergei Vassilvitskii. Revenue Optimization with Approximate Bid Predictions. *arXiv:1706.04732 [cs]*, November 2017. URL http://arxiv.org/abs/1706.04732. arXiv: 1706.04732.

Michael Mitzenmacher. A Model for Learned Bloom Filters, and Optimizing by Sandwiching. *arXiv:1901.00902 [cs, stat]*, January 2019a. URL http://arxiv.org/abs/1901.00902. arXiv: 1901.00902.

Michael Mitzenmacher. Scheduling with Predictions and the Price of Misprediction. *arXiv:1902.00732 [cs]*, May 2019b. URL `http://arxiv.org/abs/1902.00732`. arXiv: 1902.00732.

Michael Mitzenmacher and Sergei Vassilvitskii. Algorithms with Predictions. In *Beyond the Worst-Case Analysis of Algorithms*, pages 646–662. Cambridge University Press, 1 edition, December 2020. ISBN 978-1-108-63743-5 978-1-108-49431-1. doi: 10.1017/9781108637435.037.

Mehryar Mohri and Scott Yang. Accelerating Optimization via Adaptive Prediction. *arXiv:1509.05760 [cs, stat]*, October 2015. URL `http://arxiv.org/abs/1509.05760`. arXiv: 1509.05760.

Manish Purohit, Zoya Svitkina, and Ravi Kumar. Improving Online Algorithms via ML Predictions. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/hash/73a427badebe0e32caa2e1fc7530b7f3-Abstract.html`.

Alexander Rakhlin and Karthik Sridharan. Online Learning with Predictable Sequences. *arXiv:1208.3728 [cs, stat]*, August 2012. URL `http://arxiv.org/abs/1208.3728`. arXiv: 1208.3728.

Alexander Rakhlin and Karthik Sridharan. Optimization, Learning, and Games with Predictable Sequences. *arXiv:1311.1869 [cs]*, November 2013. URL `http://arxiv.org/abs/1311.1869`. arXiv: 1311.1869.

Dhruv Rohatgi. Near-Optimal Bounds for Online Caching with Machine Learned Advice. *arXiv:1910.12172 [cs]*, October 2019. URL `http://arxiv.org/abs/1910.12172`. arXiv: 1910.12172.

Tim Roughgarden, editor. *Beyond the Worst-Case Analysis of Algorithms*. Cambridge University Press, Cambridge, 2021. ISBN 978-1-108-49431-1. doi: 10.1017/9781108637435. URL `https://www.cambridge.org/core/books/beyond-the-worstcase-analysis-of-algorithms/8A8128BBF7FC2857471E9CA52E69AC21`.

Andrei A. Rusu, Mel Vecerik, Thomas Rothörl, Nicolas Heess, Razvan Pascanu, and Raia Hadsell. Sim-to-Real Robot Learning from Pixels with Progressive Nets. *arXiv:1610.04286 [cs]*, May 2018. URL `http://arxiv.org/abs/1610.04286`. arXiv: 1610.04286.

Max Simchowitz and Kevin Jamieson. Non-Asymptotic Gap-Dependent Regret Bounds for Tabular MDPs. *arXiv:1905.03814 [cs, math, stat]*, October 2019. URL `http://arxiv.org/abs/1905.03814`. arXiv: 1905.03814.

Satinder Pal Singh. Transfer of learning by composing solutions of elemental sequential tasks. *Machine Learning*, 8(3):323–339, May 1992. ISSN 1573-0565. doi: 10.1007/BF00992700. URL `https://doi.org/10.1007/BF00992700`.

Jacob Steinhardt and Percy Liang. Adaptivity and Optimism: An Improved Exponentiated Gradient Algorithm. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1593–1601. PMLR, June 2014. URL https://proceedings.mlr.press/v32/steinhardtb14.html. ISSN: 1938-7228.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

F. Tanaka and M. Yamamura. Multitask reinforcement learning on the distribution of MDPs. In *Proceedings 2003 IEEE International Symposium on Computational Intelligence in Robotics and Automation. Computational Intelligence in Robotics and Automation for the New Millennium (Cat. No.03EX694)*, volume 3, pages 1108–1113 vol.3, July 2003. doi: 10.1109/CIRA.2003.1222152.

Matthew E. Taylor and Peter Stone. Transfer Learning for Reinforcement Learning Domains: A Survey. *Journal of Machine Learning Research*, 10(56):1633–1685, 2009. ISSN 1533-7928. URL http://jmlr.org/papers/v10/taylor09a.html.

Matthew E Taylor, Peter Stone, and Yaxin Liu. Transfer Learning via Inter-Task Mappings for Temporal Difference Learning. *Journal of Machine Learning Research*, 8:2125–2167, 2009.

Andrea Tirinzoni, Matteo Pirotta, and Alessandro Lazaric. A Fully Problem-Dependent Regret Lower Bound for Finite-Horizon MDPs. *arXiv:2106.13013 [cs]*, June 2021. URL http://arxiv.org/abs/2106.13013. arXiv: 2106.13013.

Volodymyr Tkachuk, Sriram Ganapathi Subramanian, and Matthew E. Taylor. The Effect of Q-function Reuse on the Total Regret of Tabular, Model-Free, Reinforcement Learning. *arXiv:2103.04416 [cs]*, March 2021.

Lisa Torrey, Trevor Walker, Jude Shavlik, and Richard Maclin. Using Advice to Transfer Knowledge Acquired in One Reinforcement Learning Task to Another. In *Machine Learning: ECML 2005*, volume 3720, pages 412–424. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005. ISBN 978-3-540-29243-2 978-3-540-31692-3. doi: 10.1007/11564096_40. URL http://link.springer.com/10.1007/11564096_40. Series Title: Lecture Notes in Computer Science.

Andrew Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond No Regret: Instance-Dependent PAC Reinforcement Learning. *arXiv:2108.02717 [cs, stat]*, August 2021. URL http://arxiv.org/abs/2108.02717. arXiv: 2108.02717.

Alexander Wei and Fred Zhang. Optimal Robustness-Consistency Trade-offs for Learning-Augmented Online Algorithms. *arXiv:2010.11443 [cs]*, October 2020. URL http://arxiv.org/abs/2010.11443. arXiv: 2010.11443.

Chen-Yu Wei, Haipeng Luo, and Alekh Agarwal. Taking a hint: How to leverage loss predictors in contextual bandits? In *Thirty-Third Annual Conference On Learning Theory*, July 2020.

Haike Xu, Tengyu Ma, and Simon S. Du. Fine-Grained Gap-Dependent Bounds for Tabular MDPs via Adaptive Multi-Step Bootstrap. *arXiv:2102.04692 [cs]*, February 2021. URL http://arxiv.org/abs/2102.04692. arXiv: 2102.04692.

Kunhe Yang, Lin F. Yang, and Simon S. Du. Q-learning with Logarithmic Regret. *arXiv:2006.09118 [cs, math, stat]*, February 2021. URL `http://arxiv.org/abs/2006.09118`. arXiv: 2006.09118.

Chao Yu, Jiming Liu, and Shamim Nemati. Reinforcement Learning in Healthcare: A Survey. *arXiv:1908.08796 [cs]*, April 2020. URL `http://arxiv.org/abs/1908.08796`. arXiv: 1908.08796.

Andrea Zanette and Emma Brunskill. Tighter Problem-Dependent Regret Bounds in Reinforcement Learning without Domain Knowledge using Value Function Bounds. *arXiv:1901.00210 [cs, stat]*, November 2019. URL `http://arxiv.org/abs/1901.00210`. arXiv: 1901.00210.

Chicheng Zhang and Zhi Wang. Provably Efficient Multi-Task Reinforcement Learning with Model Transfer. *arXiv:2107.08622 [cs]*, July 2021. URL `http://arxiv.org/abs/2107.08622`. arXiv: 2107.08622.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost Optimal Model-Free Reinforcement Learning via Reference-Advantage Decomposition. *arXiv:2004.10019 [cs, stat]*, June 2020. URL `http://arxiv.org/abs/2004.10019`. arXiv: 2004.10019.

Zhuangdi Zhu, Kaixiang Lin, and Jiayu Zhou. Transfer Learning in Deep Reinforcement Learning: A Survey. *arXiv:2009.07888 [cs, stat]*, March 2021. URL `http://arxiv.org/abs/2009.07888`. arXiv: 2009.07888.

Appendix A. Algorithm overview

A.1. Algorithm description

Our algorithm, `QLearningPreds` (Algorithm 1) used in Theorems 8 and 9, is based loosely off of the Q-learning algorithm (Jin et al., 2018), and incorporates numerous additional aspects (including several ideas from Xu et al. (2021)) to effectively use the predictions $\widetilde{Q}_h(x,a)$. In this section we describe the main ideas of the algorithm. At each episode k, the algorithm maintains upper and lower bounds on the Q-value and V-value functions, denoted $\overline{Q}_h^k(x,a), \overline{V}_h^k(x,a)$ and $\underline{Q}_h^k(x,a), \underline{V}_h^k(x,a)$, respectively. Unlike previous versions of Q-learning, our algorithm makes use of additional functions, denoted $\overline{R}_h^k(x,a), \widetilde{Q}_h^k(x,a), \widetilde{V}_h^k(x)$, which may be interpreted as follows:

- $\widetilde{Q}_h^k(x,a)$ is a refinement of the predictions $\widetilde{Q}_h(x,a)$ given to the algorithm as input; $\widetilde{Q}_h^1$ is set to equal $\widetilde{Q}_h$ (step 1), and $\widetilde{Q}_h^k$ is refined over time as the algorithm collects trajectories.
- The values $\overline{R}_h^k(x,a)$ are used in the process of refining $\widetilde{Q}_h^k(x,a)$; $\overline{R}_h^k(x,a)$ represents an approximate upper bound on $Q_h^\star(x,a)$, assuming that the prediction $\widetilde{Q}$ is an ϵ-approximate distillation (Definition 5).
- $\widetilde{V}_h^k(x)$ is an upper estimate for the V-value function at a state x that makes use of the refined predictions $\widetilde{Q}_h^k(x,a)$.

`QLearningPreds` additionally employs the technique of *action elimination*, maintaining sets $A_h^k(x)$ (defined in step 2c) which for each x, h, k contain the actions a which could plausibly be optimal at the beginning of episode k ($A_h^1(x)$ is initialized to all of $\mathcal{A}$ in step 1). Action elimination has previously been used in bandit learning and reinforcement learning when one must be robust to adversarial corruptions (Even-Dar et al., 2006; Lykouris et al., 2018, 2020), as well as to obtain gap-based regret bounds (Xu et al., 2021; Lykouris et al., 2018, 2020). In our algorithm, the sets $A_h^k(x)$ are again used for each of these purposes (where the robustness is with respect to the possible inaccuracy of the predictions $\widetilde{Q}_h$). For convenience, we set $\mathcal{G}_h^k$ to denote the set of states x for which $|A_h^k(x)| = 1$ (meaning all but one action at x has been eliminated at the beginning of episode k; see step 2f).

After being initialized in step 1 of `QLearningPreds`, the values $\overline{Q}_h^k, \underline{Q}_h^k, \overline{V}_h^k, \underline{V}_h^k, \widetilde{Q}_h^k, \widetilde{V}_h^k, \overline{R}_h^k$ are updated in `QLearningPreds` in steps 2b and 2d according to established updating procedures, namely using exploration bonuses of $b_n = C_0 \cdot \sqrt{H^3 \iota / n}$ (for some constant C_0) and a learning rate of $\alpha_n = \frac{H+1}{H+n}$, for $n \in \mathbb{N}$ (Jin et al., 2018; Xu et al., 2021). In particular, $\overline{V}_h^k, \underline{V}_h^k, \overline{Q}_h^k, \underline{Q}_h^k$ are updated in step 2b according to the adaptive multi-step bootstrap technique of Xu et al. (2021), which uses sequences of multiple rewards (namely, at contiguous sequences of states in which the optimal action has been determined) to perform the Bellman update. Our updates differ slightly from those in previous works in that we also maintain supplementary estimates $\overline{q}_h^k, \underline{q}_h^k$ (steps 2(b)iv and 2(b)vi) to ensure that $\overline{Q}_h^k, \overline{V}_h^k$ are non-increasing with respect to k, and $\underline{Q}_h^k, \underline{V}_h^k$ are non-decreasing with respect to k (Lemma 15).

The purpose of maintaining $\overline{V}_h^k, \underline{V}_h^k, \overline{Q}_h^k, \underline{Q}_h^k$ is primarily to obtain the robustness regret bounds (i.e., of $\frac{1}{\lambda} \cdot \mathscr{C}_{M,T,\lambda}$) in Theorems 8 and 9. On the other hand, the values $\widetilde{Q}_h^k, \widetilde{V}_h^k, \overline{R}_h^k$, which are updated in step 2d of `QLearningPreds`, are used to obtain improved regret bounds in the presence of accurate predictions. The updates here only use a single step to perform the Bellman update, as in the standard Q-learning algorithm (Jin et al., 2018).

For future reference we define the following learning rate parameters used in the algorithm's analysis: for $n \geq i \geq 1$, set

$$\alpha_0^0 := 1, \qquad \alpha_n^0 := 0, \qquad \alpha_n^i := \alpha_i \prod_{j=i+1}^{n} (1 - \alpha_j). \tag{10}$$

Intuitively, α_n^i denotes the impact of an update made the ith time a state-action pair (x, a, h) is visited on the value of any value function (e.g., $\overline{Q}_h^k, \underline{Q}_h^k$, etc.) when (x, a, h) is visited for the nth time. In the remainder of the section, we describe how `QLearningPreds` chooses its policies (step 2a); the challenge of doing so leads to most of the algorithmic novelties in `QLearningPreds`.

State-specific exploration & exploitation phases At each episode k, `QLearningPreds` chooses a policy π^k by using the functions $\overline{Q}_h^k, \underline{Q}_h^k, \widetilde{Q}_h^k$ in the `PolicySelection` subroutine (Algorithm 2). A key challenge addressed in this step is that of obtaining a "best of both worlds" guarantee which improves upon the minimax regret guarantee of $\widetilde{O}(\sqrt{SATH^{O(1)}})$ (or, in the gap-based case, $\mathrm{poly}(H) \cdot \widetilde{O}\left(\sum_{(x,a,h)} \frac{1}{\Delta_h(x,a)} + \frac{|\mathcal{A}^{\mathrm{mul}}|}{\Delta_{\min}}\right)$) in the case that the predictions $\widetilde{Q}$ are accurate, but still manages to obtain sublinear regret when $\widetilde{Q}$ is arbitrarily inaccurate. `QLearningPreds` overcomes this challenge by dividing the set of episodes in which we visit each state x at each step h into two phases:

- In the first phase, we employ *exploration*: whenever (x, h) is visited during an episode k in this phase, the policy π^k takes an action $a \in A_h^k(x)$ which maximizes the gap between $\overline{Q}_h^k(x, a)$ and $\underline{Q}_h^k(x, a)$ (this approach is slightly different from the more standard UCB approach which chooses a to maximize $\overline{Q}_h^k(x, a)$ (Jin et al., 2018), but was used in Xu et al. (2021) to obtain gap-based bounds; it is used in `QLearningPreds` for the same reason).

- After a certain number of episodes, `QLearningPreds` will decide it has sufficiently explored at the state (x, h), and thus, when visiting (x, h), it will choose an action $\hat{a} \in A_h^k(x)$ which maximizes the refined predictions $\widetilde{Q}_h^k(x, \hat{a})$[10]. This second phase may be seen as a *constrained exploitation* phase: it attempts to exploit the predictions $\widetilde{Q}_h$, but the action $\hat{a}$ is constrained to lie in the action set $A_h^k(x)$. As explained below, any action a' at x which is very suboptimal will be removed from $A_h^k(x)$ after a bounded number of episodes, which limits the impact of inaccurate predictions.

We emphasize that the partition into the two phases is *state-specific*; namely, at any given episode, some states may be in their exploration phase whereas others may be in their exploitation phase. Notice that there is a tradeoff between the lengths of the two phases: if the first phase, which does not make use of the predictions $\widetilde{Q}_h$ and thus cannot outperform the minimax bounds, is too long, then if the predictions $\widetilde{Q}$ are accurate we will not improve sufficiently upon the minimax regret guarantee. On the other hand, if the first phase is too short (or nonexistent), the following may occur: suppose that the predictions $\widetilde{Q}$ are inaccurate in that for some state x, step h, and suboptimal action a, $\widetilde{Q}_h(x, a)$ is large, but $\widetilde{Q}_h(x, a^\star)$ is small, where $a^\star \neq a$ is the unique optimal action at (x, h) and satisfies $V_h^\star(x) = Q_h^\star(x, a^\star) \gg Q_h^\star(x, a)$. Suppose for simplicity that $\mathcal{A} = \{a, a^\star\}$ and that $\widetilde{Q}_h \equiv \widetilde{Q}_h^k$ (which can approximately hold). Ideally the first phase should be long enough to eliminate a from $A_h^k(x)$; this will happen when $\underline{Q}_h^k(x, a^\star)$ grows sufficiently to be greater than $\overline{Q}_h^k(x, a)$. However, if the first phase ends before this happens, then at the beginning of the second phase, $A_h^k(x) = \mathcal{A}$, and so $\pi_h^k(x)$ will be set to a in step 2 of `PolicySelection`. Thus `QLearningPreds` would suffer linear regret.

An adaptive exploration-exploitation cutoff `QLearningPreds` trades off the lengths of the exploration and exploitation phases described above according to the input parameter λ (or, in the case of Theorem 9, $\widehat{\lambda}$ as determined by $\mathscr{R}$). To describe how `QLearningPreds` makes this tradeoff, we begin by defining the *Q- and V-range functions* (following the presentation of Xu et al. (2021)). First, we make a few additional definitions: for $(k, h) \in [K] \times [H]$, for which $x_h^k \notin \mathcal{G}_h^k$, set $h'(k, h) \in [H + 1]$ to be the first step h' after step h for which $x_{h'}^k \notin \mathcal{G}_h^k$ (if such $h' \leq H$ does not exist, then set $h'(k, h) = H + 1$). Next, for $n \in \mathbb{N}$, define the following parameters β_n, which may be viewed as aggregated versions of the exploration bonuses $b_i = C_0\sqrt{H^3\iota/i}$ (recall the definition of α_n^i in (10)):

$$\beta_0 := 0, \qquad \beta_n = 2\sum_{i=1}^{n} \alpha_n^i \cdot b_i. \tag{11}$$

[10] For technical reasons, $\hat{a}$ is actually chosen to maximize $\max\{\widetilde{Q}_h^k(x, \hat{a}), \underline{Q}_h^k(x, \hat{a})\}$

Definition 10 (Range function) *For $(x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ for which $x \notin \mathcal{G}_h^k$ and $a \in A_h^k(x)$, define the* range Q-function *as follows: set $\Delta Q_h^0(x, a) = H$, and*

$$\Delta Q_h^k(x,a) := \min\left\{ \Delta Q_h^{k-1}(x,a),\ \alpha_n^0 H + \beta_n + \sum_{i=1}^n \alpha_n^i \cdot \Delta V_{h'(k_h^i,h)}^{k_h^i}\big(x_{h'(k_h^i,h)}^{k_h^i}\big) \right\}$$
$$\textit{where} \quad n = N_h^k(x,a),\ k_h^i = k_h^i(x,a)\ \forall i \in [n].$$

Moreover, for $(x, h, k) \in \mathcal{S} \times [H] \times [K]$ for which $x \notin \mathcal{G}_h^k$, define the range V-function *as follows: set $\Delta V_h^0(x) = H$, and*

$$\Delta V_h^k(x) := \min\{\Delta V_h^{k-1}(x),\ \Delta Q_h^k(x, a^\star)\} \quad \textit{for} \quad a^\star = \arg\max_{a' \in A_h^k(x)} \overline{Q}_h^k(x, a') - \underline{Q}_h^k(x, a').$$

Finally, define $\Delta Q_{H+1}^k(x, a) = \Delta V_{H+1}^k(x) = 0$ for all x, a, k.

The functions $\Delta Q_h^k, \Delta V_h^k$ should be interpreted as upper bounds on the gap between the upper and lower Q, V values; note that they satisfy a similar recursion to $\overline{Q}_h^k - \underline{Q}_h^k$ and $\overline{V}_h^k - \underline{V}_h^k$ (see (24)). Indeed, in Lemma 16 below we show that $\Delta Q_h^k, \Delta V_h^k$ are upper bounds on $\overline{Q}_h^k - \underline{Q}_h^k, \overline{V}_h^k - \underline{V}_h^k$, respectively.

Now that we have defined the range functions, the choice of policy at each (x, h) (equivalently, the choice of "exploration" and "constrained exploitation" phases described above) is simple to state: `QLearningPreds` maintains a parameter $\widehat{\Delta}^k$ at each episode k, which represents a "target error bound" that `QLearningPreds` hopes to obtain. The parameter $\widehat{\Delta}^k$ adapts to the input parameter $\mathscr{R}$ (or λ) as well as the gap-based complexity of the given MDP. Given $\widehat{\Delta}^k$ at episode k, the policy π_h^k at each step h is specified in (12) in the algorithm `PolicySelection`: following our terminology above, a state (x, h) is declared to be in the "exploration" phase if $\Delta \check{V}_h^k(x) > \varphi_h(\widehat{\Delta}^k)$[11] (meaning there is still much uncertainy about the optimal value at (x, h) relative to $\widehat{\Delta}^k$), and is defined to be in the "constrained exploitation" phase otherwise (i.e., $\Delta \check{V}_h^k(x) \le \varphi_h(\widehat{\Delta}^k)$). We will show (in Lemmas 19 and 20) that $\Delta \check{V}_h^k(x)$ is nonincreasing with respect to k and $\varphi_h(\widehat{\Delta}^k)$ is nondecreasing with respect to k; thus, each state can only move from the "exploration" to "constrained exploitation" phase.

A.2. How to choose $\widehat{\Delta}^k$

As we discussed in the previous section, $\widehat{\Delta}^k$ is chosen to adapt to the input parameter $\mathscr{R}$ or λ. In the setting of Theorem 9, where the user inputs a parameter $\mathscr{R}$ representing the target worst-case regret bound, the choice of $\widehat{\Delta}^k$ is extremely simple (Algorithm 3, `DeltaConst`): for all k, we set $\widehat{\Delta}^k := \mathscr{R}/(KH)$. In the setting of Theorem 8, where the user inputs a parameter λ specifying a trade-off between the worst-case and ideal-case settings, $\widehat{\Delta}^k$ is set (in Algorithm 4, `DeltaIncr`) to a more complex expression which is a surrogate for the worst-case regret bound $\frac{1}{\lambda}\mathscr{C}_{M,T,\lambda}$ (thus overcoming the challenge that the algorithm does not know $\mathscr{C}_{M,T,\lambda}$). This surrogate uses the *frozen range function* (defined in Definition 12), denoted $\Delta \mathring{Q}_h^k(x, a)$, as a proxy for the action-value gaps $\Delta_h(x, a)$, for all $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. We refer the reader to Section B.5 for further details.

[11] Recall that for $a > 0$ and some constant C_1, we have defined $\varphi_h(a) = C_1 \cdot \left(1 + \frac{1}{H}\right)^{4(H+1-h)} \cdot a$ in `QLearningPreds`. Thus $\varphi_h(\widehat{\Delta}^k) = \Theta(\widehat{\Delta}^k)$; the function $\varphi_h(\cdot)$ is introduced for technical considerations in the proof.

```
Algorithm 1: QLearningPreds
Input: State space $\mathcal{S}$, action space $\mathcal{A}$, horizon $H$, number of episodes $K$, predictions $\widetilde{Q}_h : \mathcal{S}\times\mathcal{A} \to [0, H]$ for all $h \in [H]$, parameter $\lambda \in [0, 1]$. For some constant $C_1 > 0$ and $1 \le h \le H+1$, set, for $a > 0$, $\varphi_h(a) = C_1 \cdot \left(1 + \frac{1}{H}\right)^{4(H+1-h)} \cdot a$.
1. Initialize $N_h^1(x, a) = 0$, $\overline{R}_h^1(x, a) = \overline{Q}_h^1(x, a) = \overline{V}_h^1(x, a) = \overline{q}_h^1(x, a) = H$, $\underline{Q}_h^1(x, a) = \underline{V}_h^1(x, a) = \underline{q}_h^1(x, a) = 0$ for all $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Also set $\widetilde{Q}_h^1 = \widetilde{Q}_h$ and $\widetilde{V}_h^1(x) = \max_{a'\in\mathcal{A}} \widetilde{Q}_h^1(x, a')$ for all $(x, h) \in \mathcal{S} \times [H]$. Set $A_h^1(x) = \mathcal{A}$ for all $(x, h) \in \mathcal{S} \times [H]$, and $\mathcal{G}_h^1 = \emptyset$ for all $h \in [H]$. Set $\widehat{\Delta}^1 \leftarrow 0$.
2. For episode $1 \le k \le K$:
   (a) Receive $\pi^k$ and the policy rollout $(x_1^k, a_1^k), \ldots, (x_H^k, a_H^k)$ from the PolicySelection algorithm.
   (b) For each $h = 1, 2, \ldots, H$ such that $x_h^k \notin \mathcal{G}_h^k$:
       i. Set $N_h^{k+1} \leftarrow N_h^k(x_h^k, a_h^k) + 1$, $n = N_h^{k+1}(x_h^k, a_h^k)$, and write $b_n = C_0\sqrt{H^3\iota/n}$.
       ii. Let $x_{h'}^k$ be the first state in the episode after $x_h^k$ so that $x_{h'}^k \notin \mathcal{G}_{h'}^k$ (if such $h'$ does not exist, set $h' = H + 1$).
       iii. Let $\widehat{r}_h^k = \sum_{h''=h}^{h'-1} r_{h''}(x_{h''}^k, a_{h''}^k)$.
       iv. Set $\overline{q}_h^{k+1}(x_h^k, a_h^k) \leftarrow (1 - \alpha_n) \cdot \overline{q}_h^k(x_h^k, a_h^k) + \alpha_n \cdot (\widehat{r}_h^k + \overline{V}_{h'}^k(x_{h'}^k) + b_n)$.
       v. Set $\overline{Q}_h^{k+1}(x_h^k, a_h^k) \leftarrow \min_{k'\le k+1} \left\{\overline{q}_h^{k'}(x_h^k, a_h^k)\right\}$.
       vi. Set $\underline{q}_h^{k+1}(x_h^k, a_h^k) \leftarrow (1 - \alpha_n) \cdot \overline{q}_h^k(x_h^k, a_h^k) + \alpha_n \cdot (\widehat{r}_h^k + \underline{V}_{h'}^k(x_{h'}^k) - b_n)$.
       vii. Set $\underline{Q}_h^{k+1}(x_h^k, a_h^k) \leftarrow \max_{k'\le k+1} \left\{\underline{q}_h^{k'}(x_h^k, a_h^k)\right\}$.
       viii. Set $\underline{V}_h^{k+1}(x_h^k) \leftarrow \max_{a'\in A_h^k(x_h^k)}\{\underline{Q}_h^{k+1}(x_h^k, a')\}$.
       ix. Set $\overline{V}_h^{k+1}(x_h) \leftarrow \max_{a'\in A_h^k(x_h)}\{\overline{Q}_h^{k+1}(x_h, a')\}$.
   (c) For all $(x, h) \in \mathcal{S} \times [H]$, set $A_h^{k+1}(x) \leftarrow \{a' \in A_h^k(x) : \overline{Q}_h^{k+1}(x, a') \ge \underline{V}_h^{k+1}(x)\}$.
   (d) For each $h = 1, 2, \ldots, H$:
       i. Set $\overline{R}_h^{k+1}(x_h^k, a_h^k) \leftarrow (1-\alpha_n)\cdot\overline{R}_h^k(x_h^k, a_h^k)+\alpha_n\cdot(r_h(x_h^k, a_h^k)+\widetilde{V}_{h+1}^k(x_{h+1})+b_n)$.
       ii. Set $\widetilde{Q}_h^{k+1}(x_h^k, a_h^k) \leftarrow \min\{\overline{R}_h^{k+1}(x_h^k, a_h^k), \widetilde{Q}_h^k(x_h^k, a_h^k), \overline{Q}_h^{k+1}(x_h^k, a_h^k)\}$.
       iii. Set $\widetilde{V}_h^{k+1}(x_h^k) \leftarrow \max_{a'\in A_h^{k+1}(x_h^k)} \max\{\widetilde{Q}_h^{k+1}(x_h^k, a'), \underline{Q}_h^{k+1}(x_h^k, a')\}$.
   (e) For all $h$ and all $(x, a) \neq (x_h^k, a_h^k)$ set $N_h^{k+1}(x, a), \overline{Q}_h^{k+1}(x, a), \underline{Q}_h^{k+1}(x, a), \overline{q}_h^{k+1}(x, a)$, $\underline{q}_h^{k+1}(x, a), \overline{V}_h^{k+1}(x), \underline{V}_h^{k+1}(x), \overline{R}_h^{k+1}(x, a), \widetilde{Q}_h^{k+1}(x, a), \widetilde{V}_h^{k+1}(x)$ equal to their values at episode $k$.
   (f) For all $h \in [H]$, set $\mathcal{G}_h^{k+1} \leftarrow \{x \in \mathcal{S} : |A_h^{k+1}(x)| = 1\}$.
   (g) Choose $\widehat{\Delta}^{k+1}$ according to either DeltaConst or DeltaIncr.
```

Algorithm 2: `PolicySelection`

Input: Internal state of the algorithm `QLearningPreds` at the beginning of episode k (including, in particular, the previous pollicy rollouts, and the functions $\widetilde{Q}_h^k, \underline{Q}_h^k, \overline{Q}_h^k$, as well as $\widehat{\Delta}^k, \mathcal{G}_h^k, A_h^k$).

1. For $h \in [H]$, construct $\Delta V_h^k(\cdot)$ per Definition 10.
2. Define the policy π^k by, for $(x, h) \in \mathcal{S} \times [H]$:

$$\pi_h^k(x) := \begin{cases} \text{The action in } A_h^k(x) & \text{if } |A_h^k(x)| = 1 \\ \arg\max_{a \in A_h^k(x)} \{\max\{\widetilde{Q}_h^k(x,a), \underline{Q}_h^k(x,a)\}\} & \text{if } \Delta V_h^k(x) \leq \varphi_h(\widehat{\Delta}^k) \\ \arg\max_{a \in A_h^k(x)} \{\overline{Q}_h^k(x,a) - \underline{Q}_h^k(x,a)\} & \text{if } \Delta V_h^k(x) > \varphi_h(\widehat{\Delta}^k) \end{cases} \tag{12}$$

3. Let $(x_1^k, a_1^k), \ldots, (x_H^k, a_H^k)$ be a policy rollout obtained by following π^k.
4. Return the policy π^k and the policy rollout $(x_1^k, a_1^k), \ldots, (x_H^k, a_H^k)$.

Algorithm 3: `DeltaConst`

Input: Episode number k, input regret bound $\mathscr{R}$ of `QLearningPreds`, and total number K of episodes.

1. Return

$$\widehat{\Delta}^{k+1} := \mathscr{R}/(KH). \tag{13}$$

Algorithm 4: `DeltaIncr`

Input: Internal state of the algorithm `QLearningPreds` at the beginning of episode $k+1$ (in particular, the necessary information to compute the frozen Q-range function), and parameters $\lambda \in [0, 1]$ and $\widetilde{\Delta}_{\text{min}} \geq 0$ (which is guaranteed to satisfy $\widetilde{\Delta}_{\text{min}} \leq \Delta_{\text{min}}$).

1. For $h \in [H]$, construct $\widetilde{\Delta}\mathring{Q}_h^{k+1}(\cdot)$ which is defined identically to $\Delta\mathring{Q}_h^{k+1}(\cdot)$ per Definition 12, except with the parameter $\widetilde{\Delta}_{\text{min}}$ replacing Δ_{min} in the clipped value functions $\Delta\check{V}_h^{k+1}, \Delta\check{Q}_h^{k+1}$.
2. Return

$$\widehat{\Delta}^{k+1} := \min\left\{ \frac{H^6 \iota^2}{\lambda \cdot K} \cdot \sum_{(x,a,h)} \frac{1}{\max\left\{\frac{\widetilde{\Delta}\mathring{Q}_h^{k+1}(x,a)}{2H}, \frac{\widetilde{\Delta}_{\text{min}}}{4H^2}\right\}}, \sqrt{\frac{SAH^8 \iota^2}{\lambda \cdot K}} \right\}. \tag{14}$$

Appendix B. Proof overview

In this section we overview the proofs of Theorems 8 and 9; we focus mainly on Theorem 9 since its proof is slightly simpler. At a high level, the key tools needed in the proof of Theorem 9 are as follows:

1. First, we need to define the the *clipped range functions* (Definition 11), as in Xu et al. (2021), which aid in proving gap-based bounds.

2. To prove the $O(\mathscr{R})$ regret bound for worst-case predictions (i.e., robustness, first item of Theorem 9), we first prove a regret decomposition (Lemma 29) showing that regret can be bounded in terms of the clipped V-range functions.

   In Lemma 27, our main technical lemma for the worst-case regret bound, we then show how to bound the clipped V-range functions in the presence of arbitrary predictions $\widetilde{Q}$ using certain *monotonicity* properties of the value functions.

3. To establish the improved regret bounds for the case that $\widetilde{Q}$ is an approximate distillation (second item of Theorem 9), we first need to bound the number of episodes during which the predictions $\widetilde{Q}$ are *not* used to choose the policy.

   Then we upper bound the value functions $\overline{R}_h^k, \widetilde{V}_h^k$ at the set of episodes k where the predictions *are* used and show that doing so suffices to bound regret.

The proof of Theorem 8 is similar to that of Theorem 9. One additional tool needed (which shows up in the algorithm `DeltaIncr`) is a variation of the clipped range functions that we call the *frozen range functions* (Definition 12).

In Sections B.1 through B.4 we expand upon the above items to overview the proof of Theorem 9. In Section B.5 we overview the changes that must be made to `QLearningPreds` and the proof to establish Theorem 8.

B.1. Clipped range functions

We begin by defining the *clipped Q-value and V-value functions*, which were originally introduced in Xu et al. (2021) to obtain gap-based bounds on the regret (they play a similar role in this paper). For real numbers x, y, define the *clip function* as follows: $\mathrm{clip}\,[x|\, y] := \mathbb{1}[x \geq y] \cdot x$.

Definition 11 (Clipped range function, Xu et al. (2021)) *For all $(x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ for which $x \notin \mathcal{G}_h^k$ and $a \in A_h^k(x)$, define the* clipped range Q-function *as follows: set $\Delta\breve{Q}_h^0(x, a) = H$, and*

$$\Delta\breve{Q}_h^k(x,a) := \min\left\{\Delta\breve{Q}_h^{k-1}(x,a),\ \alpha_n^0 H + \mathrm{clip}\left[\beta_n \middle| \frac{\Delta_{\min}}{4H^2}\right] + \sum_{i=1}^n \alpha_n^i \cdot \Delta\breve{V}_{h'(k_h^i,h)}^{k_h^i}(x_{h'(k_h^i,h)}^{k_h^i})\right\} \tag{15}$$

$$\text{where} \quad n = N_h^k(x,a),\ k_h^i = k_h^i(x,a)\ \forall i \in [n].$$

Moreover, for $(x, h, k) \in \mathcal{S} \times [H] \times [K]$ for which $x \notin \mathcal{G}_h^k$, define the clipped range V-function *as follows: set $\Delta\breve{V}_h^0(x) = H$, and*

$$\Delta\breve{V}_h^k(x) := \min\{\Delta\breve{V}_h^{k-1}(x),\ \Delta\breve{Q}_h^k(x,a^\star)\} \quad \textit{for} \quad a^\star = \operatorname*{arg\,max}_{a' \in A_h^k(x)} \overline{Q}_h^k(x,a') - \underline{Q}_h^k(x,a').$$

Finally, define $\Delta\breve{V}_{H+1}^k(x) = \Delta\breve{Q}_{H+1}^k(x,a) = 0$ *for all* x, a, k.

The clipped range functions $\Delta\breve{Q}_h^k(x,a), \Delta\breve{V}_h^k(x)$ are defined to satisfy a similar recursion as the quantities $\overline{Q}_h^k(x,a) - \underline{Q}_h^k(x,a)$ and $\overline{V}_h^k(x) - \underline{V}_h^k(x)$ (see (24)). Unlike in (24), in the definition of $\Delta\breve{Q}_h^k, \Delta\breve{V}_h^k$, the bonuses β_n are clipped, leading $\Delta\breve{V}_h^k, \Delta\breve{Q}_h^k$ to be smaller than their unclipped counterparts (Lemma 18), which aids in obtaining gap-based regret bounds. Despite this clipping, the combination of Lemmas 16 and 17 shows that, with high probability, for all x, a, h, k, the clipped range functions are still approximately lower bounded by the gap between the upper and lower Q, V-values, as follows:

$$\Delta\breve{Q}_h^k(x,a) \geq \overline{Q}_h^k(x,a) - \underline{Q}_h^k(x,a) - \frac{\Delta_{\min}}{4H}, \qquad \Delta\breve{V}_h^k(x) \geq \overline{V}_h^k(x) - \underline{V}_h^k(x) - \frac{\Delta_{\min}}{4H}. \tag{16}$$

B.2. Worst-case regret bound

In this section we overview the proof that `QLearningPreds` achieves regret $O(\mathscr{R})$ for arbitrary predictions $\widetilde{Q}$ in the setting of Theorem 9. For all (h,k) for which $x_h^k \notin \mathcal{G}_h^k$, define $\breve{\delta}_h^k = \Delta\breve{V}_h^k(x_h^k)$. Using (16), the following regret decomposition is straightforward to prove (it is similar to that in Lemma B.6 of Xu et al. (2021)).

Lemma 29 (Regret decomposition; abbreviated) *There is an event* $\mathcal{E}^{\mathrm{wc}}$ *that occurs with probability at least* $1 - 1/(H^2K)$ *so that the regret of* `QLearningPreds` *may be bounded as follows:*

$$\sum_{k=1}^K \mathbb{E}\left[V_1^\star(x_1^k) - V_1^{\pi^k}(x_1^k)\right] \leq 1 + 4 \cdot \mathbb{E}\left[\sum_{(k,h): a_h^k \notin \mathcal{A}_{h,0}^{\mathrm{opt}}(x_h^k)} \breve{\delta}_h^k \middle| \mathcal{E}^{\mathrm{wc}}\right].$$

(The right-hand side of the above expression makes sense since under the event $\mathcal{E}^{\mathrm{wc}}$*, it turns out that for all* (k,h) *so that* $a_h^k \notin \mathcal{A}_{h,0}^{\mathrm{opt}}(x_h^k)$, $x_h^k \notin \mathcal{G}_h^k$*, i.e.,* $\breve{\delta}_h^k$ *is well-defined.)*

Lemma 29 reduces the problem of bounding the regret to bounding the clipped value functions $\breve{\delta}_h^k$ for h, k such that $x_h^k \notin \mathcal{G}_h^k$. In turn, we bound $\breve{\delta}_h^k$ in Lemma 27, of which a simplified version combining it with Lemma 30 is presented below:

Lemma 27 (Abbreviated & combined with Lemma 30) *Fix any* $h \in [H]$*, any set* $\mathcal{W} \subset [K]$ *so that for all* $k \in \mathcal{W}$, $x_h^k \notin \mathcal{G}_h^k$*, and any* $k^\star \geq \max_{k\in\mathcal{W}}\{k\}$*. Then*

$$\sum_{k\in\mathcal{W}} \breve{\delta}_h^k \leq |\mathcal{W}| \cdot \varphi_h(\widehat{\Delta}^{k^\star}) + O\left(\min\left\{\sqrt{H^5SA|\mathcal{W}|\iota},\ H^5\iota \cdot \left(\sum_{(x,a,h')} \frac{1}{\Delta_{h'}(x,a)} + \frac{|\mathcal{A}^{\mathrm{mul}}|}{\Delta_{\min}}\right)\right\}\right). \tag{17}$$

The starting point for the proof of Lemma 27 is to use the definition of the clipped value functions $\Delta\breve{V}_h^k$ together with reverse induction on h (i.e., bounding the values $\breve{\delta}_h^k$ in terms of $\breve{\delta}_{h'}^k$ for $h' > h$) in a similar manner as was done in Lemma B.8 of Xu et al. (2021). However, the proof of Lemma 27 must depart from that of (Xu et al., 2021, Lemma B.8) since in the `PolicySelection` subroutine of `QLearningPreds`, we do not always choose the action $a \in A_h^k(x)$ maximizing the *confidence interval*, i.e., maximizing $\overline{Q}_h^k(x,a) - \underline{Q}_h^k(x,a)$. Typically such an action choice maximizing the confidence interval is necessary to upper bound the values $\breve{\delta}_h^k$. We are able to nevertheless bound $\breve{\delta}_h^k$ using the fact that for steps (h, k) for which we do *not* choose the action a maximizing the confidence interval (i.e., we are in the constrained exploitation phase), it must hold that $\breve{\delta}_h^k = \Delta \breve{V}_h^k(x_h^k) \leq \Delta V_h^k(x_h^k) \leq \varphi_h(\widehat{\Delta}^k)$. This observation leads to the quantity $|\mathcal{W}| \cdot \varphi_h(\widehat{\Delta}^{k^\star})$ on the right-hand side of (17).

The proof of Lemma 27 is made somewhat more complex by the fact that the choice of action at step h affects $\breve{\delta}_{h'}^k$ for $h' < h$ and various k (via the reverse induction argument), and without care one will end up with a multiplier of $\varphi_h(\widehat{\Delta}^{k^\star})$ in (17) that is much larger than $|\mathcal{W}|$. To avoid this complication, we must carefully account for the effect the values $\breve{\delta}_h^k$ have on the bounds we can prove on $\breve{\delta}_{h'}^k$ for $h' < h$. To do so we make use of a monotonicity propery of the clipped value functions (Lemma 19, showing that $\Delta \breve{V}_h^k(x)$ is non-increasing with k) and introduce the notion of *level-h* sets (Definition 24) which are intermediate sets $\mathcal{W}'$ of tuples (k', h') for which we need to bound $\sum_{(k',h') \in \mathcal{W}'} \breve{\delta}_{h'}^{k'}$ in the course of the induction.

Completing the worst-case regret bound The proof of item 1 of Theorem 9 is fairly straightforward given the above components; the details are worked out in Lemma 31. The dominant term in the bound (17) turns out to be $|\mathcal{W}| \cdot \varphi_h(\widehat{\Delta}^{k^\star})$, which due to the choice $\widehat{\Delta}^k = \mathscr{R}/(HK)$ and the bound $|\mathcal{W}| \leq K$, leads to the bound $O(\mathscr{R})$ on regret.

B.3. Exploration-constrained exploitation cutoffs

Before discussing the proof of the improved regret bound for the case of $\widetilde{Q}$ being an approximate distillation, we introduce the following notation relating to the exploration and constrained exploitation phases in `QLearningPreds` that we discussed above. For $(k, h) \in [K] \times [H]$, define $\tau_h^k \in \{0, 1\}$ as follows:

$$\tau_h^k = \begin{cases} 0 & \text{if } x_h^k \in \mathcal{G}_h^k \text{ or } \Delta V_h^k(x_h^k) \leq \varphi_h(\widehat{\Delta}^k) \\ 1 & \text{otherwise.} \end{cases} \tag{18}$$

The parameter τ_h^k is the indicator of whether `QLearningPreds` is in the exploration or constrained exploitation step at step h of episode k: if $\tau_h^k = 0$, then we have either determined the optimal action at x_h^k (i.e., $x_h^k \in \mathcal{G}_h^k$), or else the range function $\Delta V_h^k(x_h^k)$ is sufficiently small, so we engage in constrained exploitation (see the choice of π_h^k in (12), which chooses $a' \in A_h^k(x_h^k)$ maximizing $\max\{\widetilde{Q}_h^k(x_h^k, a'), \underline{Q}_h^k(x_h^k, a')\}$), and otherwise, if $\tau_h^k = 1$, we use optimistic exploration, choosing $a' \in A_h^k(x_h^k)$ to maximize the confidence interval.

Note that the parameters τ_h^k depend on the *unclipped* range functions ΔV_h^k; as we have discussed above, in order obtain our gap-based bounds, it is necessary to bound the *clipped* range functions $\Delta \breve{V}_h^k$. Therefore, when reasoning about the exploration and constrained exploitation phases, we will additionally introduce the parameters $\sigma_h^k \in \{0, 1\}$ (for $(k, h) \in [K] \times [H]$), which are defined similarly to τ_h^k except with respect to $\Delta \breve{V}_h^k$:

$$\sigma_h^k = \begin{cases} 0 & \text{if } x_h^k \in \mathcal{G}_h^k \text{ or } \Delta \breve{V}_h^k(x_h^k) \leq \frac{1}{1+\frac{1}{H}} \cdot \varphi_h(\widehat{\Delta}^k) \\ 1 & \text{otherwise.} \end{cases} \tag{19}$$

The parameters σ_h^k can be thought of as a proxy for the true exploration parameters τ_h^k. As discussed in the following section, in order to establish improved regret bounds for the case that $\widetilde{Q}$ is an approximate distillation, we need to, loosely speaking, upper bound the number of episodes in which we engage in exploration (i.e., in which the predictions $\widetilde{Q}$ are *not* used). For technical reasons, it turns out to be more convenient to bound the number of (k,h) so that $\sigma_h^k = 1$ (as opposed to bounding the number of (k,h) so that $\tau_h^k = 1$).

B.4. Proofs for $\widetilde{Q}$ an approximate distillation

Now we discuss the proof of item 2 of Theorem 9; the proof of item 2 of Theorem 8 is very similar (see Section D for the full proof). As discussed in the previous section, the first step is to bound the number of episodes for which we do *not* engage in constrained exploitation; in particular, for each h, we bound the number of k for which $\sigma_h^k = 1$:

Lemma 33 *Suppose* `QLearningPreds` *is run with* `DeltaConst` *to choose the values $\widehat{\Delta}^k$. Then for all $h \in [H]$, the number of episodes $k \in [K]$ for which $\sigma_h^k = 1$ is at most* $\max\{SAH^3, \widehat{\lambda} \cdot K\}$. *(Recall that $\widehat{\lambda}$ is chosen so that $\mathscr{R} = \frac{1}{\widehat{\lambda}} \cdot \mathscr{C}_{M,T,\widehat{\lambda}}$.)*

We write $\widehat{\Delta} = \widehat{\Delta}^k$ (as all $\widehat{\Delta}^k$ are equal). Also, for any $h \in [H]$, write $\mathcal{Y}_h := \{k : \sigma_h^k = 1\}$. The main tool in the proof of Lemma 33 is Lemma 27, which upper bounds $\sum_{k \in \mathcal{Y}_h} \breve{\delta}_h^k$ by the sum of $|\mathcal{Y}_h| \cdot (1 + 1/H)^2 \cdot \varphi_{h+1}(\widehat{\Delta})$ and some additional terms. On the other hand, that $\sigma_h^k = 1$ implies that $\breve{\delta}_h^k \geq \frac{1}{1+1/H} \cdot \varphi_h(\widehat{\Delta})$. These facts (together with the fact that $\varphi_h(\widehat{\Delta})$ is greater than $\varphi_{h+1}(\widehat{\Delta})$ by a factor of $(1+1/H)^4$) allow us to upper bound $|\mathcal{Y}_h|$ in terms of an expression which ultimately simplifies to $\max\{SAH^3, \widehat{\lambda} \cdot K\}$.

Regret decomposition and induction Given Lemma 33, we proceed to complete the proof of item 2 of Theorem 9. The first step is the following regret decomposition (stated in (89)), which follows from the fact that $\widetilde{Q}$ is an ϵ-approximate distillation as well as the definition of $\widetilde{V}_h^k$ in `QLearningPreds`: for any $\epsilon' > 0$, we have

$$\begin{aligned}
&\sum_{k=1}^{K} \mathbb{E}\left[V_1^\star(x_1^k) - V_1^{\pi^k}(x_1^k)\right] \\
\leq & O(KH(\epsilon H + \epsilon')) + \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}(1-\tau_h^k)\cdot \mathbb{1}[a_h^k \not\in \mathcal{A}_{h,\epsilon'}^{\mathrm{opt}}(x_h^k)] \cdot (\overline{R}_h^k(x_h^k, a_h^k) - Q_h^\star(x_h^k, a_h^k)) + \sum_{k=1}^{K}\sum_{h=1}^{H} 4\sigma_h^k \breve{\delta}_h^k \middle| \mathcal{E}^{\mathrm{wc}}\right].
\end{aligned} \tag{20}$$

The above regret decomposition reduces bounding the regret to bounding the following two types of quantities (under the event $\mathcal{E}^{\mathrm{wc}}$):

1. The quantity $\overline{R}_h^k(x_h^k, a_h^k) - Q_h^\star(x_h^k, a_h^k)$, for (k,h) satisfying $\tau_h^k = 0$ and $a_h^k \not\in \mathcal{A}_{h,\epsilon'}^{\mathrm{opt}}(x_h^k)$;
2. The quantity $\breve{\delta}_h^k$ for (k,h) satisfying $\sigma_h^k = 1$.[12]

The latter of these quantities (i.e., item 2) is straightforward to control: for each $h \in [H]$, we use Lemma 27 with the set $\mathcal{W}$ equal to the set of k so that $\sigma_h^k = 1$ and $k^\star = K$. Crucially, the conclusion of Lemma 33 above gives that $|\mathcal{W}| \leq \max\{SAH^3, \widehat{\lambda} \cdot K\}$, which, together with the inequality (17) of Lemma 27, gives us that $\sum_{k=1}^{K}\sum_{h=1}^{H} \sigma_h^k \breve{\delta}_h^k$ may be bounded by $O(\mathscr{C}_{M,T,\widehat{\lambda}})$. This argument is carried out formally in Lemma 44.

[12] Note that $\sigma_h^k = 1$ implies that $x_h^k \not\in \mathcal{G}_h^k$, which implies that $\breve{\delta}_h^k$ is indeed well-defined.

Bounding $\overline{R}_h^k, \widetilde{V}_h^k$ on non-exploratory episodes We next describe how the quantity in item 1 above is bounded. Our general strategy is to use the definition of $\overline{R}_h^k$ in terms of $\widetilde{V}_{h+1}^k$ (step 2(d)i of `QLearningPreds`) to bound the gaps $\overline{R}_h^k(x_h^k, a_h^k) - Q_h^\star(x_h^k, a_h^k)$ at step h in terms of the gaps $\widetilde{V}_{h+1}^{k'}(x_{h+1}^{k'}) - V_{h+1}^\star(x_{h+1}^{k'})$ at step $h+1$, for appropriate choices of k'. In turn, we will bound the gaps $\widetilde{V}_{h+1}^{k'}(x_{h+1}^{k'}) - V_{h+1}^\star(x_{h+1}^{k'})$ in terms of the gaps $\overline{R}_{h+1}^{k'}(x_{h+1}^{k'}, a_{h+1}^{k'}) - Q_{h+1}^\star(x_{h+1}^{k'}, a_{h+1}^{k'})$, completing the inductive step. When proving these bounds, we must take care to meet our goal of obtaining a regret bound (see (9)) that only has terms corresponding to tuples (x, a, h) belonging to the fooling set $\mathcal{F}(\epsilon(H+1), \epsilon')$. To do so, we use the following claim:

Claim 43 *For any (k, h) satisfying $\tau_h^k = 0$, if either*

1. *$a_h^k \notin \mathcal{A}_{h,\epsilon'}^{\mathrm{opt}}(x_h^k)$; or*
2. *$(\widetilde{V}_h^k - V_h^\star)(x_h^k) > \epsilon'$,*

then under the event $\mathcal{E}^{\mathrm{wc}}$ it holds that $(x_h^k, a_h^k, h) \in \mathcal{F}(\epsilon(H+1), \epsilon')$.

Claim 43 allows us to upper bound the term $\sum_{k=1}^K \sum_{h=1}^H (1-\tau_h^k) \cdot \mathbb{1}[a_h^k \notin \mathcal{A}_{h,\epsilon'}^{\mathrm{opt}}(x_h^k)] \cdot (\overline{R}_h^k(x_h^k, a_h^k) - Q_h^\star(x_h^k, a_h^k))$ in the regret decomposition (20) with a sum of $\overline{R}_h^k(x, a) - Q_h^\star(x, a)$ over only those $(x, a, h) \in \mathcal{F}(\epsilon(H+1), \epsilon')$. In turn, for such tuples (x, a, h), it is possible to upper bound $\overline{R}_h^k(x, a) - Q_h^\star(x, a)$ in terms of the sum of β_n (for $n = N_h^k(x, a)$) and a weighted sum of $(\widetilde{V}_{h+1}^{k'} - V_{h+1}^\star)(x_{h+1}^{k'})$ for certain values of k' (see Lemma 38). The terms β_n in this sum form the main contribution to the regret bound (9); crucially we use the fact that we only have such terms for $(x, a, h) \in \mathcal{F}(\epsilon(H+1), \epsilon')$.

Finally, when completing the inductive step by bounding the gaps $\widetilde{V}_{h+1}^{k'}(x_{h+1}^{k'}) - V_{h+1}^\star(x_{h+1}^{k'})$, we again have to ensure that we only use terms of the form $(\overline{R}_{h+1}^{k'}(x, a) - Q_{h+1}^\star(x, a))$ in our upper bound for which $(x, a, h+1) \in \mathcal{F}(\epsilon(H+1), \epsilon')$. For this we again use Claim 43 (with the second option). We refer the reader to Section D.4 for further details.

B.5. Proof of Theorem 8: implicit-λ bound

The proof of Theorem 8 is similar to that of Theorem 9. The main difference is that, because the algorithm is not given as input the target worst-case regret bound $\mathscr{R}$ (which in turn is used to choose $\widehat{\Delta}^k$ in `DeltaConst` for the proof of Theorem 9), it must construct a proxy value to assign to $\widehat{\Delta}^k$. This proxy is constructed in `DeltaIncr` (Algorithm 4): for each episode k, $\widehat{\Delta}^k$ is set in (14) to equal an expression which resembles the definition of $\frac{1}{\lambda} \cdot \mathscr{C}_{M,T,\lambda}$ in (7), except that (a) the minimum gap $\Delta_{\min}$ is replaced with the provided lower bound $\widetilde{\Delta}_{\min}$, and (b) the gaps $\Delta_h(x, a)$ are replaced the the *frozen range function* $\Delta \mathring{Q}_h^k(x, a)$, defined below:

Definition 12 (Frozen range function) *For all $(x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, define the* frozen *$Q$-function, $\Delta \mathring{Q}_h^k(x, a)$, as follows: given (x, a, h), choose $k' \leq k$ as large as possible so that $(x_h^{k'}, a_h^{k'}) = (x, a)$ and $\tau_h^{k'} = 1$ (if no such k' exists, set $k' = 1$). Then set $\Delta \mathring{Q}_h^k(x, a) = \Delta \check{Q}_h^{k'}(x, a)$.*

In Lemma 30 we show, roughly speaking, that the frozen range function at the final episode, namely $\Delta \mathring{Q}_h^K(x, a)$, is still lower bounded by the gap $\Delta_h(x, a)$, justifying its use a surrogate for the gaps. The main challenge in the proof of Theorem 8, beyond those from Theorem 9, is the fact that

$\widehat{\Delta}^k$ changes as k increases (in fact, as shown in Lemma 20, $\widehat{\Delta}^k$ is non-decreasing with k). Most notably, this affects the proof of our bound on the number of episodes k for which $\sigma_h^k = 1$ (Lemma 35; the analogous lemma for `DeltaConst` is Lemma 33). To prove Lemma 35, we partition $[K]$ into $O(\iota \cdot H)$ contiguous intervals so that inside each interval, $\widehat{\Delta}^k$ increases by a factor of at most $1 + 1/H$. For each such interval $I \subset [K]$, we bound the number of $k \in I$ so that $\sigma_h^k = 1$; this leads to an increase in our regret bounds by a factor of $O(\iota H)$.

Appendix C. Proofs for worst-case result

In this section we establish the robustness upper bounds of Theorems 8 and 9, giving a regret bound for `QLearningPreds` when the user provides *arbitrary* predictions $\widetilde{Q}_h$.

C.1. Bounds on confidence intervals

We begin by establishing various basic guarantees on the bounds $\overline{Q}_h^k, \underline{Q}_h^k, \overline{V}_h^k, \underline{V}_h^k$ maintained by `QLearningPreds`. The first such result is Lemma 13, which establishes that, with high probability, $\overline{Q}_h^k$ is an upper bound on $Q_h^\star$, $\underline{Q}_h^k$ is a lower bound on $Q_h^\star$, and similarly for $\overline{V}_h^k, \underline{V}_h^k$ (with respect to $V_h^\star$). Before stating it, we introduce the following notation: for each $k \in [K]$, let $\mathcal{H}_k$ denote the σ-algebra generated by all random variables up to step H of episode k, and $\mathcal{H}_{k,h}$ denote the σ-algebra generated by all random variables up to (and including) step h of episode $k+1$. For each $k \in [K]$ as well as $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, define the quantities $Q_h^{\star,k,b}(x,a)$ and $Q_h^{\star,k,r}(x,a)$ (as in Xu et al. (2021)) as follows: suppose we start in state x at level h, and follow the optimal policy $\pi^\star$, generating the (random) trajectory $x_h = x, x_{h+1}, \ldots, x_H$. Choose $h' \geq h+1$ as small as possible so that $x_{h'} \notin \mathcal{G}_{h'}^k$, and write

$$Q_h^{\star,k,b}(x,a) := \mathbb{E}\left[\sum_{\ell=h}^{h'-1} r_\ell(x_\ell, \pi_\ell^\star(x_\ell)) | \mathcal{H}_{k-1,h}\right], \qquad Q_h^{\star,k,r}(x,a) := \mathbb{E}[V_{h'}^\star(x_{h'}) | \mathcal{H}_{k-1,h}]. \quad (21)$$

(Note that $\mathcal{G}_{h'}^k$ is $\mathcal{H}_{k-1}$-measurable, and thus $\mathcal{H}_{k-1,h'}$-measurable for all h' and k.) It is immediate that

$$Q_h^\star(x,a) = Q_h^{\star,k,b}(x,a) + Q_h^{\star,k,r}(x,a).$$

As in Xu et al. (2021), we use the quantities $\widehat{r}_h^k$ as an unbiased estimate of $Q_h^{\star,k,b}(x_h^k, a_h^k)$. Recall that for some constant $C_0 > 1$, we use exploration bonuses $b_n = C_0\sqrt{H^3\iota/n}$, and recall the definition of the aggregated bonuses β_n in (11). Notice that item 1 of Lemma 46 gives that

$$2C_0\sqrt{H^3\iota/n} \leq \beta_n \leq 4C_0\sqrt{H^3\iota/n}. \quad (22)$$

For future reference, we will also define the constants

$$C_2 = 8C_0, \qquad C_1 = 56e^2C_2^2. \quad (23)$$

Lemma 13 *Set $p = 1/(H^2K)$. For a sufficiently large choice of the constant C_0, there is an event $\mathcal{E}^{\mathrm{wc}}$ occurring with probability $1-p$ so that the following holds under the event $\mathcal{E}^{\mathrm{wc}}$, for all episodes $k \in [K]$:*

1. *For any $(x,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ so that $x \notin \mathcal{G}_h^k$ and $a \in A_h^k(x)$, suppose the episodes k' in which (x,a) as previously taken at step h are denoted $k^1, \ldots, k^n \leq k$. Then the following inequalities hold:*

$$\overline{Q}_h^{k+1}(x,a) - \underline{Q}_h^{k+1}(x,a) \leq \alpha_n^0 \cdot H + \sum_{i=1}^n \alpha_n^i \cdot \left((\overline{V}_{h'(k^i,h)}^{k^i} - \underline{V}_{h'(k^i,h)}^{k^i})(x_{h'(k^i,h)}^{k^i}) \right) + \beta_n \tag{24}$$

$$\overline{Q}_h^{k+1}(x,a) \geq Q_h^\star(x,a) \geq \underline{Q}_h^{k+1}(x,a) \tag{25}$$

$$\overline{V}_h^{k+1}(x) \geq V_h^\star(x) \geq \underline{V}_h^{k+1}(x). \tag{26}$$

2. *Second, for all $(x,h) \in \mathcal{S} \times [H]$ all optimal actions a (i.e., those a satisfying $\Delta_h(x,a) = 0$) are in $A_h^{k+1}(x)$. In particular, for all $x \in \mathcal{G}_h^{k+1}$, $A_h^{k+1}(x)$ contains the unique optimal action at x.*

Proof For $k \in [K]$, we let $\mathcal{E}_k^{\mathrm{wc}}$ denote the event that items 1 and 2 of the lemma statement hold for all episodes $j \leq k$. We wish to show that $\Pr[\mathcal{E}_K^{\mathrm{wc}}] \geq 1 - p$.

We use induction on k to show that for all k, $\Pr[\mathcal{E}_k^{\mathrm{wc}}] \geq 1 - pk/K$. The base case $k = 0$ (i.e., $k+1 = 1$) follows from the fact that $\alpha_0^0 = 1$, $\overline{Q}_h^1(x,a) = H$, $\underline{Q}_h^1(x,a) = 0$, and that for any choice of (x,a,h) we necessarily have $n = 0$ (in particular, $\Pr[\mathcal{E}_0^{\mathrm{wc}}] = 1$). So choose any $k \geq 1$, and assume that $\Pr[\mathcal{E}_{k-1}^{\mathrm{wc}}] \geq 1 - p(k-1)/K$.

By the algorithm's update rule in steps 2(b)iv and 2(b)vi, it holds that, for all $(x,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ so that $x \notin \mathcal{G}_h^k$, letting $n = N_h^{k+1}(x,a)$,

$$\overline{q}_h^{k+1}(x,a) = \begin{cases} (1-\alpha_n) \cdot \overline{q}_h^k(x,a) + \alpha_n \cdot \left(\widehat{r}_h^k + \overline{V}_{h'(k,h)}^k(x_{h'(k,h)}^k) + b_n \right) & : \quad (x,a) = (x_h^k, a_h^k) \\ \overline{q}_h^k(x,a) \quad : \quad \text{otherwise} \end{cases}$$

$$\underline{q}_h^{k+1}(x,a) = \begin{cases} (1-\alpha_n) \cdot \underline{q}_h^k(x,a) + \alpha_n \cdot \left(\widehat{r}_h^k + \underline{V}_{h'(k,h)}^k(x_{h'(k,h)}^k) - b_n \right) & : \quad (x,a) = (x_h^k, a_h^k) \\ \underline{q}_h^k(x,a) \quad : \quad \text{otherwise.} \end{cases}$$

By iterating the above, we obtain that

$$\overline{q}_h^{k+1}(x,a) = \alpha_n^0 \cdot H + \sum_{i=1}^n \alpha_n^i \cdot \left(\widehat{r}_h^{k^i} + \overline{V}_{h'(k^i,h)}^{k^i}(x_{h'(k^i,h)}^{k^i}) + b_n \right) \tag{27}$$

$$\underline{q}_h^{k+1}(x,a) = \sum_{i=1}^n \alpha_n^i \cdot \left(\widehat{r}_h^{k^i} + \underline{V}_{h'(k^i,h)}^{k^i}(x_{h'(k^i,h)}^{k^i}) - b_n \right), \tag{28}$$

where $k^1, \ldots, k^n \leq k$ denote all previous episodes during which (x,a,h) has been visited.

To see that (24) holds, we first take the difference of (27) and (28) and use the definition of β_n in (11) to get that

$$\overline{q}_h^{k+1}(x,a) - \underline{q}_h^{k+1}(x,a) = \alpha_n^0 \cdot H + \beta_n + \sum_{i=1}^n \alpha_n^i \cdot \left((\overline{V}_{h'(k^i,h)}^{k^i} - \underline{V}_{h'(k^i,h)}^{k^i})(x_{h'(k^i,h)}^{k^i}) \right).$$

Now (24) follows by noting that $\overline{Q}_h^{k+1}(x,a) \leq \overline{q}_h^{k+1}(x,a)$ and $\underline{Q}_h^{k+1}(x,a) \geq \underline{q}_h^{k+1}(x,a)$ (note in particular that (24) holds with probability 1).

We proceed to analyze the event under which (25) and (26) hold. We may compute

$$\begin{aligned}
&\overline{q}_h^{k+1}(x,a) - Q_h^\star(x,a) \\
=&\alpha_n^0 \cdot H + \sum_{i=1}^n \alpha_n^i \cdot \left(\widehat{r}_h^{k^i} + \overline{V}^{k^i}_{h'(k^i,h)}(x^{k^i}_{h'(k^i,h)}) + b_n \right) - Q_h^\star(x,a) \\
=&\alpha_n^0 \cdot (H - Q_h^\star(x,a)) + \sum_{i=1}^n \alpha_n^i \cdot \left(\widehat{r}_h^{k^i} + \overline{V}^{k^i}_{h'(k^i,h)}(x^{k^i}_{h'(k^i,h)}) - Q_h^\star(x,a) \right) + \sum_{i=1}^n \alpha_n^i \cdot b_n \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Using item 4 of Lemma 46 and } b_0 = 0\text{)} \\
=&\alpha_n^0 \cdot (H - Q_h^\star(x,a)) + \beta_n/2 + \sum_{i=1}^n \alpha_n^i \cdot \left(\widehat{r}_h^{k^i} - Q_h^{\star,k^i,b}(x,a) \right) \\
&+ \sum_{i=1}^n \alpha_n^i \cdot \left(\overline{V}^{k^i}_{h'(k^i,h)}(x^{k^i}_{h'(k^i,h)}) - V^\star_{h'(k^i,h)}(x^{k^i}_{h'(k^i,h)}) \right) + \sum_{i=1}^n \alpha_n^i \cdot \left(V^\star_{h'(k^i,h)}(x^{k^i}_{h'(k^i,h)}) - Q_h^{\star,k^i,r}(x,a) \right).
\end{aligned} \tag{29}$$

In a similar manner, we have

$$\begin{aligned}
&\underline{q}_h^{k+1}(x,a) - Q_h^\star(x,a) \\
=&-\alpha_n^0 \cdot Q_h^\star(x,a) - \beta_n/2 + \sum_{i=1}^n \alpha_n^i \cdot \left(\widehat{r}_h^{k^i} - Q_h^{\star,k^i,b}(x,a) \right) \\
&+ \sum_{i=1}^n \alpha_n^i \cdot \left(\underline{V}^{k^i}_{h'(k^i,h)}(x^{k^i}_{h'(k^i,h)}) - V^\star_{h'(k^i,h)}(x^{k^i}_{h'(k^i,h)}) \right) + \sum_{i=1}^n \alpha_n^i \cdot \left(V^\star_{h'(k^i,h)}(x^{k^i}_{h'(k^i,h)}) - Q_h^{\star,k^i,r}(x,a) \right).
\end{aligned} \tag{30}$$

Claim 14 *There is an event $\mathcal{E}_k \subset \mathcal{E}_{k-1}^{\rm wc}$ so that $\Pr[\mathcal{E}_k] \geq 1 - pk/K$ and the following holds under $\mathcal{E}_k$: for all $h \in [H]$, all $x \in \mathcal{S} \backslash \mathcal{G}_h^k$, and all $a \in A_h^k(x)$, letting $n = N_h^{k+1}(x,a)$ and $k^1, \ldots, k^n \leq k$ denote all the previous episodes in which (x,a,h) was previously visited,*

$$\left| \sum_{i=1}^n \alpha_n^i \cdot \left(\widehat{r}_h^{k^i} - Q_h^{\star,k^i,b}(x,a) \right) \right| \leq \sqrt{\frac{H^3}{n} \cdot \log\left(\frac{4SAHK}{p} \right)} \tag{31}$$

$$\left| \sum_{i=1}^n \alpha_n^i \cdot \left(V^\star_{h'(k^i,h)}(x^{k^i}_{h'(k^i,h)}) - Q_h^{\star,k^i,r}(x,a) \right) \right| \leq \sqrt{\frac{H^3}{n} \cdot \log\left(\frac{4SAHK}{p} \right)}. \tag{32}$$

Proof [Proof of Claim 14] For $1 \leq k' \leq k$, we let $\mathcal{H}_{k'}$ denote the σ-algebra generated by all random variables up to (step H of) episode k'. It is evident that $\mathcal{E}_{k'}^{\rm wc}$ is $\mathcal{H}_{k'}$-measurable for all $k' \leq k$. Moreover let $\mathcal{H}_{k',h} \supset \mathcal{H}_{k'}$ denote the σ-algebra generated by all random variables up to (and including) step h of episode $k'+1$. Notice that $k^1 - 1, \ldots, k^n - 1$ are all stopping times with respect to the filtration $(\mathcal{H}_{k',h})_{k' \leq k}$. For $1 \leq i \leq n$, define the filtration $\mathcal{F}_i$ by $\mathcal{F}_i := \mathcal{H}_{k^i-1,h}$. Moreover, for $i \in [n]$, define

$$M_i := \alpha_n^i \cdot \left(\widehat{r}_h^{k^i} - Q_h^{\star,k^i,b}(x,a) \right) \cdot \mathbb{1}[\mathcal{E}_{k^i-1}^{\rm wc}].$$

Since $k^{i+1} - 1 \geq k^i$, M_i is $\mathcal{F}_{i+1}$-measurable for each i (as a matter of convention we set $k^{n+1} = k+1$, so M_i is $\mathcal{F}_{i+1}$-measurable even for $i = n$). Moreover, we claim that for each i,

$$\mathbb{E}[M_i | \mathcal{F}_i] = \alpha_n^i \cdot \mathbb{E}\left[\left(\widehat{r}_h^{k^i} - Q_h^{\star, k^i, b}(x,a) \right) \cdot \mathbb{1}[\mathcal{E}^{\mathrm{wc}}_{k^i - 1}] | \mathcal{F}_i \right] = 0. \tag{33}$$

To see that (33) holds, first note that conditioned on $\mathcal{F}_i$, $\mathbb{1}[\mathcal{E}^{\mathrm{wc}}_{k^i-1}] \cdot \widehat{r}_h^{k^i}$ is distributed identically to $\mathbb{1}[\mathcal{E}^{\mathrm{wc}}_{k^i-1}] \cdot \sum_{\ell=h}^{h'-1} r_\ell(x_\ell, \pi^\star_\ell(x_\ell))$ where x_ℓ is the sequence of states visited starting at $x_h = x$ and following the optimal policy $\pi^\star$ and h' is as small as possible so that $x_{h'} \not\in \mathcal{G}_{h'}^{k^i}$ (this holds since item 2 at episode $k^i - 1$ gives that under $\mathcal{E}^{\mathrm{wc}}_{k^i-1}$, the unique action in $A_\ell^{k^i}(x_\ell^{k^i})$, for $h \leq \ell < h'(k^i, h)$ is the optimal action, namely $\pi^\star_\ell(x_\ell^{k^i})$). Recall from (21) and the fact that $k^i - 1$ is a stopping time with respect to the filtration $\mathcal{H}_{k',h}$ that $\mathbb{E}\left[\sum_{\ell=h}^{h'-1} r_\ell(x_\ell, \pi^\star_\ell(x_\ell)) - Q_h^{\star,k^i,b}(x,a) | \mathcal{F}_i \right] = 0$; then the fact that $\mathcal{E}^{\mathrm{wc}}_{k^i-1}$ is $\mathcal{F}_i$-measurable gives (33).

Next, for $i \in [n]$, define

$$N_i := \alpha_n^i \cdot \left(V^\star_{h'(k^i,h)}(x^{k^i}_{h'(k^i,h)}) - Q_h^{\star,k^i,r}(x,a) \right) \cdot \mathbb{1}[\mathcal{E}^{\mathrm{wc}}_{k^i-1}].$$

Since $k^{i+1} - 1 \geq k^i$, N_i is $\mathcal{F}_{i+1}$-measurable for each i. Moreover, we claim that for each i,

$$\mathbb{E}[N_i | \mathcal{F}_i] = \alpha_n^i \cdot \mathbb{E}\left[\left(V^\star_{h'(k^i,h)}(x^{k^i}_{h'(k^i,h)}) - Q_h^{\star,k^i,r}(x,a) \right) \cdot \mathbb{1}[\mathcal{E}^{\mathrm{wc}}_{k^i-1}] | \mathcal{F}_i \right] = 0. \tag{34}$$

The validity of (34) is verified in the same way as that of (33): conditioned on $\mathcal{F}_i$, $\mathbb{1}[\mathcal{E}^{\mathrm{wc}}_{k^i-1}] \cdot V^\star_{h'(k^i,h)}(x^{k^i}_{h'(k^i,h)})$ is distributed identically to $\mathbb{1}[\mathcal{E}^{\mathrm{wc}}_{k^i-1}] \cdot V^\star_{h'}(x_{h'})$, where $x_h, \ldots, x_{h'}$ is defined as above, namely it is the sequence of states visited starting at $x_h = x$ and following the optimal policy $\pi^\star$, and h' is as small as possible so that $x_{h'} \not\in \mathcal{G}_{h'}^{k^i}$ (again we use that item 2 holds at episode $k^i - 1$ under $\mathcal{E}^{\mathrm{wc}}_{k^i-1}$). Now (21) gives that $\mathbb{E}[V^\star_{h'}(x_{h'}) - Q_h^{\star,k^i,r}(x,a) | \mathcal{F}_i] = 0$ and using this together with the fact that $\mathcal{E}^{\mathrm{wc}}_{k^i-1}$ is $\mathcal{F}_i$-measurable gives (34).

Equations (33) and (34) give that M_i and N_i are martingales with respect to the filtration $\mathcal{F}_{i+1}$. The fact that $\sum_{i=1}^n (\alpha_n^i)^2 \leq \frac{2H}{n}$ (item 2 of Lemma 46) together with the Azuma-Hoeffding inequality then gives that, for fixed x, a, h, with probability $1 - p/(SAHK)$, both of the below inequalities hold:

$$\left| \sum_{i=1}^n \alpha_n^i \cdot \left(\widehat{r}_h^{k^i} - Q_h^{\star,k^i,b}(x,a) \right) \cdot \mathbb{1}[\mathcal{E}^{\mathrm{wc}}_{k^i-1}] \right| \leq \sqrt{\frac{H^3 \cdot \log\left(\frac{4SAHK}{p} \right)}{n}} \tag{35}$$

$$\left| \sum_{i=1}^n \alpha_n^i \cdot \left(V^\star_{h'(k^i,h)}(x^{k^i}_{h'(k^i,h)}) - Q_h^{\star,k^i,r}(x,a) \right) \cdot \mathbb{1}[\mathcal{E}^{\mathrm{wc}}_{k^i-1}] \right| \leq \sqrt{\frac{H^3 \cdot \log\left(\frac{4SAHK}{p} \right)}{n}}. \tag{36}$$

Let $\mathcal{E}_k$ denote the intersection of the probability $1 - p/K$-event that both (35) and (36) hold for all x, a, h and the event $\mathcal{E}^{\mathrm{wc}}_{k-1}$. Then using the inductive hypothesis that $\Pr[\mathcal{E}^{\mathrm{wc}}_{k-1}] \geq 1 - p(k-1)/K$, we get that $\Pr[\mathcal{E}_k] \geq 1 - pk/K$. Thus, under the event $\mathcal{E}_k$, we have that

$$\left|\sum_{i=1}^n \alpha_n^i \cdot \left(\widehat{r}_h^{k^i} - Q_h^{\star,k^i,b}(x,a)\right)\right| \leq \sqrt{\frac{H^3}{n} \cdot \log\left(\frac{4SAHK}{p}\right)}$$

$$\left|\sum_{i=1}^n \alpha_n^i \cdot \left(V^\star_{h'(k^i,h)}(x^{k^i}_{h'(k^i,h)}) - Q_h^{\star,k^i,r}(x,a)\right)\right| \leq \sqrt{\frac{H^3}{n} \cdot \log\left(\frac{4SAHK}{p}\right)},$$

completing the proof of the claim. ■

Next we show that, on the event $\mathcal{E}_k$, both (25) and (26) hold at episode k, for all x, a, h. Note that, by (22), for all n,

$$\beta_n/4 \geq \frac{C_0}{2} \cdot \sqrt{\frac{H^3 \iota}{n}} \geq \sqrt{\frac{H^3}{n} \cdot \log\left(\frac{4SAHK}{p}\right)}$$

as long as the constant C_0 is chosen to be large enough. Thus, by (29) and Claim 14, under the event $\mathcal{E}_k$, we have that

$$\begin{aligned}
&\overline{q}_h^{k+1}(x,a) - Q_h^\star(x,a) \\
\geq &\beta_n/2 - \beta_n/4 + \sum_{i=1}^n \alpha_n^i \cdot \left(\overline{V}^{k^i}_{h'(k^i,h)}(x^{k^i}_{h'(k^i,h)}) - V^\star_{h'(k^i,h)}(x^{k^i}_{h'(k^i,h)})\right) - \beta_n/4 \\
= &\sum_{i=1}^n \alpha_n^i \cdot \left(\overline{V}^{k^i}_{h'(k^i,h)}(x^{k^i}_{h'(k^i,h)}) - V^\star_{h'(k^i,h)}(x^{k^i}_{h'(k^i,h)})\right) \geq 0, \qquad (37)
\end{aligned}$$

where the final inequality follows from the fact that $\mathcal{E}_k \subset \mathcal{E}_{k-1}^{\mathrm{wc}}$ and under $\mathcal{E}_{k-1}^{\mathrm{wc}}$, (26) holds (in particular, at step $h'(k^i, h)$ for state $x^{k^i}_{h'(k^i,h)}$). Using the fact that $\overline{Q}_h^{k+1}(x,a) = \min_{k' \leq k+1}\left\{\overline{q}_h^{k'}(x,a)\right\}$ (step 2(b)v of `QLearningPreds`) together with the fact that $\overline{Q}_h^k(x,a) \geq Q_h^\star(x,a)$ under $\mathcal{E}_{k-1}^{\mathrm{wc}}$, we see from (37) that $\overline{Q}_h^{k+1}(x,a) \geq Q_h^\star(x,a)$ under the event $\mathcal{E}_k$ (for all $a \in A_h^k(x)$). Since $\overline{V}_h^{k+1}(x) = \max_{a \in A_h^k(x)} \overline{Q}_h^{k+1}(x,a)$ (step 2(b)ix of `QLearningPreds`), it follows that $\overline{V}_h^{k+1}(x) \geq V_h^\star(x)$ under the event $\mathcal{E}_k$.

Thus we have verified the first inequality in each of (25) and (26) at episode k. The proof of the second inequality in each follows identically: (30) together with Claim 14 gives that under the event $\mathcal{E}_k$, we have that $\underline{q}_h^{k+1}(x,a) - Q_h^\star(x,a) \leq 0$ for all x, a, h. Then it follows that $\underline{Q}_h^{k+1}(x,a) \leq Q_h^\star(x,a)$, and using the fact that $\underline{V}_h^{k+1}(x) = \max_{a \in A_h^k(x)} \underline{Q}_h^{k+1}(x,a)$ (step 2(b)viii of `QLearningPreds`), it follows that $\underline{V}_h^{k+1}(x) \leq V_h^\star(x)$ under the event $\mathcal{E}_k$. Thus we have verified that (24), (25), and (26) hold (for any choice of x, a, h with $x \not\in \mathcal{G}_h^k$ and $a \in A_h^k(x)$) at episode k, under the event $\mathcal{E}_k$.

Finally we verify that item 2 holds at episode k under the event $\mathcal{E}_k$. Suppose, to the contrary, that there were some optimal action $a^\star$ for some state (x,h) so that $a^\star \not\in A_h^{k+1}(x)$. Since $\mathcal{E}_k \subset \mathcal{E}_{k-1}^{\mathrm{wc}}$, we have that $a^\star \in A_h^k(x)$, meaning that by the definition of $A_h^{k+1}(x)$ in step 2c of `QLearningPreds`, we must have that $\overline{Q}_h^{k+1}(x, a^\star) < \underline{V}_h^{k+1}(x)$. But we have just shown that under the event $\mathcal{E}_k$, weh have that $\underline{V}_h^{k+1}(x) \leq V_h^\star(x)$ and $\overline{Q}_h^{k+1}(x, a^\star) \geq Q_h^\star(x, a^\star)$, which implies that $Q_h^\star(x, a^\star) < V_h^\star(x)$, contradicting the fact that $a^\star$ is an optimal action at (x, h).

Thus we have shown that all statements in items 1 and 2 for episode k hold under the event $\mathcal{E}_k$, and $\mathcal{E}_k \subset \mathcal{E}_{k-1}^{\text{wc}}$ as well as $\Pr[\mathcal{E}_k] \geq 1 - pk/K$. Thus $\mathcal{E}_k^{\text{wc}} \supset \mathcal{E}_k$, meaning that $\Pr[\mathcal{E}_k^{\text{wc}}] \geq 1 - pk/K$, which completes the proof of the inductive step. ■

The following lemma shows that the upper and lower confidence bounds satisfy a monotonicity property with respect to the number of episodes k that have elapsed: in particular, the upper confidence bounds on $V_h^\star, Q_h^\star$ maintained by `QLearningPreds` are non-increasing, and the lower confidence bounds on $V_h^\star, Q_h^\star$ are non-decreasing. Note that in many previous works studying Q-learning algorithms (such as Jin et al. (2018); Xu et al. (2021)), these monotonicity properties do not necessarily hold – it is necessary to modify the Q- and V-value updates in `QLearningPreds` appropriately to ensure that Lemma 15 holds.

Lemma 15 *For all $k \in [K]$, the following inequalities hold for all $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$:*

$$\overline{Q}_h^{k+1}(x,a) \leq \overline{Q}_h^k(x,a) \tag{38}$$
$$\underline{Q}_h^{k+1}(x,a) \geq \underline{Q}_h^k(x,a) \tag{39}$$
$$\overline{V}_h^{k+1}(x) \leq \overline{V}_h^k(x) \tag{40}$$
$$\underline{V}_h^{k+1}(x) \geq \underline{V}_h^k(x) \tag{41}$$
$$\widetilde{Q}_h^{k+1}(x,a) \leq \widetilde{Q}_h^k(x,a). \tag{42}$$

Proof Fix any $k \in [K]$ and $h \in [H]$. First note that step 2e of `QLearningPreds` verifies (38) through (42) for all $(x, a) \neq (x_h^k, a_h^k)$. So it remains to to consider the case that $x = x_h^k$ and $a = a_h^k$.

First note that (38) and (39) are directly verified by steps 2(b)v and 2(b)vii, respectively, of `QLearningPreds`. To verify (40), note that $A_h^k(x_h^k) \subset A_h^{k-1}(x_h^k)$, meaning that

$$\overline{V}_h^{k+1}(x_h^k) = \max_{a' \in A_h^k(x_h^k)} \{\overline{Q}_h^{k+1}(x_h^k, a')\} \leq \max_{a' \in A_h^{k-1}(x_h^k)} \{\overline{Q}_h^{k+1}(x_h^k, a')\} \leq \max_{a' \in A_h^{k-1}(x_h^k)} \{\overline{Q}_h^k(x_h^k, a')\} = \overline{V}_h^k(x_h^k),$$

where the second inequality uses (38) and the last equality uses steps 2(b)viii and 2e of `QLearningPreds` (in particular, note that $\overline{V}_h^k(x_h^k)$ and $\overline{Q}_h^k(x_h^k, \cdot)$ remain unchanged from the previous episode before k at which x_h^k was visited).

Next we verify (41); choose $a \in A_h^k(x_h^k)$ so that $\underline{V}_h^{k+1}(x_h^k) = \underline{Q}_h^{k+1}(x_h^k, a)$, and $a' \in A_h^{k-1}(x_h^k)$ so that $\underline{V}_h^k = \underline{Q}_h^k(x_h^k, a')$. If (41) did not hold, we would have that $\underline{Q}_h^{k+1}(x_h^k, a') \geq \underline{Q}_h^k(x_h^k, a') > \underline{Q}_h^{k+1}(x_h^k, a)$, which must mean that $a' \notin A_h^k(x_h^k)$. But this is impossible since $\overline{Q}_h^{k+1}(x_h^k, a') \geq \underline{Q}_h^{k+1}(x_h^k, a') = \underline{V}_h^{k+1}(x_h^k)$, so by step 2c of `QLearningPreds` a' must belong to $A_h^k(x_h^k)$.

Finally, (42) is verified by step 2(d)ii of `QLearningPreds`. ■

C.2. Range functions and clipped range functions

Recall the definition of the range functions $\Delta V_h^k, \Delta Q_h^k$ in Definition 10, as well as the clipped range functions $\Delta\check{V}_h^k, \Delta\check{Q}_h^k$ in Definition 11. In this section we establish some basic guarantees of these functions. The first such result, Lemma 16, shows that the range functions are upper bounds on the gap between the upper and lower estimates for the Q- and V-value functions.

Lemma 16 *For all $(x, h, k, a) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ for which $x \not\in \mathcal{G}_h^k$ and $a \in A_h^k(x)$, the range functions satisfy the following under the event $\mathcal{E}^{\mathrm{wc}}$:*

$$\begin{aligned}
\Delta Q_h^k(x,a) \geq & \overline{Q}_h^k(x,a) - \underline{Q}_h^k(x,a) \\
\Delta V_h^k(x) \geq & \overline{V}_h^k(x) - \underline{V}_h^k(x).
\end{aligned}$$

Proof The proof closely follows that of (Xu et al., 2021, Lemma B.3). We use reverse induction on h. The base case $h = H+1$ is immediate since $\overline{Q}_{H+1}^k, \underline{Q}_{H+1}^k, \overline{V}_{H+1}^k, \underline{V}_{H+1}^k$ are defined to be identically 0 for all $k \in [K]$. Next fix $h \leq H$ and suppose that the statement of the lemma holds for all (x, h', k, a) for which $h' > h$. For any (x, k, a) for which $x \not\in \mathcal{G}_h^k$ and $a \in A_h^k(x)$, note that, for $n = N_h^k(x,a)$ and $k_h^i = k_h^i(x,a)$,

$$\begin{aligned}
\overline{Q}_h^k(x,a) - \underline{Q}_h^k(x,a) \leq & \alpha_n^0 \cdot H + \sum_{i=1}^n \alpha_n^i \cdot (\overline{V}^{k_h^i}_{h'(k_h^i,h)} - \underline{V}^{k_h^i}_{h'(k_h^i,h)})(x^{k_h^i}_{h'(k_h^i,h)}) + \beta_n && (43) \\
\leq & \alpha_n^0 \cdot H + \sum_{i=1}^n \alpha_n^i \cdot \Delta V^{k_h^i}_{h'(k_h^i,h)}(x^{k_h^i}_{h'(k_h^i,h)}) + \beta_n && (44) \\
= & \Delta Q_h^k(x,a). && (45)
\end{aligned}$$

where (43) follows from (24) of Lemma 13 (in particular, we use the validity of (24) for episode $k-1$) and (44) uses the inductive hypothesis (since $h'(k_h^i, h) > h$). Thus the inductive step for ΔQ_h^k is verified. To lower bound ΔV_h^k, we first note that by Definition 10 for any $x \not\in \mathcal{G}_h^k$, there is some $k' \leq k$ so that $\Delta V_h^k(x) = \Delta Q_h^{k'}(x, a^\star)$ for $a^\star = \arg\max_{a' \in A_h^{k'}(x)} \{ \overline{Q}_h^{k'}(x,a') - \underline{Q}_h^{k'}(x,a') \}$. Then

$$\begin{aligned}
\overline{V}_h^k(x) - \underline{V}_h^k(x) \leq & \overline{V}_h^{k'}(x) - \underline{V}_h^{k'}(x) && \text{(Using (40) of Lemma 15)} \\
= & \left(\max_{a' \in A_h^{k'}(x)} \overline{Q}_h^{k'}(x,a') \right) - \left(\max_{a' \in A_h^{k'}(x)} \underline{Q}_h^{k'}(x,a') \right) && (46) \\
\leq & \max_{a' \in A_h^{k'}(x)} \left\{ \overline{Q}_h^{k'}(x,a') - \underline{Q}_h^{k'}(x,a') \right\} \\
\leq & \overline{Q}_h^{k'}(x,a^\star) - \underline{Q}_h^{k'}(x,a^\star) \\
& && \text{(For } a^\star = \arg\max_{a' \in A_h^{k'}(x)} \{ \overline{Q}_h^{k'}(x,a') - \underline{Q}_h^{k'}(x,a') \}) \\
\leq & \Delta Q_h^{k'}(x,a^\star) && \text{(Using (45))} \\
= & \Delta V_h^k(x), && (47)
\end{aligned}$$

where (46) follows from steps 2(b)viii and 2(b)ix of `QLearningPreds` and the final equality (47) follows from the definition of k'. ■

Lemma 17 below shows that despite the clipping of the bonus β_n in the definition of the clipped range function (see (15)), the clipped range functions $\Delta \check{Q}_h^k, \Delta \check{V}_h^k$ remain approximate upper bounds on the range functions $\Delta Q_h^k, \Delta V_h^k$.

Lemma 17 *For all $(x, h, k, a) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ for which $x \notin \mathcal{G}_h^k$ and $a \in A_h^k(x)$, the partially-clipped range functions satisfy the following:*

$$\begin{aligned} \Delta\breve{Q}_h^k(x,a) \geq & \Delta Q_h^k(x,a) - \frac{\Delta_{\min}}{4H} \\ \Delta\breve{V}_h^k(x) \geq & \Delta V_h^k(x) - \frac{\Delta_{\min}}{4H}. \end{aligned}$$

Proof The lemma follows in a similar manner to (Xu et al., 2021, Proposition B.5). We prove by reverse induction on h and forward induction on k that

$$\Delta\breve{Q}_h^k(x,a) \geq \Delta Q_h^k(x,a) - \frac{(H+1-h)}{H} \cdot \frac{\Delta_{\min}}{4H} \tag{48}$$

and

$$\Delta\breve{V}_h^k(x) \geq \Delta V_h^k(x) - \frac{(H+1-h)}{H} \cdot \frac{\Delta_{\min}}{4H}. \tag{49}$$

The base case $h = H+1$ is immediate since $\Delta\breve{Q}_{H+1}^k, \Delta Q_{H+1}^k, \Delta\breve{V}_{H+1}^k, \Delta V_{H+1}^k$ are identically 0. The base case $k = 0$ is also immediate since $\Delta\breve{Q}_h^0(x,a) = \Delta\breve{V}_h^0(x) = \Delta Q_h^0(x,a) = \Delta V_h^0(x) = H$ for all $x, a, h \in \mathcal{S} \times \mathcal{A} \times [H]$. To establish the inductive step, note that, for any (x, a, h, k) for which $x \notin \mathcal{G}_h^k$ and $a \in A_h^k(x)$, letting $n = N_h^k(x,a)$ and $k_h^i = k_h^i(x,a)$ for $i \in [n]$, we have

$$\begin{aligned} \Delta\breve{Q}_h^k(x,a) =& \min\left\{\Delta\breve{Q}_h^{k-1}(x,a),\ \alpha_n^0 H + \operatorname{clip}\left[\beta_n \middle| \frac{\Delta_{\min}}{4H^2}\right] + \sum_{i=1}^n \alpha_n^i \cdot \Delta\breve{V}_{h'(k_h^i,h)}^{k_h^i}(x_{h'(k_h^i,h)}^{k_h^i})\right\} \\ \geq& \min\left\{\Delta Q_h^{k-1}(x,a) - \frac{(H+1-h)\Delta_{\min}}{4H^2},\right. \\ & \left.\alpha_n^0 H + \beta_n + \sum_{i=1}^n \alpha_n^i \cdot \Delta\breve{V}_{h'(k_h^i,h)}^{k_h^i}(x_{h'(k_h^i,h)}^{k_h^i}) - \frac{\Delta_{\min}}{4H^2}\right\} \\ \geq& \min\left\{\Delta Q_h^{k-1}(x,a) - \frac{(H+1-h)\Delta_{\min}}{4H^2},\right. \\ & \left.\alpha_n^0 H + \beta_n + \sum_{i=1}^n \alpha_n^i \cdot \Delta V_{h'(k_h^i,h)}^{k_h^i}(x_{h'(k_h^i,h)}^{k_h^i}) - \frac{(H-h)\Delta_{\min}}{4H^2} - \frac{\Delta_{\min}}{4H^2}\right\} \qquad (50) \\ =& \Delta Q_h^k(x,a) - \frac{(H+1-h)\Delta_{\min}}{4H^2}, \end{aligned}$$

where (50) used the inductive hypothesis (in particular, (49) at steps $h' > h$). This establishes the inductive step for (48).

We proceed to lower bound $\Delta\breve{V}_h^k(x,a)$. For fixed x, a, set $a^\star = \arg\max_{a' \in A_h^k(x)}\{\overline{Q}_h^k(x,a') - \underline{Q}_h^k(x,a')\}$. Using Definition 11, the inductive hypothesis, and the validity of (48) for step h at episode k, we have

$$\begin{aligned} \Delta\breve{V}_h^k(x) =& \min\{\Delta\breve{V}_h^{k-1}(x), \Delta\breve{Q}_h^k(x,a^\star)\} \\ \geq& \min\{\Delta V_h^{k-1}(x), \Delta Q_h^k(x,a^\star)\} - \frac{(H+1-h)\Delta_{\min}}{4H^2}, \end{aligned}$$

which completes the inductive step for (49). ■

The following straightforward lemma shows that the clipped range functions are smaller than the range functions.

Lemma 18 *For all $(x, h, k, a) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ for which $x \notin \mathcal{G}_h^k$ and $a \in A_h^k(x)$, it holds that*

$$\Delta\check{Q}_h^k(x, a) \leq \Delta Q_h^k(x, a) \qquad \text{and} \qquad \Delta\check{V}_h^k(x) \leq \Delta V_h^k(x).$$

Proof The lemma is a straightforward consequence of Definitions 10 and 11 and induction on h, k (in particular, forward induction on k and reverse induction on h): in particular, for any h, k, having established the statement for all (h', k') with either $h' > h$ or $k' < k$, we have that $\Delta\check{Q}_h^k(x, a) \leq \Delta Q_h^k(x, a)$ since

$$\begin{aligned} &\alpha_n^0 H + \mathrm{clip}\left[\beta_n \middle| \frac{\Delta_{\min}}{4H^2}\right] + \sum_{i=1}^n \alpha_n^i \cdot \Delta\check{V}_{h'(k_h^i(x,a),h)}^{k_h^i(x,a)}(x_{h'(k_h^i(x,a),h)}^{k_h^i(x,a)}) \\ \leq &\alpha_n^0 H + \beta_n + \sum_{i=1}^n \alpha_n^i \cdot \Delta\check{V}_{h'(k_h^i(x,a),h)}^{k_h^i(x,a)}(x_{h'(k_h^i(x,a),h)}^{k_h^i(x,a)}). \end{aligned}$$

It then follows immediately that $\Delta\check{V}_h^k(x) \leq \Delta V_h^k(x)$. ■

Lemma 19 establishes some monotonicity (with respect to k) properties of the range functions, analogously to Lemma 15.

Lemma 19 *For all (x, a, h, k) so that $x \notin \mathcal{G}_h^{k+1}$ and $a \in A_h^{k+1}(x)$, the following inequalities hold true:*

$$\begin{aligned} \Delta\check{Q}_h^{k+1}(x, a) \leq &\Delta\check{Q}_h^k(x, a) && (51) \\ \Delta\check{V}_h^{k+1}(x) \leq &\Delta\check{V}_h^k(x) \\ \Delta Q_h^{k+1}(x, a) \leq &\Delta Q_h^k(x, a) \\ \Delta V_h^{k+1}(x) \leq &\Delta V_h^k(x). && (52) \end{aligned}$$

Moreover, for all $(x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, it holds that

$$\Delta\mathring{Q}_h^{k+1}(x, a) \leq \Delta\mathring{Q}_h^k(x, a). \tag{53}$$

Proof The first four inequalities are immediate from Definitions 10 and 11. The final inequality follows from Definition 12 and (51). ■

Lemma 20 establishes some further monotonicity properties for `QLearningPreds`.

Lemma 20 *The following statements hold true:*

1. *When `QLearningPreds` is run with either `DeltaConst` or `DeltaIncr`, for all $k \in [K]$, it holds that $\widehat{\Delta}^{k+1} \geq \widehat{\Delta}^k$.*

2. *For any $h \in [H]$ and $k < k'$ for which $x_h^k = x_h^{k'}$, we have $\tau_h^{k'} \leq \tau_h^k$.*

Proof We begin with the first statement; it is immediate for `DeltaConst`. In the case of `DeltaIncr`, we note that by (53) of Lemma 19, $\Delta \mathring{Q}_h^k(x,a)$ is non-increasing as a function of k for all x, a, h. It is clear that the same is true of $\widetilde{\Delta} \mathring{Q}_h^k(x,a)$ (defined in step 1 of Algorithm 4). Thus the expression in (14) is non-decreasing as a function of k.

To see the second statement, note that if $\tau_h^k = 0$, then either $|A_h^k(x_h^k)| = 1$, in which case it will hold that $|A_h^{k'}(x_h^k)| = 1$ (and so $\tau_h^{k'} = 0$), or $\Delta V_h^k(x_h^k) \leq \varphi_h(\widehat{\Delta}^k)$, in which case it holds that $\Delta V_n^{k'}(x_h^k) \leq \varphi_h(\widehat{\Delta}^k) \leq \varphi_h(\widehat{\Delta}^{k'})$ (and so $\tau_h^{k'} = 0$), by (52) and the first item of this lemma. ■

Recall that the clip function is defined as follows: for real numbers x, y, we have $\mathrm{clip}\,[x |\, y] = x \cdot \mathbb{1}[x \geq y]$. We next state some lemmas establishing useful properties of the clip function in Lemmas 21, 22, and 23 below.

Lemma 21 (Claim A.8, Xu et al. (2021)) *For any positive integers a, b, c so that $a + b \geq c$ and any $x \in (0,1)$, it holds that*

$$a + b \leq \mathrm{clip}\left[a \middle|\, \frac{xc}{2}\right] + (1+x)b.$$

Lemma 22 (Claim A.13, Xu et al. (2021)) *For any $c, \epsilon > 0$, it holds that*

$$\sum_{n=1}^{\infty} \mathrm{clip}\left[\frac{c}{\sqrt{n}} \middle| \epsilon\right] \leq \frac{4c^2}{\epsilon}.$$

Lemma 23 *Fix some $c > 0$ and $h \in [H]$. For $n \in \mathbb{N}$, write $\gamma_n = c/\sqrt{n}$. Then for any function $\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$, and any subset $\mathcal{W} \subset [K]$ of size $M := |\mathcal{W}|$, it holds that*

$$\sum_{k \in \mathcal{W}} \mathrm{clip}\left[\gamma_{n_h^k} \middle| \theta(x_h^k, a_h^k)\right] \leq \min\left\{2c\sqrt{SAM}, \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \frac{4c^2}{\theta(x,a)}\right\},$$

where we recall that $n_h^k = N_h^k(x_h^k, a_h^k)$.

Proof For $(x,a) \in \mathcal{S} \times \mathcal{A}$, let $\mathcal{W}_{x,a} := \{k \in \mathcal{W} : (x_h^k, a_h^k) = (x,a)\}$. Then, on the one hand, we have

$$\begin{aligned}\sum_{k \in \mathcal{W}} \mathrm{clip}\left[\gamma_{n_h^k} \middle| \theta(x_h^k, a_h^k)\right] &= \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{k \in \mathcal{W}_{x,a}} \mathrm{clip}\left[\frac{c}{\sqrt{n_h^k}} \middle| \theta(x,a)\right] \\ &\leq \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \frac{4c^2}{\theta(x,a)},\end{aligned}$$

where the inequality uses Lemma 22. On the other hand, we have

$$\begin{aligned}\sum_{k \in \mathcal{W}} \operatorname{clip}\left[\gamma_{n_h^k} \middle| \theta(x_h^k, a_h^k)\right] &= \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{k \in \mathcal{W}_{x,a}} \operatorname{clip}\left[\frac{c}{\sqrt{n_h^k}} \middle| \theta(x,a)\right] \\ &\leq \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{i=1}^{|\mathcal{W}_{x,a}|} \frac{c}{\sqrt{i}} \\ &\leq \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} 2c\sqrt{|\mathcal{W}_{x,a}|} \\ &\leq 2c\sqrt{SAM},\end{aligned}$$

where the final inequality follows since $\sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} |\mathcal{W}_{x,a}| = M$. ■

C.3. Bounding the clipped range functions

For all $(h, k) \in [H] \times [K]$ so that $x_h^k \notin \mathcal{G}_h^k$ (so that $\Delta V_h^k(x_h^k)$ is defined), write

$$\breve{\delta}_h^k := \Delta \breve{V}_h^k(x_h^k).$$

Since $a_h^k \in A_h^k(x_h^k)$, $\Delta \breve{Q}_h^k(x_h^k, a_h^k)$ is defined, and we may thus further write

$$\breve{\theta}_h^k := \Delta \breve{Q}_h^k(x_h^k, a_h^k).$$

Per the regret decomposition in Lemma 29, we will bound the regret $\mathbb{E}\left[\sum_{k=1}^K (V_1^\star(x_1) - V_1^{\pi^k}(x_1))\right]$ by the quantity $\sum_{(k,h): a_h^k \notin \mathcal{A}_{h,0}^{\mathrm{opt}}(x_h^k)} \breve{\delta}_h^k$ (conditioned on the high-probability event $\mathcal{E}^{\mathrm{wc}}$). In this section we prove an upper bound on this latter quantity. In fact, we prove a more general result which upper bounds $\sum_{(k,h) \in \mathcal{W}} \breve{\delta}_h^k$ for various sets $\mathcal{W} \subset [K] \times [H]$; we will need this more general result in order to establish improved regret bounds in the case when the predictions $\widetilde{Q}_h$ are accurate.

We begin by defining the type of set $\mathcal{W} \subset [K] \times [H]$ for which we obtain such an upper bound, namely *level-h sets*.

Definition 24 *Fix $h \in [H]$. We say that a subset $\mathcal{W} \subset [K] \times [H]$ is a* level-h *set if the following conditions hold:*

1. *For each $k \in [K]$, there is at most one element $(\widetilde{k}, \widetilde{h}) \in \mathcal{W}$ so that $\widetilde{k} = k$.*
2. *For all $(\widetilde{k}, \widetilde{h}) \in \mathcal{W}$, it holds that both $\tau_{\widetilde{h}}^{\widetilde{k}} = 1$ and $\widetilde{h} \geq h$.*
3. *For each $(\widetilde{k}, \widetilde{h}) \in \mathcal{W}$ for which $\tilde{h} > h$, for $h \leq h' < \tilde{h}$, it holds that $x_{h'}^{\tilde{k}} \in \mathcal{G}_{h'}^{\tilde{k}}$.*

For a level-h set, $\mathcal{W}$, we next define its *reduction*, which replaces each element $(\tilde{k}, \tilde{h}) \in \mathcal{W}$ with $\tilde{h} = h$ with another element (k', h), where $k' \leq \tilde{k}$ is as small as possible subject to $(x_h^{\tilde{k}}, a_h^{\tilde{k}}) = (x_h^{k'}, a_h^{k'})$, and to the constraint that all elements in $\mathcal{W}$ are distinct. The reason that we will want to perform this operation is that our bounds on confidence intervals (in particular, (24), which manifests in Definitions 10 and 11) are given in terms of the *first* n times a particular state-action pair (x, a, h) is visited. The reduction $\mathcal{R}_h(\mathcal{W})$ has the property that for any (x, a), if there are m elements $(\tilde{k}, h) \in \mathcal{R}_h(\mathcal{W})$ with $(x_h^{\tilde{k}}, a_h^{\tilde{k}}) = (x, a)$, then those values $\tilde{k}$ represent the *first* m episodes at which (x, a, h) is visited.

Definition 25 *Fix $h \in [H]$, and consider a level-h set $\mathcal{W} \subset [K] \times [H]$. The* level-$h$ reduction *of $\mathcal{W}$, denoted $\mathcal{R}_h(\mathcal{W})$, is defined as follows: starting with $\mathcal{W}$, perform the following procedure:*

- *For each $(x, a) \in \mathcal{S} \times \mathcal{A}$, let $\mathcal{S}(x, a)$ denote the set of elements $(\tilde{k}, h) \in \mathcal{W}$ for which $(x_h^{\tilde{k}}, a_h^{\tilde{k}}) = (x, a)$. Remove the elements of $\mathcal{S}(x, a)$ from $\mathcal{W}$, and insert the elements*

  $$(k_h^1(x, a), h), (k_h^2(x, a), h), \ldots, (k_h^{|\mathcal{S}(x,a)|}(x, a), h) \tag{54}$$

  into $\mathcal{W}$. (Recall that, for any $s > 0$, $k_h^1(x, a), \ldots, k_h^s(x, a)$ are the smallest s positive integers $\tilde{k}$ so that $(x_h^{\tilde{k}}, a_h^{\tilde{k}}) = (x, a)$.)

Note that the level-h reduction satisfies the following inequality:

$$\max_{(\tilde{k}, \tilde{h}) \in \mathcal{W}} \{\tilde{k}\} \geq \max_{(\tilde{k}, \tilde{h}) \in \mathcal{R}_h(\mathcal{W})} \{\tilde{k}\}. \tag{55}$$

The following lemma shows that the level-h reduction of a level-h set is a level-h set.

Lemma 26 *Suppose that $\mathcal{W} \subset [K] \times [H]$ is a level-h set for some $h \in [H]$. Then the level-h reduction $\mathcal{R}_h(\mathcal{W})$ is also a level-h set.*

Proof We first verify that $\mathcal{R}_h(\mathcal{W})$ satisfies property 1 of Definition 24. Using the notation of Definition 25, we must check that, for each $(x, a) \in \mathcal{S} \times \mathcal{A}$, for $1 \leq i \leq |\mathcal{S}(x, a)|$, there is no $\tilde{h} > h$ so that $(k_h^i(x, a), \tilde{h}) \in \mathcal{W}$. However, if this were the case for some i and $\tilde{h}$, since $\mathcal{W}$ is a level-h set, item 3 of Definition 24 gives us that $x = x_h^{k_h^i(x,a)} \in \mathcal{G}_h^{k_h^i(x,a)}$. But then we must have $\tau_h^{k_h^i(x,a)} = 0$, which contradicts the fact that for some $\tilde{k} \geq k_h^i(x, a)$ so that $x_h^{\tilde{k}} = x$, $\tau_h^{\tilde{k}} = 1$ and Lemma 20.

That the conditions of item 2 of Definition 24 hold for $\mathcal{R}_h(\mathcal{W})$ follows directly from Lemma 20, and $\mathcal{R}_h(\mathcal{W})$ satisfies the conditions of item 3 since $\mathcal{W}$ does. ■

We are now ready to state and prove Lemma 27, which is the main technical component of the worst-case (i.e., robustness) regret bounds in item 1 of Theorem 8 and item 1 of Theorem 9. The first part (item 1) of Lemma 27 bounds $\sum_{(k,\tilde{h})} \breve{\delta}_{\tilde{h}}^k$ for any level-h set $\mathcal{W}$ (for any $h \in [H]$), via a quantity (denoted by f below) that depends on $|\mathcal{W}|$, the step index h, and the largest episode number in $\mathcal{W}$. The second part (item 2) of the lemma then extends this upper bound to a somewhat more general family of subsets $\mathcal{W} \subset [K] \times [H]$.

Lemma 27 *For all $h \in [H]$, the following statements hold:*

1. *For any level-h set $\mathcal{W} \subset [K] \times [H]$ and any $k^\star$ satisfying $k^\star \geq k$ for all $(k, \tilde{h}) \in \mathcal{W}$, it holds that $\sum_{(k,\tilde{h}) \in \mathcal{W}} \breve{\delta}_{\tilde{h}}^k \leq f(|\mathcal{W}|, h, k^\star)$, where for $M \in \mathbb{N},\ h \in [H]$,*

$$\begin{aligned} f(M, h, k^\star) :=& M \cdot \left(1 + \frac{1}{H}\right)^2 \cdot \varphi_{h+1}(\widehat{\Delta}^{k^\star}) \\ & + \sum_{h'=h}^{H} \left(1 + \frac{1}{H}\right)^{2(H-h)} \left(SAH + \min \left\{ C_2 \sqrt{H^3 SAM\iota}, \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \frac{C_2^2 H^3 \iota}{\max\left\{ \frac{\Delta \mathring{Q}_{h'}^{k^\star}(x,a)}{2H}, \frac{\Delta_{\min}}{4H^2} \right\}} \right\} \right. \\ \leq & M \cdot \varphi_h(\widehat{\Delta}^{k^\star}) + e^2 SAH^2 + \min \left\{ e^2 C_2 \sqrt{H^5 SAM\iota}, \sum_{(x,a,h') \in \mathcal{S} \times \mathcal{A} \times [H]} \frac{e^2 C_2^2 H^3 \iota}{\max\left\{ \frac{\Delta \mathring{Q}_{h'}^{k^\star}(x,a)}{2H}, \frac{\Delta_{\min}}{4H^2} \right\}} \right\} \end{aligned} \tag{56}$$

for the constant $C_2 = 8C_0$.

2. *Fix any set $\mathcal{W} \subset [K] \times [H]$ (not necessarily a level-h set), so that for all $(k, \tilde{h}) \in \mathcal{W}$, $\tilde{h} = h$ and $x_h^k \notin \mathcal{G}_h^k$. For any $k^\star$ so that $k^\star \geq k$ for all $(k, h) \in \mathcal{W}$, it holds that $\sum_{(k,\tilde{h}) \in \mathcal{W}} \breve{\delta}_{\tilde{h}}^k \leq f(|\mathcal{W}|, \tilde{h}, k^\star)$.*

Proof In the proof of the lemma we will often use the following fact: for all $(k, h) \in [K] \times [H]$ for which $\tau_h^k = 1$, by Definition 11 and the choice of a_h^k, it holds that

$$\breve{\delta}_h^k = \Delta \breve{V}_h^k(x_h^k) \leq \Delta \breve{Q}_h^k(x_h^k, a_h^k) = \breve{\theta}_h^k. \tag{57}$$

We will use reverse induction on h to prove the statement of the lemma. The base case $h = H+1$ is immediate from the convention that $\breve{\delta}_{H+1}^k = 0$ for all $k \in [K]$.

Now we treat the inductive case. Fix $h \leq H$, and suppose that the lemma statement holds for all $h' > h$. For $(x, a) \in \mathcal{S} \times \mathcal{A}$, let $\mathcal{Z}_h(x, a)$ denote the set of all episodes $k \in [K]$ for which $(x_h^k, a_h^k) = (x, a)$ and $\tau_h^k = 1$. For a positive integer m, let $\mathcal{Z}_h^m(x, a)$ denote the set consisting of the m smallest elements of $\mathcal{Z}_h(x, a)$ (or all of $\mathcal{Z}_h(x, a)$, if $m > |\mathcal{Z}_h(x, a)|$).

Fix any k, h so that $\tau_h^k = 1$, and write $k_h^i := k_h^i(x_h^k, a_h^k)$. Then, by Definition 11 and (57),

$$\breve{\delta}_h^k \leq \breve{\theta}_h^k = \alpha_{n_h^k}^0 \cdot H + \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \cdot \left(\Delta V_{h'(k_h^i,h)}^{k_h^i}(x_{h'(k_h^i,h)}^{k_h^i}) \right) + \operatorname{clip}\left[\beta_n | \frac{\Delta_{\min}}{4H^2} \right] \tag{58}$$

$$= \alpha_{n_h^k}^0 \cdot H + \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \cdot \left(\breve{\delta}_{h'(k_h^i,h)}^{k_h^i} \right) + \operatorname{clip}\left[\beta_n | \frac{\Delta_{\min}}{4H^2} \right]. \tag{59}$$

Fix any $m \in \mathbb{N}$, as well as any $(x, a) \in \mathcal{S} \times \mathcal{A}$. As before we abbreviate $k_h^i = k_h^i(x, a)$. We next work towards an upper bound on $\sum_{k \in \mathcal{Z}_h^m(x,a)} \breve{\delta}_h^k$, using (59) for each $k \in \mathcal{Z}_h^m(x, a)$. We first sum the first term of (59) over all $k \in \mathcal{Z}_h^m(x, a)$:

$$\sum_{k \in \mathcal{Z}_h^m(x,a)} H \cdot \alpha_{n_h^k}^0 \leq H \sum_{k \in \mathcal{Z}_h(x,a)} \mathbb{1}[n_h^k = 0] \leq H, \tag{60}$$

where the first inequality follows since $\alpha_0^0 = 1$ and $\alpha_t^0 = 0$ for $t > 0$, and the second inequality follows since for all $k \in \mathcal{Z}_h(x, a)$, we have $(x_h^k, a_h^k) = (x, a)$ and there can only be a single episode in $\mathcal{Z}_h(x, a)$ during which we first visit (x, a).

The sum of the second and third terms of (59) may be bounded as follows: if $n_h^k > 0$,

$$
\begin{aligned}
&\sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \cdot \breve{\delta}_{h'(k_h^i,h)}^{k_h^i} + \mathrm{clip}\left[\beta_{n_h^k} \middle| \frac{\Delta_{\min}}{4H^2}\right] \\
\le& \mathrm{clip}\left[\mathrm{clip}\left[\beta_{n_h^k} \middle| \frac{\Delta_{\min}}{4H^2}\right] \middle| \frac{\breve{\theta}_h^k}{2H}\right] + \left(1 + \frac{1}{H}\right) \cdot \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \cdot \breve{\delta}_{h'(k_h^i,h)}^{k_h^i} && (61) \\
\le& \mathrm{clip}\left[\beta_{n_h^k} \middle| \max\left\{\frac{\Delta_{\min}}{4H^2}, \frac{\breve{\theta}_h^k}{2H}\right\}\right] + \left(1 + \frac{1}{H}\right) \cdot \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \cdot \breve{\delta}_{h'(k_h^i,h)}^{k_h^i}, && (62)
\end{aligned}
$$

where (61) follows from Lemma 21 and (59) as well as the fact that $\alpha_{n_h^k}^0 = 0$ as $n_h^k > 0$. In the case that $n_h^k = 0$, we have that $\sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \cdot \breve{\delta}_{h'(k_h^i,h)}^{k_h^i} + \mathrm{clip}\left[\beta_{n_h^k} \middle| \frac{\Delta_{\min}}{4H}\right] = 0$ since $\beta_0 = 0$ by definition (see (11)).

Next, summing (59) over all $k \in \mathcal{Z}_h^m(x, a)$, and using (60) and (62), we see that

$$
\begin{aligned}
\sum_{k \in \mathcal{Z}_h^m(x,a)} \breve{\delta}_h^k \le& H + \sum_{k \in \mathcal{Z}_h^m(x,a)} \left(\mathrm{clip}\left[\beta_{n_h^k} \middle| \max\left\{\frac{\Delta_{\min}}{4H^2}, \frac{\breve{\theta}_h^k}{2H}\right\}\right] + \left(1 + \frac{1}{H}\right) \cdot \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \cdot \breve{\delta}_{h'(k_h^i,h)}^{k_h^i}\right) \\
\le& H + \sum_{k \in \mathcal{Z}_h^m(x,a)} \mathrm{clip}\left[\beta_{n_h^k} \middle| \max\left\{\frac{\Delta_{\min}}{4H^2}, \frac{\breve{\theta}_h^k}{2H}\right\}\right] + \left(1 + \frac{1}{H}\right) \cdot \sum_{k' \in \mathcal{Z}_h^m(x,a)} \breve{\delta}_{h'(k',h)}^{k'} \sum_{t=n_h^{k'}}^{\infty} \alpha_t^{n_h^{k'}} && (63) \\
=& H + \sum_{k \in \mathcal{Z}_h^m(x,a)} \mathrm{clip}\left[\beta_{n_h^k} \middle| \max\left\{\frac{\Delta_{\min}}{4H^2}, \frac{\breve{\theta}_h^k}{2H}\right\}\right] + \left(1 + \frac{1}{H}\right)^2 \cdot \sum_{k' \in \mathcal{Z}_h^m(x,a)} \breve{\delta}_{h'(k',h)}^{k'} && (64) \\
\le& H + \sum_{k \in \mathcal{Z}_h^m(x,a)} \mathrm{clip}\left[\beta_{n_h^k} \middle| \max\left\{\frac{\Delta_{\min}}{4H^2}, \frac{\breve{\theta}_h^k}{2H}\right\}\right] \\
&+ \left(1 + \frac{1}{H}\right)^2 \cdot \left(\sum_{k' \in \mathcal{Z}_h^m(x,a):\ \tau_{h'(k',h)}^{k'} = 0} \varphi_{h'(k',h)}(\widehat{\Delta}^{k'}) + \sum_{k' \in \mathcal{Z}_h^m(x,a):\ \tau_{h'(k',h)}^{k'} = 1} \breve{\delta}_{h'(k',h)}^{k'}\right), && (65)
\end{aligned}
$$

where (63) follows from exchanging the order of summation, (64) uses item 3 of Lemma 46, and (65) uses the fact that for all $k' \in \mathcal{Z}_h^m(x, a)$ for which $\tau_{h'(k',h)}^{k'} = 0$, we have that $\breve{\delta}_{h'(k',h)}^{k'} \le \Delta V_{h'(k',h)}^{k'}(x_{h'(k',h)}^{k'}) \le \varphi_{h'(k',h)}(\widehat{\Delta}^{k'})$ (by Lemma 18), since by definition of $h'(k', h)$, either $h'(k', h) = H + 1$ or else $x_{h'(k',h)}^{k'} \notin \mathcal{G}_{h'(k',h)}^{k'}$.

Consider any level-h set $\mathcal{W} \subset [K] \times [H]$, and consider any $k^\star \geq \max_{(\tilde{k},\tilde{h}) \in \mathcal{W}}\{\tilde{k}\}$. For each $(x,a) \in \mathcal{S} \times \mathcal{A}$, let $m(x,a)$ denote the number of elements $(\widetilde{k}, h) \in \mathcal{W}$ for which $(x_h^{\widetilde{k}}, a_h^{\widetilde{k}}) = (x,a)$. Let M_1 be the number of $(\widetilde{k}, \widetilde{h}) \in \mathcal{W}$ so that either $\widetilde{h} > h$ or $\widetilde{h} = h$ and $\tau^{\widetilde{k}}_{h'(\widetilde{k},h)} = 1$, M_0 be the number of $(\widetilde{k}, \widetilde{h}) \in \mathcal{W}$ so that $\widetilde{h} = h$ and $\tau^{\widetilde{k}}_{h'(\widetilde{k},h)} = 0$, and $M := M_0 + M_1 = |\mathcal{W}|$. Then

$$\begin{aligned}
\sum_{(\tilde{k},\tilde{h}) \in \mathcal{W}} \breve{\delta}_{\tilde{h}}^{\tilde{k}} \leq & \sum_{(\widetilde{k},\widetilde{h}) \in \mathcal{R}_h(\mathcal{W})} \breve{\delta}_{\widetilde{h}}^{\widetilde{k}} && \text{(By Lemma 19)} \\
= & \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{i=1}^{m(x,a)} \breve{\delta}_h^{k_h^i(x,a)} + \sum_{(\tilde{k},\tilde{h}) \in \mathcal{R}_h(\mathcal{W}) : \tilde{h} > h} \breve{\delta}_{\tilde{h}}^{\tilde{k}} \\
\leq & SAH + \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{k \in \mathcal{Z}_h^{m(x,a)}(x,a)} \operatorname{clip}\left[\beta_{n_h^k} \middle| \max\left\{ \frac{\Delta_{\min}}{4H^2}, \frac{\breve{\theta}_h^k}{2H} \right\} \right] + \sum_{(\tilde{k},\tilde{h}) \in \mathcal{R}_h(\mathcal{W}) : \tilde{h} > h} \breve{\delta}_{\tilde{h}}^{\tilde{k}} \\
& + \left(1 + \frac{1}{H}\right)^2 \cdot \left(M_0 \cdot \varphi_{h+1}(\widehat{\Delta}^{k^\star}) + \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{k' \in \mathcal{Z}_h^{m(x,a)}(x,a) : \tau^{k'}_{h'(k',h)} = 1} \breve{\delta}^{k'}_{h'(k',h)} \right) \\
& && \text{(By (65))} \\
\leq & SAH + \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{k \in \mathcal{Z}_h^{m(x,a)}(x,a)} \operatorname{clip}\left[\beta_{n_h^k} \middle| \max\left\{ \frac{\Delta_{\min}}{4H^2}, \frac{\Delta \mathring{Q}_h^{k^\star}(x,a)}{2H} \right\} \right] \\
& + \left(1 + \frac{1}{H}\right)^2 \cdot M_0 \cdot \varphi_{h+1}(\widehat{\Delta}^{k^\star}) + \left(1 + \frac{1}{H}\right)^2 \cdot \sum_{(\tilde{k},\tilde{h}) \in \mathcal{W}'} \breve{\delta}_{\tilde{h}}^{\tilde{k}}, && (66)
\end{aligned}$$

where

$$\mathcal{W}' := \left\{ (\tilde{k}, \tilde{h}) \in \mathcal{R}_h(\mathcal{W}) : \tilde{h} > h \right\} \cup \left\{ (\tilde{k}, h'(\tilde{k}, h)) : (\tilde{k}, h) \in \mathcal{R}_h(\mathcal{W}),\ \tau^{\tilde{k}}_{h'(\tilde{k},h)} = 1 \right\},$$

so that $|\mathcal{W}'| = M_1$ and $\max_{(\tilde{k},\tilde{h}) \in \mathcal{W}'}\{\tilde{k}\} \leq k^\star$. Moreover, in (66), we have used that by Lemma 19, $\breve{\theta}_h^k = \Delta \breve{Q}_h^k(x_h^k, a_h^k) \geq \Delta \mathring{Q}_h^{k^\star}(x_h^k, a_h^k)$ for all $k \in \mathcal{Z}_h^{m(x,a)}$ (since $\tau_h^k = 1$ for all $k \in \mathcal{Z}_h^{m(x,a)}(x,a)$). We claim that $\mathcal{W}'$ is a level-$(h+1)$ set. For any $k \in [K]$, if $(k, \tilde{h}) \in \mathcal{R}_h(\mathcal{W})$ for some $\tilde{h} > h$, then since $\mathcal{R}_h(\mathcal{W})$ is a level-h set (Lemma 26), it must hold that $x_h^k \in \mathcal{G}_h^k$, meaning that $\tau_h^k = 0$, and thus $(k,h) \notin \mathcal{R}_h(\mathcal{W})$. This verifies that $\mathcal{W}'$ satisfies condition 1 of Definition 24. It is immediate that for all $(\tilde{k}, \tilde{h}) \in \mathcal{W}'$, we have $\tau_{\tilde{h}}^{\tilde{k}} = 1$ and $\tilde{h} \geq h+1$ (condition 2), and condition 3 follows from the corresponding condition for $\mathcal{R}_h(\mathcal{W})$ as well as the fact that for all $(\tilde{k}, h) \in \mathcal{R}_h(\mathcal{W})$, for all h' satisfying $h + 1 \leq h' < h'(\tilde{k}, h)$, we have $x_{h'}^{\tilde{k}} \in \mathcal{G}_{h'}^{\tilde{k}}$.

Thus, we may apply the inductive hypothesis for the set $\mathcal{W}'$, which gives, together with (66) and Lemma 23, with $\theta(x,a) = \max\left\{ \frac{\Delta \mathring{Q}_h^{k^\star}(x,a)}{2H}, \frac{\Delta_{\min}}{4H^2} \right\}$ and the set $\mathcal{W}$ in Lemma 23 set to $\{k :$

$\exists (x,a) \in \mathcal{S} \times \mathcal{A}$ s.t. $k \in \mathcal{Z}_h^{m(x,a)}(x,a)\}$,

$$\begin{aligned}
\sum_{(\tilde{k},\tilde{h}) \in \mathcal{W}} \breve{\delta}_{\tilde{h}}^{\tilde{k}} \leq & SAH + \min\left\{ 8C_0\sqrt{H^3SAM\iota}, \sum_{(x,a)\in\mathcal{S}\times\mathcal{A}} \frac{64C_0^2H^3\iota}{\max\left\{\frac{\Delta\mathring{Q}_h^{k^\star}(x,a)}{2H}, \frac{\Delta_{\min}}{4H^2}\right\}} \right\} \\
& + M_0 \cdot \left(1+\frac{1}{H}\right)^2 \cdot \varphi_{h+1}(\widehat{\Delta}^{k^\star}) + \left(1+\frac{1}{H}\right)^2 \cdot f(M_1, h+1, k^\star) \\
\leq & M \cdot \left(1+\frac{1}{H}\right)^2 \cdot \varphi_{h+1}(\widehat{\Delta}^{k^\star}) \\
& + \sum_{h'=h}^{H} \left(1+\frac{1}{H}\right)^{2(H-h)} \cdot \left(SAH + \min\left\{ C_2\sqrt{H^3SAM\iota}, \sum_{(x,a)\in\mathcal{S}\times\mathcal{A}} \frac{C_2^2H^3\iota}{\Delta\mathring{Q}_{h'}^{k^\star}(x,a)} \right\}\right) \\
= & f(M, h, k^\star),
\end{aligned}$$

thus establishing item 1 of the lemma.

Next we establish item 2 of the lemma. Fix any set $\mathcal{W} \subset [K] \times [H]$ so that for all $(k, \tilde{h}) \in \mathcal{W}$, $\tilde{h} = h$ and $x_h^k \not\in \mathcal{G}_h^k$. Suppose further that $k^\star$ satisfies $k^\star \geq k$ for all $(k,h) \in \mathcal{W}$. Thus, for all $(k,h) \in \mathcal{W}$, either $\tau_h^k = 1$ or $\breve{\delta}_h^k = \Delta\breve{V}_h^k(x_h^k) \leq \varphi_h(\widehat{\Delta}^k)$. Note also that $\mathcal{W}' := \{(k,h) \in \mathcal{W} : \tau_h^k = 1\}$ is a level-h set. Then, using item 1 on the set $\mathcal{W}'$,

$$\begin{aligned}
\sum_{(k,h)\in\mathcal{W}} \breve{\delta}_h^k \leq & \sum_{(k,h)\in\mathcal{W}:\, \tau_h^k=0} \varphi_h(\widehat{\Delta}^k) + \sum_{(k,h)\in\mathcal{W}'} \breve{\delta}_h^k \\
\leq & |\mathcal{W} \backslash \mathcal{W}'| \cdot \varphi_h(\widehat{\Delta}^{k^\star}) + f(|\mathcal{W}'|, h, k^\star) \\
\leq & f(|\mathcal{W}|, h, k^\star),
\end{aligned}$$

as desired. ■

C.4. Establishing the robustness regret bounds

In this section we prove a regret decomposition in Lemma 29 and combine it with Lemma 27, which will suffice for proving the robustness regret bounds in Theorems 8 and 9. Lemma 28 below is needed to prove the regret decomposition bound. It states that the loss incurred by choosing any non-optimal action a_h^k at a state x_h^k may be bounded by the clipped value function $\breve{\delta}_h^k$; the statement (and proof) is similar to that of in Lemma 4.4 of Xu et al. (2021).

Lemma 28 *For all $(h,k) \in [H] \times [K]$ for which $x_h^k \not\in \mathcal{G}_h^k$ and $a_h^k \not\in \mathcal{A}_{h,0}^{\mathrm{opt}}(x_h^k)$, it holds, under the event $\mathcal{E}^{\mathrm{wc}}$, that*

$$V_h^\star(x_h^k) - Q_h^\star(x_h^k, a_h^k) \leq 4 \cdot \breve{\delta}_h^k.$$

Proof We assume throughout the proof that the event $\mathcal{E}^{\mathrm{wc}}$ holds (in particular, this allows us to apply Lemma 16). Since $a_h^k \in A_h^k(x_h^k)$, we have $\overline{Q}_h^k(x_h^k, a_h^k) \geq \underline{V}_h^k(x_h^k)$, and so

$$V_h^\star(x_h^k) - Q_h^\star(x_h^k, a_h^k) \leq \overline{V}_h^k(x_h^k) - \underline{V}_h^k(x_h^k) + \overline{Q}_h^k(x_h^k, a_h^k) - \underline{Q}_h^k(x_h^k, a_h^k).$$

We may bound $\overline{V}_h^k(x_h^k) - \underline{V}_h^k(x_h^k)$ as follows:

$$\begin{aligned}
&\overline{V}_h^k(x_h^k) - \underline{V}_h^k(x_h^k) \\
\leq &\Delta V_h^k(x_h^k) && \text{(By Lemma 16 and since } x_h^k \not\in \mathcal{G}_h^k) \\
\leq &\Delta \breve{V}_h^k(x_h^k) + \frac{\Delta_{\min}}{4} = \breve{\delta}_h^k + \frac{\Delta_{\min}}{4}. && \text{(By Lemma 17)}
\end{aligned}$$

By Definition 11, there is some $k' \leq k$ so that $\breve{\delta}_h^k = \Delta \breve{V}_h^k(x_h^k) = \Delta \breve{Q}_h^{k'}(x_h^{k'}, a^\star)$ for $a^\star = \arg\max_{a' \in A_h^{k'}(x_h^k)} \overline{Q}_h^{k'}(x, a') - \underline{Q}_h^{k'}(x, a')$. Now we have

$$\begin{aligned}
&\overline{Q}_h^k(x_h^k, a_h^k) - \underline{Q}_h^k(x_h^k, a_h^k) \\
\leq &\overline{Q}_h^{k'}(x_h^k, a_h^k) - \underline{Q}_h^{k'}(x_h^k, a_h^k) && \text{(By Lemma 15)} \\
\leq &\overline{Q}_h^{k'}(x_h^k, a^\star) - \underline{Q}_h^{k'}(x_h^k, a^\star) && \text{(Since } a_h^k \in A_h^k(x_h^k) \subseteq A_h^{k'}(x_h^k)) \\
\leq &\Delta Q_h^{k'}(x_h^k, a^\star) && \text{(By Lemma 16 and since } x_h^k \not\in \mathcal{G}_h^{k'}) \\
\leq &\Delta \breve{Q}_h^{k'}(x_h^k, a^\star) + \frac{\Delta_{\min}}{4} && \text{(By Lemma 17)} \\
= &\breve{\delta}_h^k + \frac{\Delta_{\min}}{4}.
\end{aligned}$$

Since $a_h^k \not\in \mathcal{A}_{h,0}^{\text{opt}}(x_h^k)$, we have that $V_h^\star(x_h^k) - Q_h^\star(x_h^k, a_h^k) \geq \Delta_{\min}$. Thus $\Delta_{\min}/2 \leq (V_h^\star(x_h^k) - Q_h^\star(x_h^k, a_h^k))/2$, meaning that $V_h^\star(x_h^k) - Q_h^\star(x_h^k, a_h^k) \leq 4 \cdot \breve{\delta}_h^k$, as desired.

■

Next we state and prove the regret decomposition bound which is used to bound the worst-case regret.

Lemma 29 (Regret decomposition for worst-case bound) *For the choice $p = 1/(H^2 K)$, the regret of* `QLearningPreds` *may be bounded as follows:*

$$\sum_{k=1}^K \mathbb{E}\left[V_1^\star(x_1^k) - V_1^{\pi^k}(x_1^k)\right] \leq 1 + 4 \cdot \mathbb{E}\left[\sum_{(k,h): a_h^k \not\in \mathcal{A}_{h,0}^{\text{opt}}(x_h^k)} \breve{\delta}_h^k \middle| \mathcal{E}^{\text{wc}}\right].$$

Proof Note that

$$\begin{aligned}
&\sum_{k=1}^K \mathbb{E}\left[(V_1^\star - V_1^{\pi^k})(x_1^k)\right] \\
=&\sum_{k=1}^K \mathbb{E}_{\pi^k}\left[\sum_{h=1}^H V_h^\star(x_h^k) - Q_h^\star(x_h^k, a_h^k)\right] \\
\leq&\sum_{k=1}^K \mathbb{E}_{\pi^k}\left[\sum_{h=1}^H \mathbb{1}[a_h^k \not\in \mathcal{A}_{h,0}^{\text{opt}}(x_h^k)] \cdot (V_h^\star(x_h^k) - Q_h^\star(x_h^k, a_h^k))\right] \\
\leq&\sum_{k=1}^K \mathbb{E}_{\pi^k}\left[\sum_{h=1}^H \mathbb{1}[a_h^k \not\in \mathcal{A}_{h,0}^{\text{opt}}(x_h^k)] \cdot (V_h^\star(x_h^k) - Q_h^\star(x_h^k, a_h^k)) \mid \mathcal{E}^{\text{wc}}\right] + KH^2 \cdot \Pr[\overline{\mathcal{E}^{\text{wc}}}].
\end{aligned}$$

Note that $\Pr[\overline{\mathcal{E}^{\mathrm{wc}}}] \leq p$, which may be bounded above by $1/(H^2K)$ if we choose $p = 1/(H^2K)$.

Now let us condition on the event $\mathcal{E}^{\mathrm{wc}}$. Since $a_h^k \in A_h^k(x_h^k)$ for all h, k, and in the event that $|A_h^k(x_h^k)| = 1$ it must be the case that $A_h^k(x_h^k)$ contains the optimal action at x_h^k (Lemma 13, item 2) under the event $\mathcal{E}^{\mathrm{wc}}$, $a_h^k \notin \mathcal{A}_{h,0}^{\mathrm{opt}}(x_h^k)$ implies that $x_h^k \notin \mathcal{G}_h^k$ under $\mathcal{E}^{\mathrm{wc}}$. Thus, conditioned on $\mathcal{E}^{\mathrm{wc}}$, using Lemma 28, we have that

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{1}[a_h^k \notin \mathcal{A}_{h,0}^{\mathrm{opt}}(x_h^k)] \cdot (V_h^\star(x_h^k) - Q_h^\star(x_h^k, a_h^k)) \leq 4 \cdot \sum_{(k,h): a_h^k \notin \mathcal{A}_{h,0}^{\mathrm{opt}}(x_h^k)} \check{\delta}_h^k.$$

This completes the proof of the lemma. ■

To combine Lemma 27 with the regret decomposition result of Lemma 29, we need a way of upper bounding the left-hand side of (56) from Lemma 27, which looks much like the gap-based bound quantity in (7) used in Theorems 8 and 9, but with the actual gaps $\Delta_h(x, a)$ replaced by the proxies $\Delta\mathring{Q}_h^{k^\star}(x, a) \geq \Delta\mathring{Q}_h^K(x, a)$. Lemma 30 below shows that the proxies $\Delta\mathring{Q}_h^K(x, a)$ are indeed upper bounds on the true gaps $\Delta_h(x, a)$.

Lemma 30 *Consider any $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, and suppose the event $\mathcal{E}^{\mathrm{wc}}$ holds. Then*

$$\max\left\{\frac{\Delta\mathring{Q}_h^K(x, a)}{2H}, \frac{\Delta_{\min}}{4H^2}\right\} \geq \max\left\{\frac{\Delta_h(x, a)}{8H}, \mathbb{1}[\mathcal{A}_{h,0}^{\mathrm{opt}}(x) = \{a\}] \cdot \frac{\Delta_{\min,h}(x)}{8H}, \frac{\Delta_{\min}}{4H^2}\right\}. \quad (67)$$

Further, for any $\widetilde{\Delta}_{\min} \leq \Delta_{\min}$, recalling the definition of $\widetilde{\Delta}\mathring{Q}_h^k$ in step 1 of Algorithm 4, it holds that

$$\max\left\{\frac{\widetilde{\Delta}\mathring{Q}_h^K(x, a)}{2H}, \frac{\widetilde{\Delta}_{\min}}{4H^2}\right\} \geq \max\left\{\frac{\Delta_h(x, a)}{8H}, \mathbb{1}[\mathcal{A}_{h,0}^{\mathrm{opt}}(x) = \{a\}] \cdot \frac{\Delta_{\min,h}(x)}{8H}, \frac{\widetilde{\Delta}_{\min}}{4H^2}\right\}. \quad (68)$$

Proof Suppose the event $\mathcal{E}^{\mathrm{wc}}$ holds (this allows us to apply Lemmas 16 and 28). By definition, there is some $k \in [K]$ so that $\Delta\mathring{Q}_h^K(x, a) = \Delta\check{Q}_h^k(x, a)$ and either $(x_h^k, a_h^k) = (x, a)$ and $\tau_h^k = 1$ or else $k = 1$. In the case $k = 1$, we have $\Delta\mathring{Q}_h^K(x, a) = H \geq \Delta_h(x, a)$. Otherwise, we consider two cases:

- Suppose $a \notin \mathcal{A}_{h,0}^{\mathrm{opt}}(x)$. Then

  $$\Delta\mathring{Q}_h^K(x, a) = \Delta\check{Q}_h^k(x_h^k, a_h^k) \geq \check{\delta}_h^k \geq \frac{1}{4} \cdot \Delta_h(x_h^k, a_h^k) = \frac{1}{4} \cdot \Delta_h(x, a),$$

  where the first inequality follows from (57) and the second inequality follows from Lemma 28.

- Suppose that a is the unique action in $\mathcal{A}_{h,0}^{\mathrm{opt}}(x)$, i.e., that $\mathcal{A}_{h,0}^{\mathrm{opt}}(x) = \{a\}$. Since $\tau_h^k = 1$, we have that $x \notin \mathcal{G}_h^k$, meaning that there is some sub-optimal action remaining in $A_h^k(x)$, which we denote by a'. Then

  $$\begin{aligned}\overline{Q}_h^k(x_h^k, a') - \underline{Q}_h^k(x_h^k, a') \leq & \overline{Q}_h^k(x_h^k, a_h^k) - \underline{Q}_h^k(x_h^k, a_h^k) \\ & \text{(Since } a_h^k \text{ maximizes the confidence interval)} \\ \leq & \Delta Q_h^k(x_h^k, a_h^k) \qquad \text{(By Lemma 16)} \\ \leq & \Delta\check{Q}_h^k(x_h^k, a_h^k) + \frac{\Delta_{\min}}{4}. \qquad \text{(By Lemma 17)}\end{aligned}$$

Moreover, as in the proof of Lemma 28, we have, by Lemmas 16 and 17 as well as (57),

$$\overline{V}_h^k(x_h^k) - \underline{V}_h^k(x_h^k) \le \Delta V_h^k(x_h^k) \le \Delta \breve{V}_h^k(x_h^k) + \frac{\Delta_{\min}}{4} \le \Delta \breve{Q}_h^k(x_h^k, a_h^k) + \frac{\Delta_{\min}}{4}.$$

Combining the above displays, we obtain

$$\begin{aligned} \Delta_{\min,h}(x) \le & \Delta_h(x, a') \\ \le & (\overline{V}_h^k(x_h^k) - \underline{V}_h^k(x_h^k)) + (\overline{Q}_h^k(x_h^k, a') - \underline{Q}_h^k(x_h^k, a')) \\ \le & 2 \cdot \Delta \breve{Q}_h^k(x_h^k, a_h^k) + \frac{\Delta_{\min}}{2}, \end{aligned}$$

which implies that $\Delta \mathring{Q}_h^K(x,a) = \Delta \breve{Q}_h^k(x_h^k, a_h^k) \ge \frac{\Delta_{\min,h}(x)}{4}$.

The above two cases imply that $\frac{\Delta \mathring{Q}_h^K(x,a)}{2H} \ge \max\left\{ \frac{\Delta_h(x,a)}{8H}, \mathbb{1}[\mathcal{A}_{h,0}^{\mathrm{opt}}(x) = \{a\}] \cdot \frac{\Delta_{\min,h}(x)}{8H} \right\}$. The first inequality, (67), follows immediately.

To establish the second inequality, (68), of the lemma, we simply note that all arguments of this lemma (including Lemmas 17 and 28) go through without modification if $\Delta_{\min}$ is replaced with any lower bound $\widetilde{\Delta}_{\min}$ of $\Delta_{\min}$ in the definitions of $\Delta \breve{V}_h^k, \Delta \breve{Q}_h^k, \Delta \mathring{Q}_h^k$. ■

The following lemma presents the worst-case regret bound for `QLearningPreds` with the sub-procedure `DeltaConst` used to choose $\widehat{\Delta}^k$.

Lemma 31 *Suppose $T \ge SAH^3$. When given as input any prediction function $\widetilde{Q}$, the regret of `QLearningPreds` (with `DeltaConst` and input parameter $\mathscr{R} \ge \max\{SAH^3, \mathscr{C}_{M,T,1}\}$) satisfies:*

$$\mathbb{E}\left[\sum_{k=1}^K (V_1^\star - V_1^{\pi^k})(x_1^k) \right] \le O(\mathscr{R}).$$

Proof We first note that the regret decomposition of Lemma 29 gives

$$\sum_{k=1}^K \mathbb{E}\left[V_1^\star(x_1^k) - V_1^{\pi^k}(x_1^k) \right] \le 1 + 4 \cdot \mathbb{E}\left[\sum_{(k,h): a_h^k \not\in \mathcal{A}_{h,0}^{\mathrm{opt}}(x_h^k)} \breve{\delta}_h^k \middle| \mathcal{E}^{\mathrm{wc}} \right].$$

Recall that under the event $\mathcal{E}^{\mathrm{wc}}$, $a_h^k \notin \mathcal{A}_{h,0}^{\mathrm{opt}}(x_h^k)$ implies that $x_h^k \notin \mathcal{G}_h^k$. Thus, conditioned on $\mathcal{E}^{\mathrm{wc}}$, we may bound $4 \sum_{(k,h): a_h^k \notin \mathcal{A}_{h,0}^{\mathrm{opt}}(x_h^k)} \breve{\delta}_h^k$ as follows:

$$\begin{aligned}
&4 \cdot \sum_{(k,h): a_h^k \notin \mathcal{A}_{h,0}^{\mathrm{opt}}(x_h^k)} \breve{\delta}_h^k \\
\leq & 4H \cdot \left(K \cdot \varphi_1(\widehat{\Delta}^K) + e^2 SAH^2 + \min\left\{ e^2 C_2 \sqrt{H^5 SAK\iota}, \sum_{(x,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \frac{e^2 C_2^2 H^3 \iota}{\max\left\{\frac{\Delta\mathring{Q}_h^K(x,a)}{2H}, \frac{\Delta_{\min}}{4H^2}\right\}} \right\} \right) \\
& \hfill \text{(Using item 2 of Lemma 27)} \\
\leq & O(\mathscr{R}) + O(SAH^3) + O\left(\min\left\{ \sqrt{H^7 SAK\iota}, H^6 \iota \cdot \left(\sum_{(x,a,h): a \notin \mathcal{A}_{h,0}^{\mathrm{opt}}(x)} \frac{1}{\Delta_h(x,a)} + \frac{|\mathcal{A}^{\mathrm{mul}}|}{\Delta_{\min}} \right) \right\} \right) \\
& \hfill \text{(By the definition of } \widehat{\Delta}^K \text{ in } \texttt{DeltaConst}\text{, Algorithm 3, and Lemma 30)} \\
\leq & O(\mathscr{R}) + O\left(\mathscr{C}_{M,T,1}\right) \leq O(\mathscr{R}),
\end{aligned}$$

where the second-to-last inequality follows from the fact that $\mathscr{R} \geq SAH^3$. ■

The following lemma presents the worst-case regret bound for `QLearningPreds` with the sub-procedure `DeltaIncr` used to choose $\widehat{\Delta}^k$.

Lemma 32 *Suppose $T \geq SAH^3$. When given as input any prediction function $\widetilde{Q}$, the regret of the `QLearningPreds` with input parameter λ (used with `DeltaIncr` and input parameter $\widetilde{\Delta}_{\min} \leq \Delta_{\min}$) satisfies:*

$$\mathbb{E}\left[\sum_{k=1}^K (V_1^\star - V_1^{\pi^k})(x_1^k)\right] \leq O\left(\min\left\{ \sqrt{\frac{SAH^9 T\iota^2}{\lambda}}, \frac{H^8 \iota^2}{\lambda} \cdot \left(\sum_{(x,a,h): a \notin \mathcal{A}_{h,0}^{\mathrm{opt}}(x)} \frac{1}{\Delta_h(x,a)} + \frac{|\mathcal{A}^{\mathrm{mul}}|}{\widetilde{\Delta}_{\min}} \right) \right\} \right).$$

Proof We first note that the regret decomposition of Lemma 29 gives

$$\sum_{k=1}^K \mathbb{E}\left[V_1^\star(x_1^k) - V_1^{\pi^k}(x_1^k)\right] \leq 1 + 4 \cdot \mathbb{E}\left[\sum_{(k,h): a_h^k \notin \mathcal{A}_{h,0}^{\mathrm{opt}}(x_h^k)} \breve{\delta}_h^k \middle| \mathcal{E}^{\mathrm{wc}} \right].$$

The guarantee that $\widetilde{\Delta}_{\min} \leq \Delta_{\min}$ gives that $\widetilde{\Delta}\mathring{Q}_h^K(x,a) \leq \Delta\mathring{Q}_h^K(x,a)$ for all x, a, h. Recall that under the event $\mathcal{E}^{\mathrm{wc}}$, $a_h^k \notin \mathcal{A}_{h,0}^{\mathrm{opt}}(x_h^k)$ implies that $x_h^k \notin \mathcal{G}_h^k$. Thus, conditioned on $\mathcal{E}^{\mathrm{wc}}$, we may bound $4\sum_{(k,h):a_h^k \notin \mathcal{A}_{h,0}^{\mathrm{opt}}(x_h^k)} \breve{\delta}_h^k$ as follows:

$$\begin{aligned}
& 4 \cdot \sum_{(k,h):a_h^k \notin \mathcal{A}_{h,0}^{\mathrm{opt}}(x_h^k)} \breve{\delta}_h^k \\
\leq & 4H \cdot \left(K \cdot \varphi_1(\widehat{\Delta}^K) + e^2 SAH^2 + \min\left\{ e^2 C_2 \sqrt{H^5 SAK\iota}, \sum_{(x,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \frac{e^2 C_2^2 H^3 \iota}{\max\left\{ \frac{\Delta \mathring{Q}_h^K(x,a)}{2H}, \frac{\Delta_{\min}}{4H^2} \right\}} \right\} \right) \\
& \text{(Using item 2 of Lemma 27)} \\
\leq & O\left(SAH^3 + KH \cdot \min\left\{ \frac{H^5 \iota^2}{\lambda K} \cdot \sum_{(x,a,h)} \frac{1}{\max\left\{ \frac{\widetilde{\Delta} \mathring{Q}_h^K(x,a)}{2H}, \frac{\widetilde{\Delta}_{\min}}{4H^2} \right\}}, \sqrt{\frac{SAH^8\iota^2}{\lambda K}} \right\} \right). \\
& \text{(By the definition of } \widehat{\Delta}^K \text{ in } \texttt{DeltaIncr}\text{, Algorithm 4 and } \widetilde{\Delta}\mathring{Q}_h^K \leq \Delta\mathring{Q}_h^k\text{)}
\end{aligned}$$

By Lemma 30 (in particular, (68)), we conclude that

$$\begin{aligned}
& \mathbb{E}\left[\sum_{k=1}^K (V_1^\star - V_1^{\pi^k})(x_1^k) \right] \\
\leq & O\left(SAH^3 + \min\left\{ \sqrt{\frac{SAH^{10}K\iota^2}{\lambda}}, \frac{H^8\iota^2}{\lambda} \cdot \left(\sum_{(x,a,h):a\notin\mathcal{A}_{h,0}^{\mathrm{opt}}(x)} \frac{1}{\Delta_h(x,a)} \right.\right.\right. \\
& \left.\left.\left. + \sum_{(x,h):|\mathcal{A}_{h,0}^{\mathrm{opt}}(x)|=1} \frac{1}{\Delta_{\min,h}(x)} + \sum_{(x,a,h)\in\mathcal{A}^{\mathrm{mul}}} \frac{1}{\widetilde{\Delta}_{\min}} \right) \right\} \right) \qquad (69) \\
\leq & O\left(SAH^3 + \min\left\{ \sqrt{\frac{SAH^9T\iota^2}{\lambda}}, \frac{H^8\iota^2}{\lambda} \cdot \left(\sum_{(x,a,h):a\notin\mathcal{A}_{h,0}^{\mathrm{opt}}(x)} \frac{1}{\Delta_h(x,a)} + \frac{|\mathcal{A}^{\mathrm{mul}}|}{\widetilde{\Delta}_{\min}} \right) \right\} \right).
\end{aligned}$$

Finally, the fact that $T \geq SAH^3$ implies that the term SAH^3 in the above expression is dominated by the second term (see also Lemma 47). ■

We remark that in the proof of Lemma 32, if $\widetilde{\Delta}_{\min} = |\mathcal{A}^{\mathrm{mul}}| = 0$, then the term $\frac{|\mathcal{A}^{\mathrm{mul}}|}{\widetilde{\Delta}_{\min}}$ can be interepreted as 0. This follows from the fact that in the inequality (69), the summation in the third term $\sum_{(x,a,h)\in\mathcal{A}^{\mathrm{mul}}} \frac{1}{\widetilde{\Delta}_{\min}}$ is over an empty set.

Appendix D. Proofs for approximate distillation bound

In this section we establish the upper bounds in item 2 of Theorem 8 and item 2 of Theorem 9, which give a regret bound for `QLearningPreds` when the predictions $\widetilde{Q}_h$ are an ϵ-approximate distillation of $Q_h^\star$.

D.1. Bounding the number of exploration episodes

A key challenge in establishing these bounds is to show that `QLearningPreds` does not spend too many episodes ignoring the predictions $\widetilde{Q}_h$ as part of the exploration phase. To this end, we bound the number of episodes k for which $\sigma_h^k = 1$ (for each $h \in [H]$). Note that this is not exactly the same as the number of episodes k for which $\tau_h^k = 1$, and that it is the parameters τ_h^k (not σ_h^k) which correspond to whether the policy π_h^k (defined in (12)) engages in exploration or constrained exploitation. We will show, however (in Claim 42), that those episodes k for which $\sigma_h^k = 0$ but $\tau_h^k = 1$ only contribute a small amount to the overall regret; this is in turn a consequence of Lemma 39, which shows that if there is a non-optimal action in $A_h^k(x)$, then $\Delta V_h^k(x)$ (which is used to define τ_h^k) and $\Delta\breve{Q}_h^k(x)$ (which is used to define σ_h^k) must be close.

Recall the definition of $\widehat{\lambda}$ in Theorem 9. Lemma 33 treats the case where `DeltaConst` is used to choose $\widehat{\Delta}^k$; it bounds, for each $h \in [H]$, the number of episodes k for which $\sigma_h^k = 1$, as a function of $\widehat{\lambda}$. The main tool in the proof is Lemma 27, which is used to show that the parameters $\breve{\delta}_h^k = \Delta\breve{V}_h^k(x_h^k)$ decrease sufficiently fast to $\Delta\breve{V}_h^k(x_h^k) \le \frac{1}{1+\frac{1}{H}} \cdot \varphi_h(\widehat{\Delta}^k)$, i.e., $\sigma_h^k = 0$, for most episodes k.

Lemma 33 *Suppose `QLearningPreds` is run with `DeltaConst` to choose the values $\widehat{\Delta}^k$. Then for all $h \in [H]$, the number of episodes $k \in [K]$ for which $\sigma_h^k = 1$ is at most* $\max\{SAH^3, \widehat{\lambda} \cdot K\}$.

Proof Per `DeltaConst`, we have that $\widehat{\Delta}^k = \mathscr{R}/(KH)$ for all $k \in [K]$. Therefore, throughout the proof of this lemma we will drop the superscript k and write $\widehat{\Delta} := \widehat{\Delta}^k$ (which holds for all $k \in [K]$).

For any $(h,k) \in [H] \times [K]$, note that $\sigma_h^k = 1$ implies that $\varphi_h(\widehat{\Delta}) < (1+1/H) \cdot \Delta\breve{V}_h^k(x_h^k) = (1+1/H) \cdot \breve{\delta}_h^k$. Write $\mathcal{Y}_h := \{k : \sigma_h^k = 1\}$. Then for each h, we have that

$$\sum_{k \in \mathcal{Y}_h} \breve{\delta}_h^k \ge \frac{1}{1+1/H} \cdot \varphi_h(\widehat{\Delta}) \cdot |\mathcal{Y}_h|.$$

Using the above inequality and item 2 of Lemma 27 with the set $\mathcal{W} = \{(k,h) : \sigma_h^k = 1\}$ (which satisfies the requirement that each $(k,h) \in \mathcal{W}$ satisfies $x_h^k \notin \mathcal{G}_h^k$), we get that

$$\begin{aligned}
&\frac{\varphi_h(\widehat{\Delta}) \cdot |\mathcal{Y}_h|}{1+1/H} \\
\le &\sum_{k \in \mathcal{Y}_h} \breve{\delta}_h^k \\
\le &|\mathcal{Y}_h| \cdot (1+1/H)^2 \cdot \varphi_{h+1}(\widehat{\Delta}) + e^2 SAH^2 + \min\left\{ e^2 C_2 \sqrt{H^5 SA |\mathcal{Y}_h| \iota}, \sum_{(x,a,h') \in \mathcal{S} \times \mathcal{A} \times [H]} \frac{e^2 C_2^2 H^3 \iota}{\max\left\{ \frac{\Delta \mathring{Q}_{h'}^K(x,a)}{2H}, \frac{\Delta_{\min}}{4H^2} \right\}} \right\}.
\end{aligned}$$

Rearranging and using the fact that $\varphi_h(\frac{\widehat{\Delta}}{1+1/H}) - (1+1/H)^2 \cdot \varphi_{h+1}(\widehat{\Delta}) \ge \frac{\varphi_{h+1}(\widehat{\Delta})}{H}$, we obtain that

$$\frac{|\mathcal{Y}_h| \cdot \varphi_{h+1}(\widehat{\Delta})}{H} \le e^2 SAH^2 + \min\left\{ e^2 C_2 \sqrt{H^5 SA |\mathcal{Y}_h| \iota}, \sum_{(x,a,h') \in \mathcal{S} \times \mathcal{A} \times [H]} \frac{e^2 C_2^2 H^3 \iota}{\max\left\{ \frac{\Delta \mathring{Q}_{h'}^K(x,a)}{2H}, \frac{\Delta_{\min}}{4H^2} \right\}} \right\}.$$

By Lemma 30, it follows that

$$\begin{aligned}
&\frac{|\mathcal{Y}_h| \cdot \varphi_{h+1}(\widehat{\Delta})}{H} \\
\leq & e^2 SAH^2 + \min\left\{ e^2 C_2 \sqrt{H^5 SA |\mathcal{Y}_h| \iota}, 16 e^2 C_2^2 H^5 \iota \cdot \left(\sum_{(x,a,h') \in \mathcal{S} \times \mathcal{A} \times [H] : a \not\in \mathcal{A}^{\mathrm{opt}}_{h',0}(x)} \frac{1}{\Delta_h(x,a)} + \frac{|\mathcal{A}^{\mathrm{mul}}|}{\Delta_{\min}} \right) \right\} \\
\leq & \min\left\{ 2e^2 C_2 \sqrt{H^5 SA |\mathcal{Y}_h| \iota}, 32 e^2 C_2^2 H^5 \iota \cdot \left(\sum_{(x,a,h') \in \mathcal{S} \times \mathcal{A} \times [H] : a \not\in \mathcal{A}^{\mathrm{opt}}_{h',0}(x)} \frac{1}{\Delta_h(x,a)} + \frac{|\mathcal{A}^{\mathrm{mul}}|}{\Delta_{\min}} \right) \right\},
\end{aligned} \tag{70}$$

where the second inequality above follows from the fact that, assuming $|\mathcal{Y}_h| \geq SAH^3$, both terms in the minimum are bounded below by $e^2 SAH^2$. Recall that $\widehat{\lambda} \geq SAH^3/K$ is defined to be as small as possible so that

$$\mathscr{R} \geq \min\left\{ \sqrt{\frac{H^9 SAK\iota}{\widehat{\lambda}}}, \frac{1}{\widehat{\lambda}} \cdot H^7 \iota \cdot \left(\sum_{(x,a,h') \in \mathcal{S} \times \mathcal{A} \times [H] : a \not\in \mathcal{A}^{\mathrm{opt}}_{h',0}(x)} \frac{1}{\Delta_h(x,a)} + \frac{|\mathcal{A}^{\mathrm{mul}}|}{\Delta_{\min}} \right) \right\},$$

and that $\widehat{\Delta} = \mathscr{R}/(KH)$ (per `DeltaConst`). We next consider two cases:

1. $\mathscr{R} \geq \sqrt{\frac{H^9 SAK\iota}{\widehat{\lambda}}}$. Then from (70),

   $$|\mathcal{Y}_h| \leq \frac{2e^2 C_2 \sqrt{H^7 SA |\mathcal{Y}_h| \iota}}{\varphi_{h+1}(\widehat{\Delta})} \leq \frac{2e^2 C_2}{C_1} \cdot \frac{\sqrt{H^7 SA |\mathcal{Y}_h| \iota}}{\sqrt{H^7 SA\iota / (\widehat{\lambda} \cdot K)}},$$

   which implies that

   $$\sqrt{|\mathcal{Y}_h|} \leq \frac{2e^2 C_2}{C_1} \cdot \sqrt{\widehat{\lambda} \cdot K},$$

   and in turn we get that $|\mathcal{Y}_h| \leq \widehat{\lambda} \cdot K$ since C_1 is chosen so that $2e^2 C_2 \leq C_1$.

2. $\mathscr{R} \geq \frac{1}{\widehat{\lambda}} \cdot H^7 \iota \cdot \left(\sum_{(x,a,h') \in \mathcal{S} \times \mathcal{A} \times [H] : a \not\in \mathcal{A}^{\mathrm{opt}}_{h',0}(x)} \frac{1}{\Delta_h(x,a)} + \frac{|\mathcal{A}^{\mathrm{mul}}|}{\Delta_{\min}} \right)$. Then from (70),

   $$\begin{aligned}
   |\mathcal{Y}_h| \leq & \frac{32 e^2 C_2^2 H^6 \iota}{\varphi_{h+1}(\widehat{\Delta})} \cdot \left(\sum_{(x,a,h') \in \mathcal{S} \times \mathcal{A} \times [H] : a \not\in \mathcal{A}^{\mathrm{opt}}_{h',0}(x)} \frac{1}{\Delta_h(x,a)} + \frac{|\mathcal{A}^{\mathrm{mul}}|}{\Delta_{\min}} \right) \\
   \leq & \frac{32 e^2 C_2^2}{C_1} \cdot K \cdot \widehat{\lambda} \leq K \cdot \widehat{\lambda},
   \end{aligned}$$

   where the final inequality follows since C_1 is chosen so that $32 e^2 C_2^2 \leq C_1$.

■

For all $h \in [H], k \in [K]$, let $\widetilde{\Delta}\breve{V}_h^k(\cdot), \widetilde{\Delta}\breve{Q}_h^k(\cdot), \widetilde{\Delta}\mathring{Q}_h^k(\cdot)$ be defined identically to $\Delta\breve{V}_h^k(\cdot), \Delta\breve{Q}_h^k(\cdot), \Delta\mathring{Q}_h^k(\cdot)$ (Definitions 11 and 12), except the parameter $\Delta_{\min}$ in the definition of these parameters is replaced with $\widetilde{\Delta}_{\min}$ (note that $\widetilde{\Delta}\mathring{Q}_h^k$ was already defined in this manner in Algorithm 4).

Lemma 34 *Suppose* `QLearningPreds` *is run with* `DeltaIncr` *(Algorithm 4) to choose the values $\widehat{\Delta}^k$. As long as $\widetilde{\Delta}_{\min} \leq \Delta_{\min}$, then for all $h \in [H],\ k \in [K],\ x \in \mathcal{S},\ a \in \mathcal{A}$, we have*

$$\max\left\{\widetilde{\Delta}\breve{Q}_h^k(x,a), \frac{\widetilde{\Delta}_{\min}}{4H^2}\right\} \leq \Delta\breve{Q}_h^k(x,a) + \frac{\Delta_{\min}}{4H}, \qquad \max\left\{\widetilde{\Delta}\breve{V}_h^k(x), \frac{\widetilde{\Delta}_{\min}}{4H^2}\right\} \leq \Delta\breve{V}_h^k(x) + \frac{\Delta_{\min}}{4H}.$$

Furthermore, the first of the above inequalities holds with $\widetilde{\Delta}\breve{Q}_h^k, \Delta\breve{Q}_h^k$ replaced by $\widetilde{\Delta}\mathring{Q}_h^k, \Delta\mathring{Q}_h^k$, respectively.

Proof We prove the result by forward induction on k and reverse induction on h, noting that the base cases $k = 1$ and $h = H + 1$ are immediate. To prove the inductive step, fix any $(h, k) \in [H] \times [K]$, and suppose that for all $h' > h,\ k' < k$, it holds that, for all x, a,

$$\max\left\{\widetilde{\Delta}\breve{Q}_{h'}^{k'}(x,a), \frac{\widetilde{\Delta}_{\min}}{4H^2}\right\} \leq \Delta\breve{Q}_{h'}^{k'}(x,a) + (H+1-h) \cdot \frac{\Delta_{\min}}{4H^2}, \tag{71}$$

$$\max\left\{\widetilde{\Delta}\breve{V}_{h'}^{k'}(x), \frac{\widetilde{\Delta}_{\min}}{4H^2}\right\} \leq \Delta\breve{V}_{h'}^{k'}(x) + (H+1-h) \cdot \frac{\Delta_{\min}}{4H^2}. \tag{72}$$

Note that for any non-negative real numbers $z, y, \widetilde{y}$ so that $y \geq \widetilde{y}$, we have $\max\{\mathrm{clip}\,[z|\, y]\,, y\} \geq \max\{\mathrm{clip}\,[z|\, \widetilde{y}]\,, \widetilde{y}\}$. Fixing some pair (x, a) and letting $n = N_h^k(x,a)$, $k_h^i = k_h^i(x,a)$ for all $i \in [n]$, we note that

$$\begin{aligned}
&\max\left\{\alpha_n^0 H + \mathrm{clip}\left[\beta_n \middle| \frac{\widetilde{\Delta}_{\min}}{4H^2}\right] + \sum_{i=1}^n \alpha_n^i \cdot \widetilde{\Delta}\breve{V}_{h'(k_h^i,h)}^{k_h^i}(x_{h'(k_h^i,h)}^{k_h^i}), \frac{\widetilde{\Delta}_{\min}}{4H^2}\right\} \\
\leq& \alpha_n^0 H + \max\left\{\mathrm{clip}\left[\beta_n \middle| \frac{\widetilde{\Delta}_{\min}}{4H^2}\right], \frac{\widetilde{\Delta}_{\min}}{4H^2}\right\} + \sum_{i=1}^n \alpha_n^i \cdot \widetilde{\Delta}\breve{V}_{h'(k_h^i,h)}^{k_h^i}(x_{h'(k_h^i,h)}^{k_h^i}) \\
\leq& \alpha_n^0 H + \max\left\{\mathrm{clip}\left[\beta_n \middle| \frac{\Delta_{\min}}{4H^2}\right], \frac{\Delta_{\min}}{4H^2}\right\} + \sum_{i=1}^n \alpha_n^i \cdot \left(\Delta\breve{V}_{h'(k_h^i,h)}^{k_h^i}(x_{h'(k_h^i,h)}^{k_h^i}) + (H-h) \cdot \frac{\Delta_{\min}}{4H^2}\right) \\
\leq& \alpha_n^0 H + \mathrm{clip}\left[\beta_n \middle| \frac{\Delta_{\min}}{4H^2}\right] + \sum_{i=1}^n \alpha_n^i \cdot \Delta\breve{V}_{h'(k_h^i,h)}^{k_h^i}(x_{h'(k_h^i,h)}^{k_h^i}) + (H+1-h) \cdot \frac{\Delta_{\min}}{4H^2}.
\end{aligned}$$

Then (71) for $(h', k') = (h, k)$ follows from the definition of $\Delta\breve{Q}_h^k$, and (72) for $(h', k') = (h, k)$ follows similarly from the definition of $\Delta\breve{V}_h^k$.

The final statement of the lemma follows since (by Definition 12) for all (x, a, h, k), there is some k' so that $\Delta\mathring{Q}_h^k(x,a) = \Delta\breve{Q}_h^{k'}(x,a)$ and $\widetilde{\Delta}\mathring{Q}_h^k(x,a) = \widetilde{\Delta}\breve{Q}_h^{k'}(x,a)$. ■

Lemma 35 establishes the same result as Lemma 33, except for the choice of `DeltaIncr` in `QLearningPreds`. The proof is more subtle, though, because of the more complex nature of the parameters $\widehat{\Delta}^k$ in `DeltaIncr`. In particular, to establish Lemma 35, we need to divide the set of episodes into different phases, so that within each phase the value of $\widehat{\Delta}^k$ only changes by a small multiplicative factor.

Lemma 35 *Suppose* `QLearningPreds` *is run with* `DeltaIncr` *(Algorithm 4) to choose the values $\widehat{\Delta}^k$. Then for all $h \in [H]$, the number of episodes $k \in [K]$ for which $\sigma_h^k = 1$ is at most* $\max\{SAH^3, \lambda \cdot K\}$.

Proof Since $\Delta\breve{Q}_h^1(x,a) = H$ for all (x,a,h), it holds that $\widehat{\Delta}^1 \geq \frac{SAH}{\lambda K}$. Also note that by definition we have $\widehat{\Delta}^k \leq \sqrt{\frac{SAH^8\iota^2}{\lambda K}}$ for all k. Note that $\sqrt{\frac{SAH^8\iota^2}{\lambda K}} \cdot \frac{\lambda K}{SAH} \leq \sqrt{\lambda K H^6 \iota^2}$. For $0 \leq i \leq \left\lceil \log_{1+\frac{1}{H}}(\sqrt{\lambda K H^6 \iota^2}) \right\rceil$, set $\omega_i := \left(1+\frac{1}{H}\right)^i \cdot \frac{SAH}{\lambda K}$.

For $(h,k) \in [H] \times [K]$, note that $\sigma_h^k = 1$ implies that $\varphi_h(\widehat{\Delta}^k) < (1+1/H) \cdot \Delta\breve{V}_h^k(x_h^k) = (1+1/H) \cdot \breve{\delta}_h^k$. For each $1 \leq i \leq \lceil \log_{1+1/H}(\sqrt{\lambda K H^6 \iota^2}) \rceil$ and $h \in [H]$ set $\mathcal{Y}_h^i := \{k \in [K] : \sigma_h^k = 1,\ \omega_{i-1} \leq \widehat{\Delta}^k \leq \omega_i\}$.

Then for each $h \in [H]$ and $0 \leq i \leq \lceil \log_{1+\frac{1}{H}}(\sqrt{\lambda K H^6 \iota^2}) \rceil$,

$$\sum_{k \in \mathcal{Y}_h^i} \breve{\delta}_h^k \geq |\mathcal{Y}_h^i| \cdot \frac{\varphi_h(\omega_{i-1})}{1+\frac{1}{H}}. \tag{73}$$

Set $\mathcal{Y}_h := \bigcup_i \mathcal{Y}_h^i$. Fix any $h \in [H]$ and i satisfying $1 \leq i \leq \lceil \log_{1+\frac{1}{H}}(\sqrt{\lambda K H^6 \iota^2}) \rceil$. Using (73) and the statement of item 2 of Lemma 27 for $\mathcal{W} = \{(k,h) : k \in \mathcal{Y}_h^i\}$, noting that for $k^\star = \max_{k \in \mathcal{Y}_h^i}\{k\}$, we have $\widehat{\Delta}^{k^\star} \leq \omega_i$, we see that

$$\begin{aligned}
&\frac{|\mathcal{Y}_h^i| \cdot \varphi_h(\omega_{i-1})}{1+1/H} \\
\leq &\sum_{k \in \mathcal{Y}_h^i} \breve{\delta}_h^k \\
\leq &|\mathcal{Y}_h^i| \cdot \left(1+\frac{1}{H}\right)^2 \cdot \varphi_{h+1}(\omega_i) + e^2 SAH^2 + \min\left\{ e^2 C_2 \sqrt{H^5 SA |\mathcal{Y}_h^i| \iota},\ \sum_{(x,a,h') \in \mathcal{S}\times\mathcal{A}\times[H]} \frac{e^2 C_2^2 H^3 \iota}{\max\left\{\frac{\Delta\mathring{Q}_{h'}^{k^\star}(x,a)}{2H}, \frac{\Delta_{\min}}{4H^2}\right\}} \right\}.
\end{aligned}$$

Rearranging and using the fact that $\frac{\varphi_h(\omega_{i-1})}{1+1/H} - \left(1+\frac{1}{H}\right)^2 \cdot \varphi_{h+1}(\omega_{i-1}) \geq \varphi_{h+1}(\omega_{i-1})/H$, we obtain that

$$\begin{aligned}
\frac{|\mathcal{Y}_h^i|}{H} \cdot \varphi_{h+1}(\omega_{i-1}) \leq &e^2 SAH^2 + \min\left\{ e^2 C_2 \sqrt{H^5 SA |\mathcal{Y}_h^i| \iota},\ \sum_{(x,a,h') \in \mathcal{S}\times\mathcal{A}\times[H]} \frac{e^2 C_2^2 H^3 \iota}{\max\left\{\frac{\Delta\mathring{Q}_{h'}^{k^\star}(x,a)}{2H}, \frac{\Delta_{\min}}{4H^2}\right\}} \right\} \\
\leq &e^2 SAH^2 + \min\left\{ e^2 C_2 \sqrt{H^5 SA |\mathcal{Y}_h^i| \iota},\ \sum_{(x,a,h') \in \mathcal{S}\times\mathcal{A}\times[H]} \frac{2e^2 C_2^2 H^4 \iota}{\max\left\{\frac{\widetilde{\Delta}\mathring{Q}_{h'}^{k^\star}(x,a)}{2H}, \frac{\widetilde{\Delta}_{\min}}{4H^2}\right\}} \right\},
\end{aligned} \tag{74}$$

where the second inequality above follows from $\widetilde{\Delta}_{\min} \leq \Delta_{\min}$ and therefore, from Lemma 34,

$$\max\left\{\frac{\widetilde{\Delta}\mathring{Q}^{k^\star}_{h'}(x,a)}{2H}, \frac{\widetilde{\Delta}_{\min}}{4H^2}\right\} \leq 2H \cdot \max\left\{\frac{\Delta\mathring{Q}^{k^\star}_{h'}(x,a)}{2H}, \frac{\Delta_{\min}}{4H^2}\right\}$$

for all x, a, h'. We now consider two cases, based on the value of $\widehat{\Delta}^{k^\star}$ (depending on which of the two terms in the minimum in (14) in the algorithm `DeltaIncr` is smaller):

1. Suppose $\widehat{\Delta}^{k^\star} = \frac{H^6\iota^2}{\lambda K} \cdot \sum_{(x,a,h)} \frac{1}{\max\left\{\frac{\widetilde{\Delta}\mathring{Q}^{k^\star}_{h}(x,a)}{2H}, \frac{\widetilde{\Delta}_{\min}}{4H^2}\right\}}$. Note that

$$\varphi_{h+1}(\omega_{i-1}) \geq C_1 \cdot \omega_{i-1} \geq C_1/(1+1/H) \cdot \omega_i \geq C_1/(1+1/H) \cdot \widehat{\Delta}^{k^\star} \geq C_1/2 \cdot \widehat{\Delta}^{k^\star}, \tag{75}$$

as well as $\lceil \log_{1+1/H}(\sqrt{\lambda K H^6 \iota^2}) \rceil \leq 8H\iota$. Then using (74), we get that

$$\begin{aligned}
|\mathcal{Y}_h^i| \leq& \frac{1}{\varphi_{h+1}(\omega_{i-1})} \cdot \left(e^2 SAH^3 + \sum_{(x,a,h')} \frac{2e^2 C_2^2 H^5 \iota}{\max\left\{\frac{\widetilde{\Delta}\mathring{Q}^{k^\star}_{h'}(x,a)}{2H}, \frac{\widetilde{\Delta}_{\min}}{4H^2}\right\}} \right) \\
\leq& \frac{1}{\varphi_{h+1}(\omega_{i-1})} \cdot \sum_{(x,a,h')} \frac{4e^2 C_2^2 H^5 \iota}{\max\left\{\frac{\widetilde{\Delta}\mathring{Q}^{k^\star}_{h'}(x,a)}{2H}, \frac{\widetilde{\Delta}_{\min}}{4H^2}\right\}} \qquad (76) \\
\leq& \frac{2 \cdot 7H\iota}{C_1 \cdot \widehat{\Delta}^{k^\star} \cdot \lceil \log_{1+1/H}(\sqrt{\lambda K H^6 \iota^2}) \rceil} \cdot \sum_{(x,a,h')} \frac{4e^2 C_2^2 H^5 \iota}{\max\left\{\frac{\widetilde{\Delta}\mathring{Q}^{k^\star}_{h'}(x,a)}{2H}, \frac{\widetilde{\Delta}_{\min}}{4H^2}\right\}} \\
=& \frac{56 e^2 C_2^2 \cdot \lambda K}{C_1 \cdot \lceil \log_{1+1/H}(\sqrt{\lambda K H^6 \iota^2}) \rceil} \leq \frac{\lambda K}{\lceil \log_{1+1/H}(\sqrt{\lambda K H^6 \iota^2}) \rceil}, \qquad (77)
\end{aligned}$$

where (76) follows since $\max\left\{\frac{\widetilde{\Delta}\mathring{Q}^{k^\star}_{h'}(x,a)}{2H}, \frac{\widetilde{\Delta}_{\min}}{4H^2}\right\} \leq H$ for all (x, a, h'), and (77) follows since C_1 is chosen so that $C_1 \geq 56e^2C_2^2$ (see (23)). Therefore, $|\mathcal{Y}_h| \leq \sum_{i=1}^{\lceil \log_{1+1/H}(\sqrt{\lambda K H^6 \iota^2}) \rceil} |\mathcal{Y}_h^i| \leq \lambda K$.

2. Otherwise, by the definition of $\widehat{\Delta}^k$ in (14), we have $\widehat{\Delta}^{k^\star} = \sqrt{\frac{SAH^8\iota^2}{\lambda K}} = \widehat{\Delta}^K$. Note that (75) still holds, and so, using (74), we get that, for each i,

$$\begin{aligned}
|\mathcal{Y}_h^i| \leq& \frac{1}{\varphi_{h+1}(\omega_{i-1})} \cdot \left(e^2 SAH^3 + e^2 C_2 \sqrt{H^7 SA |\mathcal{Y}_h^i| \iota} \right) \\
\leq& \frac{1}{\varphi_{h+1}(\omega_{i-1})} \cdot 2e^2 C_2 \sqrt{H^7 SA |\mathcal{Y}_h^i| \iota} \qquad (78) \\
\leq& \frac{\sqrt{8H\iota}}{C_1 \cdot \widehat{\Delta}^{k^\star} \cdot \sqrt{\lceil \log_{1+1/H}(\sqrt{\lambda K H^7 \iota^2}) \rceil}} \cdot 2e^2 C_2 \sqrt{H^7 SA |\mathcal{Y}_h^i| \iota},
\end{aligned}$$

which implies that

$$\sqrt{|\mathcal{Y}_h^i|} \leq \frac{6e^2 C_2}{C_1} \cdot \frac{1}{\sqrt{\lceil \log_{1+1/H}(\sqrt{\lambda K H^7 \iota^2}) \rceil}} \cdot \sqrt{\lambda K},$$

and since C_1 is chosen so that $C_1 \geq 6e^2 C_2$, we get that $|\mathcal{Y}_h| \leq \lambda K$, as desired.

Thus, in both cases, we obtain that $|\mathcal{Z}_h| \leq \max\{SAH^3, \lambda K\}$, completing the proof of the lemma. ■

D.2. Bounding the value functions $\overline{R}_h^k, \widetilde{Q}_h^k, \widetilde{V}_h^k$

In this section we establish some basic bounds on the value functions $\overline{R}_h^k, \widetilde{Q}_h^k, \widetilde{V}_h^k$ maintained by `QLearningPreds` to refine the predictions $\widetilde{Q}_h$. Many of the results are analogous to the bounds on $\overline{Q}_h^k, \underline{Q}_h^k, \overline{V}_h^k, \underline{V}_h^k$ proven in Section C.1. However, since the updating procedures are distinct from those used to update the upper and lower Q- and V-value functions (in particular, we do not use the multi-step bootstrap of Xu et al. (2021) to update $\widetilde{Q}_h^k, \widetilde{V}_h^k$), we cannot derive the results in this section directly from those in Section C.1.

The first result, Lemma 36, is a straightforward consequence of the updates to $\overline{R}_h^k$ in `QLearningPreds`.

Lemma 36 *For any $(x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, suppose the episodes in which (x, a) was previously taken at step h are denoted $k^1, \ldots, k^n < k$ (in particular, $k^i = k_h^i(x, a)$ and $n = N_h^k(x, a)$). Then the following identity holds:*

$$(\overline{R}_h^k - Q_h^\star)(x, a) = \alpha_n^0 (H - Q_h^\star(x, a)) + \sum_{i=1}^n \alpha_n^i \cdot \left((\widetilde{V}_{h+1}^{k^i} - V_{h+1}^\star)(x_{h+1}^{k^i}) + \left((\hat{\mathbb{P}}_h^{k^i} - \mathbb{P}_h) V_{h+1}^\star \right)(x, a) + b_i \right).$$

Proof Note that $\overline{R}_h^k(x, a)$ is updated as follows:

$$\overline{R}_h^{k+1}(x, a) = \begin{cases} (1 - \alpha_n) \cdot \overline{R}_h^k(x, a) + \alpha_n \cdot [r_h(x, a) + \widetilde{V}_{h+1}^k(x_{h+1}^k) + b_n] : & (x, a) = (x_h^k, a_h^k), \\ \overline{R}_h^k(x, a) : & \text{else}, \end{cases}$$

where $n = N_h^{k+1}(x, a)$ in the first case above. Iterating the above, we obtain that for any (x, a, h, k), letting $n = N_h^k(x, a)$,

$$\overline{R}_h^k(x, a) = \alpha_n^0 \cdot H + \sum_{i=1}^n \alpha_n^i \cdot \left(r_h(x, a) + \widetilde{V}_{h+1}^{k^i}(x_{h+1}^{k^i}) + b_i \right). \tag{79}$$

Using the Bellman optimality equation $Q_h^\star(x, a) = r_h(x, a) + \mathbb{P}_h V_{h+1}^\star(x, a)$ together with the fact that $\sum_{i=0}^n \alpha_n^i = 1$ and the notation $(\hat{\mathbb{P}}_h^{k^i} V_{h+1})(x, a) = V_{h+1}(x_{h+1}^{k^i})$ for $(x, a) = (x_h^{k^i}, a_h^{k^i})$, we see that, for $n = N_h^k(x, a)$,

$$Q_h^\star(x, a) = \alpha_n^0 \cdot Q_h^\star(x, a) + \sum_{i=1}^n \alpha_n^i \cdot \left(r_h(x, a) + (\mathbb{P}_h - \hat{\mathbb{P}}_h^{k^i}) V_{h+1}^\star(x, a) + V_{h+1}^\star(x_{h+1}^{k^i}) \right). \tag{80}$$

Subtracting (80) from (79) gives the desired result. ■

The following straightforward lemma, which generalizes item 2 of Lemma 13, shows that any *approximately* optimal action a at any state (x, h) either remains in $A_h^k(x)$ at each episode k or else there is some other action in $A_h^k(x)$ with smaller sub-optimality than a.

Lemma 37 *Under the event $\mathcal{E}^{\mathrm{wc}}$, for any $\epsilon > 0$ and every $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, if it holds that $\Delta_h(x, a) \leq \epsilon$, then for each $k \in [K]$, at least one of the following must hold true:*

- *$a \in A_h^k(x)$; or*
- *For some $a^\star \in A_h^k(x)$ (in particular, we may choose $a^\star \in A_h^k(x)$ maximizing $\underline{Q}_h^k(x, a^\star)$), $\Delta_h(x, a^\star) \leq V_h^\star(x) - \underline{Q}_h^k(x, a^\star) < \epsilon$.*

Proof If $a \notin A_h^k(x)$, then it must be the case that for some $k' \leq k$, $\overline{Q}_h^{k'}(x, a) < \underline{V}_h^{k'}(x)$; by Lemma 15, we have $\underline{V}_h^k(x) \geq \underline{V}_h^{k'}(x)$, and so some action $a^\star \in A_h^k(x)$ must satisfy $Q_h^\star(x, a^\star) \geq \underline{Q}_h^k(x, a^\star) = \underline{V}_h^k(x) > \overline{Q}_h^{k'}(x, a) \geq Q_h^\star(x, a)$. Hence $\Delta_h(x, a^\star) = V_h^\star(x) - Q_h^\star(x, a^\star) \leq V_h^\star(x) - \underline{Q}_h^k(x, a^\star) < V_h^\star(x) - Q_h^\star(x, a) = \Delta_h(x, a) \leq \epsilon$. ■

The next lemma, Lemma 38, uses Lemmas 36 and 37 above together with a martingale concentration inequality to show bounds on $\widetilde{Q}_h^k, \widetilde{V}_h^k$ that hold with high probability. We note that an additional necessary ingredient is the assumption that the input predictions $\widetilde{Q}_h$ are an ϵ-approximate distillation of $Q_h^\star$; this is used to show that for all $k \in [K]$, $\widetilde{Q}_h^k$ is also an approximate distillation with high probability (item 3), which in turn is used to show that $\widetilde{V}_h^k$ is approximately lower bounded by $V_h^\star$ (item 4).

Lemma 38 *Set $p = 1/(H^2 K)$. Suppose that $\widetilde{Q}$ is an ϵ-approximate distillation on the optimal value function $Q_h^\star$. Then, there is an event $\mathcal{E}^{\mathrm{pred}}$ with $\Pr[\mathcal{E}^{\mathrm{pred}}] \geq 1 - p$ so that the following hold under $\mathcal{E}^{\mathrm{pred}} \cap \mathcal{E}^{\mathrm{wc}}$:*

1. *For $n \in \mathbb{N}$, recall that $\beta_n = 2 \sum_{i=1}^n \alpha_n^i b_i$. Then for any $(x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, it holds that, for $n = N_h^k(x, a)$,*

$$(\overline{R}_h^k - Q_h^\star)(x, a) \leq \alpha_n^0 H + \sum_{i=1}^n \alpha_n^i \cdot (\widetilde{V}_{h+1}^{k_h^i(x,a)} - V_{h+1}^\star)(x_{h+1}^{k_h^i(x,a)}) + \beta_n.$$

2. *For all $(x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, it holds that $\overline{R}_h^k(x, a) \geq Q_h^\star(x, a) - \epsilon \cdot (H + 1 - h)$.*

3. *For all $(x, h, k) \in \mathcal{S} \times [H] \times [K]$, there is some $\bar{a} \in \mathcal{A}$ so that $\Delta_h(x, \bar{a}) \leq \epsilon$ and $\widetilde{Q}_h^k(x, \bar{a}) \geq Q_h^\star(x, \bar{a}) - \epsilon \cdot (H + 1 - h)$. In particular, $\widetilde{Q}^k$ is an $\epsilon \cdot (H + 2 - h)$-approximate distillation on $Q^\star$.*

4. *For all $(x, h, k) \in \mathcal{S} \times [H] \times [K]$, it holds that $\widetilde{V}_h^k(x) \geq V_h^\star(x) - \epsilon \cdot (H + 2 - h)$.*

5. *For any* $(h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ *so that* $\tau_h^k = 0$*, it holds that, for* $n = N_h^k(x_h^k, a_h^k)$*,*

$$(\overline{R}_h^k - Q_h^\star)(x_h^k, a_h^k) \le \alpha_n^0 H + \mathrm{clip}\left[\beta_n | \frac{[\Delta_h(x_h^k, a_h^k) - 2\epsilon \cdot (H+1)]_+}{2H}\right] + \left(1 + \frac{1}{H}\right) \cdot \sum_{i=1}^n \alpha_n^i \cdot (\widetilde{V}_{h+1}^{k_h^i(x_h^k, a_h^k)} - V_{h+1}^\star)(x_{h+1}^{k_h^i(x_h^k, a_h^k)}).$$

6. *For any* $(h, k) \in [H] \times [K]$ *so that* $\tau_h^k = 0$*, it holds that* $\widetilde{V}_h^k(x_h^k) \le \overline{R}_h^k(x_h^k, a_h^k) + \epsilon \cdot H$*.*

Proof Fix any $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Set $k^0 = 0$ and for $i \ge 1$,

$$k^i := \min\left(\left\{k \in [K] : k > k^{i-1} \text{ and } (x_h^k, a_h^k) = (x, a)\right\} \cup \{K+1\}\right).$$

Let $\mathcal{H}_k$ denote the σ-field generated by all random variables up to and including episode k, step H; the random variable k^i is a stopping time of the filtration $(\mathcal{H}_k)_{k \ge 0}$. Let $\mathcal{F}_i$, $i \ge 0$ be the filtration given by $\mathcal{F}_i = \mathcal{H}_{k^i}$. Then $\left(\mathbb{1}[k^i \le K] \cdot [(\hat{\mathbb{P}}_h^{k^i} - \mathbb{P}_h)V_{h+1}^\star](x, a)\right)_{i=1}^K$ is a martingale difference sequence adapted to the filtration $\mathcal{F}_i$. By the Azuma-Hoeffding inequality and a union bound over all $m \in [K]$, it holds that, for some constant $C_0 > 0$, with probability at least $1 - p/(SAH)$,

$$\forall m \in [K] : \quad \left|\sum_{i=1}^m \alpha_m^i \cdot \mathbb{1}[k^i \le K] \cdot [(\hat{\mathbb{P}}_h^{k^i} - \mathbb{P}_h)V_{h+1}^\star](x, a)\right| \le \frac{C_0 H}{4}\sqrt{\sum_{i=1}^m (\alpha_m^i)^2 \cdot \iota} \le \frac{C_0}{2}\sqrt{\frac{H^3 \iota}{m}}, \tag{81}$$

where the final inequality follows from item 2 of Lemma 46. Taking a union bound over all $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, we get that with probability $1 - p$, for all $(x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$,

$$\left|\sum_{i=1}^n \alpha_n^i \left[(\hat{\mathbb{P}}_h^{k_h^i(x,a)} - \mathbb{P}_h)V_{h+1}^\star\right](x, a)\right| \le \frac{C_0}{2}\sqrt{\frac{H^3 \iota}{n}} \quad \text{where } n = N_h^k(x, a). \tag{82}$$

(Here we have applied (81) with $m = N_h^k(x, a) \le K$, and used the fact that $\mathbb{1}[k_h^i(x, a) \le K] = 1$ for $i \le N_h^k(x, a)$.) Let $\mathcal{E}^{\text{pred}}$ denote the probability $1 - p$ event under which (82) holds. From (22) we have that $\beta_n/2 \ge C_0\sqrt{H^3\iota/n}$. Then item 1 of the lemma follows from Lemma 36 and (82).

To establish the remaining items of the lemma statement, we use reverse induction on h. The base case $h = H+1$ is immediate since as a matter of convention, all of $\overline{R}_{H+1}^k, Q_{H+1}^\star, \widetilde{Q}_{H+1}^k \widetilde{V}_{H+1}^k, V_{H+1}^\star$ are identically 0. Assuming that items 2, 3, and 4 hold for step $h+1$, (82) and Lemma 36 give that, under the event $\mathcal{E}^{\text{pred}}$, for each $(x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, for $n = \widetilde{N}_h^k(x, a)$,

$$\begin{aligned}(\overline{R}_h^k - Q_h^\star)(x, a) \ge & \sum_{i=1}^n \alpha_n^i \cdot \left((\widetilde{V}_{h+1}^{\widetilde{k}_h^i(x,a)} - V_{h+1}^\star)(x_{h+1}^{\widetilde{k}_h^i(x,a)}) + \left((\hat{\mathbb{P}}_h^{\widetilde{k}_h^i(x,a)} - \mathbb{P}_h)V_{h+1}^\star\right)(x, a) + b_i\right) \\ \ge & -\epsilon \cdot (H + 1 - h) + \beta_n/2 - \frac{C_0}{2}\sqrt{H^3\iota/n} \\ \ge & -\epsilon \cdot (H + 1 - h),\end{aligned}$$

thus establishing item 2 of the lemma at step h.

To establish item 3 at step h, we use increasing induction on k. The base case $k = 1$ follows from the fact that, by assumption, $\widetilde{Q}^1$ is an ϵ-approximate distillation on $Q^\star$. To establish the inductive step, we note that by construction, $\widetilde{Q}^k_h(x,a) = \min\{\overline{R}^k_h(x,a), \widetilde{Q}^{k-1}_h(x,a), \overline{Q}^k_h(x,a)\}$ for all $(x,a) \in \mathcal{S} \times \mathcal{A}$. By the inductive hypothesis (on k), for each $x \in \mathcal{S}$, there is some $\bar{a} \in \mathcal{A}$ so that $\Delta_h(x,\bar{a}) \leq \epsilon$ and $\widetilde{Q}^{k-1}_h(x,\bar{a}) \geq Q^\star_h(x,\bar{a}) - \epsilon \cdot (H+1-h)$. Under the event $\mathcal{E}^{\mathrm{wc}}$ have that $\overline{Q}^k_h(x,\bar{a}) \geq Q^\star_h(x,\bar{a})$, and we have already established (item 2) that $\overline{R}^k_h(x,\bar{a}) \geq Q^\star_h(x,\bar{a}) - \epsilon \cdot (H+1-h)$, which implies that $\widetilde{Q}^k_h(x,\bar{a}) \geq Q^\star_h(x,\bar{a}) - \epsilon \cdot (H+1-h)$.

Finally we establish item 4 at step h. Again we use increasing induction on k, noting that the base case $k = 1$ follows from the fact that, for all $x \in \mathcal{S}$, $\widetilde{V}^1_h(x) = \max_{a \in \mathcal{A}} \widetilde{Q}^1_h(x,a) \geq V^\star_h(x) - \epsilon$, using that $\widetilde{Q}^k$ is an ϵ-approximate distillation on $Q^\star$. To establish the inductive step (i.e., at episode k, assuming that item 4 holds at episode $k-1$ and step h), note that for any $x \in \mathcal{S}$,

$$\widetilde{V}^k_h(x) = \max_{a' \in A^k_h(x)} \left\{ \max\{\widetilde{Q}^k_h(x,a'), \underline{Q}^k_h(x,a')\} \right\} \tag{83}$$

Moreover, since $\widetilde{Q}^k_h$ is an $\epsilon \cdot (H+2-h)$-approximate distillation on $Q^\star_h$ (item 3 at step h), for any x, there is some $\bar{a} \in \mathcal{A}$ so that $\Delta_h(x,\bar{a}) + [Q^\star_h(x,\bar{a}) - \widetilde{Q}^k_h(x,\bar{a})]_+ \leq \epsilon \cdot (H+2-h)$. By Lemma 37 with $(x,a,h) = (x,\bar{a},h)$, since $\Delta_h(x,\bar{a}) \leq \epsilon \cdot (H+2-h)$, it holds that either $\bar{a} \in A^k_h(x)$ or else there is some $a^\star \in A^k_h(x)$ so that $V^\star_h(x) - \underline{Q}^k_h(x,a^\star) < \epsilon$. If $\bar{a} \in A^k_h(x)$, then

$$\max_{a' \in A^k_h(x)} \left\{ \max\{\widetilde{Q}^k_h(x,a'), \underline{Q}^k_h(x,a')\} \right\} \geq \widetilde{Q}^k_h(x,\bar{a}) \geq V^\star_h(x) - \epsilon \cdot (H+2-h).$$

Otherwise, we have

$$\max_{a' \in A^k_h(x)} \left\{ \max\{\widetilde{Q}^k_h(x,a^k_h), \underline{Q}^k_h(x,a^k_h)\} \right\} \geq \underline{Q}^k_h(x,a^\star) > V^\star_h(x) - \epsilon \cdot (H+2-h).$$

Thus, by (83) and the inductive hypothesis (on k), it holds that $\widetilde{V}^k_h(x) \geq V^\star_h(x) - \epsilon \cdot (H+2-h)$, as desired.

Next we establish items 5 and 6 of the lemma. By item 2 of the lemma, under the event $\mathcal{E}^{\mathrm{wc}}$, we have, for all $(h,k) \in [H] \times [K]$,

$$\overline{R}^k_h(x^k_h,a^k_h) \geq Q^\star_h(x^k_h,a^k_h) - \epsilon \cdot H \geq \underline{Q}^k_h(x^k_h,a^k_h) - \epsilon \cdot H.$$

Further, note that $\widetilde{Q}^k_h(x^k_h,a^k_h) \leq \overline{R}^k_h(x^k_h,a^k_h)$ by the definition of $\widetilde{Q}^k_h$ (step 2(d)ii of the algorithm). Then by (83) and the definition of a^k_h (using that $\tau^k_h = 0$), we have that

$$\begin{aligned}
\overline{R}^k_h(x^k_h,a^k_h) &\geq \max\{\underline{Q}^k_h(x^k_h,a^k_h), \widetilde{Q}^k_h(x^k_h,a^k_h)\} - \epsilon \cdot H \\
&= \max_{a' \in A^k_h(x^k_h)} \left\{ \max\{\underline{Q}^k_h(x^k_h,a'), \widetilde{Q}^k_h(x^k_h,a')\} \right\} - \epsilon \cdot H \\
&= \widetilde{V}^k_h(x^k_h) - \epsilon \cdot H,
\end{aligned}$$

establishing item 6. Further, by item 4 of the lemma, we have that $\overline{R}^k_h(x^k_h,a^k_h) \geq V^\star_h(x^k_h) - 2\epsilon \cdot (H+1)$. Hence

$$\overline{R}^k_h(x^k_h,a^k_h) - Q^\star_h(x^k_h,a^k_h) \geq V^\star_h(x^k_h) - Q^\star_h(x^k_h,a^k_h) - 2\epsilon \cdot (H+1) \geq \Delta_h(x^k_h,a^k_h) - 2\epsilon \cdot (H+1).$$

The statement of item 5 then follows from item 1 and Lemma 21. ■

D.3. Additional bounds on Q- and V-value functions

In this section we prove some additional bounds on $\overline{V}_h^k, \underline{V}_h^k, \Delta V_h^k, \Delta \breve{V}_h^k$.

Recall from Lemma 18 that for any (x, h, k, a) for which $x \notin \mathcal{G}_h^k$, we have $\Delta \breve{V}_h^k(x) \leq \Delta V_h^k(x)$. The next lemma (in particular, (85)) shows that a reverse inequality holds up to a factor of $1 + 1/H$, *if there is a non-optimal action in* $A_h^k(x)$. This fact formalizes the intuition that the purpose of defining the clipped value functions $\Delta \breve{V}_h^k, \Delta \breve{Q}_h^k$ is to avoid paying for the case when only optimal actions remain in $A_h^k(x)$ (in which case one should not suffer any regret no matter which action is taken at x).

Lemma 39 *For any $(x, h, k) \in \mathcal{S} \times [H] \times [K]$, if there is a non-optimal action in $A_h^k(x)$, then, under the event $\mathcal{E}^{\text{wc}}$, it holds that*

$$\Delta \breve{V}_h^k(x) \geq \frac{\Delta_{\min,h}(x)}{4} \tag{84}$$

$$\overline{V}_h^k(x) - \underline{V}_h^k(x) \leq \Delta V_h^k(x) \leq \left(1 + \frac{1}{H}\right) \cdot \Delta \breve{V}_h^k(x). \tag{85}$$

Proof We begin by verifying (84). Fix x, h, k and set $a^\star = \arg\max_{a \in A_h^k(x)} \{\overline{Q}_h^k(x, a) - \underline{Q}_h^k(x, a)\}$. Then by Definition 11, we have that $\Delta \breve{V}_h^k(x) = \min\left\{\Delta \breve{V}_h^{k-1}(x),\ \Delta \breve{Q}_h^k(x, a^\star)\right\}$. If $\Delta \breve{V}_h^k(x) = \Delta \breve{V}_h^{k-1}(x)$, then we may replace k with $k-1$, noting the existence of a non-optimal action in $A_h^k(x)$ implies the existence of a non-optimal action in $A_h^{k-1}(x)$ (continuing this process may eventually lead to the case $k = 0$, for which (84) and (85) hold by the definition $\Delta \breve{V}_h^0(x) = H$). So we may assume that $\Delta \breve{V}_h^k(x) = \Delta \breve{Q}_h^k(x, a^\star)$.

Let a' denote some sub-optimal action in $A_h^k(x)$. Then under the event $\mathcal{E}^{\text{wc}}$, we must have that $x \notin \mathcal{G}_h^k$, meaning that

$$\begin{aligned}
\overline{Q}_h^k(x, a') - \underline{Q}_h^k(x, a') \leq & \overline{Q}_h^k(x, a^\star) - \underline{Q}_h^k(x, a^\star) && \text{(Since } a^\star \text{ maximizes the confidence interval)} \\
\leq & \Delta Q_h^k(x, a^\star) && \text{(By Lemma 16)} \\
\leq & \Delta \breve{Q}_h^k(x, a^\star) + \frac{\Delta_{\min}}{4}. && \text{(By Lemma 17)}
\end{aligned}$$

Moreover, as in the proof of Lemma 28, we have, by Lemmas 16 and 17,

$$\overline{V}_h^k(x) - \underline{V}_h^k(x) \leq \Delta V_h^k(x) \leq \Delta \breve{V}_h^k(x) + \frac{\Delta_{\min}}{4H} = \Delta \breve{Q}_h^k(x, a^\star) + \frac{\Delta_{\min}}{4H}. \tag{86}$$

Combining the above displays, we obtain that under the event $\mathcal{E}^{\text{wc}}$,

$$\begin{aligned}
\Delta_{\min,h}(x) \leq & \Delta_h(x, a') \\
\leq & (\overline{V}_h^k(x) - \underline{V}_h^k(x)) + (\overline{Q}_h^k(x, a') - \underline{Q}_h^k(x, a')) \\
\leq & 2 \cdot \Delta \breve{Q}_h^k(x, a^\star) + \frac{\Delta_{\min}}{2},
\end{aligned}$$

which implies that $\Delta \breve{V}_h^k(x) = \Delta \breve{Q}_h^k(x, a^\star) \geq \frac{\Delta_{\min,h}(x)}{4}$. This verifies (84).

To verify (85), we use the first two inequalities in (86) to get

$$\overline{V}_h^k(x) - \underline{V}_h^k(x) \leq \Delta V_h^k(x) \leq \Delta \check{V}_h^k(x) + \frac{\Delta_{\min}}{4H} \leq \Delta \check{V}_h^k(x) + \frac{\Delta_{\min,h}(x)}{4H} \leq \left(1 + \frac{1}{H}\right) \cdot \Delta \check{V}_h^k(x),$$

where the final inequality follows from (84). ■

The following simple lemma shows that $\widetilde{V}_h^k$ is bounded above by $\overline{V}_h^k$, which is an immediate consequence of the definition of $\widetilde{V}_h^k$ in `QLearningPreds`.

Lemma 40 *For all $x, a, h, k \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, it holds that $\widetilde{V}_h^k(x) \leq \overline{V}_h^k(x)$.*

Proof The definition of $\widetilde{Q}_h^k$ at step 2(d)ii ensures that for all x, a, j, k, we have that $\widetilde{Q}_h^k(x, a) \leq \overline{Q}_h^k(x, a)$. The conclusion of the lemma follows from the fact that $\overline{V}_h^k(x) = \max_{a' \in A_h^k(x)}\{\overline{Q}_h^k(x, a')\}$ and $\widetilde{V}_h^k(x) \leq \max_{a' \in A_h^k(x)}\{\widetilde{Q}_h^k(x, a')\}$. ■

D.4. Regret bounds for approximate distillation

The below lemma, the main result of this section, shows that in the case that the provided predictions $\widetilde{Q}$ are an ϵ-approximate distillation of the true value function $Q^\star$, then we may bound the regret of `QLearningPreds` by a quantity that in general will be smaller than generic worst-case regret bounds. In particular, the set of states and actions $\mathcal{S} \times \mathcal{A} \times [H]$ is replaced with the fooling set $\mathcal{F}(\epsilon'/2, \epsilon')$, which will be significantly smaller if the predictions $\widetilde{Q}_h$ are very accurate.

Lemma 41 *Suppose the event $\mathcal{E}^{\mathrm{wc}} \cap \mathcal{E}^{\mathrm{pred}}$ holds, and set $\epsilon' := 4\epsilon \cdot (H+1)$. If $\widetilde{Q}$ is an ϵ-approximate distillation on $Q^\star$ and either $\widetilde{Q}$ lacks ϵ'-fooling optimal actions (Definition 7). Then the following regret bounds hold:*

$$\begin{aligned}
&\sum_{k=1}^{K} (V_1^\star - V_1^{\pi^k})(x_1^k) \\
\leq &O((\epsilon H + \epsilon') \cdot TH) + O\left(H \cdot \sum_{h=1}^{H} \sum_{k=1}^{K} \sigma_h^k \check{\delta}_h^k\right) \\
&+ O\left(\min\left\{\sqrt{H^6 K \iota \cdot |\mathcal{F}(\epsilon(H+1), \epsilon')|}, \sum_{(x,a,h) \in \mathcal{F}(\epsilon(H+1), \epsilon')} \frac{H^4 \iota}{[\Delta_h(x,a) - 2\epsilon \cdot (H+1)]_+}\right\}\right).
\end{aligned} \tag{87}$$

Proof Recall from `QLearningPreds` that the values $\tau_h^k \in \{0, 1\}$ are defined as follows: $\tau_h^k = 0$ if $x_h^k \in \mathcal{G}_h^k$ or $\Delta V_h^k(x_h^k) \leq \varphi_h(\widehat{\Delta}^k)$, and $\tau_h^k = 1$ otherwise. Also recall that we defined values $\sigma_h^k \in \{0, 1\}$ for all $(h, k) \in [H] \times [K]$ as follows: $\sigma_h^k = 0$ if $x_h^k \in \mathcal{G}_h^k$ or $\Delta \check{V}_h^k(x_h^k) \leq \frac{1}{1+\frac{1}{H}} \cdot \varphi_h(\widehat{\Delta}^k)$, and $\sigma_h^k = 1$ otherwise.

For any $h \in [H]$, let $\mathcal{W}_h^\sigma \subset [K]$ denote the set of episodes k for which $\sigma_h^k = 1$. Similarly, let $\mathcal{W}_h^\tau \subset [K]$ denote the set of episodes k for which $\tau_h^k = 1$.

Set $\epsilon' = 4\epsilon \cdot (H+1)$, and for each $(x,h) \in \mathcal{S} \times [H]$, let $\mathcal{A}^{\mathrm{opt}}_{h,\epsilon'}(x)$ denote the set of actions $a' \in \mathcal{A}$ so that $V^\star_h(x) - Q^\star_h(x,a') \le \epsilon'$. Next, for any $k \in [K]$, we have that, under the event $\mathcal{E}^{\mathrm{wc}}$,

$$\begin{aligned}
&\sum_{h=1}^H V^\star_h(x_h^k) - Q^\star_h(x_h^k, a_h^k) \\
\le& \epsilon' H + \sum_{h=1}^H \mathbb{1}[a_h^k \notin \mathcal{A}^{\mathrm{opt}}_{h,\epsilon'}(x_h^k)] \cdot (V^\star_h(x_h^k) - Q^\star_h(x_h^k,a_h^k)) \\
\le& (\epsilon \cdot (H+1) + \epsilon')H + \sum_{h=1}^H (1-\sigma_h^k) \cdot \mathbb{1}[a_h^k \notin \mathcal{A}^{\mathrm{opt}}_{h,\epsilon'}(x_h^k)] \cdot (\widetilde{V}_h^k(x_h^k) - Q^\star_h(x_h^k,a_h^k)) + 4\sigma_h^k \cdot \breve{\delta}_h^k \qquad (88) \\
\le& (2\epsilon \cdot (H+1) + 2\epsilon')H + \sum_{h=1}^H (1-\tau_h^k) \cdot \mathbb{1}[a_h^k \notin \mathcal{A}^{\mathrm{opt}}_{h,\epsilon'}(x_h^k)] \cdot (\overline{R}_h^k(x_h^k,a_h^k) - Q^\star_h(x_h^k,a_h^k)) + 4\sigma_h^k \cdot \breve{\delta}_h^k, \qquad (89)
\end{aligned}$$

where (88) follows from item 4 of Lemma 38 and Lemma 28, and (89) follows from item 6 of Lemma 38 and Claim 42 below.

Claim 42 *For any $(h,k) \in [H] \times [K]$ so that $\tau_h^k = 1$, at least one of the following statements holds true under the event $\mathcal{E}^{\mathrm{wc}}$:*

- $\widetilde{V}_h^k(x_h^k) \le Q^\star_h(x_h^k,a_h^k) + \epsilon'$.
- $\Delta V_h^k(x_h^k) \le \left(1 + \frac{1}{H}\right) \cdot \Delta \breve{V}_h^k(x_h^k)$ *and* $\sigma_h^k = 1$.

Proof Suppose that the second statement does not hold true. Then either $\Delta V_h^k(x_h^k) > (1+1/H) \cdot \Delta\breve{V}_h^k(x_h^k)$ or $\sigma_h^k = 0$. First suppose that $\sigma_h^k = 0$. Since $\tau_h^k = 1$, we have $x_h^k \notin \mathcal{G}_h^k$, meaning that $\Delta\breve{V}_h^k(x_h^k) \le \frac{1}{1+1/H} \cdot \varphi_h(\widehat{\Delta}^k)$. But $\tau_h^k = 1$ also implies that $\Delta V_h^k(x_h^k) > \varphi_h(\widehat{\Delta}^k)$, which implies that $\Delta V_h^k(x_h^k) > (1+1/H) \cdot \Delta\breve{V}_h^k(x_h^k)$.

Thus we may assume from here on that $\Delta V_h^k(x_h^k) > (1+1/H) \cdot \Delta\breve{V}_h^k(x_h^k)$. By Lemma 39, under the event $\mathcal{E}^{\mathrm{wc}}$, $A_h^k(x_h^k)$ must consist of only optimal actions. By definition of $\widetilde{V}_h^k$, there is some $a \in A_h^k(x_h^k)$ so that $\widetilde{V}_h^k(x_h^k) = \max\{\widetilde{Q}_h^k(x_h^k,a), \underline{Q}_h^k(x_h^k,a)\}$. We know that a must be an optimal action, i.e., $\Delta_h(x_h^k,a) = 0$. Since the input predictions $\widetilde{Q}$ lack ϵ'-fooling optimal actions (Definition 7),[13] it holds that $\widetilde{Q}_h(x_h^k,a) \le V^\star_h(x_h^k) + \epsilon'$. Therefore,

$$\widetilde{V}_h^k(x_h^k) = \widetilde{Q}_h^k(x_h^k,a) \le \widetilde{Q}_h(x_h^k,a) \le V^\star_h(x_h^k) + \epsilon'.$$

Moreover, since $a_h^k \in A_h^k(x_h^k)$ (and therefore is an optimal action), we have that $V^\star_h(x_h^k) = Q^\star_h(x_h^k,a_h^k)$, meaning that $\widetilde{V}_h^k(x_h^k) - Q^\star_h(x_h^k,a_h^k) \le \epsilon'$, as desired. ■

We next need the following claim:

Claim 43 *For any (k,h) satisfying $\tau_h^k = 0$, if either*

[13] We remark that this is the only place in the proof where we use that the predictions $\widetilde{Q}$ lack ϵ'-fooling optimal actions.

1. *$a_h^k \notin \mathcal{A}_{h,\epsilon'}^{\mathrm{opt}}(x_h^k)$; or*

2. *$(\widetilde{V}_h^k - V_h^\star)(x_h^k) > \epsilon'$,*

then we have that, under the event $\mathcal{E}^{\mathrm{wc}}$, $(x_h^k, a_h^k, h) \in \mathcal{F}(\epsilon \cdot (H+1), \epsilon')$.

Proof [Proof of Claim 43] For the entirety of the proof of the claim we assume that $\mathcal{E}^{\mathrm{wc}}$ holds. We first suppose that $a_h^k \notin \mathcal{A}_{h,\epsilon'}^{\mathrm{opt}}(x_h^k)$. Notice that $\Delta_h(x_h^k, a_h^k) > \epsilon'$ since $a_h^k \notin \mathcal{A}_{h,\epsilon'}^{\mathrm{opt}}(x_h^k)$. By item 4 of Lemma 38 and the choice of a_h^k when $\tau_h^k = 0$,

$$\max\{\widetilde{Q}_h^k(x_h^k, a_h^k), \underline{Q}_h^k(x_h^k, a_h^k)\} = \max_{a' \in A_h^k(x_h^k)} \left\{ \max\{\widetilde{Q}_h^k(x_h^k, a'), \underline{Q}_h^k(x_h^k, a')\} \right\} = \widetilde{V}_h^k(x_h^k) \geq V_h^\star(x_h^k) - \epsilon \cdot (H+1) \tag{90}$$

If $\underline{Q}_h^k(x_h^k, a_h^k) \geq V_h^\star(x_h^k) - \epsilon \cdot (H+1)$, then it holds that $Q_h^\star(x_h^k, a_h^k) \geq V_h^\star(x_h^k) - \epsilon \cdot (H+1) > V_h^\star(x_h^k) - \epsilon'$ (since $\epsilon' > \epsilon \cdot (H+1)$), which contradicts $a_h^k \notin \mathcal{A}_{h,\epsilon'}^{\mathrm{opt}}(x_h^k)$. Hence $\widetilde{Q}_h^k(x_h^k, a_h^k) \geq V_h^\star(x_h^k) - \epsilon \cdot (H+1)$, meaning that $\widetilde{Q}_h^k(x_h^k, a_h^k) - Q_h^\star(x_h^k, a_h^k) \geq \Delta_h(x_h^k, a_h^k) - \epsilon \cdot (H+1)$. Since $\widetilde{Q}_h(x_h^k, a_h^k) = \widetilde{Q}_h^1(x_h^k, a_h^k) \geq \widetilde{Q}_h^k(x_h^k, a_h^k)$, we get that $(x_h^k, a_h^k, h) \in \mathcal{F}(\epsilon \cdot (H+1), \epsilon')$.

Next suppose that $(\widetilde{V}_h^k - V_h^\star)(x_h^k) > \epsilon'$. Then, again using the choice of a_h^k when $\tau_h^k = 0$,

$$\max\{\widetilde{Q}_h^k(x_h^k, a_h^k), \underline{Q}_h^k(x_h^k, a_h^k)\} \geq \widetilde{V}_h^k(x_h^k) > V_h^\star(x_h^k) + \epsilon'.$$

Since $\underline{Q}_h^k(x_h^k, a_h^k) \leq Q_h^\star(x_h^k, a_h^k) \leq V_h^\star(x_h^k)$ under the event $\mathcal{E}^{\mathrm{wc}}$, we must have $\widetilde{Q}_h^k(x_h^k, a_h^k) > V_h^\star(x_h^k) + \epsilon'$, which implies that $\widetilde{Q}_h^1(x_h^k, a_h^k) > V_h^\star(x_h^k) + \epsilon'$, meaning that $(x_h^k, a_h^k, h) \in \mathcal{F}(\epsilon \cdot (H+1), \epsilon')$.
■

Next for any $h \in [H]$, we compute

$$\begin{aligned}
&\sum_{k \notin \mathcal{W}_h^\tau} \mathbb{1}[a_h^k \not\in \mathcal{A}_{h,\epsilon'}^{\text{opt}}(x_h^k)] \cdot (\overline{R}_h^k(x_h^k, a_h^k) - Q_h^\star(x_h^k, a_h^k)) \\
\leq & \sum_{(x,a,h) \in \mathcal{F}_h(\epsilon \cdot (H+1), \epsilon')} \sum_{\substack{i \in [N_h^{K+1}(x,a)]: \\ k_h^i(x,a) \not\in \mathcal{W}_h^\tau}} (\overline{R}_h^{k_h^i(x,a)}(x,a) - Q_h^\star(x,a)) && (91) \\
\leq & H \cdot |\mathcal{F}_h(\epsilon(H+1), \epsilon')| + \sum_{(x,a,h) \in \mathcal{F}_h(\epsilon(H+1), \epsilon)} \left(\sum_{i=1}^{N_h^{K+1}(x,a)} \left(1 + \frac{1}{H}\right) \sum_{t=1}^{i} \alpha_i^t \cdot (\widetilde{V}_{h+1}^{k_h^t(x,a)} - V_{h+1}^\star)(x_{h+1}^{k_h^t(x,a)}) \right. \\
& \left. + \text{clip}\left[\beta_i \middle| \frac{[\Delta_h(x,a) - 2\epsilon \cdot (H+1)]_+}{4H} \right] \right) && (92) \\
\leq & H \cdot |\mathcal{F}_h(\epsilon(H+1), \epsilon')| + \sum_{(x,a,h) \in \mathcal{F}_h(\epsilon(H+1), \epsilon')} \left(\left(1 + \frac{1}{H}\right) \sum_{t=1}^{N_h^{K+1}(x,a)} (\widetilde{V}_{h+1}^{k_h^t(x,a)} - V_{h+1}^\star)(x_{h+1}^{k_h^t(x,a)}) \cdot \sum_{i=t}^{\infty} \alpha_i^t \right. \\
& \left. + \sum_{i=1}^{N_h^{K+1}(x,a)} \text{clip}\left[\beta_i \middle| \frac{[\Delta_h(x,a) - 2\epsilon \cdot (H+1)]_+}{4H} \right] \right) \\
\leq & H \cdot |\mathcal{F}_h(\epsilon(H+1), \epsilon')| + \sum_{(x,a,h) \in \mathcal{F}_h(\epsilon(H+1), \epsilon')} \min\left\{ 8C_0 \sqrt{H^3 \iota \cdot N_h^{K+1}(x,a)}, \frac{64 C_0^2 H^4 \iota}{[\Delta_h(x,a) - 2\epsilon \cdot (H+1)]_+} \right\} \\
& (1 + 1/H)^2 \cdot \sum_{k \in [K]} (\widetilde{V}_{h+1}^k - V_{h+1}^\star)(x_{h+1}^k), && (93)
\end{aligned}$$

where (91) uses Claim 43, (92) uses item 5 of Lemma 38 and the fact that $(x_h^{k_h^i(x,a)}, a_h^{k_h^i(x,a)}) = (x,a)$, and the final inequality (93) uses item 3 of Lemma 46 and Lemma 22. Moreover, we have that

$$\begin{aligned}
&\sum_{k \in [K]} (\widetilde{V}_{h+1}^k - V_{h+1}^\star)(x_{h+1}^k) \\
\leq & K \cdot \epsilon' + \sum_{k \in [K]} \mathbb{1}[\sigma_{h+1}^k = 1 \text{ and } \Delta V_{h+1}^k(x_{h+1}^k) \leq (1 + 1/H) \breve{\delta}_{h+1}^k] \cdot (\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(x_{h+1}^k) \\
& + \sum_{k \not\in \mathcal{W}_{h+1}^\tau} (\widetilde{V}_{h+1}^k - V_{h+1}^\star)(x_{h+1}^k) && \text{(Using Claim 42 and Lemma 40)} \\
\leq & (2\epsilon' + \epsilon \cdot H) \cdot K + \sum_{k=1}^{K} 2\sigma_{h+1}^k \cdot \breve{\delta}_{h+1}^k + \sum_{k \not\in \mathcal{W}_{h+1}^\tau} \mathbb{1}[(\widetilde{V}_{h+1}^k - V_{h+1}^\star)(x_{h+1}^k) > \epsilon'] \cdot \left((\overline{R}_{h+1}^k - Q_{h+1}^\star)(x_{h+1}^k, a_{h+1}^k) \right) \\
& && \text{(Using item 6 of Lemma 38 and Lemma 16)} \\
\leq & (2\epsilon' + \epsilon H) \cdot K + 2 \sum_{k=1}^{K} \sigma_{h+1}^k \breve{\delta}_{h+1}^k + \sum_{(x,a,h+1) \in \mathcal{F}_{h+1}(\epsilon(H+1), \epsilon')} \sum_{i=1}^{N_{h+1}^{K+1}(x,a)} (\overline{R}_{h+1}^{k_{h+1}^i(x,a)}(x,a) - Q_{h+1}^\star(x,a)), \\
& && (94)
\end{aligned}$$

where (94) follows from Claim 43 (in particular, if $\tau_{h+1}^k = 0$ and $(\widetilde{V}_{h+1}^k - V_{h+1}^\star)(x_{h+1}^k) > \epsilon'$, then $(x_{h+1}^k, a_{h+1}^k, h+1) \in \mathcal{F}_{h+1}(\epsilon(H+1), \epsilon'))$.

Combining (93) and (94), and iterating for $h, h+1, \ldots, H$, we see that

$$\begin{aligned}
&\sum_{(x,a,h) \in \mathcal{F}_h(\epsilon(H+1),\epsilon')} \sum_{i=1}^{N_h^{K+1}(x,a)} (\overline{R}_h^{k_h^i(x,a)} - Q_h^\star(x_h^k, a_h^k)) \\
\leq & e^2 KH \cdot (2\epsilon' + \epsilon H) + e^2 H \cdot |\mathcal{F}(\epsilon(H+1),\epsilon')| + 2e^2 \sum_{h'=h+1}^{H} \sum_{k=1}^{K} \sigma_{h'}^k \breve{\delta}_{h'}^k \\
& + e^2 \sum_{h'=h}^{H} \sum_{(x,a,h') \in \mathcal{F}_{h'}(\epsilon(H+1),\epsilon')} \min \left\{ 8C_0 \sqrt{H^3 \iota \cdot N_{h'}^{K+1}(x,a)}, \frac{64 C_0^2 H^4 \iota}{[\Delta_{h'}(x,a) - 2\epsilon \cdot (H+1)]_+} \right\} \\
\leq & e^2 KH \cdot (2\epsilon' + \epsilon H) + e^2 H \cdot |\mathcal{F}(\epsilon(H+1),\epsilon')| + 2e^2 \sum_{h'=h+1}^{H} \sum_{k=1}^{K} \sigma_{h'}^k \breve{\delta}_{h'}^k \\
& + e^2 \sum_{h'=h}^{H} \min \left\{ 8C_0 \sqrt{H^3 K \iota \cdot |\mathcal{F}_{h'}(\epsilon(H+1),\epsilon')|}, \sum_{(x,a,h') \in \mathcal{F}_{h'}(\epsilon(H+1),\epsilon')} \frac{64 C_0^2 H^4 \iota}{[\Delta_{h'}(x,a) - 2\epsilon \cdot (H+1)]_+} \right\} \\
\leq & e^2 KH \cdot (2\epsilon' + \epsilon H) + e^2 H \cdot |\mathcal{F}(\epsilon(H+1),\epsilon')| + 2e^2 \sum_{h'=h+1}^{H} \sum_{k=1}^{K} \sigma_{h'}^k \breve{\delta}_{h'}^k \\
& + \min \left\{ 8C_0 e^2 \sqrt{H^4 K \iota \cdot |\mathcal{F}(\epsilon(H+1),\epsilon')|}, \sum_{(x,a,h) \in \mathcal{F}(\epsilon(H+1),\epsilon')} \frac{64 e^2 C_0^2 H^4 \iota}{[\Delta_h(x,a) - 2\epsilon \cdot (H+1)]_+} \right\}.
\end{aligned}$$

Combining this with (89) and (91) gives that

$$\begin{aligned}
&\sum_{k \in [K]} (V_1^\star - V_1^{\pi^k})(x_1) \\
\leq & HK(2\epsilon(H+1) + 2\epsilon') + HK \cdot \Pr[\overline{\mathcal{E}^{\mathrm{wc}} \cap \mathcal{E}^{\mathrm{pred}}}] + e^2 KH^2 \cdot (2\epsilon' + \epsilon H) + e^2 H^2 \cdot |\mathcal{F}(\epsilon(H+1),\epsilon')| \\
& + \min \left\{ 8C_0 e^2 \sqrt{H^6 K \iota \cdot |\mathcal{F}(\epsilon(H+1),\epsilon')|}, \sum_{(x,a,h) \in \mathcal{F}(\epsilon(H+1),\epsilon')} \frac{64 e^2 C_0^2 H^5 \iota}{[\Delta_h(x,a) - 2\epsilon \cdot (H+1)]_+} \right\} \\
& + 2e^2 H \sum_{h=1}^{H} \sum_{k=1}^{K} \sigma_h^k \breve{\delta}_h^k \\
\leq & O((\epsilon H + \epsilon') \cdot TH) + O\left(\min \left\{ \sqrt{H^6 K \iota \cdot |\mathcal{F}(\epsilon(H+1),\epsilon')|}, \sum_{(x,a,h) \in \mathcal{F}(\epsilon(H+1),\epsilon')} \frac{H^4 \iota}{[\Delta_h(x,a) - 2\epsilon \cdot (H+1)]_+} \right\} \right) \\
& + 2e^2 H \sum_{h=1}^{H} \sum_{k=1}^{K} \sigma_h^k \breve{\delta}_h^k,
\end{aligned}$$

where in the final inequality we use that $H^2 \cdot |\mathcal{F}(\epsilon(H+1),\epsilon')| \leq \sqrt{H^6 K \iota \cdot |\mathcal{F}(\epsilon(H+1),\epsilon')|}$ as long as $K \geq |\mathcal{F}(\epsilon(H+1),\epsilon')|$. (For $K < |\mathcal{F}(\epsilon(H+1),\epsilon')|$ the trivial regret bound of KH is bounded above by $\sqrt{H^6 K\iota \cdot |\mathcal{F}(\epsilon(H+1),\epsilon')|}$.) Moreover, we also use that $H^2 \cdot |\mathcal{F}(\epsilon(H+1),\epsilon')| \leq \sum_{(x,a,h)\in\mathcal{F}(\epsilon(H+1),\epsilon')} \frac{H^4\iota}{[\Delta_h(x,a)-2\epsilon\cdot(H+1)]_+}$ in the final line. This verifies the statement (87) of the theorem.

■

Note that Lemma 41 does not quite establish the guarantee of the improved regret bounds of Theorems 8 or 9 when the predictions $\widetilde{Q}$ are an approximate distillation of $Q^\star$. In particular, we have not yet shown how to bound the term $\sum_{h=1}^H \sum_{k=1}^K \sigma_h^k \breve{\delta}_h^k$. We do so in Lemmas 44 and 45 below; the first treats the case where `QLearningPreds` uses `DeltaConst`, and the second treats the case where `QLearningPreds` uses `DeltaIncr`.

Lemma 44 *For any prediction function $\widetilde{Q}$, the algorithm `QLearningPreds` (with `DeltaConst`) has the following guarantee under the event $\mathcal{E}^{\mathrm{wc}}$:*

$$\sum_{h=1}^H \sum_{k=1}^K \sigma_h^k \cdot \breve{\delta}_h^k \leq O\left(\min\left\{\sqrt{\widehat{\lambda}\cdot H^8 SAT\iota}, H^7\iota \cdot \left(\sum_{(x,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]: a\notin\mathcal{A}_{h,0}^{\mathrm{opt}}(x)} \frac{1}{\Delta_h(x,a)} + \frac{|\mathcal{A}^{\mathrm{mul}}|}{\Delta_{\min}}\right)\right\}\right).$$

Proof Suppose that $\mathcal{E}^{\mathrm{wc}}$ holds. Recall that $\breve{\delta}_h^k$ is only defined for h,k so that $x_h^k \notin \mathcal{G}_h^k$; but if $x_h^k \in \mathcal{G}_h^k$, then $\sigma_h^k = 0$, meaning that the sum $\sum_{h=1}^H \sum_{k=1}^K \sigma_h^k \cdot \breve{\delta}_h^k$ is well-defined.

Further, recall that $\widehat{\lambda}$ is chosen so that $\frac{1}{\widehat{\lambda}} \cdot \mathscr{C}_{M,T,\widehat{\lambda}} = \mathscr{R}$. We claim that $\widehat{\lambda} \geq SAH^3/K$; to see this, note that $\frac{1}{\lambda}\cdot\mathscr{C}_{M,T,\lambda}$ is a decreasing function of λ, and that for the choice $\lambda_0 = SAH^3/K = SAH^4/T$,

$$\frac{1}{\lambda_0}\mathscr{C}_{M,T,\lambda_0} \geq \min\left\{TH^2, \frac{H^7 T}{SAH^4}\cdot\frac{SAH}{2H}\right\} \geq T/2 \geq \mathscr{R}.$$

Finally note that $\widehat{\Delta}^K = \frac{\mathscr{R}}{KH} = \frac{\mathscr{C}_{M,T,\widehat{\lambda}}}{\widehat{\lambda}\cdot KH}$.

For each $h \in \mathcal{H}$, set $\mathcal{Y}_h := \{k : \sigma_h^k = 1\}$. Lemma 33 gives that $|\mathcal{Y}_h| \leq \max\{SAH^3, \widehat{\lambda} \cdot K\} \leq \widehat{\lambda} \cdot K$, where we use that $\widehat{\lambda}$ is chosen so that $\widehat{\lambda} K \geq SAH^3$. By item 2 of Lemma 27,

$$\begin{aligned}
&\sum_{h=1}^{H} \sum_{k=1}^{K} \sigma_h^k \breve{\delta}_h^k \\
=&\sum_{h=1}^{H} \sum_{k \in \mathcal{Y}_h} \breve{\delta}_h^k \\
\leq&\sum_{h=1}^{H} \left(|\mathcal{Y}_h| \cdot \varphi_h(\widehat{\Delta}) + e^2 SAH^2 + \min\left\{ e^2 C_2 \sqrt{H^5 SA |\mathcal{Y}_h| \iota}, \sum_{(x,a,h') \in \mathcal{S} \times \mathcal{A} \times [H]} \frac{e^2 C_2^2 H^3 \iota}{\max\left\{ \frac{\Delta \mathring{Q}_{h'}^K(x,a)}{2H}, \frac{\Delta_{\text{min}}}{4H^2} \right\}} \right\} \right) \\
\leq& e^2 SAH^3 + \min\left\{ e^2 C_2 \sqrt{H^7 SA \widehat{\lambda} K \iota}, \sum_{(x,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]} \frac{e^2 C_2^2 H^4 \iota}{\max\left\{ \frac{\Delta \mathring{Q}_{h}^K(x,a)}{2H}, \frac{\Delta_{\text{min}}}{4H^2} \right\}} \right\} \\
&+ O(\widehat{\lambda} \cdot K) \cdot \min\left\{ \sqrt{\frac{H^7 SA \iota}{\widehat{\lambda} \cdot K}}, \frac{1}{\widehat{\lambda} \cdot K} \cdot H^6 \iota \cdot \left(\sum_{(x,a,h') \in \mathcal{S} \times \mathcal{A} \times [H] : a \notin \mathcal{A}_{h',0}^{\text{opt}}(x)} \frac{1}{\Delta_h(x,a)} + \frac{|\mathcal{A}^{\text{mul}}|}{\Delta_{\text{min}}} \right) \right\}.
\end{aligned}$$

By Lemma 30 (which we can use since $\mathcal{E}^{\text{wc}}$ holds) and Lemma 47 (together with $\widehat{\lambda} \geq SAH^3/K$), it therefore follows that

$$\begin{aligned}
&\sum_{h=1}^{H} \sum_{k=1}^{K} \sigma_h^k \breve{\delta}_h^k \\
\leq& O\left(\min\left\{ \sqrt{\widehat{\lambda} \cdot H^8 SAT\iota}, H^7 \iota \cdot \left(\sum_{(x,a,h) \in \mathcal{S} \times \mathcal{A} \times [H] : a \notin \mathcal{A}_{h,0}^{\text{opt}}(x)} \frac{1}{\Delta_h(x,a)} + \frac{|\mathcal{A}^{\text{mul}}|}{\Delta_{\text{min}}} \right) \right\} \right).
\end{aligned}$$

■

Lemma 45 *For any prediction function $\widetilde{Q}$, the algorithm `QLearningPreds` given some parameter $\lambda \geq SAH^3/K$ (with `DeltaIncr` and some input parameter $\widetilde{\Delta}_{\text{min}} \leq \Delta_{\text{min}}$) has the following guarantee under the event $\mathcal{E}^{\text{wc}}$:*

$$\sum_{h=1}^{H} \sum_{k=1}^{K} \sigma_h^k \breve{\delta}_h^k \leq O\left(\min\left\{ \sqrt{\lambda \cdot SAH^9 T \iota^2}, H^9 \iota^2 \cdot \left(\sum_{(x,a,h) : a \notin \mathcal{A}_{h,0}^{\text{opt}}(x)} \frac{1}{\Delta_h(x,a)} + \frac{|\mathcal{A}^{\text{mul}}|}{\widetilde{\Delta}_{\text{min}}} \right) \right\} \right).$$

Proof For each $h \in \mathcal{H}$, set $\mathcal{Y}_h := \{k : \sigma_h^k = 1\}$. Lemma 35 gives that $|\mathcal{Y}_h| \leq \max\{SAH^3, \lambda \cdot K\}$. Recall that the input parameter λ was assumed to satisfy $\lambda \geq SAH^3/K$, meaning that $|\mathcal{Y}_h| \leq \lambda K$.

By item 2 of Lemma 27,

$$
\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}\sigma_h^k\breve{\delta}_h^k \\
=&\sum_{h=1}^{H}\sum_{k\in\mathcal{Y}_h}\breve{\delta}_h^k \\
\leq&\sum_{h=1}^{H}\left(|\mathcal{Y}_h|\cdot\varphi_h(\widehat{\Delta}^K)+e^2SAH^2+\min\left\{e^2C_2\sqrt{H^5SA|\mathcal{Y}_h|\iota},\ \sum_{(x,a,h')\in\mathcal{S}\times\mathcal{A}\times[H]}\frac{e^2C_2^2H^3\iota}{\max\left\{\frac{\Delta\mathring{Q}_{h'}^K(x,a)}{2H},\frac{\Delta_{\min}}{4H^2}\right\}}\right\}\right) \\
\leq& e^2SAH^3+\min\left\{e^2C_2\sqrt{H^7SA\lambda K\iota},\ \sum_{(x,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]}\frac{e^2C_2^2H^4\iota}{\max\left\{\frac{\Delta\mathring{Q}_{h}^K(x,a)}{2H},\frac{\Delta_{\min}}{4H^2}\right\}}\right\} \\
&+O(1)\cdot H\cdot\lambda K\cdot\min\left\{\frac{H^6\iota^2}{\lambda\cdot K}\cdot\sum_{(x,a,h)}\frac{1}{\max\left\{\frac{\widetilde{\Delta}\mathring{Q}_{h}^K(x,a)}{2H},\frac{\widetilde{\Delta}_{\min}}{4H^2}\right\}},\sqrt{\frac{SAH^8\iota^2}{\lambda\cdot K}}\right\}.
\end{aligned}
$$

By Lemma 30 (which we can apply since $\mathcal{E}^{\mathrm{wc}}$ holds) and Lemma 47 (using that $\lambda \geq SAH^3/K$), it follows that

$$
\sum_{h=1}^{H}\sum_{k=1}^{K}\sigma_h^k\breve{\delta}_h^k \leq O\left(\min\left\{\sqrt{\lambda\cdot SAH^9T\iota^2},H^9\iota^2\cdot\left(\sum_{(x,a,h):a\notin\mathcal{A}_{h,0}^{\mathrm{opt}}(x)}\frac{1}{\Delta_h(x,a)}+\frac{|\mathcal{A}^{\mathrm{mul}}|}{\widetilde{\Delta}_{\min}}\right)\right\}\right).
$$

■

Appendix E. Proofs of main theorems

We begin by proving Theorem 9, restated below for convenience.

Theorem 9 (Restated) *The algorithm* `QLearningPreds` *with the* `DeltaConst` *subroutine satisfies the following two guarantees, when given as input a parameter $\mathscr{R} \in [SAH^3, \frac{T}{SA}]$ and predictions $\widetilde{Q}$:*

1. *If $\mathscr{R} \geq \mathscr{C}_{M,T,1}$, then for an* arbitrary *choice of input predictions $\widetilde{Q}$, the regret of* `QLearningPreds` *is $O(\mathscr{R})$.*
2. *Fix any $\epsilon > 0$, and set $\epsilon' = 4\epsilon\cdot(H+1)$. When the input predictions $\widetilde{Q}$ are an ϵ-approximate distillation of $Q^\star$ (Definition 5) and lack ϵ'-fooling optimal actions (Definition 7), the regret of* `QLearningPreds` *is*

$$
O\left(\mathscr{C}_{M,T,\widehat{\lambda}}+\epsilon'TH+\min\left\{\sqrt{H^5T\iota\cdot|\mathcal{F}(\epsilon'/2,\epsilon')|},\ \sum_{(x,a,h)\in\mathcal{F}(\epsilon'/2,\epsilon')}\frac{H^4\iota}{[\Delta_h(x,a)-\epsilon'/2]_+}\right\}\right), \tag{95}
$$

where $\widehat{\lambda} \in (0,1)$ is chosen so that $\frac{1}{\widehat{\lambda}} \cdot \mathscr{C}_{M,T,\widehat{\lambda}} = \mathscr{R}$.

Proof We begin with the proof of item 1. It is without loss to assume $T \geq SAH^3$; otherwise, by similar reasoning to that in Lemma 47, the trivial regret bound of T suffices. Now, item 1 is an immediate consequence of Lemma 31.

We next prove item 2. The event $\mathcal{E}^{\mathrm{wc}} \cap \mathcal{E}^{\mathrm{pred}}$ does not hold with probability at most $2p = 2/(H^2K)$, which adds at most $T \cdot 2p = O(1)$ to the regret bound. Thus it suffices to bound the regret conditioned on $\mathcal{E}^{\mathrm{wc}} \cap \mathcal{E}^{\mathrm{pred}}$. Then item 2 is an immediate consequence of Lemmas 41 and 44. ■

We next prove Theorem 8; below we present the version of the theorem which does not require that each state has a unique optimal action. In this more general setting, the subroutine `DeltaIncr` of `QLearningPreds` requires as input a parameter $\widetilde{\Delta}_{\mathrm{min}}$ which is guaranteed to be a lower bound on Δ_{min}. The resulting regret bounds will depend on a modified version of the λ-complexity $\mathscr{C}_{M,T,\lambda}$ (see (7)) with the parameter $\widetilde{\Delta}_{\mathrm{min}}$ replacing Δ_{min}; more precisely, we define

$$\mathscr{C}_{M,T,\lambda,\widetilde{\Delta}_{\mathrm{min}}} := \min\left\{\sqrt{\lambda \cdot TSAH^8\iota},\ H^7\iota \cdot \left(\sum_{(x,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]:a\notin\mathcal{A}^{\mathrm{opt}}_{h,0}(x)} \frac{1}{\Delta_h(x,a)} + \frac{|\mathcal{A}^{\mathrm{mul}}|}{\widetilde{\Delta}_{\mathrm{min}}}\right)\right\} \tag{96}$$

In the special case that $|\mathcal{A}^{\mathrm{mul}}| = 0$ (i.e., each state has a unique optimal action) and $\widetilde{\Delta}_{\mathrm{min}} = 0$, the quantity $\frac{0}{0}$ in (96) is to be interpreted as 0.

Theorem 8 (Full version) *Suppose we run algorithm `QLearningPreds` (Algorithm 1) with input parameter $\lambda \in [0,1]$, together with the `DeltaIncr` subroutine (Algorithm 4) with parameter $\widetilde{\Delta}_{\mathrm{min}}$ which is guaranteed to satisfy $\widetilde{\Delta}_{\mathrm{min}} \leq \Delta_{\mathrm{min}}$. Then, when given predictions $\widetilde{Q}$, the algorithms satisfy the following guarantees:*

1. *Suppose $\lambda \geq \frac{SAH^4}{T}$. Then for an arbitrary choice of input predictions $\widetilde{Q}$, the regret of `QLearningPreds` is*
$$O\left(\frac{\iota}{\lambda} \cdot \mathscr{C}_{M,T,\lambda,\widetilde{\Delta}_{\mathrm{min}}}\right).$$

2. *Fix any $\epsilon > 0$, and set $\epsilon' = 4\epsilon \cdot (H+1)$. When the input predictions $\widetilde{Q}$ are an ϵ-approximate distillation of $Q^\star$ (Definition 5) and lack ϵ'-fooling actions (Definition 7), the regret of `QLearningPreds` is*

$$O\left(H^2\iota \cdot \mathscr{C}_{M,T,\lambda,\widetilde{\Delta}_{\mathrm{min}}} + \epsilon' TH + \min\left\{\sqrt{H^5 T\iota \cdot |\mathcal{F}(\epsilon'/2,\epsilon')|},\ \sum_{(x,a,h)\in\mathcal{F}(\epsilon'/2,\epsilon')} \frac{H^4\iota}{[\Delta_h(x,a) - \epsilon'/2]_+}\right\}\right). \tag{97}$$

Proof We begin with the proof of item 1. As in the proof of Theorem 9, it is without loss to assume that $T \geq SAH^3$, as otherwise the trivial regret bound of T suffices. Now, item 1 is an immediate consequence of Lemma 32.

We next prove item 2. The event $\mathcal{E}^{\text{wc}} \cap \mathcal{E}^{\text{pred}}$ does not hold with probability at most $2p = 2/(H^2K)$, which adds at most $T \cdot 2p = O(1)$ to the regret bound. Thus it suffices to bound the regret conditioned on $\mathcal{E}^{\text{wc}} \cap \mathcal{E}^{\text{pred}}$. Then item 2 is an immediate consequence of Lemmas 41 and 45. ■

Appendix F. Miscellaneous lemmas

The following simple lemma establishes some properties of the parameters α_n^i (defined in (10)).

Lemma 46 (Lemma 4.1, Jin et al. (2018)) *The real numbers α_n^i satisfy the following properties:*

1. *For every $n \geq 1$, $\frac{1}{\sqrt{n}} \leq \sum_{i=1}^n \frac{\alpha_n^i}{\sqrt{i}} \leq \frac{2}{\sqrt{n}}$.*
2. *For every $n \geq 1$, $\max_{i \in [n]} \alpha_n^i \leq \frac{2H}{n}$ and $\sum_{i=1}^n (\alpha_n^i)^2 \leq \frac{2H}{n}$.*
3. *For every $i \geq 1$, $\sum_{n=i}^\infty \alpha_n^i = 1 + \frac{1}{H}$.*
4. *For every $n \geq 1$, it holds that $\sum_{i=1}^n \alpha_n^i = 1$.*

Recall the definition of $\mathscr{C}_{M,T,\lambda}$ in (7)

Lemma 47 *For any $\lambda \geq \frac{SAH^3}{K}$, it holds that*

$$\mathscr{C}_{M,T,\lambda} \geq SAH^6/2.$$

Proof Since $\lambda \geq SAH^3/K = SAH^4/T$, it holds that $\sqrt{\lambda \cdot TSAH^8\iota} \geq SAH^6$.

Next, it is evident that

$$\sum_{(x,a,h): a \notin \mathcal{A}_{h,0}^{\text{opt}}(x)} \frac{1}{\Delta_h(x,a)} + \frac{|\mathcal{A}^{\text{mul}}|}{\Delta_{\min}} \geq \frac{SAH}{2H} \geq SA/2,$$

since for each (x,h), all but one of the actions a in $\mathcal{A}$ are either counted in the form of $\frac{1}{\Delta_h(x,a)} \geq 1/H$ or $1/\Delta_{\min} \geq 1/H$.

Putting the above statements together gives the desired bound. ■

Appendix G. Proof for bandit case

In this section we prove Proposition 4, which specializes our main results to the case of multi-armed bandits. Though our bounds for multi-armed bandits are superseded by our regret bounds for online learning in MDPs, we present a separate proof for the bandit case to provide intuition about our techniques. Following Definition 5, we say that a prediction function $\widetilde{Q} : \mathcal{A} \to \mathbb{R}$ is an *ϵ-approximate distillation (in the bandit setting)* if there is some arm $\widetilde{a} \in \mathcal{A}$ so that

$$(Q^\star(a^\star) - Q^\star(\widetilde{a})) + [Q^\star(\widetilde{a}) - \widetilde{Q}(\widetilde{a})]_+ \leq \epsilon, \tag{98}$$

where $a^\star$ denotes the optimal arm (assumed to be unique).

```
Algorithm 5: BanditPreds
Input: Action space $\mathcal{A}$, number of time steps $T$, predictions $\widetilde{Q} : \mathcal{A} \to [0, 1]$, parameter $\lambda \in [0, 1]$
       and $\delta > 0$.
1. For each $a \in \mathcal{A}$, initialize $\overline{Q}^1(a) = \infty$, $\underline{Q}^1(a) = -\infty$, $N^1(a) = 0$, and $\widetilde{Q}^1(a) = \widetilde{Q}(a)$.
2. For $1 \leq t \leq T$:
    (a) If $t \leq \lambda \cdot T$:
        i. Select action $a^t := \arg\max_{a \in \mathcal{A}} \{\overline{Q}^t(a)\}$.
    (b) Else (i.e., if $t > \lambda \cdot T$):
        i. Select action $a^t := \arg\max_{a \in \mathcal{A}} \left\{\widetilde{Q}^t(a)\right\}$.
    (c) For each action $a \in \mathcal{A}$, let $N^{t+1}(a)$ denote the number of times $a$ was taken up to (and
        including) step $t$.
    (d) For each action $a \in \mathcal{A}$, let $\hat{\mu}^{t+1}(a)$ denote the mean of all rewards received when
        taking $a$ up to step $t$ (if $N^{t+1}(a) = 0$, set $\hat{\mu}^{t+1}(a) = 0$).
    (e) Update the $Q$-value functions as follows: for each $a \in \mathcal{A}$, set
            $\overline{Q}^{t+1}(a) := \hat{\mu}^{t+1}(a) + \sqrt{\frac{2 \log 1/\delta}{N^{t+1}(a)}}$
            $\underline{Q}^{t+1}(a) := \hat{\mu}^{t+1}(a) - \sqrt{\frac{2 \log 1/\delta}{N^{t+1}(a)}}$
            $\widetilde{Q}^{t+1}(a) := \max\left\{\underline{Q}^{t+1}(a), \min\left\{\overline{Q}^{t+1}(a), \widetilde{Q}(a)\right\}\right\}.$
```

Proposition 4 *There is an algorithm (`BanditPreds`, Algorithm 5) which satisfies the following two guarantees, when given as input a parameter $\lambda \in \left(\frac{A}{T}, 1\right)$ and predictions $\widetilde{Q}$:*

1. *Fix any $\epsilon > 0$. If the predictions $\widetilde{Q}$ are an ϵ-approximate distillation of $Q^\star$, then the regret is $\widetilde{O}(\epsilon T + \sqrt{|\mathcal{G}| \cdot T} + \sqrt{\lambda \cdot AT})$, where*
$$\mathcal{G} := \left\{ a \in \mathcal{A} \backslash \{a^\star\} : \widetilde{Q}(a) \geq Q^\star(a^\star) - \epsilon \right\}.$$
2. *For an arbitrary choice of $\widetilde{Q}$, the regret is $\widetilde{O}\left(\sqrt{\frac{TA}{\lambda}}\right)$.*

For convenience, for each $t \leq T$, we define

$$\widetilde{V}^t := \max_{a \in \mathcal{A}} \left\{ \widetilde{Q}^t(a) \right\} \tag{99}$$

By construction of the algorithm `BanditPreds`, note that $\widetilde{Q}^t(a^t) = \widetilde{V}^t$. We define the following "good event" $\mathcal{E}_0$:

$$\mathcal{E}_0 = \left\{ \forall t \leq T,\ \forall a \in \mathcal{A},\ \left| Q^\star(a) - \hat{\mu}^t(a) \right| \leq \sqrt{\frac{2 \log 1/\delta}{N^t(a)}} \right\}.$$

Note that under the event $\mathcal{E}_0$, $Q^\star(a) \in [\underline{Q}^t(a), \overline{Q}^t(a)]$ for all $a \in \mathcal{A}$.

Lemma 48 *Suppose the event $\mathcal{E}_0$ holds. Then for any sub-optimal action $a \neq a^\star$:*

1. *If a is taken at step $t \leq \lambda T$, then $N^t(a) \leq \frac{8 \log 1/\delta}{\Delta(a)^2}$.*
2. *If a is taken at step $t > \lambda T$ and $\widetilde{Q}$ is an ϵ-approximate distillation of $Q^\star$, and if $\Delta(a) > \epsilon$, then $\widetilde{Q}(a) \geq Q^\star(a^\star) - \epsilon$. In such a case, we have $N^t(a) \leq \frac{8 \log 1/\delta}{(\Delta(a) - \epsilon)^2}$.*

Proof If a is taken when $t \leq \lambda T$, we must have that $\overline{Q}^t(a) \geq \overline{Q}^t(a^\star) \geq Q^\star(a^\star)$, meaning that

$$Q^\star(a) \geq \overline{Q}^t(a) - \sqrt{\frac{8 \log 1/\delta}{N^t(a)}} \geq Q^\star(a^\star) - \sqrt{\frac{8 \log 1/\delta}{N^t(a)}},$$

from which the first point follows.

Choose $\widetilde{a}$ satisfying (98). Then for any t, under the event $\mathcal{E}_0$,

$$\widetilde{Q}^t(\widetilde{a}) \geq \min\{Q^\star(\widetilde{a}), \widetilde{Q}(\widetilde{a})\} \geq Q^\star(a^\star) - \epsilon,$$

where the final inequality follows from (98).

If a is taken when $t > \lambda T$, then we must have that $\widetilde{Q}^t(a) \geq \widetilde{Q}^t(\widetilde{a}) \geq Q^\star(a^\star) - \epsilon$. Since $\Delta(a) > \epsilon$, we have that $\underline{Q}^t(a) \leq Q^\star(a) < Q^\star(a^\star) - \epsilon$, meaning that $\widetilde{Q}^t(a) \leq \widetilde{Q}(a)$, and hence $\widetilde{Q}(a) \geq Q^\star(a^\star) - \epsilon$. In such a case, we have $\overline{Q}^t(a) \geq \widetilde{Q}^t(a) \geq Q^\star(a^\star) - \epsilon$, meaning that

$$Q^\star(a) \geq \overline{Q}^t(a) - \sqrt{\frac{8 \log 1/\delta}{N^t(a)}} \geq Q^\star(a^\star) - \sqrt{\frac{8 \log 1/\delta}{N^t(a)}} - \epsilon,$$

from which the desired inequality follows. ■

Lemma 49 *Suppose the event $\mathcal{E}_0$ holds. Then if $\Delta > 0$ satisfies $\frac{16A\log 1/\delta}{\Delta^2} \leq \lambda T$, in the first $\frac{16A\log 1/\delta}{\Delta^2}$ steps, some action a with $\Delta(a) \leq \Delta$ has been taken at least $\frac{8\log 1/\delta}{\Delta^2}$ times.*

Proof By item 1 of Lemma 48, under the event $\mathcal{E}_0$, each action $a \in \mathcal{A}$ is taken at most $\frac{8\log 1/\delta}{\Delta(a)^2}$ time steps in time steps $t \leq \lambda T - 1$. Let A_Δ denote the number of actions a with $\Delta(a) \leq \Delta$. Then by the pigeonhole principle, in the first

$$\sum_{a\in\mathcal{A}:\Delta(a)>\Delta} \frac{8\log 1/\delta}{\Delta(a)^2} + \frac{8A_\Delta \log 1/\delta}{\Delta^2} \leq \frac{8A\log 1/\delta}{\Delta^2} \leq \lambda T - 1$$

steps, some action a with $\Delta(a) \leq \Delta$ has been taken at least $\frac{8\log 1/\delta}{\Delta^2}$ times. ■

Proof [Proof of Proposition 4] We set $\delta = 1/(AT^2)$ in `BanditPreds` (Algorithm 5). It is straightforward from a Chernoff bound and union bound that $\Pr(\mathcal{E}_0) \geq 1 - \delta \cdot TA$. Therefore, the expected regret is bounded above by

$$\begin{aligned}\mathbb{E}\left[\sum_{a\in\mathcal{A}} \Delta(a)\cdot N^{T+1}(a)\right] \leq& \delta TA \cdot T + \mathbb{E}\left[\mathbb{1}[\mathcal{E}_0]\cdot\sum_{a\in\mathcal{A}}\Delta(a)\cdot N^{T+1}(a)\right]\\ \leq& 1 + \mathbb{E}\left[\sum_{a\in\mathcal{A}}\Delta(a)\cdot N^{T+1}(a)|\mathcal{E}_0\right],\end{aligned}$$

where the second inequality uses $\delta \leq 1/(AT^2)$.

We begin by bounding the regret in the event that $\widetilde{Q}$ is an ϵ-approximate distillation of $Q^\star$ (i.e., item 1 of Proposition 4). By item 2 of Lemma 48, any arm a with $\Delta(a) \geq 2\epsilon$ that is pulled at some step $t > \lambda T$ must satisfy $\widetilde{Q}(a) \geq Q^\star(a^\star) - \epsilon$, i.e., $a \in \mathcal{G}$. Moreover, for such arms a, we have that $N^{T+1}(a) \leq \frac{32\log 1/\delta}{\Delta(a)^2}$. Therefore, under the event $\mathcal{E}_0$, we have

$$\begin{aligned}\sum_{a\in\mathcal{A}}\Delta(a)\cdot N^{T+1}(a) \leq& 2\epsilon T + \sum_{a\in\mathcal{G}}\Delta(a)\cdot N^{T+1}(a) + \sum_{a\in\mathcal{A}\backslash\mathcal{G}}\Delta(a)\cdot N^{\lambda T+1}(a)\\ \leq& 2\epsilon T + \sqrt{\lambda TA\log 1/\delta} + \sum_{a\in\mathcal{A}\backslash\mathcal{G}:\Delta(a)>\sqrt{A\log 1/\delta/(\lambda T)}} \frac{8\log 1/\delta}{\Delta(a)}\\ &+ \mathbb{1}[|\mathcal{G}|>0]\cdot\left(\sqrt{T|\mathcal{G}|\log 1/\delta} + \sum_{a\in\mathcal{G}:\Delta(a)>\sqrt{|\mathcal{G}|\log 1/\delta/T}}\frac{32\log 1/\delta}{\Delta(a)}\right)\\ \leq& O\left(\epsilon T + \sqrt{\lambda TA\log 1/\delta} + \sqrt{T|\mathcal{G}|\log 1/\delta}\right).\end{aligned}$$

Next we prove item 2 (the robustness claim) of the proposition. Set $\Delta_\lambda := \sqrt{\frac{16A\log 1/\delta}{T\lambda}}$. Then by Lemma 49, in the first λT steps of `BanditPreds`, some action $\bar{a}$ with $\Delta(\bar{a}) \leq \Delta_\lambda$ has been taken at least $\frac{8\log 1/\delta}{\Delta_\lambda^2} = \frac{T\lambda}{2A}$ times. Hence

$$\underline{Q}^{\lambda T+1}(\bar{a}) \geq Q^\star(a^\star) - \Delta(\bar{a}) - \sqrt{\frac{16A\log 1/\delta}{T\lambda}} \geq Q^\star(a^\star) - 2\Delta_\lambda.$$

Therefore, for all $t > \lambda T$, $\max_{a \in \mathcal{A}}\{\widetilde{Q}^t(a)\} \geq \max_{a \in \mathcal{A}}\{\underline{Q}^t(a)\} \geq Q^\star(a^\star) - 2\Delta_\lambda$. Hence, for any action $a \in \mathcal{A}$ that is taken at step $t > \lambda T$ and satisfies $\Delta(a) > 2\Delta_\lambda$, we have that $\overline{Q}^t(a) \geq Q^\star(a^\star) - 2\Delta_\lambda$, meaning that

$$Q^\star(a) \geq \overline{Q}^t(a) - \sqrt{\frac{8\log 1/\delta}{N^t(a)}} \geq Q^\star(a^\star) - \sqrt{\frac{8\log 1/\delta}{N^t(a)}} - 2\Delta_\lambda,$$

meaning that $N^t(a) \leq \frac{8\log 1/\delta}{(\Delta(a)-2\Delta_\lambda)^2}$. Hence, under the event $\mathcal{E}_0$, we have

$$\sum_{a \in \mathcal{A}} \Delta(a) \cdot N^{T+1}(a) \leq 4\Delta_\lambda \cdot T + \sum_{a \in \mathcal{A}: \Delta(a) > 4\Delta_\lambda} \frac{32\log 1/\delta}{\Delta(a)} \leq O\left(\sqrt{\frac{TA\log 1/\delta}{\lambda}}\right).$$

■